The London School of Economics and Political Science

Essays in Labour Economics and Innovation

Gaia Dossi

A thesis submitted to the Department of Economics

for the degree of Doctor of Philosophy

September 2024

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 55,884 words.

Statement of conjoint work

I confirm that Chapter 2 was jointly coauthored with Enrico Berkes, Davide Coluccia, and Mara Squicciarini. Chapter 3 was jointly coauthored with Davide Coluccia.

Abstract

This thesis consists of three chapters that examine how scientists and inventors shape the rate, direction, and diffusion of science and innovation.

Chapter 1 studies the consequences of the Black-white gap among scientists and inventors. Using data on US patents, medical research articles, and clinical trials linked to the racial distribution of last names in the US population, I find that the racial and ethnic composition of scientists has important implications for the direction and the rate of medical research and innovation.

Chapter 2 studies how societies react to adverse events. Following the 1918 Influenza Pandemic in the US, we document an increase in country-level religiosity and innovation. Within counties, individuals from more religious backgrounds become more religious, while those from less religious backgrounds are more likely to pursue scientific occupations. Facing adversity widens the distance in religiosity between science-oriented individuals and the rest of the population, and it increases the polarization of religious beliefs.

Chapter 3 studies how human mobility affects the production and diffusion of innovation. Using full-count census data for the UK and the US and newly-digitized UK patent data, we document that out-migration promotes the diffusion of new technologies from the country of destination to the country of origin of migrants. While physical return migration is an important driver of this "return innovation" effect, the interactions between emigrants and their origin communities promote technology diffusion even without return migration.

Acknowledgements

My supervisors John Van Reenen, Oriana Bandiera, and Xavier Jaravel have provided invaluable guidance and support throughout my PhD. I am incredibly grateful to them for their mentorship, their generosity with their time and insights, and for giving me opportunities I never could have dreamed of. They have been, and will continue to be, an inspiration to do groundbreaking research and provide outstanding mentorship to young researchers.

I have also benefited from advice and interactions with many other members of the community at the LSE, especially Nava Ashraf, Tim Besley, Maarten De Ridder, Jeremiah Dittmar, Matthias Doepke, Ethan Ilzetzki, Camille Landais, Steve Machin, Alan Manning, Ralf Martin, Guy Michaels, Steve Pischke, Johannes Spinnewijn, Anna Valero, and the amazing research community at the CEP and in the Economics Department.

I am also incredibly grateful for the mentorship and support I have received from Paola Giuliano and Paola Sapienza since before starting the PhD. I have learned so much from them, and their advice and encouragement have been invaluable.

I wish to thank Geri Miric, Lubala Chibwe, Linda Cleavely, Martin Hannon, Charles Tocock, Tosin Lamidi, Emma Taverner, Hitesh Patel, Michael Rose, and Nick Warner for their outstanding help with administrative tasks and IT.

This thesis would have been impossible without a very special group of people, my coauthors and "almost-coauthors" Livia Alfonsi, Virginia Minni, Marta Morando, Mara Squicciarini, Miguel Acosta, Enrico Berkes, Louise Guillouët, Lucas Husted, and Lorenzo

Pessina. I feel incredibly lucky to have gone through this journey with them.

I am incredibly grateful to my friends in London and abroad and to my family for providing constant support during good and difficult times.

This thesis is dedicated to the memory of my grandmothers, Alba and Ernesta.

Contents

1 Race and Science **25**

1.1 Introduction 25

1.2 Data 32

1.2.1 Measuring Scientific Production 32

1.2.2 Measuring Patenting Activity 33

1.2.3 Measuring Race and Ethnicity 34

1.2.4 Matching Procedure 35

1.2.5 Validation of Black-sounding Last Names as a Proxy for Race . . 36

1.3 Race and the Direction of Science and Innovation 37

1.3.1 The Black-white Gap in Medical Research and Innovation 37

1.3.2 Measuring the Direction of Science and Innovation 39

1.3.3 Demographics-based Approach: Results 42

1.3.4 Frequency-based Approach: Results 45

1.3.5 Robustness Checks 46

1.3.6 Impact 49

1.4 Mechanisms 49

1.4.1 Race or Socio-economic Status? Ancestry Variation 49

1.4.2 A Shock to Relative Mortality: The Introduction of HAART . . . 53

1.5 Race and the Quantity and Quality of Innovation 56

1.5.1 Research Design 56

1.5.2 Results 57

1.6 Equalizing Access to Innovation and Science 59

1.6.1 Model Outline 59

1.6.2 Estimation 61

1.6.3 Policy Counterfactuals 61

1.7 Discussion 62

1.8 Conclusion 64

1.9 Appendix: Figures and Tables 89

1.9.1 Appendix Figures 89

1.9.2 Appendix Tables 98

1.10 Appendix: Model 133

1.10.1 Workers 133

1.10.2 Occupational choice 135

1.10.3 Workers' equilibrium 136

1.10.4 Firms 137

1.10.5 Equilibrium 138

1.10.6 Estimation 138

2 Dealing With Adversity: Religiosity or Science? Evidence From the Great Influenza Pandemic 144

2.1 Introduction 144

2.2 Historical Background 150

2.2.1 The Great Influenza Pandemic 150

2.2.2 The Pandemic and Religion 151

2.2.3 The Pandemic and Science 152

2.3 Data 153

2.3.1 Religiosity Measure 153

2.3.2 Measuring Scientific Progress 157

2.3.3 Exposure to the Great Influenza Pandemic 158

2.4 Main Results: County-Level Analysis 159

2.4.1 Empirical Strategy 159

2.4.2 The Effect of the Influenza Pandemic on Religiosity 161

2.4.3 The Effect of the Influenza Pandemic on Scientific Progress 164

2.4.4 Joint Dynamics of Religiosity and Innovation 168

2.5 Mechanisms: Individual-Level Analysis 169

2.5.1 Turning to Religion or Turning to Science 170
2.5.2 Science-Oriented Individuals Became Less Religious 172
2.5.3 Polarization of Religious Beliefs 173
2.6 Discussion: Interpretation and Limitations of the Results 175
2.7 Conclusions 177
2.8 Figures 186
2.9 Tables 191
2.10 Appendix: Data 196
2.10.1 Names 196
2.10.2 Religious Affiliations 196
2.10.3 Patents 197
2.10.4 Occupational Structure 198
2.10.5 Controls & Mortality Statistics 198
2.10.6 Other Data 199
2.10.7 Boundary Harmonization 199
2.10.8 Details on Sample Construction 199
2.11 Appendix: Figures and Tables 202
2.11.1 Figures 202
2.11.2 Tables 209

3 Return Innovation: The Knowledge Spillovers of the British Migration

to the United States, 1870-1940 **228**

3.1 Introduction 228

3.2 Historical and Institutional Background 238

3.2.1 The English and Welsh Migration to the United States 238

3.2.2 Intellectual Property Protection in the US and the UK 242

3.2.3 Anecdotal Evidence of Return Innovation 244

3.3 Data 246

3.3.1 Migration Data 246

3.3.2 Patent Data 249

3.3.3 Other Variables 250

3.4 Empirical Strategy 252

3.4.1 Baseline Methodology 252

3.4.2 Threats to Identification 254

3.4.3 Shift-Share Instrumental Variable Strategy 255

3.4.4 Shock Propagation Difference-in-Differences Strategy 257

3.5 Empirical Results 260

3.5.1 Exposure to US Innovation Shapes Innovation in the UK 260

3.5.2 Innovation Shocks in the US Diffuse to the UK 264

3.5.3 Technology Transfer and Spillovers: A Text-Based Approach . . . 267

3.6 Potential Mechanisms and Discussion 268

3.6.1 Is Return Innovation Return Migration? 269
3.6.2 Return Innovation Through Interactions 270
3.6.3 Return Innovation Through Economic Integration 274
3.6.4 Potential Additional Mechanisms 280
3.6.5 Discussion 282
3.7 Conclusions 283
3.8 Figures 303
3.9 Tables 310
3.10 Appendix: Data Sources and Methods 318
3.10.1 Summary of Data Sources 318
3.10.2 Geo-referenced Census Records 322
3.10.3 Linked Inventor Sample 324
3.10.4 Figures 328
3.10.5 Tables 334
3.11 Appendix: Novel Patent Data 337
3.11.1 Sources and Digitization 337
3.11.2 External Validation 340
3.11.3 Measuring Pairwise Similarity Between US and UK Patents . . . 341
3.11.4 Summary Statistics and Stylized Facts 343
3.11.5 Figures 345

3.11.6 Tables 350

3.12 Appendix: Linked International Migrants Sample 351

3.12.1 Sources and Linking Algorithm 351

3.12.2 Internal and External Validation 354

3.12.3 Return Migration Data 358

3.12.4 Summary Statistics and Stylized Facts 359

3.12.5 Figures 361

3.12.6 Tables 367

3.13 Appendix: Additional Results 368

3.13.1 Trade-Induced Technology Transfer 368

3.13.2 Selection of British Migrants 370

3.13.3 Long-Run Effect of Return Innovation 371

3.13.4 Further Additional Results 372

3.13.5 Figures 376

3.13.6 Tables 379

3.14 Appendix: Robustness Analysis 389

3.14.1 Alternative Baseline Specifications 389

3.14.2 Instrumental Variable Strategy 391

3.14.3 Shock Propagation 394

3.14.4 Figures 400

3.14.5 Tables 411

List of Figures

1.1 Pr(Black | Last Name) in the US Population, 2010 Census 73

1.2 Pr(Inventor) and Pr(Scientist) by Quintiles of Pr(Black | Last Name) . . 74

1.3 Causes of Death Ranked by Relative Mortality 75

1.4 Demographics-based Approach, Research on Other Demographic Groups 76

1.5 Impact of HAART on Likelihood that an HIV-related Paper has "Black or African American" Among its MeSH Terms 77

1.6 Impact of HAART on Match between Black-sounding Researchers and HIV Research 78

1.7 Equalizing Access to Innovation and Science, % of Inventors and Scientists 79

1.8 Pr(Black | Last name), US Census vs Florida 89

1.9 Pr(Black | Last name), US Census vs Florida, Scientists and Inventors . . 90

1.10 Correlation Between Pr(Black | Last name) of First and Last Author . . 91

1.11 Example of a PubMed Article with Black or African American Among its MeSH Codes, but Not Classified with a Dictionary Approach 92

1.12 Example of a PubMed Article Classified as Addressing Black or African Americans Using a Dictionary Classification, but Not Classified Using MeSH Codes 93

1.13 Demographics-based Approach, Additional Split by Publication Type . . 94

1.14 Relative Mortality, Weighted by Total Number of Deaths 95

1.15 Incidence of Sickle Cell Anemia by Region of the World 96

1.16 Incidence of Melanoma by Region of the World 96

1.17 Mortality Rates of White and Black or African Americans, Melanoma . . 97

1.18 Mortality Rates of White and Black or African Americans, Thyroid Neoplasm 97

2.1 Estimated Religiosity Scores by Name 186

2.2 Spatial Distribution of Excess Mortality During the Great Influenza Pandemic 187

2.3 Impact of the Influenza on Religiosity 188

2.4 Impact of the Influenza on Innovation 189

2.5 Impact of the Influenza on the Polarization of Religious Beliefs 190

2.6 Estimated Names Religiosity Scores, by Confession 202

2.7 In-sample and Out-of-sample Fit of the Religiosity Measure 203

2.8 Example of Pharmaceutical Patent 204

2.9 Correlation Between WW1 Deaths and Excess Deaths 205

2.10 Correlation Between Abramitzky, Boustan and Eriksson (2016) Religiosity and Baseline Religiosity 206

2.11 Correlation Between Religiosity and Science 207

2.12 Distribution of Cities in the Alternative Sample 208

3.1 Number of US Immigrants from the UK and Linked US-UK Migrants, 1840–1930 303

3.2 Spatial Distribution of US Migrants Across British Districts 304

3.3 OLS and Reduced-Form Association between Knowledge Exposure and Innovation 305

3.4 Heterogeneous Effects of Return Innovation Across Technology Classes . 306

3.5 Effect of Exposure to US Innovation Shocks on UK Innovation 307

3.6 Effect of Exposure to US Shocks on the UK-US Patent Similarity 308

3.7 Effect of the Transatlantic Telegraph Cable on Innovation 309

3.8 Spatial Distribution of the Share of Geo-coded Addresses in the UK Population Censuses, 1851–1911 328

3.9 Distribution of the Share of Geo-coded Addresses by Census 329

3.10 Number of Active Newspapers Over the Period 1880–1940, by District . . 330

3.11 Distribution of Districts Connected to the UK Telegraph Network in 1862 331

3.12 Matching Rate of the Linked Inventors-Census Sample, 1881–1911 332

3.13 Distribution of Inventors Across UK Districts, 1881–1911 333

3.14 Sample Annotated Patent Documents: the Bessemer Process and the First Modern Safety Bicycle 345

3.15 Confusion Matrix of the Technology Sector Classifier 346

3.16 Composition of Total Patents Granted in the United Kingdom Across Technology Classes, 1880–1939 347

3.17 Distribution of Patents and Patents per Capita Across Districts, 1880–1939 348

3.18 Distribution of the Total Number of Patents Granted Across Districts and Selected Technology Classes, 1880–1939 349

3.20 Quality of Matches in the Complete Linked Sample: Names and Surnames 361

3.21 Comparison Between Linked Data and Estimates from Baines 362

3.19 Share of British Immigrants in the US Census Matched to the UK Census 363

3.22 Falsification Exercise of the Intergenerational Linked Sample 364

3.23 Comparison Between Linked Samples: Full and Restricted 365

3.24 Distribution of Emigration Rate Across Districts, 1871–1930 366

3.25 Flexible Difference-in-Differences Estimated Effect of Tariff Reform on Innovation 376

3.26 Flexible Triple Differences Estimated Effect of the Influenza Pandemic on US Innovation 377

3.27 Long-Run Association Between Knowledge Exposure and Subsequent Innovation Activity, 1940–2020 378

3.29 First Stage Binned Scatter Plot 400

3.30 Shock-Level Balance Tests for Instrumental Variable Validity 401

3.31 District-Level Balance Tests for Instrumental Variable Validity 402

3.32 Effect of Synthetic Innovation Shocks Across Technology Classes 403

3.33 Effect of the Influenza Shock Across Technology Classes 404

3.34 Alternative Staggered Triple Differences Estimators for the Effect of US Synthetic Shocks on Innovation 405

3.28 Alternative S.E. Estimators of the Return Innovation Result 406

3.35 Flexible Triple Differences Effect of Family Member Out-Migration on Innovation 407

3.36 Alternative Staggered Double Differences Estimators for the Effect of Neighborhood Out-Migration on Innovation 408

3.37 Flexible Difference-in-Differences Effect of Neighborhood-Level Out-Migration on Innovation 409

3.38 Co-variate Balance for Individual-Level Design 410

List of Tables

1.1 Sample Statistics, Research Articles and Patents 80

1.2 Demographics-based Approach 81

1.3 Frequency-based Approach 82

1.4 Ancestry Variation: Sickle Cell Anemia 83

1.5 Ancestry Variation: Melanoma 84

1.6 DD Around the Introduction of HAART, NIH Grants 85

1.7 DD Around the Introduction of HAART, PubMed Articles 86

1.8 Inventors with Black-sounding Name Are Less Likely to be Granted a Patent 87

1.9 Patents Granted to Inventors with Black-sounding Name Have Higher Impact 88

1.10 Extract of Racial Probabilities by Last Name, 2010 Census 98

1.11 Validation of Black-sounding Names as Predictor of Race, Florida 99

1.12 Differences in Observables Between Matched and Unmatched Scientists . 100

1.13 Differences in Observables Between Matched and Unmatched Inventors . 101

1.14 Demographics-based Approach, Dictionary Classification 102

1.15 Demographics-based Approach, Logit Model 103

1.16 Demographics-based Approach, Controlling for Gender of First Author . 104

1.17 Demographics-based Approach, Controlling for Journal FE, State FE, Affiliation FE 105

1.18 Demographics-based Approach, Alternative Definitions of Black-sounding Last Name 106

1.19 Demographics-based Approach, Including All Articles (Not Only Top 1,000 Journals by Journal Commercialization Impact Factor (JCIF) 107

1.20 Demographics-based Approach, Split by Below/Above Median Journal Commercialization Impact Factor 108

1.21 Demographics-based Approach, Articles Linked to a Patent 109

1.22 Demographics-based Approach, Controlling for MeSH Codes Linked to Article 110

1.23 Demographics-based Approach, Research on Other Demographic Groups 111

1.24 Demographics-based Approach: Evidence from Ongoing Clinical Trials . 113

1.25 Frequency-based Approach, Dictionary Classification 114

1.26 Frequency-based Approach (Logit Model) 115

1.27 Frequency-based Approach, Controlling for Gender of First Author . . . 116

1.28 Frequency-based Approach, Controlling for Journal FE, State FE, First Author Affiliation FE 117

1.29 Frequency-based Approach, Alternative Definitions of Black-sounding Last Name 118

1.30 Frequency-based Approach, Including All Articles (Not Only Top 1,000 Journals by Journal Commercialization Impact Factor (JCIF) 120

1.31 Frequency-based Approach, Split by Below/Above Journal Commercialization Impact Factor (JCIF) 121

1.32 Frequency-based Approach: Articles Linked to a Patent 122

1.33 Frequency-based Approach: Threshold Set at 1.3 Higher Mortality 123

1.34 Frequency-based Approach: Threshold Set at 1.7 Higher Mortality 124

1.35 Frequency-based Approach: Alternative Definitions of Frequency 125

1.36 Demographics-based Approach, Total Citations 126

1.37 Demographics-based Approach, Citations (Relative Citation Ratio) . . . 127

1.38 Frequency-based Approach, Total Citations 128

1.39 Frequency-based Approach, Citations (Relative Citation Ratio) 129

1.40 Demographics-based Approach, Split by Below/Above median Journal Impact Factor (JIF) 130

1.41 Frequency-based Approach, Split by Below/Above Median Journal Impact Factor (JIF) 131

1.42 Sickle Cell Anemia, by MeSH Code 132

2.1 The Impact of the Influenza on Religiosity 191

2.2 The Impact of the Influenza on Innovation 192

2.3 Impact of the Influenza on Occupational Choice 193

2.4 Religious Background, Religiosity, and STEM Occupations 194

2.5 Effect of the Influenza on Individual Religiosity: STEM and Non-STEM . 195

2.6 Summary Statistics 209

2.7 STEM Professions 210

2.8 Balance Checks Regressions 211

2.9 Impact of the Influenza on Religiosity: Weighted Regressions 212

2.10 Impact of the Influenza on Religiosity 213

2.11 Impact of the Influenza on Religiosity: Accounting for Migration 214

2.12 Impact of the Influenza on Religiosity: City-Level Analysis 215

2.13 Impact of the Influenza on Religiosity: Names Scores without Fixed Effects 216

2.14 Impact of the Influenza on Religiosity: Alternative Thresholds 217

2.15 Impact of the Influenza on the Concentration of Names 218

2.16 Impact of the Influenza on Religiosity measured with Saint and Biblical Names 219

2.17 Impact of the Influenza on Innovation: Weighted Regressions 220

2.18 Impact of the Influenza on Innovation: Robustness Regressions 221

2.19 Impact of the Influenza on Innovation: Alternative Measures of Overall Innovation 222

2.20 Impact of the Influenza on Innovation: Alternative Measures of Pharmaceutical Innovation 223

2.21 Impact of the Influenza on Patent Importance 224

2.22 Impact of the Influenza on Innovation: Accounting for Migration 225

2.23 Impact of the Influenza on Innovation: City-Level Analysis 226

2.24 Religiosity and the Intensity of Innovation by Exposure to the Influenza . 227

3.1 Descriptive Statistics of Selected Variables 310

3.2 Effect of Exposure to US Technology on Innovation in Great Britain . . . 311

3.3 Effect of Exposure to US Innovation Shocks on UK Innovation 312

3.4 Association Between UK Innovation and Exposure to US Technology Through Overall and Return Emigration 313

3.5 Effect of Family Member Emigration on Innovation Produced by Relatives of the Emigrant in the UK 314

3.6 Effect of Emigration on Innovation Produced by Former Neighbors of the Emigrant in the UK 315

3.7 The Transatlantic Telegraph Cable and Innovation in the UK 316

3.8 Effect of US Emigration on Newspaper Coverage of US-related News . . 317

3.9 Descriptive Statistics on Newspapers and Newspaper Coverage in the UK 334

3.10 Correlation Between Inventor Characteristics and N. Matches 335

3.11 List of Industries By Tariff Rate, 1925–1935 336

3.12 Total Number of Patents Granted in the UK: Comparison Across Three Datasets 350

3.13 Correlation Between Number of Matches and Observable Characteristics 367

3.14 Zero-Stage Regressions Between Immigrant Shares and Railway Access . 379

3.15 British Immigrants Assortative Matching Across US Counties 380

3.16 Double and Triple Differences Effect of The Smoot-Hawley Act on Innovation 381

3.17 Selection of US Emigrants Compared to the Rest of the British Population 382

3.18 Selection of British Immigrants Compared to the Rest of the US Population 383

3.19 Association Between Out-Migration and the Volume of Innovation 384

3.20 Estimated Effect of the Influenza Pandemic on US Innovation 385

3.21 Double and Triple Differences Effect of Synthetic Innovation Shocks on Subsequent US Innovation 386

3.22 Long-Run Sector Correlation Between Knowledge Exposure and Innovation 387

3.23 Heterogeneity Analysis of the Effect of Neighborhood Out-Migration on Innovation 388

3.24 Knowledge Exposure and Innovation: Alternative Dependent Variables . 411

3.25 Knowledge Exposure and Innovation: Alternative Measures of Knowledge Exposure 412

3.26 Knowledge Exposure and Innovation: Alternative Sets of Fixed Effects . 413

3.27 Return Innovation Accounting for Patent Quality 414

3.28 Return Innovation Accounting for Patent Quality of US Patents 415

3.29 Effect of Exposure to US Technology on Innovation in Great Britain: Patents with Firm Assignee 416

3.30 First Stage of the Instrumental Variable Estimation 417

3.31 Return Innovation Result Using the Leaveout Instrument 418

3.32 Return Innovation Result Using the Modified Leaveout Instruments . . . 419

3.33 Robustness Analysis on the Effect of Exposure to US Technology on the Similarity and Originality of Innovation in Great Britain 420

3.34 Triple Differences Estimated Effect of US Synthetic Shocks on UK Innovation: Alternative Thresholds 421

3.35 Double and Triple Differences Estimated Effect of the Great Influenza Pandemic in the US on Innovation in the UK: Robustness Analysis . . . 422

3.36 Effect of Exposure to US Technology on the Similarity and Originality of Innovation in Great Britain 423

3.37 Triple Differences Effect of Exposure to US Shocks on the Similarity Between UK and US Innovation 424

3.38 Difference-in-Differences Effect of Neighborhood-Level Out-Migration on Innovation: Alternative Proximity Threshold 425

Chapter 1

Race and Science

1.1 Introduction

Black or African Americans are underrepresented in every scientific and innovation field. Compared to white Americans, they are one-fourth as likely to be scientists (NSF, 2015), and one-eighth as likely to hold a patent (Akcigit and Goldschlag, 2023). The underrepresentation of minorities in innovation is a salient issue in US policy. In 2022, Congress passed the "Unleashing American Innovators Act" encouraging the United States Patent and Trademark Office (USPTO) to implement outreach initiatives aimed at diversifying the inventor pool. However, to date, we know surprisingly little about the economic and distributional implications of this large gap.

The potential costs of minority gaps among scientists and inventors fall into two categories. First, private costs may emerge if the benefits of patent ownership are not shared equally among the population. Second, societal costs may arise if the lack of minority inventors not only leads to unequal opportunity, but also to economic and welfare losses for society at large. Recent research suggests that reducing barriers to occupational choice, and therefore minority gaps, anywhere in the economy will yield positive implications for growth (Hsieh, Hurst, Jones and Klenow; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019; 2019).

In this paper, I provide novel evidence on the implications of the racial composition of scientists and inventors for the production of science. I document that it has implications for the direction of research and innovation, as well as its quantity and quality.

Addressing this question poses several challenges. The first is lack of data, as the demographic characteristics of inventors are not collected by patent offices, nor do inventors fall under any standard occupational category in labor survey data. The second challenge lies in identifying race separately from socio-economic status, as the two concepts are highly intertwined in the United States (Rose, 2023). The third challenge lies in studying whether innovation produced by Black or African Americans is the same as the one produced by white Americans. Previous research has found that female, wealthy, and older inventors tend to target their research towards their own demographic group. To document this pattern, these studies classify innovations by studying exclusively female diseases (Koning, Samila and Ferguson; Koning, Samila and Ferguson, 2020; 2021), consumption patterns by gender, age, and socio-economic status (Einiö, Feng and Jaravel, 2023). However, race is a social construct, not a biological one. For this reason, defining a measure of innovation specifically addressing Black or African Americans is challenging: classifying diseases based on race is not obvious, and consumption patterns may be driven by the social environment, rather than by race-specific factors.

In this paper, I measure race using information on the aggregate frequency of last names by race and ethnicity from the 2010 US Census (Comenetz, 2016). I merge this data with medical research articles from the PubMed database, ongoing clinical trials from the web portal `ClinicalTrials.gov`, research grants awarded by the National Institutes of Health (NIH), and applications and granted patents by the United States Patent and Trademark Office (USPTO). Because the focus of this paper is racial inequality in the United States and uses a proxy for race that relies on racial frequencies across US residents, I restrict the sample to researchers and Principal Investigators whose main affiliation is with a US institution and to inventors residing in the US.

I begin by documenting that having a Black-sounding name is correlated with a lower probability of being a scientist or inventor. An individual with a Black-sounding name

is five times less likely to be a scientist or inventor compared to someone with a white-sounding name. How does this gap affect the production of science?

In the first part of the paper, I document that the racial composition of inventors shapes the direction of research and innovation. I focus on medical innovation, measured through research articles and the patents linked to them, to construct two measures of *Innovation directed towards Black or African Americans*. The first measure focuses on research and innovation that explicitly addresses Black or African Americans (I label this the "Demographics-based" approach). The second measure focuses on research on diseases with relatively higher incidence among Black or African Americans compared to white Americans (I label this the "Frequency-based" approach). These two measures provide complementary ways to classify the content of scientific production. The first is more direct, but more narrow, as only 1% of all articles in the dataset mention any demographic group. The second measure is less direct, but broader: over one third of all articles in the dataset can be classified using this approach.

To construct the first measure, I focus on research explicitly addressing Black or African Americans. I document that researchers with a Black-sounding last name are more than twice as likely compared to those with White-sounding last name to produce research directed towards Black or African Americans. This is largely driven by Black-sounding researchers publishing more articles reporting the results of clinical trials that include Black or African Americans. Turning to data on *ongoing* clinical trials, a similar pattern emerges: compared to scientists with a White-sounding last name, those with a Black-sounding one are almost four times as likely to include Black or African Americans in trials. I run placebo exercises and I find that Black-sounding researchers are not more likely than white-sounding ones to conduct research addressing other demographic groups (such as Hispanic or Latinos, or Asian Americans), and they are significantly less likely to publish articles on demographic groups other than their own. This finding provides a novel explanation for the lack of inclusion of Black or African Americans in medical research and complements existing findings that point to the role of low enrolment in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022), and low research funds

dedicated to typically Black or African American diseases (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

To construct the second measure, I build an index of relative mortality of Black or African Americans compared to white Americans using administrative data from the Center for Disease Prevention (CDC). A relative mortality index equal to 1 means that a Black or African American and a white American are equally likely to die of that disease. I find that Black scientists are more likely to research and commercialize research on diseases with higher mortality among Black or African Americans. Symmetrically, white scientists are more likely to focus on diseases with higher mortality among white Americans. The magnitudes are large: compared to white researchers, Black or African American researchers focus on diseases that have a relative incidence that is 12% higher on average.

After documenting that Black scientists are more likely to conduct research directed towards Black or African Americans, I use two research designs to identify the drivers of these patterns. With the first design, I draw a causal link between race and the direction of research. This allows for disentangling other factors correlated with race such as the local environment or socio-economic status. With the second design, I exploit an exogenous shift in HIV-related mortality among Black or African Americans compared to white Americans, and I draw a causal link between race-specific mortality and the match between researchers and the content of their innovation.

With the first research design, I develop a novel approach building on findings in medicine and anthropology. I study conditions that gave an evolutionary advantage to ancestors and persist today, even if the advantage no longer holds. More specifically, I focus on genetic conditions that provided an advantage in the environment where ancestors were located. I focus on two case studies: sickle cell anemia and melanoma. Sickle cell anemia is more common among individuals of African ancestry because being a carrier of the condition protects against malaria. Melanoma is more common among white individuals because this condition is triggered by exposure to UV light, from which Black or African Americans are more shielded thanks to the darker skin color. I find that scientists are

more likely to focus on diseases more common in their own demographic group, while I find no correlation with research in similar diseases with balanced incidence among demographic groups. This association is large: across all research articles, scientists with a Black-sounding last name are 50% more likely compared to those with a White-sounding last name to research sickle cell anemia. Within specific subfield, they are over twice as likely to research this disease compared to white scientists. Assuming that the incidence of these diseases varies with ancestry but not with current socio-economic status, this approach allows to draw a causal link between race and the content of the research (not only via factors linked to socio-economic status, or the local environment).

With the second research design, I identify the impact of relative disease incidence on the direction of innovation by studying a shock to the relative mortality rate of Black or African Americans compared to white Americans. I build on evidence documenting that Human Immunodeficiency Virus (HIV) became a higher contributor to the mortality of Black or African Americans (compared to white Americans) after the introduction of the Highly Active Antiretroviral Therapy (HAART) in 1996 (Levine, Briggs, Kilbourne, King, Fry-Johnson, Baltrus, Husaini and Rust, 2007). The evidence suggests that when HIV becomes a "Relatively More Black" disease, then the relative share of scientists with a Black-sounding last name researching the disease should go up. I test this hypothesis using data on the universe of research grants awarded by the National Institutes of Health between 1988 and 2012. Grant data is especially suited for studying how researchers react to policy shocks because it more immediately reflects research activities. First, I find that HIV-related articles become five times more likely to mention Black or African Americans compared to the pre-period. I interpret this finding as a "first-stage" showing that HIV became a "relatively Black" disease. Turning to the direction of research, I find that Black researchers are 50% more likely to focus on HIV compared to whites, while there was no such pattern in the period before the introduction of HAART. A similar pattern holds when looking at research publications.

In the second part of the paper, I study how the racial composition of scientists and inventors affects the quantity and quality of scientific production. The large racial gaps suggest

that there may be many "missing" scientists and inventors, thus affecting the quantity of the overall production of science and innovation. (Hsieh, Hurst, Jones and Klenow; Bell, Chetty, Jaravel, Petkova and Van Reenen; Bloom, Van Reenen and Williams, 2019; 2019; 2019). Looking within the patenting process, I find evidence that racial gaps may also affect the quality of science and innovation. I find that patent applications from Black inventors are 6% less likely to be granted than those from white inventors, both unconditionally and after controlling for detailed technology class fixed effects. Approximately half of this gap is explained by differential access to resources. After controlling for assignee fixed effects, state of residence fixed effects, and proxies of lawyer quality, the residual gap is equal to 3%. Looking at granted patents, the gap is reversed. Patents from Black-sounding inventors have a higher impact (as measured by forward citations) compared to white-sounding ones. These findings suggest that minority gaps among scientists and inventors may result not only in lower quantity, but also in lower quality of innovation.

In the final part of the paper, I quantify how closing racial gaps would affect the quantity and quality of innovation through the lens of the general equilibrium Roy model developed by Hsieh, Hurst, Jones and Klenow (2019). Closing minority gaps may have general equilibrium effects for at least two reasons. First, if minorities anticipate discrimination, they may choose different careers. Second, as more individuals enter an occupation, the labor demand for that occupation will change, causing equilibrium wages to decrease. In the model, individuals are either white or Black and choose among a set of occupations based on heterogeneous preferences, heterogeneous talent, and group-specific preferences. The economy is composed of the home sector, standard market occupations, and the "science and innovation" sector. Black individuals face a "minority tax" in the acquisition of human capital and another tax in the labor market. I estimate the model using moments from the Current Population Survey (CPS), and externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019) to quantify the magnitude of this loss. I find that equalizing access to science and innovation will lead to an overall increase of 1 p.p., corresponding to 10% over baseline, in the number of scientists and inventors. This increase is driven by a higher number of Black inventors and scientists selecting into these

occupations. The share of white inventors and scientists experiences a decrease, albeit quantitatively small (less than 1% over baseline).

Related Literature This paper is closely related to three strands of literature. First, it contributes to the literature showing that inventor demographics matter for the direction of innovation. Einiö, Feng and Jaravel (2023) document a match between gender, socio-economic status and age of inventors and entrepreneurs and their innovation. Koning, Samila and Ferguson; Koning, Samila and Ferguson (2020; 2021) document that female researchers are more likely to carry out research on female diseases. A growing body of literature shows that interaction with diverse students peers (Truffa and Wong, 2022), representation in clinical trials (Michelman and Msall, 2023), access to data (Nagaraj, Shears and de Vaan, 2020), and geography (Moscona and Sastry; Fry, 2022; 2023) increase the representation in the focus of scientific research. Additionally, a growing body of evidence shows the impact of the location of inventors on the direction of their research. For example, Fry (2023) shows that the outbreak of a health crisis sparks research from local researchers on those diseases. I contribute to this literature by documenting a link between the race of an inventor (scientist) and the content of their innovation (research).

Second, this paper contributes to the literature on the costs of talent misallocation. Hsieh, Hurst, Jones and Klenow (2019) show that lower barriers in accessing the labor market for minorities explain 40% of GDP growth between 1960 and 2010. Chetty, Dossi, Smith, Van Reenen, Zidar and Zwick (2023) uses a similar model to quantify the welfare gains from equalizing the access to entrepreneurship for women in the United States in the last two decades. Ashraf, Bandiera, Minni and Quintas-Martınez (2023) find that the misallocation of female talent across countries harms the productivity of firms. I contribute to this literature by showing that inequality in the direction of innovation and science is an additional consequence of misallocation.

Third, this paper contributes to the literature on the underprovision of medical research and innovation benefiting Black or African Americans. A set of studies has shown that low

representation in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams; Alsan, Campbell, Leister and Ojo, 2022; 2023) and low funding (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020) are drivers of this underprovision. I add to this literature by showing that the racial composition of scientists and inventors contributes to racial gap in medical research and innovation.

The rest of the paper is organized as follows. In section 1.2, I describe the data. In section 1.3, I study the link between the racial composition of scientists and inventors and the direction of research and innovation. In section 1.4, I study the mechanisms driving the match between researchers and the content of their research and innovation. In section 1.5, I study the link between the racial composition of scientists and inventors and the quantity and quality of research and innovation. In section 1.6, I estimate the labor market effects of equalizing access to innovation and science using a Roy model of occupational choice. In section 1.7, I discuss these findings. In section 1.8, I conclude.

1.2 Data

In this section, I describe the data sources and the construction of the variables used in the empirical analysis. In section 1.2.1, I discuss data on scientific publications and research scientists. In section 1.2.2, I discuss data on patents and inventors. In section 1.2.3, I discuss how I proxy the race of scientists and inventors. In section 1.2.4, I discuss the matching procedure of scientists and inventors with the measure of Black-sounding name. In section 1.2.5, I discuss how I validate the measure of Black-sounding name as proxy of ethnicity. Sample statistics are reported in Table 1.1.

1.2.1 Measuring Scientific Production

To measure scientific production, I rely on three data sources. The first is research articles from the PubMed database. PubMed is a free search engine accessing the MEDLINE database of articles on life sciences and biomedical topics. It includes bibliographic

information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. It also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution. For each article, this dataset provides information on title, abstract, MeSH terms, journal, year of publication, publication type, institutional affiliation of first author. I complement this data with information on citations from the iCite database of the National Institutes of Health (NIH) (Hutchins, Yuan, Anderson and Santangelo, 2016). I link this data to the patents that cite these papers (Marx and Fuegi; Marx and Fuegi, 2020; 2022). Following Koning, Samila and Ferguson (2021), I restrict the main analysis to research articles in the top 1,000 journals in terms of their commercialization impact factor (JCIF) (Marx and Fuegi; Marx and Fuegi, 2020; 2022). For all datasets, I restrict the sample to years 2002 to 2018. Additionally, I restrict the data to researchers affiliated with an institution located in the United States.[1] Second, I use data on ongoing clinical trials from `ClinicalTrials.gov`, a registry run by the United States National Library of Medicine (NLM) at the National Institutes of Health. The third dataset is the universe of research grants awarded by the National Institutes of Health (NIH) between 1985 and 2012. Data on the universe of National Institutes of Health grants awarded between 1985 and 2012 comes from the NIH database.[2]

1.2.2 Measuring Patenting Activity

To measure patenting activity, I use data on the universe of patent applications filed in the United States between 2001 and 2018. The data were extracted from the Patent Examination Research dataset (PatEx) merged with the PatentsView dataset. The first dataset (PatEx) is maintained by the USPTO and contains detailed information on all patent applications filed since January 2001 (Graham, Marco and Miller, 2015). The second (PatentsView) is also maintained by the USPTO, but focuses on data linking

[1]Throughout, I restrict the sample to the first author of the article, grant, or clinical trial. For this reason, I impose that affiliation of the *first* author must be with a U.S. institution.

[2]The NIH is the largest funder of medical research in the United States after the U.S. government. The National Science Foundation (NSF) documented in a 2015 report that, of the 86 billion spent on basic research in 2015, 44% came from federal agencies, 28% came from the National Institutes of Health, 29% came from pharmaceutical companies, 19% came from biotechnology companies.

inventors, their organizations, and locations. To construct the final sample, I include only applications filed between January 2001 and December 2018. Patents before 2001 are excluded because data on applications is not available before then.[3] I stop at December 2018 to allow enough time for grant decisions to be made, and I further exclude cases on which decisions have not been made, therefore excluding provisional, reissue, and Patent Cooperation Treaty applications. In this paper, I focus on the United States, and for this reason I restrict my sample to inventors who are resident in the US at time of filing.[4]

1.2.3 Measuring Race and Ethnicity

I proxy the race and ethnicity of scientists and inventors using their last names. Inferring unobserved characteristics through names is a procedure widely used in research on the ethnicity of inventors (Kerr; Gaule and Piacentini, 2008; 2013) and in the economic history literature.[5] In this paper, I rely on last names, rather than first names, for two reasons. First, because they are more stable over time for the same individual, as well as more stable across cohorts. This is a crucial aspect as the age distribution of inventors is different from the one of the wider population. Second, for data availability: thanks to a recent initiative of the U.S. Census Bureau, we now have available data on the population frequency of each last name in the United States. For first names, much smaller and less representative datasets are instead available.[6] This approach follows best practices on how to infer ethnicity for the U.S. population (Kozlowski, Murray, Bell, Hulsey, Larivière, Monroe-White and Sugimoto, 2022).

I use aggregated data published by the Census Bureau (Comenetz, 2016) on the frequen-

[3]The American Inventors Protection Act passed in 1999 made it mandatory to publish all non-provisional patent applications after December 2000.

[4]Information on the place of residence is listed on the patent, along with the name and last name of all inventors, the name of patent attorneys, and the name of the patent examiner who was assigned the patent application.

[5]Names have been used, among others, to measure race and ethnicity (Abramitzky, Boustan and Connor; Fouka, 2020; 2020), individualism (Bazzi, Fiszbein and Gebresilasse, 2020), socioeconomic background (Olivetti, Paserman, Salisbury and Weber, 2020), and religiosity (Berkes, Coluccia, Dossi and Squicciarini, 2023).

[6]To the best of my knowledge, the most comprehensive data on the racial and ethnic distribution of first names in the U.S. comes from Tzimioukis (2015) and is constructed on a sample of 8 million mortgage applicants.

cies of last names by race from the 2010 U.S. population Census. These data report the probability that a given last name in the U.S. population is associated with a given race. It is provided for the entire population excluding those individuals whose last name appears in fewer than 100 records. This is the Bayes Optimal Classifier, that is the solution that minimizes the chance of misclassification of the race variable in the population. If you picked a random individual with last name *Washington* from the U.S. in 2010 and asked someone to guess this person's race (without additional information), the best guess would be based on what is available from the aggregated Census file.

In the Census dataset, each last name is linked to six racial probabilities: white, Black, Hispanic, Asian or Pacific Islander, Native American, Two or more races. Empirically, these probabilities do not always add up to 1 as the U.S. Census Bureau censors all cells with fewer than 100 last names. For each last name, I re-standardize probabilities such that they add up to 1 and I split the "Two or more races" category equally across the other racial probabilities.

In the main analysis, I assign race or ethnicity based on the first researcher listed on the paper, or on the first inventor listed on the patent. In robustness checks (reported in the Appendix), I verify that results hold using alternative measures built leveraging data on all members of the team, or of both the first and the last listed authors.

1.2.4 Matching Procedure

To build the final dataset, I match data on scientists described in section 1.2.1 and on inventors described in section 1.2.2 with the distribution of last names by race described in section 1.2.3. To avoid false positives, I adopt a conservative approach in conducting the merge, keeping only *exact* matches. With this approach, I merge 78% of all research articles and 82% of all patents. The match is performed based on the last name of the first author. Sample statistics for the final sample are detailed in Table 1.1.

In Appendix Table 1.12 and 1.13, I report a test of balancedness for the match. Matched

and unmatched observations are remarkably balanced across all demographics except geography: matched observations (both for scientists and for inventors) are significantly less likely to be located in California compared to the matched sample (25% compared to 31%). This is consistent with the hypothesis that California may have relatively more foreign-born inventors, whose last name may not be recorded among Census records.

In Figure 1.1, I plot the distribution of the US population over Pr(Black | Last name). Over 55% of individuals have a last name with a probability between 0 and 0.1 of being held by a Black or African American individual. In contrast, when examining the distribution of scientists and inventors (shown in grey), this fraction increases to 70%. This large difference is more clearly illustrated in Figure 1.2. In this plot, I divide the population into quintiles based on Pr(Black | Last name) and normalize them by the first quintile. If scientists and inventors were represented proportionally to the overall population, all dots would lie on the red line. However, this figure tells a different story. Individuals in the top quintile of Pr(Black | Last name) are over five times less likely to be scientists or inventors.

1.2.5 Validation of Black-sounding Last Names as a Proxy for Race

To validate this data, I use a dataset of voter registration data matched to race of individuals who are registered to vote in the state of Florida from Dossi and Morando (2023). In Appendix Table 1.11, I report the individual-level correlation of Black-sounding name with a dummy = 1 if the individuals self-reported being Black or African American in the Voter Registration data. In colum (2), I show that the correlation is similar (and the R-squared as well) when including the vector of other racial probabilities.[7] In columns (3), I include a control for median income in zip code of residence. While the correlation decreases, it remains statistically significant and high. This result suggests that the vector of racial probabilities correlated with race even when conditioning for an individual-level

[7]White-sounding name is the omitted category.

measure of socio-economic status. In columns (4) and (5) I control for average median income by last name and its standard deviation. The association between Black-sounding name and self-reported race remains positive, significant, and close to 1.In Appendix Figure 1.8, I use the full Voter Florida data which consists of 9 million individuals to show that Pr(Black | Last name) in Florida has a 1:1 correlation to the one in the United States. I obtain a similar picture when building a last name-level vector of racial probabilities and correlate the two. This exercises lends confidence to the validity of Pr(Black | Last name) as a predictor of race even in subsamples of the United States (i.e., a single state such as Florida). Second, I test this correlation on the sample of scientists and inventors. I construct a vector of racial probabilities by last name in the subsample of, respectively, scientists and inventors. In Appendix Figure 1.9, I show the correlation of Pr(Black | Last name, scientist) and Pr(Black | Last name, inventor) with Pr(Black | Last name). The relationship is liner. However, the slope is lower than 1: this is consistent with the fact that the overall share of Black or African Americans is lower in these subsamples.

1.3 Race and the Direction of Science and Innovation

In this section, I document that the race of scientists correlates with the content of their research. In section 1.3.1, I summarize the evidence on research addressing Black or African Americans from the existing literature. In section 1.3.2, I describe how I construct the indicators of innovation directed towards Black or African Americans. In section 1.3.3, I show results using the demographics-based approach, and in section 1.3.4, I show results using the frequency-based approach.

1.3.1 The Black-white Gap in Medical Research and Innovation

There is a large gap in health outcomes and life expectancy between Black or African Americans and white Americans.[8] Social determinants of health, such as education, in-

[8]In 2018, the gap in life expectancy between Black or African Americans and white Americans was 3.6 years (Schwandt, Currie, Bär, Banks, Bertoli, Bütikofer, Cattan, Chao, Costa, González et al., 2021).

come, and access to the healthcare system, are documented drivers of these disparities (Cutler, Deaton and Lleras-Muney; Cutler, Lleras-Muney and Vogl; Chetty, Stepner, Abraham, Lin, Scuderi, Turner, Bergeron and Cutler; Schwandt, Currie, Bär, Banks, Bertoli, Bütikofer, Cattan, Chao, Costa, González et al., 2006; 2008; 2016; 2021). In addition to these factors, a growing body of evidence documents that research and innovations in healthcare may create or perpetuate these inequalities.

Do research and medical advances benefit Black or African Americans less often compared to white Americans? Answering this question is complicated because it is hard to determine what the optimal rate of innovation should be, and who in the population it should target (Bryan and Williams, 2021). In this section, I summarize some of the existing evidence suggesting that medical research and innovation benefits Black or African Americans less often compared to other groups in the population. As benchmark rates, I rely on the proportionality approach often used in public health, according to which it is optimal to target research efforts in a way proportional to the size of the population affected (unconditional of the demographic group).[9][10]

First, a large body of evidence documents that Black or African Americans are underrepresented in clinical trials compared to population size. The lack of diversity can significantly affect public health as the results of these studies may not accurately represent the broader population. This can result in drugs and treatments that are less effective or even harmful to certain groups of people. One example includes the use of certain cardiovascular medications, which have been found to produce clinical differences between white individuals of European ancestry and individuals of African ancestry (Johnson, 2008).[11]

Second, Black or African Americans are underrepresented in observational studies, both in

This difference has widened after the COVID-19 pandemic.

[9]Importantly, this approach may lead in itself to health inequality due to the smaller share of Black or African Americans compared to white Americans (Cutler, Meara and Richards-Shubik, 2012).

[10]For example, the NIH allocates research funds in accordance with disease burden (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

[11]Several clinical trials have shown varying responses to the blood pressure-lowering effects of beta-blockers and ACE inhibitors.

research articles and in medical textbooks. This has real implications for health outcomes. For example, Louie and Wilkes (2018) finds that no textbook has images of six common skin cancers in skin of color. In turn, a recent study among medical students found a very high rate of underdiagnosis of skin conditions among Black or African Americans (Fenton, Elliott, Shahbandi, Ezenwa, Morris, McLawhorn, Jackson, Allen and Murina, 2020), plausibly due to lack of training and resources to learn how to diagnose these conditions.[12]

Third, diseases with higher incidence among Black or African Americans compared to whites are under-researched compared to similar diseases. Making this statement is complicated by the lack of systematic data on the incidence of diseases. However, suggesting (yet compelling) evidence on this comes from analyzing a genetic disease more frequent among Black or African Americans: sickle cell anemia. The number of publications on sickle cell anemia proportional to the affected population is almost five times smaller compared to cystic fibrosis, a genetic disease comparable to Sickle Cell Anemia but more frequent among whites (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

Fourth, recent evidence shows that artificial intelligence and healthcare algorithms are biased against Black patients (Obermeyer, Powers, Vogeli and Mullainathan, 2019). This is both a *consequence* of lower research output on Black patients, hence a smaller training set compared to white ones and a *cause* of health inequality. However, it represents an additional way through which innovation in healthcare leads to inequality.

1.3.2 Measuring the Direction of Science and Innovation

In the rest of this section, I test the hypothesis that Black scientists and inventors located in the United States are more likely to produce medical research and innovation that benefits Black or African Americans.

[12]Another example is the utilization of the pulse oximeter. This is a medical device that measures blood oxygen saturation by passing light through the skin, typically on a fingertip. This device overestimates oxygen saturation in individuals with darker skin tone, leading to under-diagnosis (Sjoding, Dickson, Iwashyna, Gay and Valley, 2020).

The first challenge in studying the correlation between race and the direction of innovation is defining which innovations benefit Black or African Americans. I propose two definitions of "Innovation directed towards Black or African Americans". The first is based on whether research explicitly addresses Black or African Americans as a demographic group. The second is based on the relative incidence of disease among Black or African Americans compared to white Americans. I focus on research articles and the patents linked to them as a measure of commercialization of the ideas introduced in these articles (Marx and Fuegi; Marx and Fuegi, 2020; 2022).

Linking race to the content of medical research is not straightforward because race is a social construct, rather than a biological one. However, it is widely documented in the medical literature that certain conditions have differential incidence conditional on race. This is driven both by environmental factors and by small genetic variations in our genetic heritage that link back to ancestors. To define diseases common among the Black or African American population, I propose two different strategies: a demographics-based approach, and a frequency-based approach. The first, which I label "demographics-based", defines research as directed to Black or African Americans if the research explicitly mentions Black or African Americans. The second, which I label "frequency-based", defines a disease as typically Black if its incidence is relatively higher among Black or African Americans than it is among white Americans. Symmetrically, I define typically-white diseases.

With the first definition, I define an indicator of innovation directed towards Black or African Americans coding a binary variable that takes value 1 if a research article is associated with the MeSH code corresponding to *Black or African American*, and takes value 0 otherwise. MeSH codes may vary over time, and are not updated retroactively. This should be a smaller concern in this context as I look at a relatively short time span. However, to address this I map articles to diseases using a dictionary approach based on the abstract of each publication. The results, shown in the Appendix, are qualitatively unchanged and quantitatively similar using the dictionary approach.[13]

[13]In Appendix Figure 1.11, I report an example of an article with Black or African American among its MeSH codes, but which is not classified as 1 by the dictionary classification. In Appendix Figure 1.12,

This approach captures a direct link between the race of the researchers and the subject of their research. However, its main disadvantage is that it is fairly narrow: only approximately 1% of all articles are explicitly linked to a demographic group.

With the second definition, I define a complementary measure that is less direct but more comprehensive. I use data on death rates by race from administrative records from the Center for Disease Control and Prevention (CDC). Because comprehensive data on the incidence of disease by race is not available (nor, for most diseases, systematically kept track of), I proxy disease incidence with disease-related mortality.[14] I use mortality rates between 1999 and 2015 as a proxy for disease incidence. I build a measure of *relative mortality* for each disease d reported in the official statistics as the cause of deaths at least 5,000 times (sum of non-Hispanic white and non-Hispanic Black or African Americans) over this period.[15] For each disease d, the measure is built as follows:

$$\text{Relative mortality}_d = \frac{(\text{N. deaths Black or African Americans due to } d)/(\text{N. Black or African Americans})}{(\text{N. deaths white Americans due to } d)/(\text{N. white Americans})} \tag{1.1}$$

This measure captures excess mortality of Black or African Americans compared to white Americans for a given disease d, and is similar to the one employed by Cutler, Meara and Richards-Shubik (2012) to study inequality in infant medical care. I classify diseases into three groups: *Typically White* if relative mortality is at least 1.5 times higher among non-Hispanic white Americans, *Typically Black* if relative mortality is at least 1.5 times higher among non-Hispanic Black or African Americans, *Similar Incidence* is defined as

I report an example of an article which is classified as 1 by the dictionary classification, but does not have Black or African American among its MeSH codes. The two measures overlap for approximately 65% of observations.

[14] I use data on Underlying Cause-of-Death by race and ethnicity. "Underlying Cause-of-Death" is defined by the World Health Organization as "the disease or injury which initiated the train of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury." It is selected from the conditions entered by the physician on the death certificate. When more than one cause or condition is listed, the underlying cause is determined by the sequence of conditions on the certificate.

[15] The relative ranking of diseases is fairly stable over time (and it is essentially identical for the period 1999 to 2001). While, magnitudes vary slightly over time, the results are robust to choosing different time spans. Mortality rates start in 1999 as this is the year when the CDC started using the International Classification of Diseases, Tenth Revision (ICD-10) to report causes of death.

the residual category. I map articles to diseases using the set of associated Medical Subject Headings (MeSH), as standard in the literature.[16] Similarly to the Demographics-based approach, I consider an article as addressing a given disease if the MeSH code linked to that disease is listed as one of the MeSH codes associated with that publication. One third of all publications are matched to at least one ICD-10 Cause of Death. Whenever more than one MeSH code is linked to a cause of death, I assign to that article an average of the relative mortality of all diseases linked to the article. Approximately 10% of all articles of the sample are linked to more than one ICD-10 Causes of Deaths.

1.3.3 Demographics-based Approach: Results

The first approach to measuring the direction of innovation is to define an indicator of innovation directed towards Black or African Americans coding a binary variable that takes value 1 if a research article is associated with the MeSH code corresponding to Black or African American, and takes value 0 otherwise. This variable captures a direct link between the race of the researchers and the subject of their research.

I estimate a model where $y = 1$ if an article addresses a given disease, 0 otherwise:

$$y_{ip} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta X_i + \Gamma Z_p + \varepsilon_{ip} \tag{1.2}$$

Where i is the first author, p denotes the article, α_t are year of publication fixed effects, X_i is a vector including Hispanic-sounding name, Asian-sounding name, American Native-sounding name, with omitted category white-sounding name. Standard errors are clustered at last the name level. β is the propensity of Black-sounding scientists to research y compared to white-sounding ones.[17]

[16] The MeSH thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine (NLM). MeSH terms are assigned manually by indexers of the NLM.

[17] β will likely underestimate the Black-white gap because, as shown in Figure 1.8, Individuals with Black-sounding name are underrepresented among scientists. This can be visualized using Bayes rule: $P(B \mid S, L) = P(I \mid B, L) \times P(B \mid L) \,/\, P(S \mid L)$, where B is a Black individual, S is a scientist, L represents a last name. For this reason, whenever I refer to "Black-sounding name", I refer to the probability that that last name is held by a Black person in the US population.

In Table 1.2, I report the results of the estimation of equation (1.2) on innovation directed towards Black or African Americans defined according to the demographics-based approach. The results in column (1) show that Black-sounding scientists are more than twice as likely compared to white-sounding ones to produce research that addresses African Americans.Next, I test whether this association is driven by observational studies, or by clinical trials. Disentangling the two is crucial because of mounting evidence that low participation of African Americans in clinical trials affects the "production" of clinical trials that include African Americans (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022). In column (2), I find that Black-sounding researchers are 2.5 times more likely to mention African Americans in papers that report the results of clinical trials. In column (3), we see that a correlation similar to the one reported in column (1) holds for all other (observational) articles. While the result in column (2) could be the result of larger participation of Black or African Americans in clinical trials when the researcher is of their own race, consistently with findings by Alsan, Campbell, Leister and Ojo (2023), the association reported in column (3) shows that the willingness to participate in trials seems not to be the only driver of the match between the race and ethnicity of scientists and the direction of their research.[18]

In columns (4) to (6), I re-estimate the results shown in columns (1) to (3) on the subsample of articles which focus on human subjects.[19] Results are similar to those in first three columns, and suggest that the positive and significant coefficients of Black-sounding last name are not simply driven by a higher propensity of Black-sounding researchers to conduct research on human subjects. To further test whether there is an association between researchers and direction of research based on their race or ethnicity, I re-run Table 1.2 on a dummy variable equal to one if the article has Hispanic or Latino among its MeSH terms, and equal to zero otherwise; and on a dummy variable equal to one if the article

[18]In Figure 1.13, I report coefficients across a more disaggregated set of publication types. "Publication type" is provided by PubMed for each research article. An article can be assigned to multiple publication types, but this is fairly rare. The split I report in this figure is virtually unchanged when excluding articles assigned to more than one publication type.

[19]More specifically, I keep only those articles whose assigned MeSH terms all fall under the Human category out of this article's MeSH terms that fall into the Human, Animal, or Molecular/Cellular Biology categories defined by Hutchins, Davis, Meseroll and Santangelo (2019).

has Asian American among its MeSH terms, equal to zero otherwise.[20] The coefficients in column (1) are reported in Figure 1.4, while the full estimation results are reported in Appendix Table 1.23. These results reveal that scientists are disproportionately more likely to include individuals of their own demographic group in their research.[21]

In Table 1.24, I look at the sample of ongoing clinical trials registered on `ClinicalTrials.gov`. This sample consists of 12,366 trials registered in the portal between 2002 and 2018 and with Principal Investigator affiliated with a US institution. I define a set of indicators for Black or African American, Hispanic or Latino, and Asian using a dictionary approach on the project description.[22] To assign race to Trial Investigators, I take the average of each racial probability across all registered Principal Investigators (as researchers are not listed in a meaningful way, e.g. by contribution, like in the case of articles or patents). In column (1), I show the results of estimating equation (1.2) on the indicator for Black or African American. The association is positive and highly statistically significant: compared to a white-sounding one, a Black-sounding team of researchers is four times more likely to run trials with Black or African Americans. In columns (2) and (3), I test whether Black-sounding teams are more likely to research diverse populations, or whether this match between scientist and content of their innovation is linked to race. I find that these teams are not significantly more likely to include Hispanic or Latinos, or Asians, in their trial description. However, Hispanic-sounding teams are more likely to include Hispanic or Latinos in their project descriptions, and Asian-sounding teams are more likely to include Asians. This finding is line with the results on articles using the demographics-based approach (Figure 1.4) and on white-sounding inventors found using the frequency-based approach: the match between race or ethnicity of researchers and the direction of their research is not specific to a given group, but rather an empirical

[20] A dummy variable for research addressing white Americans cannot be built because the corresponding MeSH term is almost never used (similarly, white ethnicity is rarely explicitly mentioned in the abstract). This reflects the fact that including white Americans in trials and observational studies in the norm.

[21] It is worth noting that in this plot I do not exclude article addressing multiple demographic groups, so the variables on the LHS are not mutually exclusive. Results are even stronger when those observations are excluded.

[22] I assume that a project description that mentions a given demographic group is designed to include that group. I implement a dictionary-based classification as MeSH terms are not available in this data.

regularity across demographic groups.[23]

This finding provides a novel explanation for the lack of inclusion of Black or African Americans in medical research and complements existing findings that point to the role of low enrolment in clinical trials (Alsan, Durvasula, Gupta, Schwartzstein and Williams, 2022), and lower funding allocated to typically Black diseases (Farooq, Mogayzel, Lanzkron, Haywood and Strouse, 2020).

1.3.4 Frequency-based Approach: Results

With the second approach, I define "Innovation benefiting Black or African Americans" as those research articles addressing diseases that have higher mortality among Black or African Americans compared to white Americans. In Table 1.3 I report the result of the estimation of equation (1.2) on relative mortality. Throughout, I control for the logarithm of total deaths among Black or African Americans and White Americans due to disease d, over the period 1999 to 2015. I estimate the following model:

$$y_{ipd} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta Z_i + \gamma \log(\text{Total n. deaths})_d + \epsilon_{ip} \tag{1.3}$$

In column (1), I estimate equation (1.3) on a dependent variable equal to the logarithm of the relative mortality of disease d among Black or African Americans compared to white Americans. In column (2), I estimate the same equation on a dependent variable that equals 1 if a disease is at least 1.5 more frequent among Black or African Americans, equals 0 otherwise. In column (3), I estimate the same equation on a dependent variable that equals 1 for diseases that have roughly balanced incidence among Black or African Americans and whites, that is those with relative mortality less than 1.5 among Black or African Americans compared to white Americans, and less than 1.5 among white Americans compared to Black or African Americans, and equals 0 otherwise. In column

[23]It is worth noting that, while the magnitudes of the match between researcher and content of research reported in this table vary by demographic group, it is hard to infer real magnitudes of the match without keeping disease frequency and incidence constant.

(4), I estimate the same equation on a dependent variable that equals 1 if a disease is at least 1.5 more frequent among white Americans, equals 0 otherwise.

Compared to white-sounding ones, Black-sounding researchers more likely to research diseases more frequent among Black or African Americans, while white-sounding researchers are more likely to research diseases more frequent among white Americans. The coefficients shown in column (1) show that, compared to white-sounding researchers, Black-sounding researchers focus on diseases that have a relative mortality that is 12% higher on average among Black or African Americans. The results in column (2), looking at relatively more white diseases, reveal that white-sounding researchers are 15% more likely to research *typically white* diseases, such as Alzheimer's diseases or melanoma. The results in column (3) show that there is no significant difference for diseases with balanced incidence among whites and Black or African Americans. Finally, column (4) reveals that Black-sounding researchers are 20% more likely to research *typically Black* diseases. Results are qualitatively and quantitatively similar when adopting different thresholds.[24] In Appendix Table 1.32, I run specifications that mimic the ones in columns (1) to (4) of Table 1.3, but on the sample of articles linked to a patent (Marx and Fuegi; Marx and Fuegi, 2020; 2022).

These findings are in line with research by Koning, Samila and Ferguson (2021) showing that women have higher probability to engage in research and innovation benefiting women and by Einiö, Feng and Jaravel (2023) showing that inventors are more likely to create new product benefiting individuals similar by gender, age, and socio-economic status.

1.3.5 Robustness Checks

I conduct a large set of robustness checks to gauge the validity of these findings.

[24]In Appendix Table 1.33 and Table 1.34, I report a version of this Table where I define the variables in column (2) through (5) setting thresholds at 1.3 and 1.7.

Demographics-based Approach: Robustness

In Table 1.14, I define the dummy "Research directed toward Black or African Americans" with a dictionary classification based on the publication abstract. As the main outcome is binary, in Table 1.15, I re-estimate Table 1.2 using a logit model to make sure results are not driven by the choice of a linear probability model. Results hold throughout.

In Table 1.16, I control for the gender of the first author using data on gender inferred from first names from Koning, Samila and Ferguson (2021). In Table 1.17, I include, respectively, journal FE, US state where the institution of affiliation of the first author is located FE, and first author affiliation FE. Results hold similarly to those shown in Table 1.2.

In Table 1.18, I re-estimate equation 1.2 using different definitions of Black-sounding name (and symmetrically defined definitions for the vector of other racial probabilities). In Panel A, I include the vector of racial probabilities of the last author of the paper. Panel B, I construct an average of Pr(Black | Last name) across all researchers listed in the publication.[25] In Panel C, I recode Pr(Black | Last name) as binary variable taking value 1 if it is $\geq .5$, $= 0$ otherwise.[26] In Panel D, I restrict thee set of last names to those last names present in the 1930 US Census among US-born individuals, and with a frequency of at least 100. This exercise aims to test if results hold in the subset of US-born scientists. Results are robust to using different definitions of Black-sounding last name.[27]

In Table 1.19, I re-estimate Table 1.2 using data on all published articles by a first author affiliated with a US institution, regardless of the Commercialization Impact Factor of

[25]Both in Panel A and in Panel B, I drop single-authored publications. For Panel A, the average is computed across all listed authors whose last name is successfully matched to a last name in the matrix of racial probabilities.

[26]I do this symmetrically for all racial probabilities, and construct an additional dummy "multiple race" taking value 1 if none of the other racial probabilities is above .5, = 0 otherwise. Only 2% of last names fall into this category.

[27]In Panel C, Column (2) and Column (5), coefficients are not significant due to lack of power: by using variation only in the set of last names with Pr(Black-sounding name $\geq .5$), variation in the RHS variable comes from very few observations in this smaller subsample of the data.

the journal. Results are consistent and larger in magnitude. Interestingly, the average Pr(Black | Last name) is higher (.072 compared to .064). In Table 1.20, I use the same sample used in the main Table and I further split it by Journal Commercialization Impact Factor (JCIF). In Panel A, I show results for articles published in journals with below-median JCIF. In panel B, in articles published in journals with above-median JCIF. The pattern is similar to the one observed in the previous table. While estimated coefficients are smaller, they are of similar magnitudes across the two tables if they are rescaled by the mean of the dependent variable, which is lower in the above-median sample. Finally, in Table 1.21 I show that similar patterns hold in the articles linked to a patent, providing an even more applied measure of research. The same patterns hold in this subsample.[28]

In Table 1.22, I re-estimate columns (1) to (3) of Table 1.2 on the sample of articles and MeSH terms linked to them. In columns (1), (3), and (5), I replicate the results in the main Table. In columns (2), (4), and (6), I control for MeSH terms FE, and the coefficients are essentially unchanged. This test suggests that even keeping the same medical conditions constant, scientists with a Black-sounding last name are more likely to include Black or African Americans in their research.

Frequency-based Approach: Robustness

In Tables 1.25 to 1.32, I implement a set of robustness checks symmetric to those shown in Tables 1.14 to 1.21 for the Demographics-based approach. Results hold throughout, and are in line with the patterns observed for the demographics-based approach.

Additionally, in Table 1.33 and Tables 1.34 I show that results are robust to choosing different thresholds to compute the variables *Typically White* and *Typically Black*. Finally, in Table 1.35, I show that results are robust to using the absolute number of deaths, rather than the relative one as in the main analysis.

[28]In Column (2) and (5), the coefficients are not significant due to the very small sample size.

1.3.6 Impact

What is the scientific impact of research conducted by scientists and benefiting their own demographic group? To study this question, in Table 1.36 and Table 1.38 I test whether these publications are cited by other publications in a differential way compared to other publications. The results reveal no statistically significant evidence suggesting this is the case. This finding is robust to using the Relative Citations Ratio, a metric that normalizes the total number of citations by the citations received by publications in the same area of research and year (Table 1.37 and Table 1.39).

Finally, I split the sample by below- and above- median Journal Impact Factor (JIF), and I re-estimate Table 1.2 and Table 1.3. The results, shown in Table 1.40 and 1.41, reveal that scientists with a Black-sounding last name are equally likely to conduct research benefiting their own demographic group across Journal Impact Factors.

1.4 Mechanisms

In this section, I study the mechanisms behind the match between the race of researchers and the direction of research and innovation documented in the previous section.

1.4.1 Race or Socio-economic Status? Ancestry Variation

One residual question from the findings shown so far is whether the match between the race of researchers and the direction of innovation is linked to race, or to other factors that correlate with race such as environmental factors of socio-economic status. The ranking of diseases I introduced with the Frequency-based approach correlates with relative incidence by race, but may also mimic their relative incidence by socio-economic status.

As a result, one argument could be that investment in preventive behaviors, as well as

higher engagement with the health system, would act as a substitute for that research. To establish a link between the race of researchers and the content of innovation, I develop an approach that relies on variations in probability of having a certain disease due to different ancestral origins. Drawing on findings in medicine and anthropology, I focus on articles that study diseases that are more common among Black or African Americans (white) for reasons linked to their ancestry, but not to environmental factors or to socioeconomic status. More specifically, I focus on a disease more frequent among Black or African Americans, sickle cell anemia, and a disease more frequent among white Americans, melanoma. The incidence of these diseases is linked to conditions that provided an evolutionary advantage to ancestors in their environment of origin and, through genetic inheritance, carry over to today. To disentangle race from socio-economic status through the ancestry-based approach, two assumptions must hold true.

The first assumption is that human mobility happens at a faster rate compared to human evolution. That is, even when humans migrate, their adaptation to the new environment will be slow. The second assumption is that the incidence of the disease is not correlated with current socio-economic status (nor socio-economic status at birth), or with other factors linked to the individual's current environment (or environment at birth). The first assumption is likely to hold in this setting as evolution is a slow process, and much slower than human migrations to the United States. The second assumption is likely to hold, as detection of this disease is nearly perfect in the US because sickle cell anemia is covered by Universal Newborn Screening. In addition, I provide falsification tests for each disease.

Sickle cell anemia is a multisystem disorder and the most common genetic disease in the United States, affecting 1 in 500 Black or African Americans. It is caused by a mutation in the hemoglobin beta chain in which glutamic acid is substituted with valine at position six on chromosome 11. Sickle cell disease is more common in individuals with African ancestry because carriers of the gene are protected from severe forms of malaria.

In columns (1) and (2) of Table 1.4, I report the results of the estimation of equation (1.2) where y is an indicator for whether a research article lists sickle cell anemia among

its MeSH codes. As a first falsification test, I study research on β-thalassemia, the other major hereditary hemoglobinopathy. Both diseases are recessive autosomal disorders. Differently from sickle cell anemia, the incidence of β thalassemia is not only high among those of African ancestry, but also among other populations. Current studies have shown that there is no strong evidence in support of the fact that the carriers are protected against malaria (Introini, Marin-Menendez, Nettesheim, Lin, Kariuki, Smith, Jean, Brewin, Rees, Cicuta et al., 2022).

In columns (3) and (4) of Table 1.4, I report the results of the estimation of equation (1.2) where y is an indicator variable taking value 1 if the article addresses thalassemia, 0 otherwise. The estimation results show that, across all articles, Black-sounding researchers are twice as likely to research sickle cell anemia compared to white ones. In the subsample of hematology articles, they are almost five times more likely to research the disease compared to white-sounding researchers. As a comparison, they are not differentially likely to research thalassemia. In terms of overall magnitudes, the number of articles on sickle cell anemia over this period is approximately 2.3 times higher compared to the ones on thalassemia, while the incidence of sickle cell anemia in the United States is estimated to be 20 times higher compared to the one of beta thalassemia.[29]

In order for sickle cell anemia and melanoma to deliver causal links between race and the direction of innovation, we need that they are uncorrelated with current socio-economic status. In the case of sickle cell anemia, I provide three pieces of evidence in support of this hypothesis. First, a similar higher incidence among Black or African Americans compared to Caucasians is found in different contexts such as the United Kingdom. Because healthcare in the UK is provided universally by the government, socio-economic status should matter less here. While we lack systematic evidence on the incidence of sickle cell anemia by socio-economic status, several studies have shown that both individuals from low socio-economic status and those from higher one may suffer from the disease.

[29]The estimated number of individuals with sickle cell anemia is 100,000 (CDC). The estimated number of individuals with beta-thalassemia is 5,000 (estimates on the incidence of beta-thalassemia are taken from NIH.gov.

Ancestry Variation: Melanoma

Melanoma, also known as malignant melanoma, is a type of cancer that develops from the pigment-producing cells melanocytes. melanoma typically occurs in the skin, and is caused both by genetic and environmental factors. Caucasian race, male sex, and older age are well-recognized factors associated with an increased risk of developing melanoma (Azoury and Lange, 2014). A key environmental trigger of melanoma is exposure to UV light. melanoma is less frequent among individuals with darker skin (such as those of African descent). This is because a darker skin tone is indication of more melanin, a molecule that protects against UV light. Therefore, Black or African Americans are far less likely to develop melanoma than non-Hispanic white Americans (at a rate of 1 per 100,000 compared to 30 per 100,000) due to the protection that melanin, the body's natural skin pigment, provides from damaging ultraviolet rays.[30]

In columns (1) and (2) of Table 1.5, I report the results of the estimation of equation (1.2) where y is a dummy equal to 1 if a research article addresses melanoma, equal to 0 otherwise. Across all articles, scientists are 50% less likely than white-sounding ones to research melanoma, and similarly among articles related to neoplasms.

Differently from the case of sickle cell anemia and thalassemia, finding a comparable disease is more challenging. I adopt a data-driven approach and study research on thyroid neoplasms. Thyroid cancer had similar mortality among Black or African Americans and among White Americans, and for Black or African Americans the mortality rate is similar to the one due to melanoma.[31].

In columns (3) and (4), I study the likelihood that a scientist with Black-sounding name engages in research on thyroid neoplasm. Both unconditionally, as shown in column

[30]Interestingly, the evolutionary reason why individuals of African ancestry have darker skin tone (i.e., more melanin in their skin), is not linked to the incidence of melanoma, but to the benefits and costs of UV light for the functioning of the human body. The leading theory in anthropology is that the levels of melanin correlated with latitude to allow the human body to absorb enough vitamin D (triggered by exposure to UV light), while preserving folate levels, which are depleted by UV light.

[31]Mortality rates by group due to each type of cancer are shown in Appendix Figure 1.17 and Appendix Figure 1.18.

(3), and within research articles on neoplasms, as shown in column (4), Black-sounding researchers are not differentially likely to research thyroid neoplasm compared to white-sounding ones.

1.4.2 A Shock to Relative Mortality: The Introduction of HAART

Today, Human Immunodeficiency Virus (HIV) is the disease with the highest relative mortality for Black or African Americans compared to white Americans: Black or African Americans are over 9 times more likely to die of HIV. However, this was not always the case. The medical literature has documented that HIV became a relatively higher contributor to the mortality of Black or African Americans (compared to whites) after the introduction of the Highly Active Antiretroviral Therapy (HAART) in 1996 (Levine, Briggs, Kilbourne, King, Fry-Johnson, Baltrus, Husaini and Rust; Levine, Rust, Pisu, Agboto, Baltrus, Briggs, Zoorob, Juarez, Hull, Goldzweig et al., 2007; 2010). HAART is a medication regimen used to manage and treat HIV and it is composed of several drugs in the antiretroviral classes of medications. Approved by the U.S. Food and Drug Administration (FDA) in 1995, HAART became widely available in 1996. HAART has dramatically decreased morbidity and mortality among HIV-infected individuals. Antiretroviral therapy delays the progression of HIV-related disease and prolongs survival, but these benefits have not been equitably distributed by race.

The evidence shown so far would predict that, when HIV becomes a relatively more "Black" disease, the relative share of scientists with Black-sounding name researching the disease should go up. I test this prediction using data on the universe of research grants awarded by the National Institutes of Health between 1988 and 2012. Using data on awarded grants is more suited to study how researchers react to this policy shock due to their more "real-time" nature. I classify grants as being related to HIV using a dictionary-based approach on the description of the research grant. Additionally, I define an indicator for "Research addressing Black or African Americans" based on a dictionary classification on the abstract. Similarly to the rest of the analysis, I assign a vector of

racial probabilities based on the last name of the first researcher listed on the article. First, I estimate a difference-in-differences equation around the introduction of HAART:

$$y_{pt} = \alpha_t + \beta_1 \text{Research about HIV}_p + \beta_2 \text{Research about HIV}_p \times \text{Post HAART} + \epsilon_{pt} \quad (1.4)$$

where p is a research grant, α_t are grant award year fixed effects, Research about HIV_i is a dummy = 1 if the research grant addresses HIV, = 0 otherwise, Post HAART is a dummy = 1 if the grant year is ≥ 2000, = 0 otherwise. Standard errors are clustered by last name.

Second, I estimate a similar differences-in-differences model where y is an indicator = 1 if a research grant addresses HIV, 0 otherwise

$$y_{ipt} = \delta_t + \gamma_1 \text{Black-s. name}_{it} + \gamma_2 \text{Black-s. name}_i \times \text{Post HAART} + \delta X_{it} + \theta X_{it} \times \text{Post HAART} + \epsilon_{ipt} \quad (1.5)$$

where i is the first researcher listed on the grant, p is a research grant, δ_t are grant award year fixed effects, Research about HIV_i is a dummy = 1 if the research grant addresses HIV, = 0 otherwise, Post HAART is a dummy = 1 if the grant year is > 1996, = 0 otherwise. X_i is a vector including Hispanic-sounding name, Asian-sounding name, American Native-sounding name, with omitted category white-sounding name. Standard errors are clustered by last name.

In column (1) of Table 1.6 I show that after the introduction of HAART, articles on research projects addressing HIV become five times more likely to mention "Black or African American" in their abstract. This result confirms that HIV becomes more associated with Black or African Americans.[32] In column (2), I test whether homophily between Black-sounding researchers changes after the introduction of HAART. After HAART, the difference in propensity to research HIV compared to white-sounding researchers increases four-fold compared to the pre-HAART period. The magnitude is similar to the increase in column (1), but the results in column (2) remain positive and

[32]The fact that the correlation was positive also before the introduction of HAART suggests that research mentioned African Americans also before HAART, but mostly jointly with mentioning white as well.

significant even restricting the sample to those grants which do not explicitely mention Black or African Americans.

I also estimate equations 1.4 and 1.5 in the sample of PubMed articles published between 1988 and 2017.[33] Results are consistent with the findings for the NIH grants, and are reported in Table XX. To investigate possible heterogeneity of treatment effects over time, I also estimate a more flexible model that, rather than interacting the vector of racial probabilities with the Post HAART indicator, interacts it with time dummies:

$$y_{ipt} = \alpha_i + \alpha_{s(i)\times t} + \sum_{k=-3}^{+7} \beta^k \times I(k \leq t - 1997 \leq k+1) \times \text{HIV}_c + \boldsymbol{x}'_{ct}\boldsymbol{\delta} + +\varepsilon_{ipt}, \quad (1.6)$$

where $I(k \leq t - 1997 \leq k+1)$ is an indicator variable that takes value one if t is in the three-year window indexed by k, and zero otherwise. $\boldsymbol{x}'_{ct}$ is a vector of year FE. y_{ipt} is an indicator taking value 1 if the paper has "Black or African American" among its MeSH codes, 0 otherwise. The results of this estimation are reported in Figure 1.5, and confirm the findings of the DiD design. After the introduction of HAART, articles on HIV become relatively more likely to mention Black or African Americans.

$$y_{ipt} = \alpha_i + \alpha_{s(i)\times t} + \sum_{k=-3}^{+7} \beta^k \times I(k \leq t - 1997 \leq k+1) \times \text{Black-sounding name}_c + \boldsymbol{x}'_{ct}\boldsymbol{\delta} + \varepsilon_{ipt}, \quad (1.7)$$

where $I(k \leq t - 1997 \leq k+1)$ is an indicator variable that takes value one if t is in the three-year window indexed by k, and zero otherwise. $\boldsymbol{x}'_{ct}$ is a vector of racial probabilities interacted with time dummies, and it includes year FE. The results of this estimation on a dummy equal to 1 if the article has "HIV" among its MeSH codes are reported in Figure 1.6. The results of the dynamic differences-in-differences design confirm the results of Table 1.6. After the introduction of HAART, compared to White-sounding scientists, Black-sounding ones become relatively more likely to research HIV.

These results establish a more direct link between the relative incidence of a disease and homophily of researchers, and they suggest that homophily is not driven by specific

[33] I begin the analysis in 1988 as this is the first year the MeSH code corresponding to HIV was introduced.

diseases, but by the incidence of these diseases across demographic groups.

1.5 Race and the Quantity and Quality of Innovation

Having established a link between the racial composition of scientists and inventors and the direction of innovation, in this section I study how this affects the quantity and quality of innovation. In section 1.5.1, I describe the research design. In section 1.5.2 I describe how this correlates with access to resources within the patenting process, and with the impact of granted patents.

1.5.1 Research Design

I test the correlation between having a Black-sounding name and the likelihood of being granted a patent, conditional on applying. I estimate the following equation on the sample of patent applications submitted to the US Patent Office between 2001 and 2018:

$$y_{ip} = \alpha_t + \beta \,\text{Black-sounding name}_i + \delta X_i + \Gamma Z_p + \epsilon_{ip} \tag{1.8}$$

Where $y = 1$ if an application is granted, $= 0$ otherwise, i is an inventor, p is a patent application, α_t is filing year FE, X_i is a vector of racial probabilities including Hispanic-sounding, Asian-sounding, American Native-sounding. The omitted category is white-sounding, and Z_p is a vector of application-level controls. Standard errors are clustered at the last name level.[34] Results are shown in Table 1.8. In Table 1.9, I report the results of the estimation of equation 1.8 with the dependent variable y equal to the total number of forward citations received by the patent, restricting the sample to granted patents. In this specification, β is the average gap in patent granting (conditional on applying) for an inventor with a Black-sounding name, compared to an inventor with a white-sounding name.

[34] I cluster errors at the last name level because that is the level of variation of the independent variable of interest (Pr(Black | Last name). In robustness checks, I cluster by patent class, or by name and patent class and significance is virtually unchanged.

1.5.2 Results

The first specification (column 1) includes only year fixed effects. The coefficient of Black-sounding last name is negative and significant: having a Black-sounding last name (compared to a white-sounding one) is associated with a 6% lower probability of being granted a patent conditional on applying. In column (2), I include USPC subclass FE. The coefficient decreases slightly but remains within the same 95% confidence interval. This suggests that the racial gap in patent-granting is not due to differential sorting of Black-sounding inventors into patent classes with lower grant rates. In column (3), I add assignee FE. The assignee is the entity (such as a firm, or a university) that holds the property right on the patent. By law, a patent must report at least one inventor but can omit the assignee. In such a case, the inventor is implicitly assumed to be the holder of the property right. In this specification, I combine assignee fixed effect with an indicator for "small entity", granted by the USPTO to universities, single inventors, and firms of up to 500 employees. As a result, all entities without assignee will have assignee equal to "small entity". I adopt this strategy to keep the sample comparable to columns (1) and (2), but the coefficients in column (3) through (5) are similar when I drop observations without assignee. Assignee fixed effects explain approximately one third of the unconditional gap shown in column (1). This suggests that Black-sounding inventors sort into entities with lower acceptance rate. However, even within the same organization (and patent subclass), they are still granted patents at lower rate conditional on applying for one. As assignees can be entities largely heterogeneous by location and size, I control for a set of additional variables to test whether, within organizations, Black-sounding inventors may be located in different areas compared to white-sounding ones, or they may have access to less support within the patenting process. In column (4), I include state of residence fixed effects. The magnitude of the gap remains almost identical compared to the one in column (3), which suggests that Black-sounding inventors do not sort into geographically different locations of the same firm (at least by state). In column (5), I include proxies for attorney quality. Patent attorneys play a key role in the patenting process as they help draft the patent document, and interact with patent examiners. I build two novel measures of attorney "quality" from the patent document. The list of

attorneys for each patent is reported on the patent application, but the number of listed attorneys is typically very large (the average number across all applications is 15). For this reason, I build the average grant rate of all attorneys listed on the patent ("average grant rate of the legal team"). Looking at the universe of patent applications between 2001 and 2018, I build for each attorney the average grant rate across all patents they have worked on.[35] Then, I compute the average of the individual-specific grant rate across all attorneys listed on the patent.[36] The coefficient on the average grant rate of the legal team is positive and significant and explains one-fifth of the Black- vs white-sounding gap shown in column (4). These results suggest that differential access to resources indeed differs across white individuals and Black individuals, but approximately half of the racial gap in patent granting remains after accounting for the available metrics of differential access to resources.

I run several robustness checks to check the validity of these results to different assumptions and metrics. First, I run the specification assigning Pr(Black | Last name) (and the vector of racial probabilities) as the average across all members of the team. Second, I restrict the analysis to single-author patents (who do not have this potential measurement issue of multiple inventors). Results are comparable throughout.

One potential explanation for the results shown in Table 1.8 is that these patents are of lower quality. However, this hypothesis does not find support in the data. More specifically, patents granted to Black-sounding inventors have higher impact compared to the ones granted to white-sounding ones. Using a Poisson count model, I estimate specification 1.8 on the subset of granted patents and where y is the number of forward citations. These trace the acknowledged contributions of prior art, and are used in the literature as an ex-post measure of patent quality and economic value. The number of citations includes any citation received by the patent in the five years after its filing.[37] The results show that Black-sounding patents are cited more compared to white-sounding

[35] I restrict the sample to those attorneys that have worked on at least 20 patents over the period, but results are consistent when I adopt different thresholds.

[36] Whenever a patent has no attorneys listed on the patent, I plug in a zero. I relax this assumption in robustness checks and similar results hold.

[37] Citations are constructed taking into account both citations from other patents, and citations from other applications.

patents. The results are shown in Table 1.9, columns (1) to (5).

These results suggest that barriers in the patenting process may prevent patents from Black-sounding inventors from being granted. What would be the impact of equalizing access to science and innovation on the overall production of science? To answer this question, in the next section I estimate a model of occupational choice (Hsieh, Hurst, Jones and Klenow, 2019).

1.6 Equalizing Access to Innovation and Science

In this section, I study the labor market consequences of equalizing access to innovation and science through a model of occupational choice. In section 1.6.1 I outline the setup of the model. In section 1.6.2, I discuss the model estimation. In section 1.6.3, I report policy counterfactuals.

1.6.1 Model Outline

To quantify the effect of equalizing access to innovation and science for Black or African Americans, I estimate a Roy model of occupational choice (Hsieh, Hurst, Jones and Klenow, 2019). There are at least two reasons why a general equilibrium model is useful to estimate the impact of removing barriers for Black or African Americans from the economy. First, if individuals anticipate discrimination, they may choose another career early on in life. Therefore, we would observe them select into different occupations where they may be less productive, or which they may enjoy less, but where barriers are lower. Second, as more individuals enter into an occupation, labor demand for that occupation will change and equilibrium wages will go down.

The economy is composed of a continuum of individuals i who are either white or Black. Their group is indexed by $g = \{\text{white, Black}\}$. Each individual chooses an occupation j to maximize their lifetime utility. Individuals choose their lifetime occupation and decide

how much time to dedicate to schooling before entering the labor market (the pre-period), and live for three periods. Their lifetime utility is equal to:

$$\log U_i = \alpha \sum_{t=1}^{3} \log \, c_{it} + \log \, (1 - s_{ij}) + \log \, z_{jg} + \log \, \mu_{ij} \tag{1.9}$$

α represents the tradeoff between utility in the pre-period and utility over the remaining of the lifetime. s is time spent in school to acquire human capital (so that $(1-s)$ is leisure), z is group-specific utility derived from working in occupation j, and μ is individual utility from working in occupation j. The parameter z_{jg} relaxes the assumption that, in the absence of barriers, all groups will select occupation j at the same rate. It can be interpreted as preferences, beliefs, or experience.[38] Individual consumption c_{it} is equal to:

$$c_{it} = (1 - \tau^w_{jg}) w_{jt} \varepsilon_{ij} h_{ij} - (1 + \tau^h_{jg}) e_{ij} \tag{1.10}$$

Where τ_{jg} is the job- and group- specific tax to work in occupation j, w_{jt} is the efficiency wage, ε_{ij} is individual productivity of working in occupation j, (i.e., their "talent" for occupation j), h_{ij} is human capital acquired to work in occupation j. Over the lifetime, individuals repay the loan they got in the first period to acquire education. They repay it equally across all three periods. In every period they have to repay $1/3 \, (1+\tau^h_{jg}) e_{ij}$, where τ^h_{jg} is the tax on acquisition of human capital.[39] In this economy, output is produced by one firm that aggregates labor inputs from J occupations through the production function:

$$Y = \left[\sum_{j=1}^{J} \left(A_j \cdot H_j \right)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \tag{1.11}$$

Where H_j is total efficiency units of labor in each occupation, σ is the elasticity of

[38] For example, Hsieh, Hurst, Jones and Klenow (2019) documents a decrease of z in the US between 1960 and 2010 a decrease women in the home sector, which they interpret as changes in social norms for women working in the market sector or changing preferences for fertility.

[39] For inventors, the barrier in human capital accumulation τ^h can be micro–founded by Bell, Chetty, Jaravel, Petkova and Van Reenen (2019), who show that exposure to more inventors in childhood leads to an increased propensity of becoming an inventor. The on-the-job barrier τ^w is motivated by the evidence shown in section 1.5. Similar evidence does not exist for scientists, but I assume it is plausible that similar frictions might be at play. In ongoing work, I re-estimate the model calibrating the barriers using these micro-founded estimated.

substitution across occupations, A_j is the exogenously-given productivity of occupation j. In equilibrium, $w_j t$ clears the labor market in each sector so that $H_{jt}^{supply} = H_{jt}^{demand}$.

1.6.2 Estimation

To estimate the model, I impose a series of additional assumptions. First, in the main estimation, I assume that individuals only select occupations based on talent (not on individual preferences). That is, that $\mu = 1$. Secondly, I assume that talent is distributed equally across groups in a given sector. Third, I assume that white individuals do not face any barriers in the acquisition of human capital or in the labor market. That is, $\tau_{white}^{h} = 0$ and $\tau_{white}^{w} = 0$ in all occupations and all periods. Fourth, I normalize preference for the home sector to be equal to 1 for all groups. Fifth, I assume that the return to experience γ is the same for all sectors, groups, and cohorts. I estimate the model using internally calibrated parameters τ^h, τ^w, and z, for which I rely on data from the first two section of the paper and from the Current Population Survey (CPS) from 1990 to 2018. I use externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019).

1.6.3 Policy Counterfactuals

In this section, I estimate the impact of equalizing access to innovation and science, i.e. lifting barriers to human capital accumulation and in the labor market, on the total share of inventors and scientists in the economy. I define as "innovation and science" an occupation combining individuals who report working in "engineering" and "natural scientists". The results of this policy counterfactual are shown in Figure 1.7. Removing barriers increases the number of inventors and scientists by 10%. This increase is driven entirely by an increase in Black or African Americans, while the number of white individuals decreases by a negligible amount. Overall, in the economy, this corresponds to an increase in 1 p.p. in the number of scientists and inventors. In the plot, I report counterfactual estimations under a policy that lifts barriers only in the innovation and science sector (light grey bars), and a policy that lifts barriers everywhere in the economy. The

broad result is consistent across the two different policies. The increase in the share of Black or African Americans is higher under the second policy. The intuition behind this result is that, at the margin, more Black or African Americans will sort into innovation if other occupations with high barriers elsewhere in the economy remain distorted (e.g., lawyers). Why does the overall number of inventors and scientists go up? This finding hinges on the assumption on the elasticity of substitution between occupations in aggregate production (σ). Following Hsieh, Hurst, Jones and Klenow (2019), I calibrate this parameter to have value 3, and experiment with different values in robustness checks. While the relative magnitudes of the increase in inventors vary with the value of σ, the core result on the increase in the number of inventors, and in the relative share of Black or African Americans among them, remains. Under any parameter value σ, the average productivity (or, quality) of scientists and inventors increases after removing frictions, as talent is now optimally allocated across occupations. This result is in line with the empirical evidence on quality of granted patents discussed in section 1.5.

1.7 Discussion

The findings reported in section 1.6 suggest that equalizing access to innovation and science would lead to an increase in the quality and quantity of innovation, and to a larger share of Black or African American scientists and inventors. Would this lead to more research benefiting Black or African Americans?

Below, I summarize two arguments that suggest that the current underrepresentation of Black or African American scientists and inventors leads to the underprovision of innovation benefiting Black or African Americans. This suggests that closing the Black-white gap in scientists and inventors will likely lead to more research benefiting Black or African Americans.

First, in section 1.3 I document that Black or African American researchers are more likely to run clinical trials with Black or African American participants. This is in line

with recent evidence on the lack in participation in clinical trials of Black or African Americans (Alsan, Durvasula, Gupta, Schwartzstein and Williams; Alsan, Campbell, Leister and Ojo, 2022; 2023). Additionally, the fact that even in *registered* trials Black or African American scientists tend to involve Black or African Americans more suggests that part of the effect is driven by researchers' decisions on which demographic groups to include in their studies. This hypothesis is supported by the fact that the coefficient in Table 1.2, Column (2) for published trials is larger compared to the one reported in Table 1.24, Column (2), which reports results on trial registration.

Second, when looking at white researchers (over-represented compared to the overall population), a similar pattern emerges. Compared to scientists with a Black-sounding last name, they are 15% more likely to research typically white diseases, and 30 to 40% more likely to research melanoma. Additionally, in articles and clinical trials I document that Hispanic-sounding researchers are more likely to design trials and publish articles with individuals of Hispanic or Latino origin, and Asian-sounding researchers with individuals of Asian origin. This suggests that, even when a group is over-represented in the pool of scientists and inventors compared to population size, the propensity to innovate for their own demographic group remains. The match between race or ethnicity and direction of research and innovation is not linked to being a minority group, but seems to be an empirical regularity across demographics.

Finally, an outstanding question is: *Why* do scientists and inventors produce research and innovation that benefits their own demographic group disproportionately more?

The match between researchers and HIV-related research suggests that this pattern is linked to demographic characteristics, rather than specific diseases. This is further supported by the finding that even conditioning on disease, Black-sounding scientists are more likely to include Black or African Americans in their research.

Differential returns are unlikely to explain per se why this match emerges. Research benefiting Black or African Americans does not have significantly different returns compared to other research in the same area. Additionally, this match holds across bins of journal

impact factors (high compared to low).

Plausible mechanisms through which the match between researcher and content of research may emerge are information asymmetries, differential expectations on market size, or preferences for innovating for one's own demographic group. One way to rationalize these findings through preferences is via intrinsic motivation: we know from seminal work by Stern (2004) that scientists are intrinsically motivated agents. The existence of a match between their race and ethnicity and the direction of their research and innovation would be consistent with intrinsic motivation being linked to a specific topic, rather than "general purpose" (i.e., scientists "paying" to be scientists *and* research a specific topic).

Understanding the relative contribution of information and preferences will have fundamental implications for the design of policies to ensure that the production of science and innovation benefits every group of society.

1.8 Conclusion

Previous literature has shown that minority gaps among inventors contribute to and perpetuate unequal opportunities within society (Bell, Chetty, Jaravel, Petkova and Van Reenen, 2019). Such disparities can have large implications for the quality and quantity of innovation (Bloom, Van Reenen and Williams; Bell, Chetty, Jaravel, Petkova and Van Reenen; Einiö, Feng and Jaravel, 2019; 2019; 2023). In this paper, I find evidence consistent with the racial gap in scientists and inventors leading to a lower volume of research and innovation via a lower overall number of scientists and inventors. Inventors and scientists are a key input in the production function of science and innovation, and race-specific barriers in the economy lead to an inefficiently low number of them. I find that this is not only driven by differential sorting into these occupations, but also by Black or African American inventors facing higher barriers within the patenting process.

What are the implications of too few Black or African American scientists and inventors? In the second part of the paper, I test the hypothesis that Black or African American scientists and inventors tend to innovate for Black or African Americans. I focus on medical innovation, a research field where innovation for Black or African Americans has been shown to be underprovided. Defining a set of novel metrics, I document that Black or African American scientists are more likely to produce medical research and innovation benefiting Black or African Americans. This finding complements existing evidence on the importance of gender, socio-economic status, and age for the direction of innovation (Koning, Samila and Ferguson; Koning, Samila and Ferguson; Einiö, Feng and Jaravel, 2020; 2021; 2023). This evidence suggests that equality among those producing science and research, as well as those commercializing these innovations, is fundamental to guarantee that all demographic groups in the population equally benefit from medical advancements.

Future research should focus on three issues of primary importance. First, on advancing our understanding of the reasons behind the racial gap in inventors and scientists, and on how to design policies and institutions aimed at closing this gap.[40] Second, on devising strategies to incentivize the production of innovation benefiting Black or African Americans in the transition to a more equal pool of researchers. This is likely to be a costly process as it takes a large amount of resources to get scientists to change the direction of their research (Myers, 2020). However, this represents an essential step to ensure that research and innovation benefit everyone in the population. Finally, an open question is whether, and to what extent, the racial and ethnic diversity of background and experience in the pool of innovators may foster, in itself, the production of novel ideas.

[40] A series of studies beginning with the seminal work by Bertrand and Mullainathan (2004), and most recently findings by Kline, Rose and Walters (2022), show that racial discrimination affects hiring decisions. Whether racial bias plays a role in patent granting decisions today is an open question. Recent findings (Jensen, Kovács and Sorenson; Avivi, 2018; 2023) show evidence consistent with gender bias at the patent office. Coluccia, Dossi and Ottinger (2023) finds that racial discrimination at the patent office harmed innovation at the beginning of the 20th century.

References

Abramitzky, Ran, Leah Platt Boustan, and Dylan Connor. 2020. "Leaving the Enclave: Historical Evidence on Immigrant Mobility from the Industrial Removal Office." National Bureau of Economic Research.

Akcigit, Ufuk, and Nathan Goldschlag. 2023. "Measuring the Characteristics and Employment Dynamics of US Inventors." National Bureau of Economic Research.

Alsan, Marcella, Maya Durvasula, Harsh Gupta, Joshua Schwartzstein, and Heidi L Williams. 2022. "Representation and Extrapolation: Evidence from Clinical Trials." National Bureau of Economic Research.

Alsan, Marcella, Romaine A Campbell, Lukas Leister, and Ayotomiwa Ojo. 2023. "Investigator Racial Diversity and Clinical Trial Participation." National Bureau of Economic Research.

Ashraf, Nava, Oriana Bandiera, Virginia Minni, and Vıctor Quintas-Martınez. 2023. "Gender Roles and the Misallocation of Labour Across Countries." *Working paper.*

Avivi, Hadar. 2023. "Are Patent Examiners Gender Neutral?" *Working paper.*

Azoury, Saïd C., and Julie R. Lange. 2014. "Epidemiology, Risk Factors, Prevention, and Early Detection of Melanoma." *Surgical Clinics of North America*, 94(5): 945–962. Melanoma.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States." *Econometrica*, 88(6): 2329–2368.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *The Quarterly Journal of Economics*, 134(2): 647–713.

Berkes, Enrico, Davide Coluccia, Gaia Dossi, and Mara Squicciarini. 2023. "Dealing with Adversity: Religiosity and Science? Evidence from the Great Influenza Pandemic." *Working paper.*

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Bloom, Nicholas, John Van Reenen, and Heidi Williams. 2019. "A Toolkit of Policies to Promote Innovation." *Journal of Economic Perspectives*, 33(3): 163–84.

Bryan, Kevin A, and Heidi L Williams. 2021. "Innovation: Market Failures and Public Policies." In *Handbook of Industrial Organization.* Vol. 5, 281–388. Elsevier.

Chetty, Raj, Gaia Dossi, Matthew Smith, John Van Reenen, Owen Zidar, and Eric Zwick. 2023. "America's Missing Entrepreneurs."

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association between Income and Life Expectancy in the United States, 2001-2014." *JAMA*, 315(16): 1750–1766.

Coluccia, Davide M, Gaia Dossi, and Sebastian Ottinger. 2023. "Racial Discrimination and Lost Innovation." *Working paper.*

Comenetz, J. 2016. "Frequently Occurring Surnames in the 2010 Census. United States Census Bureau."

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The Determinants of Mortality." *Journal of Economic Perspectives*, 20(3): 97–120.

Cutler, David M, Adriana Lleras-Muney, and Tom Vogl. 2008. "Socioeconomic Status and Health: Dimensions and Mechanisms."

Cutler, David M, Ellen Meara, and Seth Richards-Shubik. 2012. "Induced Innovation and Social Inequality: Evidence from Infant Medical Care." *Journal of Human Resources*, 47(2): 456–492.

Dossi, Gaia, and Marta Morando. 2023. "Political Ideology and Innovation."

Einiö, Elias, Josh Feng, and Xavier Jaravel. 2023. "Social Push and the Direction of Innovation." *Working Paper*.

Farooq, Faheem, Peter J Mogayzel, Sophie Lanzkron, Carlton Haywood, and John J Strouse. 2020. "Comparison of US Federal and Foundation Funding of Research for Sickle Cell Disease and Cystic Fibrosis and Factors Associated with Research Productivity." *JAMA Network Open*, 3(3): e201737–e201737.

Fenton, Anne, Erika Elliott, Ashkan Shahbandi, Ekene Ezenwa, Chance Morris, Justin McLawhorn, James G Jackson, Pamela Allen, and Andrea Murina. 2020. "Medical Students' Ability to Diagnose Common Dermatologic Conditions in Skin of Color." *Journal of the American Academy of Dermatology*, 83(3): 957–958.

Fouka, Vasiliki. 2020. "Backlash: The Unintended Effects of Language Prohibition in US Schools After World War I." *The Review of Economic Studies*, 87(1): 204–239.

Fry, Caroline Viola. 2023. "Crisis and the Trajectory of Science: Evidence from the 2014 Ebola Outbreak." *The Review of Economics and Statistics*, 105(4): 1028–1038.

Gaule, Patrick, and Mario Piacentini. 2013. "Chinese Graduate Students and US Scientific Productivity." *Review of Economics and Statistics*, 95(2): 698–701.

Graham, Stuart JH, Alan C Marco, and Richard Miller. 2015. "The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination." *Georgia Tech Scheller College of Business Research Paper No. WP*, 43.

Hsieh, Chang-Tai, Erik Hurst, Charles I Jones, and Peter J Klenow. 2019. "The Allocation of Talent and US Economic Growth." *Econometrica*, 87(5): 1439–1474.

Hutchins, B Ian, Matthew T Davis, Rebecca A Meseroll, and George M Santangelo. 2019. "Predicting Translational Progress in Biomedical Research." *PLoS biology*, 17(10): e3000416.

Hutchins, B Ian, Xin Yuan, James M Anderson, and George M Santangelo. 2016. "Relative Citation Ratio (RCR): A New Metric that Uses Citation Rates to Measure Influence at the Article Level." *PLoS biology*, 14(9): e1002541.

Introini, Viola, Alejandro Marin-Menendez, Guilherme Nettesheim, Yen-Chun Lin, Silvia N Kariuki, Adrian L Smith, Letitia Jean, John N Brewin, David C Rees, Pietro Cicuta, et al. 2022. "The Erythrocyte Membrane Properties of Beta Thalassaemia Heterozygotes and their Consequences for Plasmodium Falciparum Invasion." *Scientific Reports*, 12(1): 8934.

Jensen, Kyle, Balázs Kovács, and Olav Sorenson. 2018. "Gender Differences in Obtaining and Maintaining Patent Rights." *Nature Biotechnology*, 36(4): 307–309.

Johnson, Julie A. 2008. "Ethnic Differences in Cardiovascular Drug Response: Potential Contribution of Pharmacogenetics." *Circulation*, 118(13): 1383–1393.

Kerr, William R. 2008. "Ethnic Scientific Communities and International Technology Diffusion." *The Review of Economics and Statistics*, 90(3): 518–537.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic Discrimination among Large US Employers." *The Quarterly Journal of Economics*, 137(4): 1963–2036.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2020. "Inventor Gender and the Direction of Invention." Vol. 110, 250–254, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2021. "Who Do We Invent for? Patents by Women Focus More on Women's Health, but Few Women Get to Invent." *Science*, 372(6548): 1345–1348.

Kozlowski, Diego, Dakota S Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R Sugimoto. 2022. "Avoiding Bias When Inferring Race Using Name-based Approaches." *PloS One*, 17(3): e0264270.

Levine, Robert S, George S Rust, Maria Pisu, Vincent Agboto, Peter A Baltrus, Nathaniel C Briggs, Roger Zoorob, Paul Juarez, Pamela C Hull, Irwin Goldzweig, et al. 2010. "Increased Black–White Disparities in Mortality After the Introduction of Lifesaving Innovations: A Possible Consequence of US Federal Laws." *American Journal of Public Health*, 100(11): 2176–2184.

Levine, Robert S, Nathaniel C Briggs, Barbara S Kilbourne, William D King, Yvonne Fry-Johnson, Peter T Baltrus, Baqar A Husaini, and George S Rust. 2007. "Black–white Mortality from HIV in the United States Before and After Introduction of Highly Active Antiretroviral Therapy in 1996." *American Journal of Public Health*, 97(10): 1884–1892.

Louie, Patricia, and Rima Wilkes. 2018. "Representations of Race and Skin Tone in Medical Textbook Imagery." *Social Science & Medicine*, 202: 38–42.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on Science: Worldwide Front-page Patent Citations to Scientific Articles." *Strategic Management Journal*, 41(9): 1572–1594.

Marx, Matt, and Aaron Fuegi. 2022. "Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-article Citations." *Journal of Economics & Management Strategy*, 31(2): 369–392.

Michelman, Valerie, and Lucy Msall. 2023. "Sex, Drugs, and RD: Missing Innovation from Regulating Female Enrollment in Clinical Trials."

Moscona, Jacob, and Karthik Sastry. 2022. "Inappropriate Technology: Evidence from Global Agriculture."

Myers, Kyle. 2020. "The Elasticity of Science." *American Economic Journal: Applied Economics*, 12(4): 103–134.

Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan. 2020. "Improving Data Access Democratizes and Diversifies Science." *Proceedings of the National Academy of Sciences*, 117(38): 23490–23498.

NSF. 2015. "Women, Minorities, and Persons with Disabilities in Science and Engineering." *National Science Foundation Techinical Report*.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, 366(6464): 447–453.

Olivetti, Claudia, M Daniele Paserman, Laura Salisbury, and E Anna Weber. 2020. "Who Married, (To) Whom, and Where? Trends in Marriage in the United States, 1850-1940." *NBER Working Paper*.

Rees, David C, Thomas N Williams, and Mark T Gladwin. 2010. "Sickle-cell Disease." *The Lancet*, 376(9757): 2018–2031.

Rose, Evan K. 2023. "A Constructivist Perspective on Empirical Discrimination research." *Journal of Economic Literature*, 61(3): 906–923.

Schadendorf, Dirk, Alexander CJ van Akkooi, Carola Berking, Klaus G Griewank, Ralf Gutzmer, Axel Hauschild, Andreas Stang, Alexander Roesch, and Selma Ugurel. 2018. "Melanoma." *The Lancet*, 392(10151): 971–984.

Schwandt, Hannes, Janet Currie, Marlies Bär, James Banks, Paola Bertoli, Aline Bütikofer, Sarah Cattan, Beatrice Zong-Ying Chao, Claudia Costa, Libertad González, et al. 2021. "Inequality in Mortality between Black and White Americans by Age, Place, and Cause and in Comparison to Europe, 1990 to 2018." *Proceedings of the National Academy of Sciences*, 118(40): e2104684118.

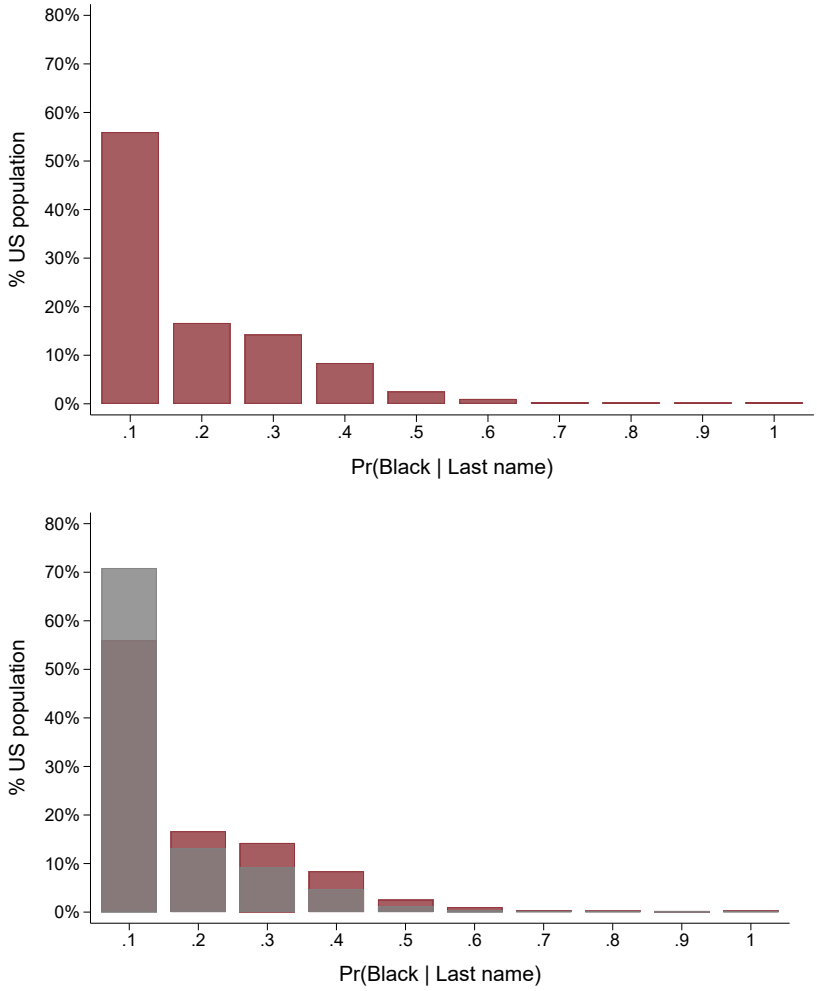
Sjoding, Michael W., Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. 2020. "Racial Bias in Pulse Oximetry Measurement." *New England Journal of Medicine*, 383(25): 2477–2478. PMID: 33326721.

Stern, Scott. 2004. "Do Scientists Pay to Be Scientists?" *Management Science*, 50(6): 835–853.

Truffa, Francesca, and Ashley Wong. 2022. "Undergraduate Gender Diversity and Direction of Scientific Research."

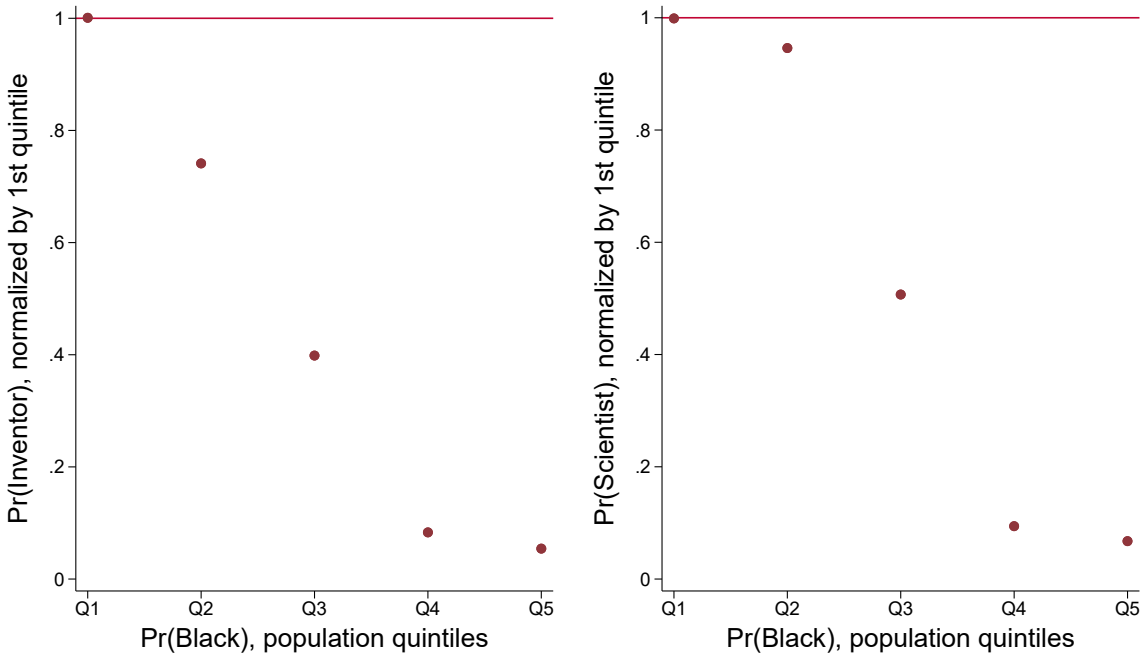
Figures

Figure 1.1: Pr(Black | Last Name) in the US Population, 2010 Census



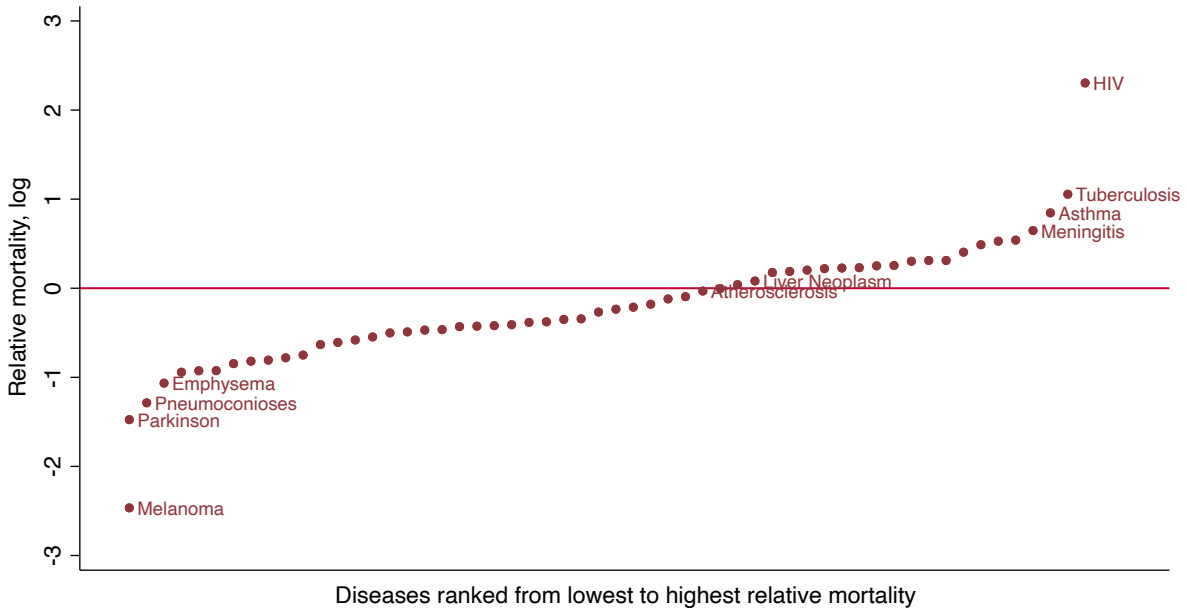
Notes. This figure plots the distribution of the US population across deciles of Pr(Black-sounding | Last name), using data from the 2010 US Census (Comenetz, 2016). One bar corresponds to the share of individuals in the US population who are in the given decile of probability of being Black or African American based on their last name.

Figure 1.2: Pr(Inventor) and Pr(Scientist) by Quintiles of Pr(Black | Last Name)



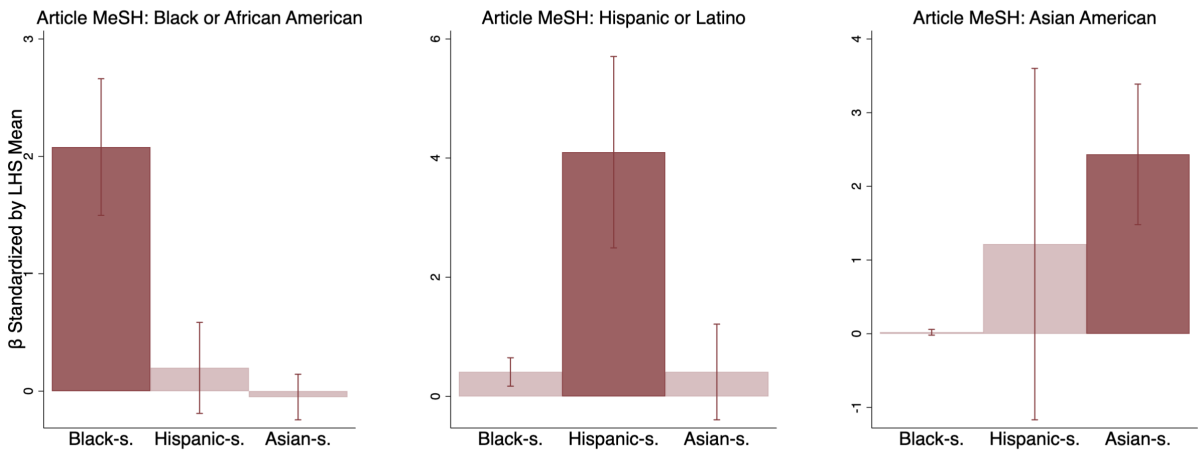
Notes. These plots report, respectively, the share of inventors (LHS) and the share of scientists (RHS), by quintile of Pr(Black | Last name). The shares are normalized by the first quintile. The share of inventors is computed on all US resident inventors listed as first author of a patent application submitted to the USPTO between January 2001 and December 2018. The share of scientists is computed on all US resident scientists listed as first author of a PubMed publication published between January 2002 and December 2018. On the x-axis, Pr(Black) refers to Pr(Black | Last name) from the 2010 US Census (Comenetz, 2016). The frequency of each last name in the US population is also from Comenetz (2016).

Figure 1.3: Causes of Death Ranked by Relative Mortality



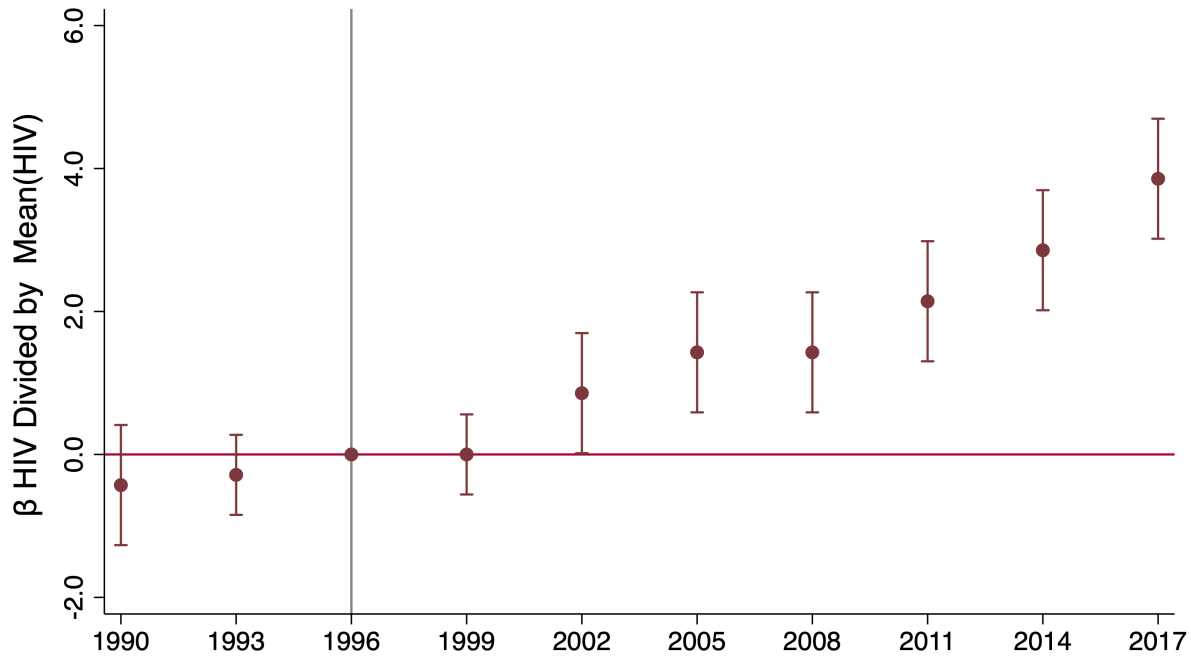
Notes. This figure displays relative mortality rate of Black individuals compared to white individuals in the United States in 2000 to 2016. One observation is one disease with at least 1,000 related deaths of white individuals and Black individuals over the full period. Data comes from the CDC. More details on the construction of this data are reported in section 1.2. On the x-axis, I rank disease by relative mortality rate of Black Americans vs white Americans. On the y-axis, I show the actual relative mortality rate of Black Americans compared to white Americans due to the given disease over the period. It is computed as total number of deaths of Black individuals reported due to the given disease divided by total number of Black individuals in the population, divided by total number of deaths of white individuals reported due to the given disease divided by total number of whites in the population, in log. The red line corresponds to equal mortality rate for Black individuals compared to white individuals (i.e. when y = 0). The sample is restricted to mortality reported between years 1999 and 2015, and includes diseases with at least 5,000 registered deaths (white Americans + Black Americans) over the period. The final sample includes 57 diseases. I report labels of the four diseases with highest relative mortality among whites, and of the four diseases with highest relative mortality among Black individuals.

Figure 1.4: Demographics-based Approach, Research on Other Demographic Groups



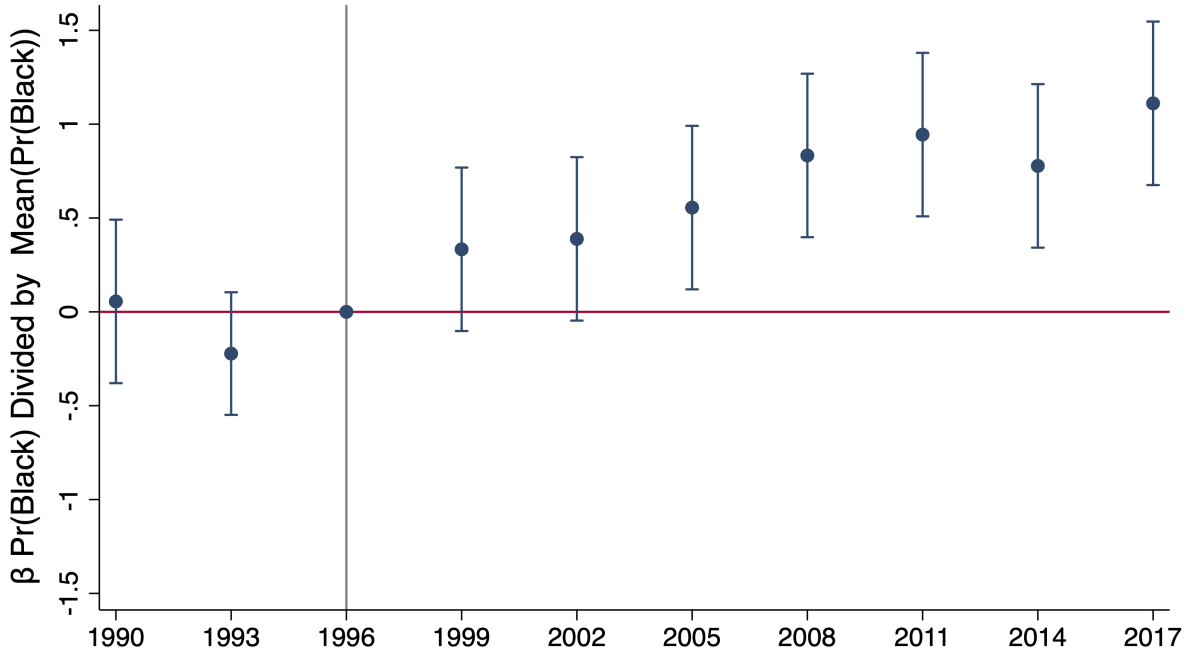
Notes. This plot reports the coefficients of column (1), Appendix Table 1.23, Panel A, Panel B, and Panel C. The bars plot 95% confidence intervals. S.e. are clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. The first plot from the left reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. The middle plot reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Hispanic or Latino" among its MeSH codes, = 0 otherwise. The plot on the right reports the result of estimating Equation 2 where the dependent variable is a dummy = 1 if the article has "Asian American" among its MeSH codes, = 0 otherwise.

Figure 1.5: Impact of HAART on Likelihood that an HIV-related Paper has "Black or African American" Among its MeSH Terms



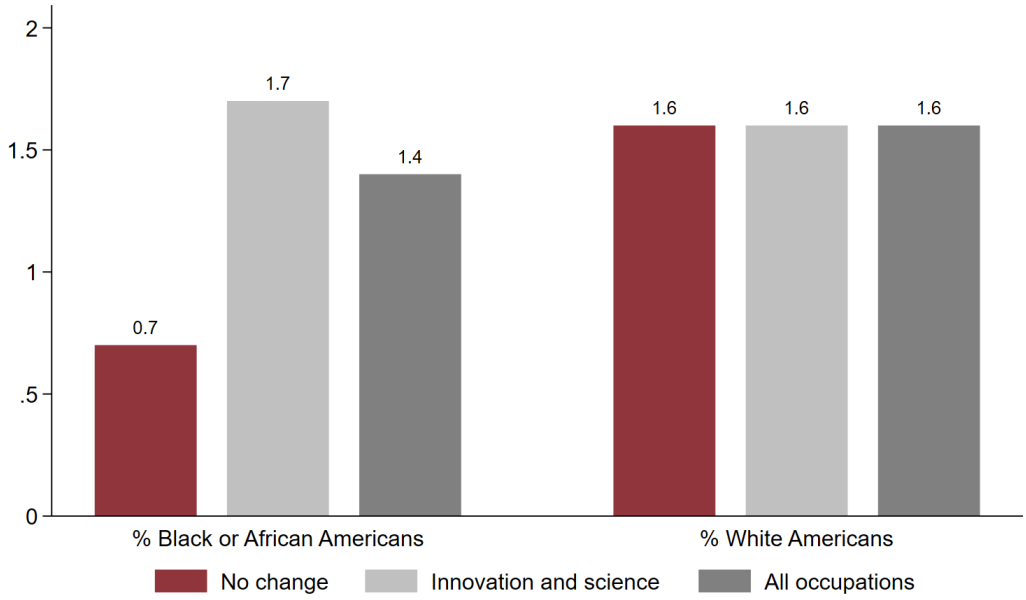
Notes. This plot reports the coefficients of an event study where the dependent variable id a dummy = 1 if the article has "Black or African American" among its MeSH terms, and = 0 otherwise. Time dummies (aggregate in three-year period) are interacted with a set of dummies = 1 if the article has "HIV" among its MeSH terms, = 0 otherwise. The sample includes all articles published by a first author residing in the US between 1988 and 2017. Standard errors are clustered at the last name level.

Figure 1.6: Impact of HAART on Match between Black-sounding Researchers and HIV Research



Notes. This plot reports the coefficients of an event study where the dependent variable id a dummy = 1 if the article has "HIV" among its MeSH terms, and = 0 otherwise. Time dummies (aggregated in three-year periods) are interacted with Pr(Black | Last name) of the first author of the article. The vector of controls includes the other racial probabilities associated with the first author's last name, i.e. Hispanic-sounding last name, Asian-sounding last name, Am. Native-sounding last name, and this vector of racial probabilities interacted with time dummies. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. The sample includes all articles published by a first author residing in the US between 1988 and 2017. Standard errors are clustered at the last name level.

Figure 1.7: Equalizing Access to Innovation and Science, % of Inventors and Scientists



Notes. This figure reports the result of estimating the model outlined in section 6. The red bar (first and fourth bar from the left) reports the share of life scientists and engineers in an economy with race-specific frictions, estimated using Current Population Survey (CPS) data for 1995 to 2018. The first bar reports the share among whites (i.e., the number of white scientists and inventors normalized by the total number of whites). The fourth bar reports the share among Black individuals (i.e., the number of Black scientists and inventors normalized by the total number of Black individuals). The light gray bars (second and fifth bars) report, respectively, the share of white scientists and inventors, and the share of Black scientists and inventors in the economy after removing race-specific barriers to science an innovation. In this policy counterfactual, race-specific barriers to human capital accumulation and within the labor market are only removed in the innovation and science sector, and remain equal to the baseline in all other occupations. The dark gray bars (third and sixth bar), show respectively the results of a policy counterfactual where race-specific barriers to human capital accumulation and within the labor market are removed everywhere in the economy. The third bar from the left show the share of inventors and scientists among whites. The sixth bar from the left shows the share of inventor and scientists among Black individuals.

Tables

Table 1.1: Sample Statistics, Research Articles and Patents

Panel A: PubMed Articles

Variable	Mean	Std. dev.	Min.	Max.	N
MeSH: Black or African American	0.007	0.086	0.000	1.000	651,253
Relative mortality, log	-0.039	0.899	-2.465	2.304	138,657
Typically White	0.331	0.470	0.000	1.000	138,657
Similar Incidence	0.488	0.500	0.000	1.000	138,657
Typically Black	0.181	0.385	0.000	1.000	138,657
MeSH: Sickle Cell Anemia	0.001	0.036	0.000	1.000	651,253
MeSH: Thalassemia	0.000	0.019	0.000	1.000	651,253
MeSH: Melanoma	0.006	0.079	0.000	1.000	651,253
MeSH: Thyroid Neoplasm	0.002	0.042	0.000	1.000	651,253
Black-sounding last name	0.063	0.115	0.000	1.000	651,253
Hispanic-sounding last name	0.046	0.151	0.000	1.000	651,253
Asian-sounding last name	0.316	0.432	0.000	1.000	651,253
American Native-sounding last name	0.005	0.013	0.000	0.981	651,253
White-sounding last name	0.571	0.407	0.000	1.000	651,253

Panel B: USPTO Patent Applications

Variable	Mean	Std. dev.	Min.	Max.	N
Granted patent	0.717	0.450	0	1.000	2,009,485
Number of total citations	23.299	263.724	0	27522.000	1,483,345
Black-sounding name	0.080	0.120	0	1.000	2,009,485
Hispanic-sounding name	0.038	0.128	0	0.996	2,009,485
Asian-sounding name	0.193	0.368	0	0.997	2,009,485
American native-sounding name	0.005	0.012	0	0.978	2,009,485
White-sounding name	0.684	0.358	0	1.000	2,009,485

Notes. In Panel A, the unit of observation is a PubMed researched article published between 2002 and 2018. In Panel B, the unit of observation is a patent application with first inventor resident in the US filed at the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, American-native sounding name, and white-sounding name refer to race or ethnicity of the first author listed on the patent application. Black-sounding name, Hispanic-sounding name, and Asian-sounding name, American-native sounding name, and white-sounding name refer to race or ethnicity of the first author listed on the publication.

Table 1.2: Demographics-based Approach

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.010***	0.025***	0.009***	0.028***	0.029***	0.028***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.009)	(0.005)
Hispanic-sounding name	0.000	0.001	0.000	0.003	0.001	0.003
	(0.001)	(0.003)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.003***	-0.002	-0.003***	0.000	-0.002	-0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.006	-0.011	-0.005	-0.015	-0.011	-0.018
	(0.007)	(0.029)	(0.007)	(0.016)	(0.031)	(0.016)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
R-squared	0.001	0.001	0.001	0.001	0.002	0.001
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.3: Frequency-based Approach

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.048***	-0.002	0.050***
	(0.034)	(0.018)	(0.017)	(0.014)
Hispanic-sounding name	0.006	-0.005	0.000	0.005
	(0.024)	(0.014)	(0.012)	(0.009)
Asian-sounding name	-0.006	-0.022***	0.044***	-0.022***
	(0.009)	(0.005)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.053	-0.021	-0.032
	(0.177)	(0.112)	(0.124)	(0.084)
Total mortality, log	-0.108***	-0.039***	0.139***	-0.100***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.015	0.164	0.144
LHS (mean)	-0.039	0.331	0.488	0.181
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2016. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 1.4: Ancestry Variation: Sickle Cell Anemia

	(1)	(2)	(3)	(4)
	MeSH: Sickle Cell Anemia		MeSH: Thalassemia	
Black-sounding name	0.001*	0.091***	-0.000	-0.011
	(0.001)	(0.034)	(0.000)	(0.010)
Hispanic-sounding name	-0.000	-0.015	-0.000	-0.011*
	(0.000)	(0.017)	(0.000)	(0.007)
Asian-sounding name	-0.000**	-0.004	-0.000	0.002
	(0.000)	(0.008)	(0.000)	(0.004)
Am. Native-sounding name	-0.006***	-0.232**	0.000	0.021
	(0.002)	(0.100)	(0.002)	(0.089)
Hemoglobin-related articles		X		X
Observations	651,253	13,297	651,253	13,297
R-squared	0.000	0.004	0.000	0.004
LHS (mean)	0.001	0.062	0.000	0.017
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation 1.2 where the dependent variable is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with hemoglobin among their MeSH codes. Column (3) reports the result of estimating equation 1.2 where the dependent variable is a dummy = 1 if the article has thalassemia among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of hemoglobin-related diseases. Column (5) reports the result of estimating equation 1.2 where the dependent variable is a dummy = 1 if the article has hemophilia among its MeSH codes, = 0 otherwise. Column (6) reports the results of estimating the same equation as in column (5), but on the subsample of hemoglobin-related diseases.

Table 1.5: Ancestry Variation: Melanoma

	(1)	(2)	(3)	(4)
	MeSH: Melanoma		MeSH: Thyroid neoplasm	
Black-sounding name	-0.002*	-0.023**	0.000	0.001
	(0.001)	(0.011)	(0.001)	(0.005)
Hispanic-sounding name	0.001	-0.004	0.000	-0.000
	(0.001)	(0.009)	(0.000)	(0.003)
Asian-sounding name	0.000	-0.005	-0.000	-0.003**
	(0.000)	(0.003)	(0.000)	(0.001)
Am. Native-sounding name	0.000	0.040	-0.005*	-0.023
	(0.006)	(0.070)	(0.003)	(0.028)
Neoplasm-related articles		X		X
Observations	651,253	67,804	651,253	67,804
R-squared	0.000	0.002	0.000	0.001
LHS (mean)	0.006	0.060	0.002	0.011
RHS (mean)	0.063	0.061	0.063	0.061

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (1.2) where the dependent variable is a dummy = 1 if the article has "Melanoma" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of neoplasm-related articles. "Neoplasm-related" articles are defined as those articles with neoplasm among their MeSH codes. Column (3) reports the result of estimating equation (1.2) where the dependent variable is a dummy = 1 if the article has "Thyroid neoplasm" among its MeSH codes, = 0 otherwise. Column (4) reports the results of estimating the same equation as in column (3), but on the subsample of neoplasm-related articles.

Table 1.6: DD Around the Introduction of HAART, NIH Grants

	(1) Research on African Americans	(2) Research on HIV
HIV	0.004***	
	(0.001)	
HIV X Post HAART	0.020***	
	(0.003)	
Black-sounding name		0.004
		(0.007)
Black-sounding name X Post HAART		0.021**
		(0.008)
Vector of racial prob.		X
Vector of racial prob. X Post HAART		X
Observations	589,277	589,277
R-squared	0.006	0.005
LHS (mean)	0.042	0.042
RHS (mean)	0.083	0.083

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research grant awarded by the National Institutes of Health between 1988 and 2017 by a first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to the last names of the first author listed on the publication. White-sounding name is the omitted category. All columns include FE for the year in which the grant was awarded. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (1.4). The dependent variable is a dummy = 1 if the research grant mentions Black or African Americans, = 0 otherwise. Post HAART is an indicator = 1 if the award grant year is > 1996, = 0 otherwise. Column (2) reports the result of estimating equation (1.5). The dependent variable is a dummy = 1 if the research grant mentions HIV, = 0 otherwise. The results in column (2) include a vector of controls including Hispanic-sounding name, Asian-sounding name, and American-Native-sounding name. It also includes these controls interacted with the Post HAART indicator. All columns include year FE.

Table 1.7: DD Around the Introduction of HAART, PubMed Articles

	(1) Research on African Americans	(2) Research on HIV
HIV	0.004***	
	(0.001)	
HIV X Post HAART	0.020***	
	(0.003)	
Black-sounding name		0.004
		(0.007)
Black-sounding name X Post HAART		0.021**
		(0.008)
Vector of racial prob.		X
Vector of racial prob. X Post HAART		X
Observations	589,277	589,277
R-squared	0.006	0.005
LHS (mean)	0.042	0.042
RHS (mean)	0.083	0.083

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 1988 and 2017, with first author affiliated with a US institution, and published in a journal. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the last names of the first author listed on the publication. White-sounding name is the omitted category. All columns include FE for the year in which the grant was awarded. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating equation (1.4). The dependent variable is a dummy = 1 if the article has "Black or African Americans" among its MeSH terms, = 0 otherwise. Post HAART is an indicator = 1 if the award grant year is > 1996, = 0 otherwise. Column (2) reports the result of estimating equation (1.5). The dependent variable is a dummy = 1 if the research has "HIV" among its MeSH terms, = 0 otherwise. The results in column (2) include a vector of controls including Hispanic-sounding name, Asian-sounding name, and American-Native-sounding name. It also includes these controls interacted with the Post HAART indicator. All columns include year FE.

Table 1.8: Inventors with Black-sounding Name Are Less Likely to be Granted a Patent

	(1)	(2)	(3)	(4)	(5)
			Pr(Granted)		
Black-sounding name	-0.060***	-0.064***	-0.037***	-0.037***	-0.029***
	(0.008)	(0.007)	(0.006)	(0.006)	(0.006)
Hispanic-sounding name	-0.052***	-0.049***	-0.033***	-0.034***	-0.031***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Asian-sounding name	0.052***	0.043***	0.010***	0.008***	0.007***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Am. Native-sounding name	-0.118**	-0.097*	-0.112**	-0.128**	-0.124***
	(0.055)	(0.051)	(0.050)	(0.050)	(0.048)
Average grant rate of legal team					0.312***
					(0.003)
Observations	2,009,485	2,009,485	2,009,485	2,009,283	2,009,283
R-squared	0.015	0.057	0.162	0.164	0.178
LHS (mean)	1.000	1.000	1.000	1.000	1.000
RHS (mean)	0.080	0.080	0.080	0.080	0.080

Notes. Robust s.e. clustered by last name in parentheses. The unit of observation is a patent with first inventor resident in the US granted by the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Column (1), I report the results of the estimation of equation (1.8) where y is a dummy = 1 if a patent is granted by the USPTO, = 0 otherwise. y is standardized to have mean 1. Its mean prior to standardization is equal to 0.718. In this specification, I control for year FE. In Column (2), I run the same specification as Column (1), but adding patent subclass FE. In Column (3), I add assignee FE. In column (4), I add state of residence FE. In column (5), I add controls for the "quality" of the legal team that handled the patent application. I control for the total number of lawyers listed on the patent, and on the average grant rate of the legal team, computed as the average of the individual average grant rate of each lawyer listed on the patent.

Table 1.9: Patents Granted to Inventors with Black-sounding Name Have Higher Impact

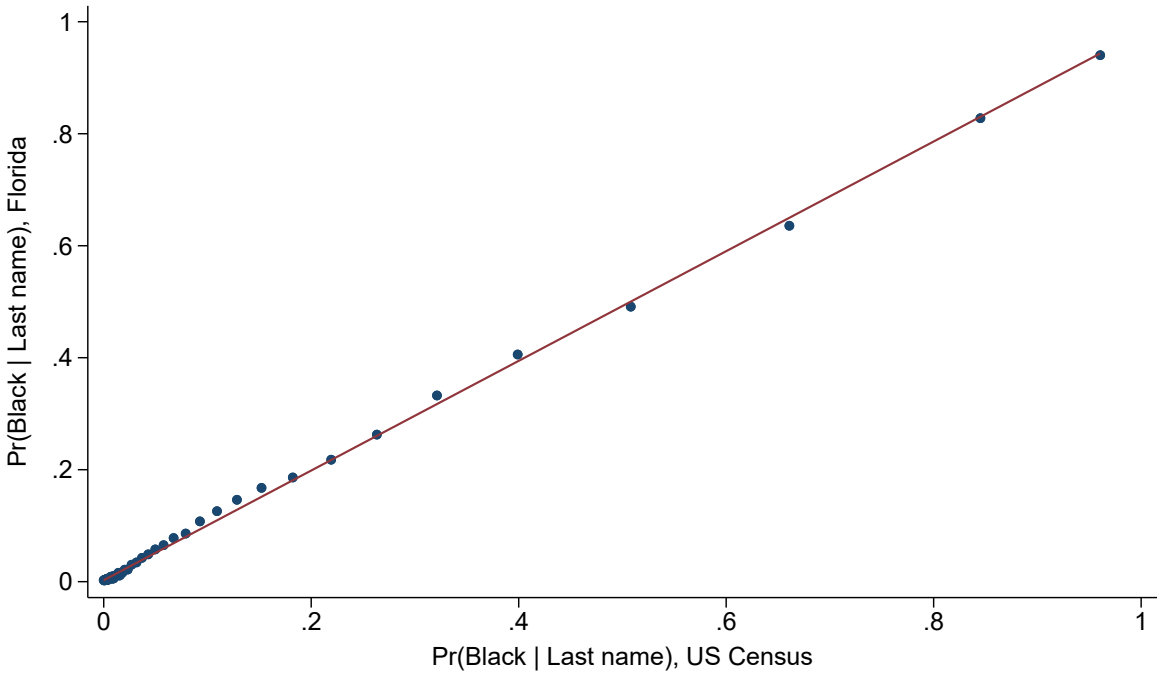
	(1)	(2)	(3)	(4)	(5)
	N. forward citations				
Black-sounding name	0.968*	0.785**	0.812**	0.812**	0.816**
	(0.542)	(0.397)	(0.374)	(0.374)	(0.378)
Hispanic-sounding name	-0.449***	-0.336***	-0.281***	-0.281***	-0.279***
	(0.154)	(0.100)	(0.097)	(0.097)	(0.099)
Asian-sounding name	-0.371***	-0.212***	-0.265***	-0.265***	-0.272***
	(0.098)	(0.048)	(0.047)	(0.047)	(0.047)
Am. Native-sounding name	-0.961	-0.145	0.056	0.056	-0.067
	(1.725)	(1.309)	(1.287)	(1.287)	(1.298)
Average grant rate of legal team					0.317***
					(0.117)
Observations	1,483,345	1,483,345	1,445,735	1,445,735	1,445,735
LHS (mean)	23.676	23.718	23.986	23.986	23.986
RHS (mean)	0.078	0.078	0.078	0.078	0.078

Notes. Robust s.e. clustered by last name in parentheses. The unit of observation is a patent with first inventor resident in the US granted by the the USPTO between 2001 and 2018. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. White-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Column (1), I report the results of the estimation of a equation (1.8) using a Poisson count model where y is the total number of forward citations received by the patent. In this specification, I control for year FE. In Column (2), I run the same specification as Column (1), but adding patent subclass FE. In Column (3), I add assignee FE. In column (4), I add state of residence FE. In column (5), I add controls for the "quality" of the legal team that handled the patent application. I control for the total number of lawyers listed on the patent, and on the average grant rate of the legal team, computed as the average of the individual average grant rate of each lawyer listed on the patent.

1.9 Appendix: Figures and Tables

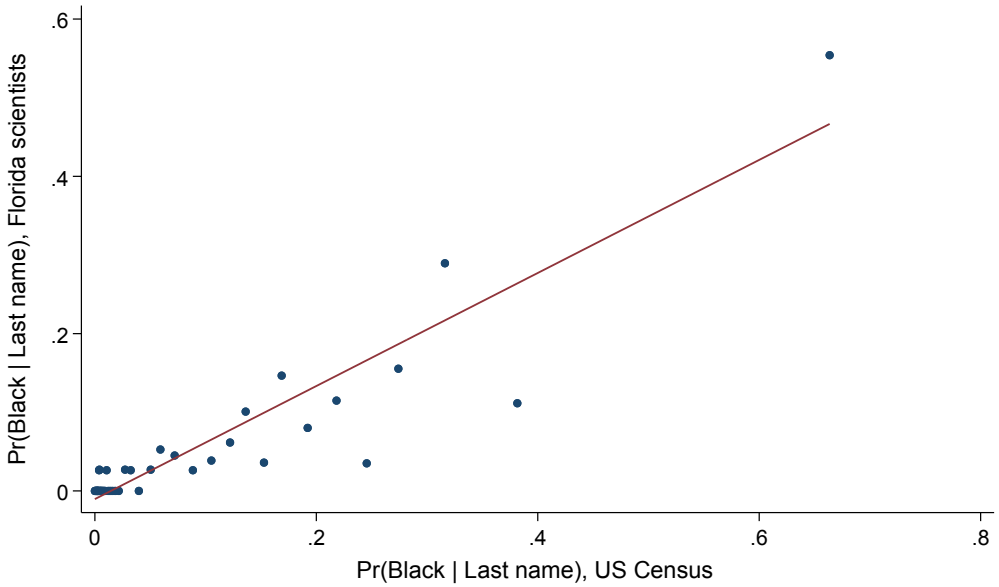
1.9.1 Appendix Figures

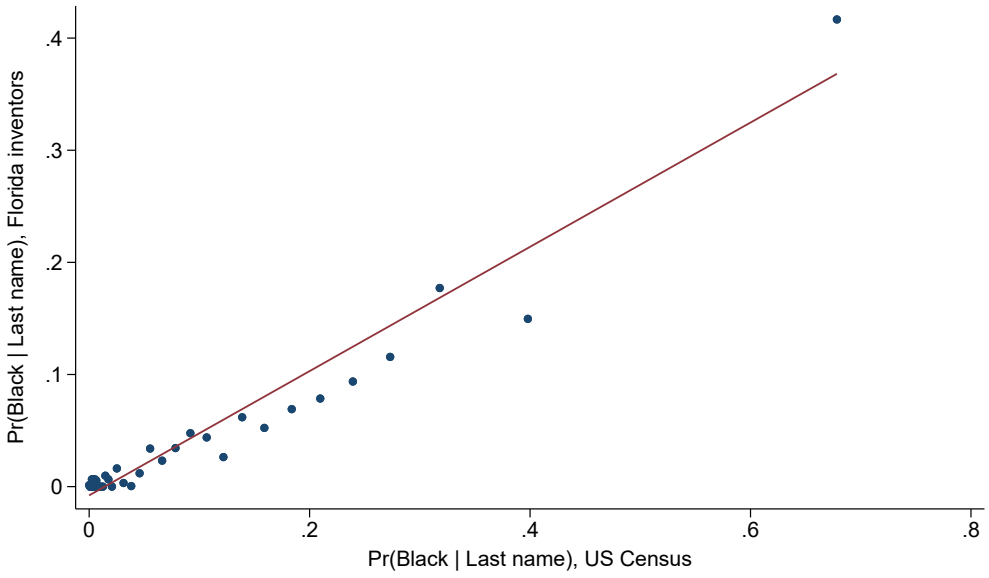
Figure 1.8: Pr(Black | Last name), US Census vs Florida



Notes. One observation is one last name. The underlying data is the sample of registered voters in Florida (Dossi and Morando, 2023) merged with information on racial frequency by last name from Comenetz (2016). This Figure reports a binned scatter plot with Pr(Black | Last name) on the x-axis, and the share of individuals who reported race = Black, by last name, on the y-axis. $\hat{\beta} = 0.979$, robust s.e.= 0.004.

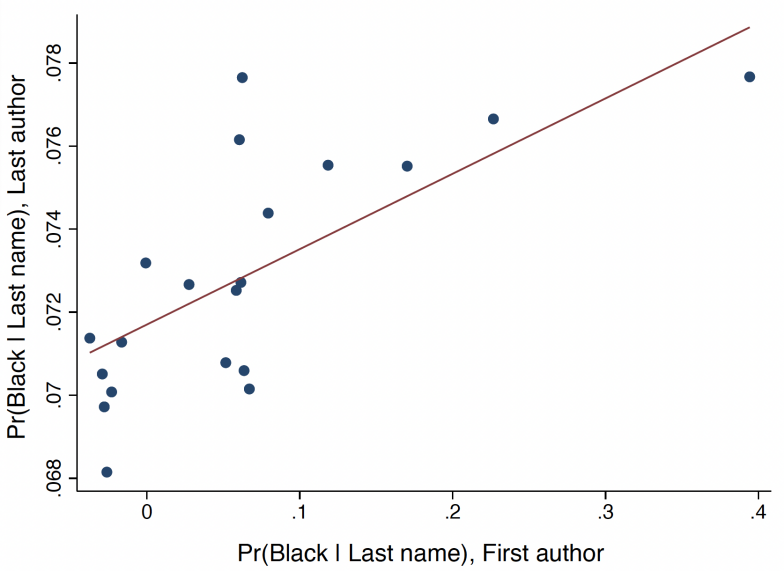
Figure 1.9: Pr(Black | Last name), US Census vs Florida, Scientists and Inventors





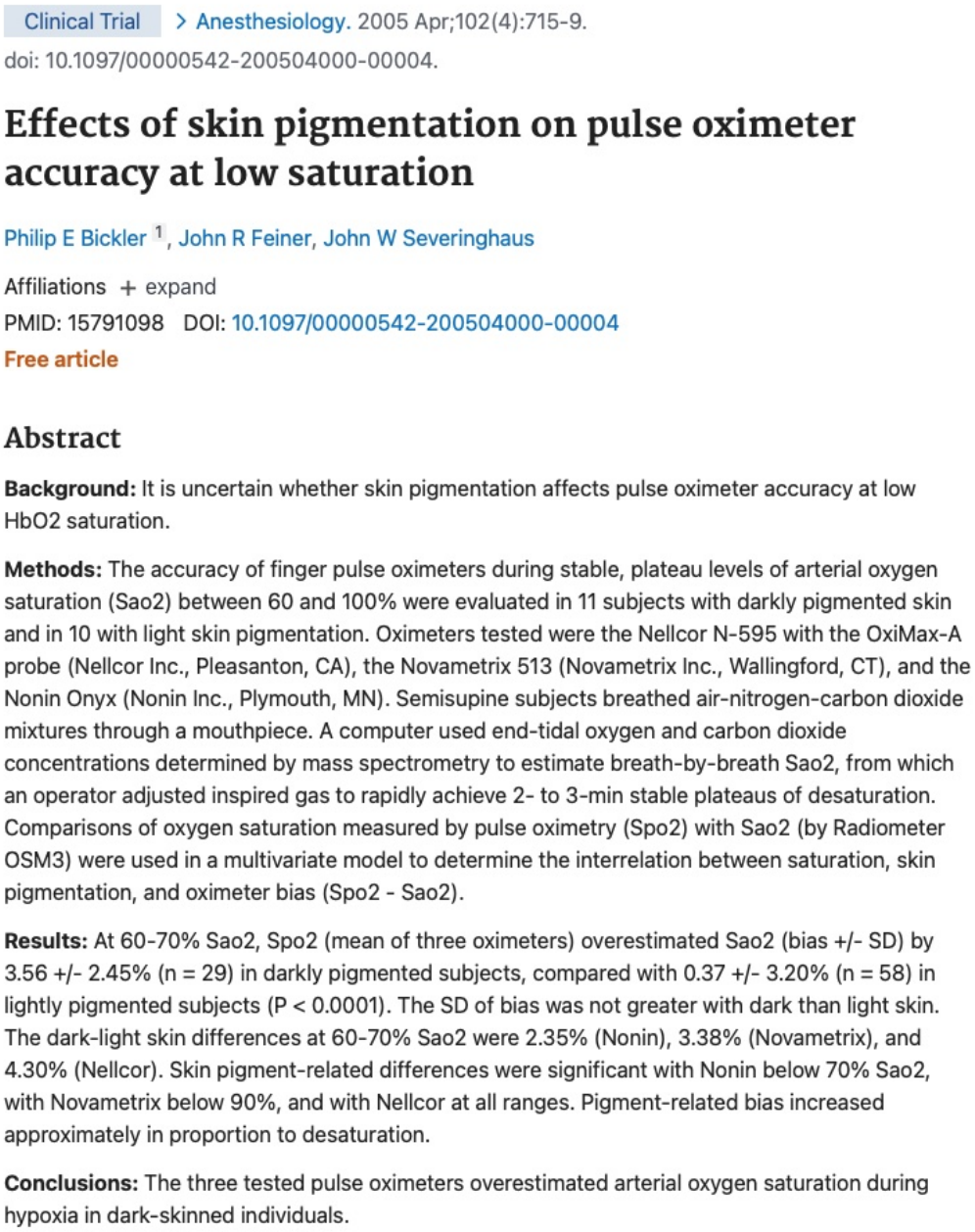
Notes. One observation is one last name. The underlying data is the sample of registered voters in Florida (Dossi and Morando, 2023) who are inventors, merged with information on racial frequency by last name from Comenetz (2016). Details on the match between voter register data and inventors is reported in Dossi and Morando (2023). This Figure reports a binned scatter plot with Pr(Black | Last name) on the x-axis, and the share of inventors who reported race = Black, by last name, on the y-axis.

Figure 1.10: Correlation Between Pr(Black | Last name) of First and Last Author



Notes. This figure displays the correlation between Pr(Black | Last name) of first authors and of last authors in the sample of PubMed publications described in Table 1. The plot includes controls for Pr(Hispanic | Last name), Pr(Asian | Last name), Pr(Am. Native | Last name) of both first and last author. The sample excludes single-authored publications.

Figure 1.11: Example of a PubMed Article with Black or African American Among its MeSH Codes, but Not Classified with a Dictionary Approach



Clinical Trial > Anesthesiology. 2005 Apr;102(4):715-9.
doi: 10.1097/00000542-200504000-00004.

Effects of skin pigmentation on pulse oximeter accuracy at low saturation

Philip E Bickler [1], John R Feiner, John W Severinghaus

Affiliations + expand
PMID: 15791098 DOI: 10.1097/00000542-200504000-00004
Free article

Abstract

Background: It is uncertain whether skin pigmentation affects pulse oximeter accuracy at low HbO2 saturation.

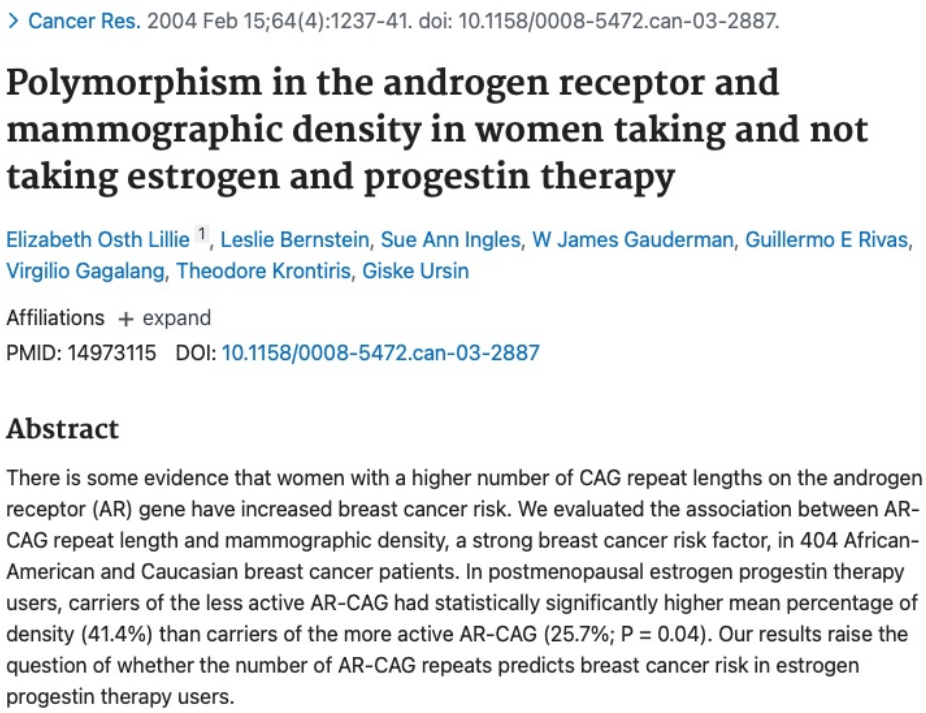
Methods: The accuracy of finger pulse oximeters during stable, plateau levels of arterial oxygen saturation (Sao2) between 60 and 100% were evaluated in 11 subjects with darkly pigmented skin and in 10 with light skin pigmentation. Oximeters tested were the Nellcor N-595 with the OxiMax-A probe (Nellcor Inc., Pleasanton, CA), the Novametrix 513 (Novametrix Inc., Wallingford, CT), and the Nonin Onyx (Nonin Inc., Plymouth, MN). Semisupine subjects breathed air-nitrogen-carbon dioxide mixtures through a mouthpiece. A computer used end-tidal oxygen and carbon dioxide concentrations determined by mass spectrometry to estimate breath-by-breath Sao2, from which an operator adjusted inspired gas to rapidly achieve 2- to 3-min stable plateaus of desaturation. Comparisons of oxygen saturation measured by pulse oximetry (Spo2) with Sao2 (by Radiometer OSM3) were used in a multivariate model to determine the interrelation between saturation, skin pigmentation, and oximeter bias (Spo2 - Sao2).

Results: At 60-70% Sao2, Spo2 (mean of three oximeters) overestimated Sao2 (bias +/- SD) by 3.56 +/- 2.45% (n = 29) in darkly pigmented subjects, compared with 0.37 +/- 3.20% (n = 58) in lightly pigmented subjects (P < 0.0001). The SD of bias was not greater with dark than light skin. The dark-light skin differences at 60-70% Sao2 were 2.35% (Nonin), 3.38% (Novametrix), and 4.30% (Nellcor). Skin pigment-related differences were significant with Nonin below 70% Sao2, with Novametrix below 90%, and with Nellcor at all ranges. Pigment-related bias increased approximately in proportion to desaturation.

Conclusions: The three tested pulse oximeters overestimated arterial oxygen saturation during hypoxia in dark-skinned individuals.

Notes. This is an example of a PubMed article of the dataset which has "Black or African American" among its MeSH codes, but would *not* be classified as such using a dictionary approach. This is because its abstract does not explicitly mention "Black American", "African American", or even "Black".

Figure 1.12: Example of a PubMed Article Classified as Addressing Black or African Americans Using a Dictionary Classification, but Not Classified Using MeSH Codes



> Cancer Res. 2004 Feb 15;64(4):1237-41. doi: 10.1158/0008-5472.can-03-2887.

Polymorphism in the androgen receptor and mammographic density in women taking and not taking estrogen and progestin therapy

Elizabeth Osth Lillie [1], Leslie Bernstein, Sue Ann Ingles, W James Gauderman, Guillermo E Rivas, Virgilio Gagalang, Theodore Krontiris, Giske Ursin

Affiliations + expand

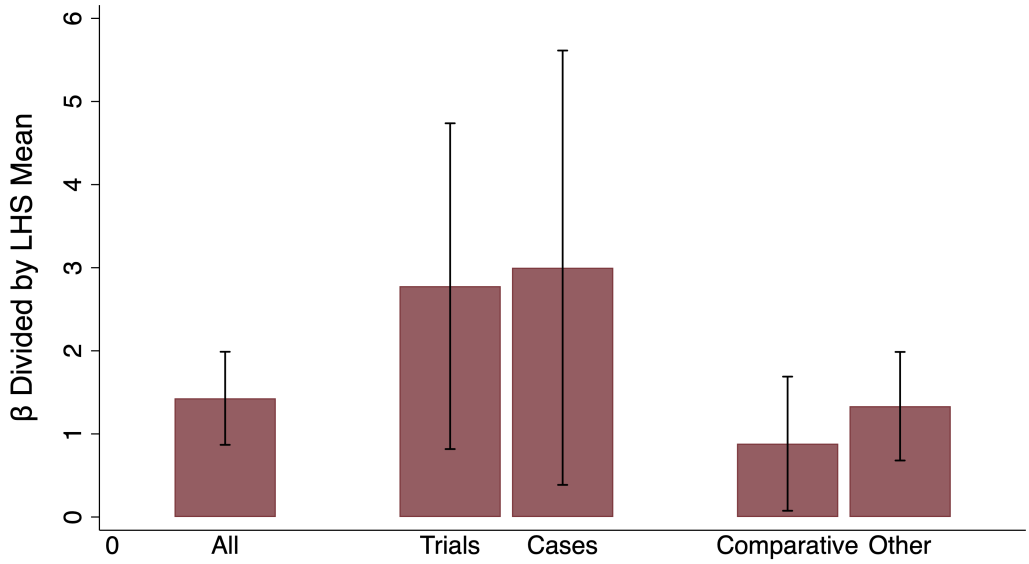
PMID: 14973115 DOI: 10.1158/0008-5472.can-03-2887

Abstract

There is some evidence that women with a higher number of CAG repeat lengths on the androgen receptor (AR) gene have increased breast cancer risk. We evaluated the association between AR-CAG repeat length and mammographic density, a strong breast cancer risk factor, in 404 African-American and Caucasian breast cancer patients. In postmenopausal estrogen progestin therapy users, carriers of the less active AR-CAG had statistically significantly higher mean percentage of density (41.4%) than carriers of the more active AR-CAG (25.7%; P = 0.04). Our results raise the question of whether the number of AR-CAG repeats predicts breast cancer risk in estrogen progestin therapy users.

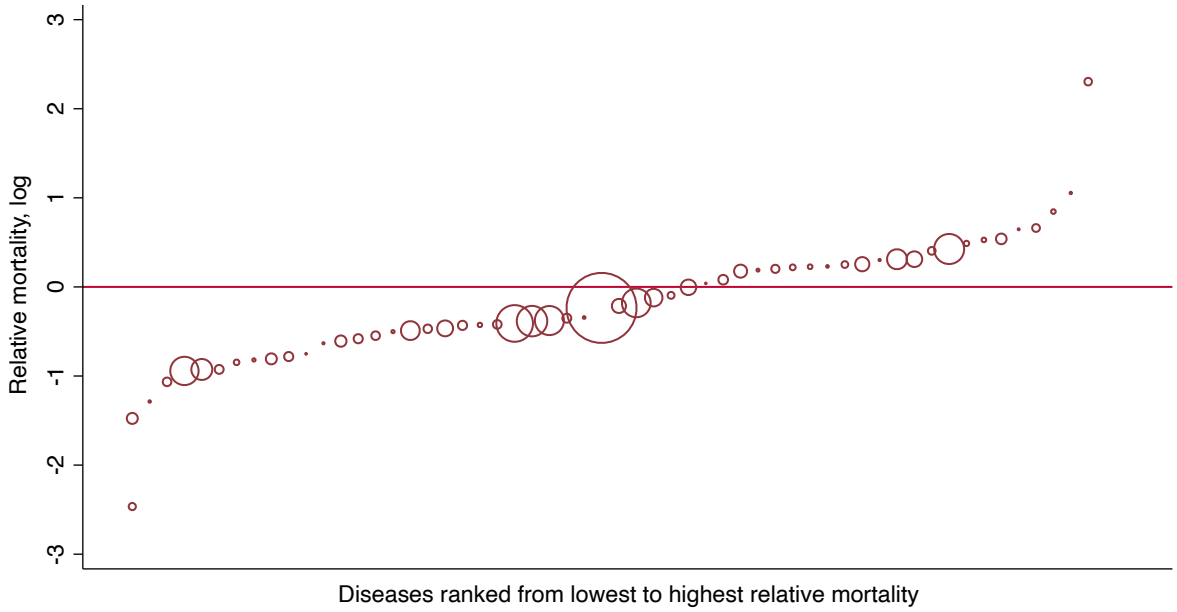
Notes. This is an example of a PubMed article of the dataset which does not have "Black or African American" among its MeSH codes. This article is classified as benefit Black or African Americans because the term "African American" appears in its abstract.

Figure 1.13: Demographics-based Approach, Additional Split by Publication Type



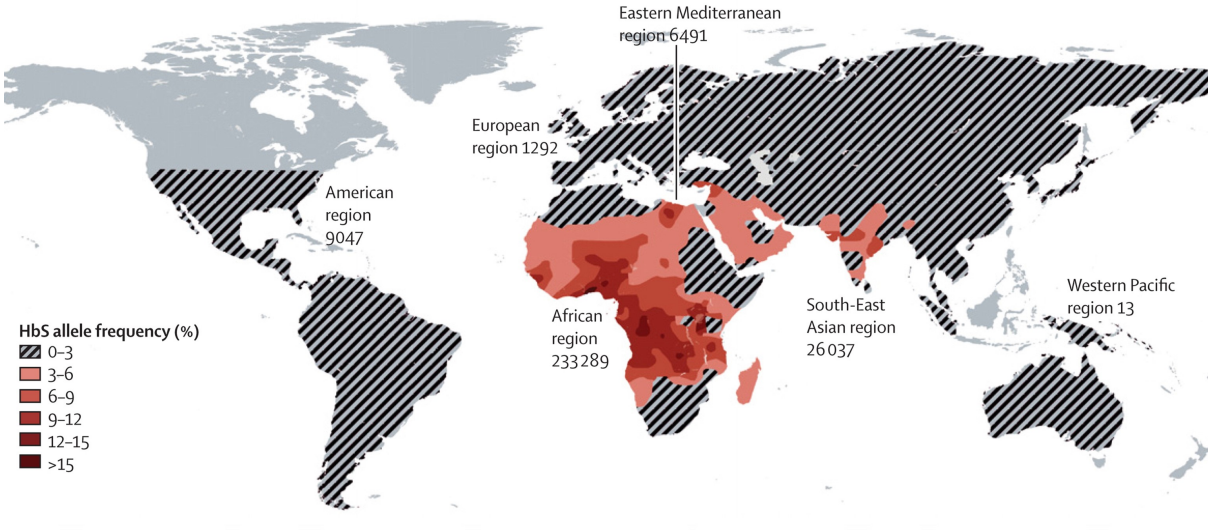
Notes. This figure reports the coefficients of Black-sounding last name for the estimation of equation 1.2 on the sample shown in column (1), Table 1.2, and split by publication type. Publication type is provided by PubMed. In this plot, the first bar on the left reports the coefficient of Black-sounding last name in column (1), Table 1.2, divided by the mean of the dependent variable. The second bar ("Trials") reports the coefficient of Black-sounding last name for equation 1.2 estimated on the subsample of articles linked to a clinical trial. The third bar ("Cases") reports the coefficient of Black-sounding last name for equation 1.2 estimated on the subsample of case reports. The fourth bar ("Comparative") reports the coefficient of Black-sounding last name for equation 1.2 estimated on the subsample of comparative studies. The fifth bar ("Other") reports the coefficient of Black-sounding last name for equation 1.2 estimated on all other articles.

Figure 1.14: Relative Mortality, Weighted by Total Number of Deaths



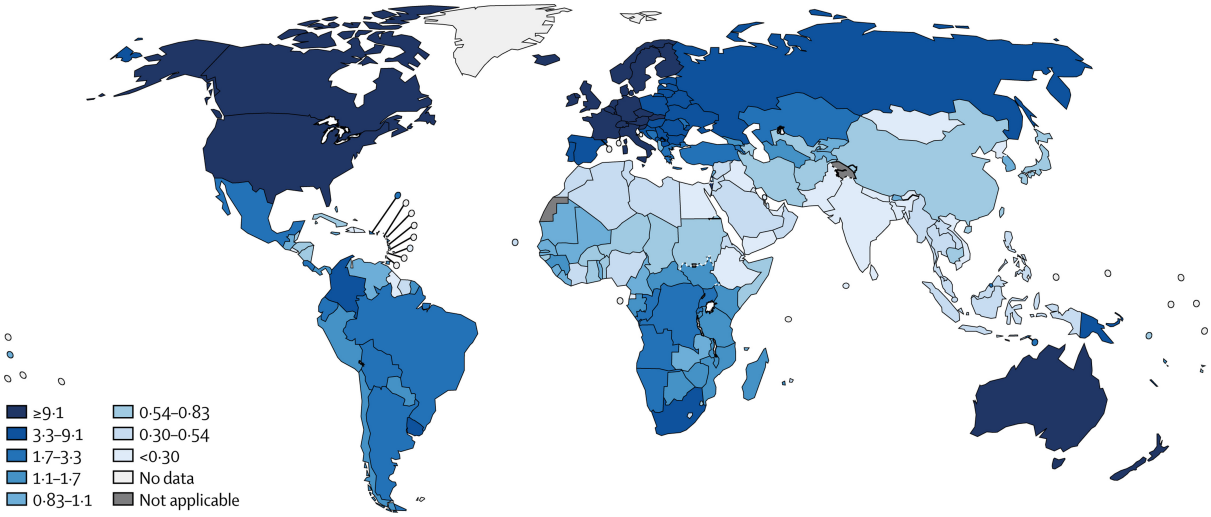
Notes. This figure displays relative mortality rate of Black individuals compared to white individuals in the United States in 2000 to 2015. It is equivalent to Figure 1.3 reported in the main text with the difference that observations are weighted by the total number of deaths by the given disease over the period. One observation is one disease with at least 1,000 related deaths of white individuals and Black individuals over the full period. Data comes from the CDC. More details on the construction of this data in section 1.3.2. On the x-axis, I rank disease by relative mortality rate of Black individuals vs white individuals. On the y-axis is the relative mortality rate of Black individuals compared to white individuals due to the given disease over the period. It is computed as total number of deaths of Black individuals reported due to the given disease divided by the total number of Black individuals in the population, divided by the total number of deaths of white individuals reported due to the given disease divided by the total number of white individuals in the population, in log. The red line corresponds to equal mortality rate for Black individuals compared to white individuals (i.e. when y = 0).

Figure 1.15: Incidence of Sickle Cell Anemia by Region of the World



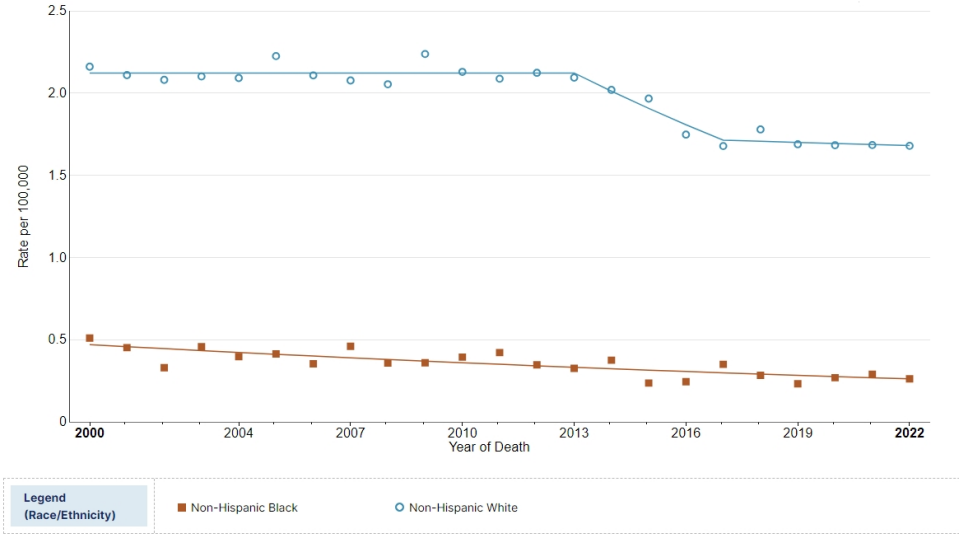
Notes. Incidence of sickle cell anemia by world region (Rees, Williams and Gladwin, 2010).

Figure 1.16: Incidence of Melanoma by Region of the World



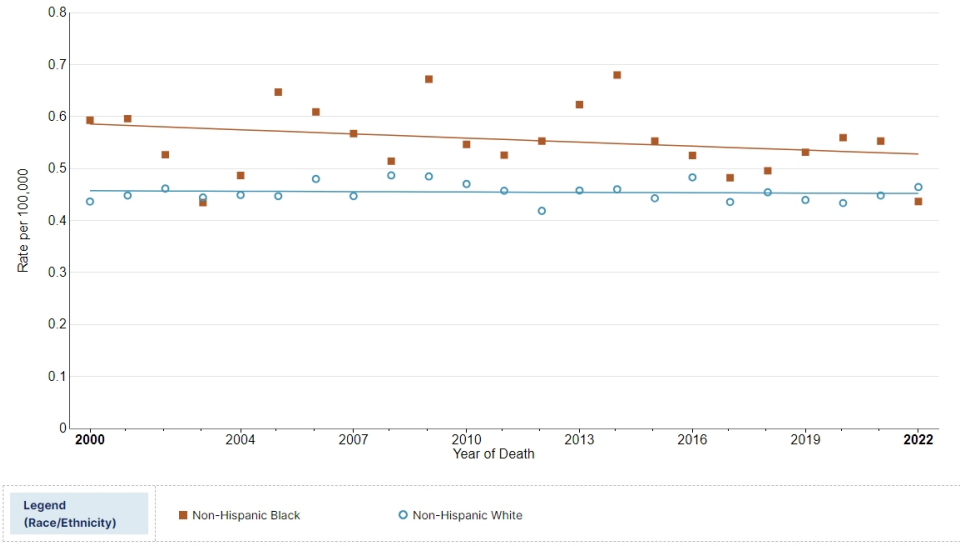
Notes. Incidence of melanoma by world region (Schadendorf, van Akkooi, Berking, Griewank, Gutzmer, Hauschild, Stang, Roesch and Ugurel, 2018).

Figure 1.17: Mortality Rates of White and Black or African Americans, Melanoma



Notes. Source: U.S. Mortality Data (1969-2022), National Center for Health Statistics, CDC.

Figure 1.18: Mortality Rates of White and Black or African Americans, Thyroid Neoplasm



Notes. Source: U.S. Mortality Data (1969-2022), National Center for Health Statistics, CDC.

1.9.2 Appendix Tables

Table 1.10: Extract of Racial Probabilities by Last Name, 2010 Census

Last name	White	Black	Hispanic	Asian	Am. Native
Top 10 last names by Pr(Black | Last name)					
WASHINGTON	5%	91%	3%	0%	1%
JEFFERSON	18%	77%	3%	0%	2%
BOOKER	29%	68%	2%	0%	0%
BANKS	40%	56%	3%	0%	0%
JOSEPH	30%	56%	3%	10%	1%
MOSLEY	42%	55%	2%	0%	1%
JACKSON	41%	55%	3%	0%	1%
CHARLES	35%	54%	8%	1%	2%
DORSEY	43%	54%	2%	0%	0%
RIVERS	42%	52%	3%	1%	2%
Top 10 last names by Pr(White | Last name)					
YODER	98%	0%	1%	0%	0%
FRIEDMAN	97%	0%	2%	1%	0%
KRUEGER	97%	0%	2%	1%	0%
SCHWARTZ	97%	0%	2%	1%	0%
SCHMITT	97%	0%	2%	1%	0%
MUELLER	97%	0%	2%	1%	0%
WEISS	96%	0%	2%	1%	0%
NOVAK	96%	0%	2%	1%	0%
OCONNELL	97%	0%	2%	1%	0%
KLEIN	97%	0%	2%	1%	0%

Notes. This table reports the vector of racial probabilities from the US Census (2010) for the 10 last names with highest probability of being Black, and the 10 last names with highest probability of being white among the 1,000 most frequent last names. *white* refers to % non-Hispanic or Latino white alone; *Black* refers to % non-Hispanic or Latino Black or African American alone; *Hispanic* refers to percent Hispanic or Latino origin; *Asian* refers to percent non-Hispanic or Latino Asian and Native Hawaiian and other Pacific Islander alone; *Am. Native* refers to percent non-Hispanic or Latino American Indian and Alaska Native alone. Probabilities do not always add up to one due to censoring of those cells with less than 100 observation in the population, and due to the residual category *two or more races*. They are rescaled to add up to 1. Data source: Comenetz (2016).

Table 1.11: Validation of Black-sounding Names as Predictor of Race, Florida

	(1)	(2)	(3)	(4)	(5)
			Race = Black		
Black-sounding name	1.021***	1.019***	0.956***	0.937***	0.919***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Asian-sounding name		0.006***	0.013***	0.015***	0.015***
		(0.001)	(0.001)	(0.001)	(0.001)
Hispanic-sounding name		-0.000	-0.027***	-0.035***	-0.046***
		(0.000)	(0.000)	(0.000)	(0.000)
Am. Native-sounding name		0.325***	0.256***	0.234***	0.223***
		(0.007)	(0.007)	(0.007)	(0.007)
Median zip code income (log)			-0.197***	-0.194***	-0.194***
			(0.000)	(0.000)	(0.000)
Average zip code income by last name (log)				-0.066***	-0.150***
				(0.002)	(0.002)
Average zip code income by last name (sd)					0.000***
					(0.000)
Observations	10,167,678	10,167,678	10,133,338	10,133,338	10,126,273
R-squared	0.245	0.246	0.276	0.276	0.276
LHS (mean)	0.141	0.141	0.141	0.141	0.141
RHS (mean)	0.132	0.132	0.132	0.132	0.132

Notes. One observation is a registered voter in the state of Florida in 2017 (Dossi and Morando, 2023). Black-sounding name, Hispanic-sounding name, and Asian-sounding name, and American Native-sounding name are the probability that the individual is Black, Hispanic, Asian, or American Native based on their last name, and is assigned based on the 2010 U.S. Census (Comenetz, 2016). white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Median zip code income (log) is the average median income in the zip code of residence of the voter. Average zip code income by last name (log) and Average zip code income by last name (sd) are, respectively, is the average median zip code income by last name, computed on the population of registered voters in Florida in 2017, and its standard deviation. The dependent variable is a dummy = 1 if the voter's reported race or ethnicity is "Black or African American", = 0 otherwise.

Table 1.12: Differences in Observables Between Matched and Unmatched Scientists

	Matched		Unmatched		Matched-Unmatched	
	Mean (1)	Standard Deviation (2)	Mean (3)	Standard Deviation (4)	Standardized Difference (5)	P-value Equivalence Test (6)
Year	2011.000	4.952	2011.000	4.897	0.000	0.000
Team size	2.723	2.018	2.781	2.022	-0.029	0.000
Granted patent	0.716	0.451	0.719	0.450	-0.007	0.000
Forward citations	23.510	262.600	20.240	141.000	0.013	0.000
Small entity	0.341	0.474	0.320	0.467	0.044	0.000
Assignee	0.293	0.455	0.297	0.457	-0.011	0.000
Female	0.067	0.250	0.081	0.273	-0.055	0.000
Male	0.811	0.391	0.792	0.406	0.049	0.000
Unassigned gender	0.059	0.235	0.074	0.261	-0.063	0.000
Unknown gender	0.063	0.243	0.054	0.225	0.040	0.000
California	0.249	0.433	0.305	0.460	-0.127	1.000
Texas	0.070	0.256	0.059	0.236	0.045	0.000
New York	0.060	0.237	0.072	0.258	-0.049	0.000
Massachusetts	0.048	0.213	0.058	0.233	-0.047	0.000
Washington state	0.042	0.202	0.043	0.203	-0.003	0.000
Michigan	0.039	0.193	0.041	0.198	-0.012	0.000
Illinois	0.036	0.186	0.037	0.189	-0.006	0.000
New Jersey	0.032	0.176	0.042	0.201	-0.056	0.000
Florida	0.031	0.174	0.030	0.171	0.007	0.000
Pennsylvania	0.031	0.173	0.030	0.172	0.004	0.000
Ohio	0.030	0.172	0.026	0.158	0.028	0.000
Minnesota	0.031	0.172	0.025	0.155	0.036	0.000
North Carolina	0.025	0.156	0.019	0.138	0.036	0.000
Colorado	0.022	0.146	0.017	0.130	0.032	0.000
Connecticut	0.018	0.134	0.019	0.138	-0.008	0.000
Georgia	0.019	0.137	0.014	0.119	0.036	0.000
Other states	0.216	0.412	0.163	0.369	0.133	1.000

Notes. One observation is a patent application filed at the USPTO between 2001 and 2018 with first inventors resident in the United States. "Matched" refers to patents whose first inventors has a match in the Census distribution of last names by race. "Unmatched" refers to patents that are not matched. Descriptive statistics (mean and standard deviation) of inventors matched to a last name from Comenetz (2016) (Columns 1 & 2) and unmatched (Columns 3 & 4). Column 5 shows the standardized difference between matched and unmatched in the full sample of inventors. Column 6 reports the largest p-value for the equivalence test of means using a two one-sided t-tests approach. The null hypothesis is that the difference is larger than 10% of a sd, or smaller than 10% of a sd. The share of matched patents is equal to 82%.

Table 1.13: Differences in Observables Between Matched and Unmatched Inventors

	Matched		Unmatched		Matched-Unmatched	
	Mean (1)	Standard Deviation (2)	Mean (3)	Standard Deviation (4)	Standardized Difference (5)	P-value Equivalence Test (6)
Year	2011.000	4.952	2011.000	4.897	0.000	0.000
Team size	2.723	2.018	2.781	2.022	-0.029	0.000
Granted patent	0.716	0.451	0.719	0.450	-0.007	0.000
Forward citations	23.510	262.600	20.240	141.000	0.013	0.000
Small entity	0.341	0.474	0.320	0.467	0.044	0.000
Assignee	0.293	0.455	0.297	0.457	-0.011	0.000
Female	0.067	0.250	0.081	0.273	-0.055	0.000
Male	0.811	0.391	0.792	0.406	0.049	0.000
Unassigned gender	0.059	0.235	0.074	0.261	-0.063	0.000
Unknown gender	0.063	0.243	0.054	0.225	0.040	0.000
California	0.249	0.433	0.305	0.460	-0.127	1.000
Texas	0.070	0.256	0.059	0.236	0.045	0.000
New York	0.060	0.237	0.072	0.258	-0.049	0.000
Massachusetts	0.048	0.213	0.058	0.233	-0.047	0.000
Washington state	0.042	0.202	0.043	0.203	-0.003	0.000
Michigan	0.039	0.193	0.041	0.198	-0.012	0.000
Illinois	0.036	0.186	0.037	0.189	-0.006	0.000
New Jersey	0.032	0.176	0.042	0.201	-0.056	0.000
Florida	0.031	0.174	0.030	0.171	0.007	0.000
Pennsylvania	0.031	0.173	0.030	0.172	0.004	0.000
Ohio	0.030	0.172	0.026	0.158	0.028	0.000
Minnesota	0.031	0.172	0.025	0.155	0.036	0.000
North Carolina	0.025	0.156	0.019	0.138	0.036	0.000
Colorado	0.022	0.146	0.017	0.130	0.032	0.000
Connecticut	0.018	0.134	0.019	0.138	-0.008	0.000
Georgia	0.019	0.137	0.014	0.119	0.036	0.000
Other states	0.216	0.412	0.163	0.369	0.133	1.000

Notes. One observation is a patent application filed at the USPTO between 2001 and 2018 with first inventors resident in the United States. "Matched" refers to patents whose first inventors has a match in the Census distribution of last names by race. "Unmatched" refers to patents that are not matched. Descriptive statistics (mean and standard deviation) of inventors matched to a last name from Comenetz (2016) (Columns 1 & 2) and unmatched (Columns 3 & 4). Column 5 shows the standardized difference between matched and unmatched in the full sample of inventors. Column 6 reports the largest p-value for the equivalence test of means using a two one-sided t-tests approach. The null hypothesis is that the difference is larger than 10% of a sd, or smaller than 10% of a sd. The share of matched patents is equal to 82%.

Table 1.14: Demographics-based Approach, Dictionary Classification

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	Black or African American (dictionary classification on abstract)					
Black-sounding name	0.008***	0.023***	0.007***	0.022***	0.027***	0.021***
	(0.002)	(0.009)	(0.001)	(0.004)	(0.009)	(0.004)
Hispanic-sounding name	-0.000	-0.003	0.000	0.003	-0.002	0.003
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)
Asian-sounding name	-0.002***	0.000	-0.002***	0.002**	0.001	0.002
	(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.002	-0.046**	0.001	-0.004	-0.036*	-0.002
	(0.006)	(0.020)	(0.007)	(0.015)	(0.019)	(0.018)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
R-squared	0.000	0.001	0.000	0.001	0.001	0.001
LHS (mean)	0.007	0.010	0.006	0.017	0.010	0.018
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article abstract contains either the word "African American", or the word "Black" jointly with "race", "racial", or "ethnic", = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.15: Demographics-based Approach, Logit Model

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	1.001***	1.875***	0.916***	1.212***	2.002***	1.094***
	(0.155)	(0.473)	(0.151)	(0.162)	(0.452)	(0.161)
Hispanic-sounding name	0.001	0.108	-0.005	0.148	0.111	0.131
	(0.119)	(0.348)	(0.123)	(0.131)	(0.388)	(0.136)
Asian-sounding name	-0.527***	-0.234	-0.535***	-0.014	-0.306	-0.038
	(0.057)	(0.178)	(0.058)	(0.059)	(0.209)	(0.061)
Am. Native-sounding name	-0.558	-0.474	-0.556	-0.739	-0.289	-0.905
	(1.095)	(3.509)	(1.076)	(1.166)	(3.439)	(1.162)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. I estimate logit models in columns (1) through (6). The table reports exp(coefficients).

Table 1.16: Demographics-based Approach, Controlling for Gender of First Author

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.011***	0.026***	0.010***	0.028***	0.029***	0.027***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.010)	(0.005)
Hispanic-sounding name	0.000	0.003	0.000	0.005*	0.003	0.005
	(0.001)	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.003***	-0.002	-0.003***	-0.001	-0.003	-0.001
	(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	(0.001)
Am. Native-sounding name	-0.007	-0.006	-0.007	-0.018	-0.008	-0.022
	(0.008)	(0.034)	(0.007)	(0.018)	(0.035)	(0.018)
Female dummy	0.004***	0.007***	0.004***	0.014***	0.008***	0.015***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Article on human subjects				X	X	X
Observations	539,707	38,032	501,675	177,170	30,777	146,393
R-squared	0.001	0.003	0.001	0.003	0.003	0.003
LHS (mean)	0.008	0.010	0.008	0.019	0.010	0.021
RHS (mean)	0.069	0.074	0.069	0.074	0.074	0.074

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Female dummy is a variable = 1 if the gender of the first author estimated from their first name is female, = 0 otherwise. The omitted category is male first author. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.17: Demographics-based Approach, Controlling for Journal FE, State FE, Affiliation FE

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Panel A: Journal FE						
Black-sounding name	0.011***	0.024***	0.010***	0.029***	0.027***	0.029***
	(0.002)	(0.009)	(0.002)	(0.004)	(0.009)	(0.005)
Observations	630,029	41,398	588,556	194,839	33,331	161,439
R-squared	0.031	0.031	0.035	0.036	0.033	0.042
LHS (mean)	0.008	0.010	0.007	0.020	0.010	0.022
RHS (mean)	0.063	0.072	0.063	0.071	0.072	0.071
Panel B: State FE						
Black-sounding name	0.009***	0.025***	0.008***	0.008***	0.029***	0.025***
	(0.002)	(0.009)	(0.002)	(0.002)	(0.009)	(0.005)
Observations	649,249	41,748	607,501	607,501	33,629	164,912
R-squared	0.002	0.005	0.002	0.002	0.006	0.006
LHS (mean)	0.007	0.010	0.007	0.007	0.010	0.021
RHS (mean)	0.063	0.072	0.063	0.063	0.072	0.071
Panel C: Affiliation of first author FE						
Black-sounding name	0.008***	0.047**	0.006***	0.031***	0.052***	0.027***
	(0.002)	(0.018)	(0.002)	(0.007)	(0.019)	(0.007)
Observations	342,843	17,353	325,379	89,198	13,659	75,425
R-squared	0.008	0.033	0.008	0.020	0.047	0.021
LHS (mean)	0.007	0.011	0.007	0.021	0.012	0.023
RHS (mean)	0.062	0.072	0.062	0.071	0.073	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include year FE and a vector of racial frequencies. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table 1.18: Demographics-based Approach, Alternative Definitions of Black-sounding Last Name

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
			MeSH: Black or African American			
Panel A: Black-sounding last name of first and last author						
Black-sounding name	0.009***	0.026**	0.008***	0.025***	0.030***	0.023***
	(0.002)	(0.011)	(0.002)	(0.005)	(0.011)	(0.005)
Black-sounding name (Last author)	0.008***	0.017***	0.008***	0.025***	0.019**	0.025***
	(0.002)	(0.006)	(0.002)	(0.004)	(0.007)	(0.005)
Article on human subjects				X	X	X
Observations	501,771	29,063	472,708	149,926	23,309	126,617
R-squared	0.001	0.002	0.001	0.001	0.003	0.001
LHS (mean)	0.007	0.009	0.007	0.019	0.010	0.021
RHS (mean)	0.064	0.072	0.064	0.073	0.073	0.073
Panel B: Average of Pr(Black | Last name) across all authors						
Black-sounding name	0.010***	0.024***	0.009***	0.029***	0.027***	0.028***
	(0.002)	(0.009)	(0.002)	(0.005)	(0.009)	(0.005)
Article on human subjects				X	X	X
Observations	621,326	40,962	580,364	189,148	32,987	156,161
R-squared	0.001	0.001	0.001	0.001	0.001	0.001
LHS (mean)	0.008	0.010	0.007	0.020	0.010	0.022
RHS (mean)	0.063	0.072	0.063	0.072	0.073	0.071
Panel C: Dummy for Pr(Black | Last name) ≥ 0.5						
Black-sounding name (Binary)	0.011***	0.012	0.011***	0.026***	0.014	0.028***
	(0.002)	(0.012)	(0.002)	(0.006)	(0.012)	(0.007)
Article on human subjects				X	X	X
Observations	638,409	41,348	597,061	195,895	33,322	162,573
R-squared	0.001	0.001	0.001	0.001	0.001	0.001
LHS (mean)	0.007	0.010	0.007	0.019	0.010	0.021
RHS (mean)	0.009	0.009	0.009	0.010	0.009	0.010
Panel D: Restricting sample to last names present in the 1930 Census						
Black-sounding name	0.009***	0.022***	0.008***	0.026***	0.027***	0.026***
	(0.002)	(0.007)	(0.002)	(0.005)	(0.008)	(0.005)
Article on human subjects				X	X	X
Observations	507,339	35,656	471,683	164,658	28,763	135,895
R-squared	0.000	0.001	0.000	0.001	0.001	0.001
LHS (mean)	0.008	0.010	0.008	0.020	0.011	0.022
RHS (mean)	0.076	0.080	0.075	0.081	0.080	0.081

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include year FE and a vector of racial frequencies. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table 1.19: Demographics-based Approach, Including All Articles (Not Only Top 1,000 Journals by Journal Commercialization Impact Factor (JCIF)

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
			MeSH: Black or African American			
Black-sounding name	0.034***	0.051***	0.033***	0.045***	0.053***	0.045***
	(0.001)	(0.006)	(0.001)	(0.002)	(0.007)	(0.002)
Hispanic-sounding name	0.004***	0.002	0.004***	0.006***	0.003	0.006***
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Asian-sounding name	-0.003***	-0.001	-0.002***	-0.002***	-0.002	-0.002***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Am. Native-sounding name	-0.006	-0.021	-0.005	-0.008	-0.022	-0.008
	(0.007)	(0.017)	(0.008)	(0.009)	(0.017)	(0.010)
Article on human subjects				X	X	X
Observations	2,274,170	99,314	2,174,856	1,624,717	95,828	1,528,889
R-squared	0.002	0.004	0.002	0.003	0.004	0.003
LHS (mean)	0.009	0.014	0.009	0.013	0.015	0.013
RHS (mean)	0.072	0.077	0.071	0.074	0.077	0.074

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.20: Demographics-based Approach, Split by Below/Above Median Journal Commercialization Impact Factor

	All	Linked to trial	Not Linked to trial	All	Linked to trial	Not Linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample						
Black-sounding name	0.015***	0.037***	0.012***	0.026***	0.040***	0.023***
	(0.003)	(0.012)	(0.003)	(0.005)	(0.011)	(0.005)
Observations	253,951	22,606	231,345	122,497	18,983	103,514
R-squared	0.001	0.003	0.001	0.001	0.003	0.001
LHS (mean)	0.010	0.010	0.010	0.018	0.010	0.019
RHS (mean)	0.067	0.072	0.067	0.072	0.072	0.072
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample						
Black-sounding name	0.007***	0.011	0.007***	0.033***	0.015	0.037***
	(0.002)	(0.008)	(0.002)	(0.007)	(0.010)	(0.008)
Observations	397,302	19,338	377,964	76,958	14,810	62,148
R-squared	0.001	0.002	0.001	0.001	0.002	0.001
LHS (mean)	0.006	0.009	0.005	0.022	0.010	0.025
RHS (mean)	0.061	0.071	0.060	0.070	0.072	0.070

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is further restricted to those articles linked to a patent (Marx and Fuegi; Marx and Fuegi, 2020; 2022). Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.21: Demographics-based Approach, Articles Linked to a Patent

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Black-sounding name	0.004**	0.008	0.004*	0.025***	0.008	0.030***
	(0.002)	(0.007)	(0.002)	(0.009)	(0.008)	(0.012)
Hispanic-sounding name	-0.001	-0.004***	-0.001	-0.005	-0.004***	-0.006
	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.006)
Asian-sounding name	-0.002***	-0.000	-0.002***	-0.001	-0.003**	-0.003
	(0.000)	(0.002)	(0.000)	(0.002)	(0.001)	(0.002)
Am. Native-sounding name	0.000	0.000	-0.000	0.004	-0.016	-0.000
	(0.007)	(0.024)	(0.007)	(0.032)	(0.018)	(0.038)
Article on human subjects				X	X	X
Observations	184,146	10,057	174,089	29,140	7,348	21,792
R-squared	0.000	0.001	0.000	0.002	0.002	0.002
LHS (mean)	0.003	0.004	0.003	0.014	0.003	0.017
RHS (mean)	0.058	0.070	0.058	0.067	0.071	0.066

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is further restricted to those articles linked to a patent (Marx and Fuegi; Marx and Fuegi, 2020; 2022). Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.22: Demographics-based Approach, Controlling for MeSH Codes Linked to Article

	All articles		Linked to trial		Not linked to trial	
	(1)	(2)	(3)	(4)	(5)	(6)
	Black or African American					
Black-sounding name	0.011***	0.010***	0.029***	0.026**	0.010***	0.008***
	(0.002)	(0.002)	(0.011)	(0.010)	(0.002)	(0.002)
Hispanic-sounding name	-0.000	0.000	0.002	0.002	-0.000	0.000
	(0.001)	(0.001)	(0.004)	(0.003)	(0.001)	(0.001)
Asian-sounding name	-0.004***	-0.001**	-0.001	-0.001	-0.004***	-0.001***
	(0.000)	(0.000)	(0.002)	(0.001)	(0.000)	(0.000)
Am. Native-sounding name	-0.005	-0.009	-0.015	-0.016	-0.004	-0.009
	(0.009)	(0.008)	(0.031)	(0.028)	(0.008)	(0.008)
MeSH code FE		X		X		X
Observations	8,668,848	8,667,886	616,278	613,189	8,052,570	8,051,509
R-squared	0.001	0.136	0.002	0.125	0.001	0.139
LHS (mean)	0.009	0.009	0.011	0.011	0.009	0.009
RHS (mean)	0.062	0.062	0.071	0.071	0.062	0.062

Notes. Robust s.e. in parentheses clustered by last name and publication id. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor, with associated MeSH code. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1) and (2) report the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Columns (3) and (4) report the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Columns (5) and (6) report the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (2), (4), and (6) include MeSH FE.

Table 1.23: Demographics-based Approach, Research on Other Demographic Groups

Panel A: Research on Black or African American Individuals

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American (no other race/ethnicity)					
Black-sounding name	0.007***	0.021**	0.006***	0.020***	0.024***	0.019***
	(0.002)	(0.009)	(0.001)	(0.004)	(0.009)	(0.004)
Hispanic-sounding name	-0.001	0.001	-0.001	-0.001	0.000	-0.001
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.002)
Asian-sounding name	-0.002***	-0.002	-0.002***	-0.001	-0.002*	-0.001
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.005	0.007	0.005	0.013	0.007	0.014

Panel B: Research on Hispanic or Latino Individuals

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	MeSH: Hispanic or Latino (no other race/ethnicity)					
Black-sounding name	0.001	0.004	0.001	0.004*	0.008	0.003
	(0.001)	(0.004)	(0.001)	(0.002)	(0.005)	(0.002)
Hispanic-sounding name	0.004***	0.005	0.004***	0.014***	0.008*	0.016***
	(0.001)	(0.004)	(0.001)	(0.003)	(0.004)	(0.003)
Asian-sounding name	-0.001***	-0.000	-0.001***	0.001	0.001	0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.003	0.006	0.003	0.008	0.006	0.009

Panel C: Research on Asian American Individuals

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	MeSH: Asian American (no other race/ethnicity)					
Black-sounding name	0.000	-0.002	0.001	-0.001	-0.001	-0.001
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.002)
Hispanic-sounding name	-0.000	-0.002	-0.000	0.001	-0.002	0.001
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Asian-sounding name	0.000*	0.009***	0.000	0.006***	0.010***	0.006***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	0.004	0.006	0.004	0.009	0.006	0.009

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns include controls for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. In Panel A, Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article if the article has "Black or African American" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Hispanic or Latino" or "Asian American" among its MeSh codes. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects. Panel B is identical to Panel A, but the dependent variable is a dummy = 1 if the article if the article has "Hispanic or Latino" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Black or African American" or "Asian American" among its MeSh codes. Panel C is identical to Panel A, but the dependent variable is a dummy = 1 if the article if the article has "Asian American" among its MeSH codes, = 0 otherwise. The dummy is = 0 if the article has either "Black or African American" or "Hispanic or Latino" among its MeSh codes.

Table 1.24: Demographics-based Approach: Evidence from Ongoing Clinical Trials

	(1) Black or African American	(2) Hispanic or Latino	(3) Asian
Black-sounding name	0.041***	0.007	0.005
	(0.015)	(0.006)	(0.004)
Hispanic-sounding name	0.003	0.024***	0.004
	(0.007)	(0.008)	(0.004)
Asian-sounding name	-0.004	-0.001	0.006***
	(0.003)	(0.002)	(0.002)
Observations	12,366	12,366	12,366
LHS (mean)	0.012	0.007	0.004
Black-sounding name (mean)	0.074	0.074	0.074
Hispanic-sounding name (mean)	0.065	0.065	0.065
Asian-sounding name (mean)	0.195	0.195	0.195

Notes. Robust s.e. in parentheses clustered by Pr(Black | Black-sounding last name). The unit of observation is an ongoing clinical trial registered between 2001 and 2018 with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to the average race and ethnicity across the names of all investigators listed on the trial. Column (1) reports the result of estimating equation (1.2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Black, = 0 otherwise. Column (2) reports the result of estimating equation (1.2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Hispanics or Latinos, = otherwise. Column (3) reports the result of estimating equation (1.2) where the dependent variable is a dummy = 1 if the description of the clinical trial mentions Asians, = 0 otherwise. This data comes from the web portal `Clinicaltrials.gov`. All columns include trial registration year FE.

Table 1.25: Frequency-based Approach, Dictionary Classification

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.181***	-0.039***	-0.014	0.052**
	(0.050)	(0.011)	(0.022)	(0.021)
Hispanic-sounding name	-0.002	0.004	-0.001	-0.003
	(0.034)	(0.009)	(0.014)	(0.013)
Asian-sounding name	-0.004	-0.001	0.016**	-0.016***
	(0.015)	(0.004)	(0.006)	(0.006)
Am. Native sounding name	-0.118	-0.073	0.058	0.015
	(0.282)	(0.054)	(0.162)	(0.146)
Observations	90,929	90,929	90,929	90,929
R-squared	0.003	0.001	0.001	0.002
LHS (mean)	0.126	0.085	0.603	0.312
RHS (mean)	0.066	0.066	0.066	0.066

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. Differently from Table 1.3 in the main text, here articles are linked to diseases through a dictionary classification on the article abstract.

Table 1.26: Frequency-based Approach (Logit Model)

	(1) Typically White	(2) Similar incidence	(3) Typically Black
Black-sounding name	0.801***	0.986	1.428***
	(0.066)	(0.077)	(0.145)
Hispanic-sounding name	0.905***	1.229***	0.859***
	(0.020)	(0.027)	(0.027)
Asian-sounding name	0.975	1.002	1.031
	(0.061)	(0.059)	(0.073)
Am. Native-sounding name	1.290	0.861	0.860
	(0.639)	(0.512)	(0.602)
Total mortality, log	0.839***	2.024***	0.477***
	(0.005)	(0.017)	(0.004)
Observations	138,657	138,657	138,657
LHS (mean)	0.331	0.488	0.181
RHS (mean)	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the results of estimating Equation 1.3 using a logit model where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (2) reports the results of estimating equation (1.3) using a logit model where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (1.3) using a logit model where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. The table reports exp(coefficients).

Table 1.27: Frequency-based Approach, Controlling for Gender of First Author

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Black-sounding name	0.136***	-0.056***	0.006	0.050***
	(0.037)	(0.020)	(0.020)	(0.014)
Hispanic-sounding name	-0.006	-0.003	0.009	-0.005
	(0.028)	(0.016)	(0.015)	(0.009)
Asian-sounding name	-0.003	-0.019***	0.039***	-0.019***
	(0.011)	(0.006)	(0.006)	(0.004)
Am. Native-sounding name	-0.269	0.098	-0.025	-0.073
	(0.191)	(0.125)	(0.123)	(0.079)
Total mortality, log	-0.076***	-0.012***	0.087***	-0.075***
	(0.004)	(0.002)	(0.002)	(0.001)
Female dummy	0.053***	-0.062***	0.055***	0.007**
	(0.009)	(0.005)	(0.005)	(0.003)
Observations	118,008	118,008	118,008	118,008
R-squared	0.011	0.004	0.040	0.063
LHS (mean)	-0.126	0.411	0.460	0.129
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Female dummy is a variable equal to 1 if the gender of the first author assigned based on their last name if female, = 0 otherwise. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 1.28: Frequency-based Approach, Controlling for Journal FE, State FE, First Author Affiliation FE

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Panel A: Journal FE				
Black-sounding name	0.083***	-0.053***	0.030**	0.023**
	(0.026)	(0.014)	(0.012)	(0.011)
Observations	138,601	138,601	138,601	138,601
LHS (mean)	-0.039	0.331	0.488	0.181
Panel B: State FE				
Black-sounding name	0.116***	-0.047***	0.002	0.045***
	(0.034)	(0.017)	(0.016)	(0.014)
Observations	138,171	138,171	138,171	138,171
LHS (mean)	-0.039	0.331	0.488	0.181
Panel C: Affiliation of first author FE				
Black-sounding name	0.095**	-0.056**	0.018	0.039**
	(0.043)	(0.023)	(0.023)	(0.017)
Observations	70,072	70,072	70,072	70,072
LHS (mean)	-0.027	0.332	0.483	0.185

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns control for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name, which refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns also control for the total of total number of deaths (Black + white Americans) due to the given disease. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the first author's affiliation. In Panel C, all columns include affiliation of first author FE.

Table 1.29: Frequency-based Approach, Alternative Definitions of Black-sounding Last Name

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Panel A: Black-sounding last name of first and last author				
Black-sounding name	0.103***	-0.039**	0.001	0.038**
	(0.037)	(0.019)	(0.018)	(0.015)
Black-sounding name (Last author)	0.102***	-0.035**	-0.034**	0.069***
	(0.034)	(0.016)	(0.016)	(0.013)
Observations	100,029	100,029	100,029	100,029
R-squared	0.030	0.015	0.161	0.142
LHS (mean)	-0.046	0.331	0.490	0.178
RHS (mean)	0.064	0.064	0.064	0.064
Panel B: Average of Pr(Black | Last name) across all authors				
Black-sounding name	0.115***	-0.045**	-0.002	0.047***
	(0.035)	(0.018)	(0.017)	(0.014)
Observations	133,465	133,465	133,465	133,465
R-squared	0.032	0.015	0.163	0.144
LHS (mean)	-0.040	0.331	0.489	0.180
RHS (mean)	0.064	0.064	0.064	0.064
Panel C: Dummy for Pr(Black | Last name) ≥ 0.5				
Black sounding name (Dummy)	0.105***	-0.050***	0.021	0.029**
	(0.032)	(0.018)	(0.020)	(0.012)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.015	0.164	0.144
LHS (mean)	-0.039	0.331	0.488	0.181
RHS (mean)	0.009	0.009	0.009	0.009
Panel D: Restricting sample to last names present in 1930 Census				
Black-sounding name	0.093**	-0.031	-0.013	0.044***
	(0.039)	(0.020)	(0.019)	(0.016)
Observations	107,551	107,551	107,551	107,551
R-squared	0.031	0.016	0.167	0.144
LHS (mean)	-0.037	0.333	0.481	0.185
RHS (mean)	0.076	0.076	0.076	0.076

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. All columns control for Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name, which refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns also control for the total of total number of deaths (Black + white Americans) due to the given disease. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. In Panel A, all columns include journal FE. In Panel B, all columns include State FE, where state refers to the US state of the affiliation of the first author. In Panel C, all columns include affiliation of first author FE.

Table 1.30: Frequency-based Approach, Including All Articles (Not Only Top 1,000 Journals by Journal Commercialization Impact Factor (JCIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar Incidence	(4) Typically Black
Black-sounding name	0.201***	-0.074***	-0.007	0.081***
	(0.029)	(0.012)	(0.012)	(0.012)
Hispanic-sounding name	-0.011	-0.003	0.007	-0.004
	(0.020)	(0.010)	(0.009)	(0.008)
Asian-sounding name	-0.093***	-0.004	0.062***	-0.058***
	(0.008)	(0.004)	(0.003)	(0.003)
Am. Native-sounding name	-0.131	0.014	0.085	-0.099
	(0.171)	(0.078)	(0.073)	(0.073)
Total mortality, log	-0.125***	-0.039***	0.144***	-0.105***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	403,985	403,985	403,985	403,985
R-squared	0.044	0.017	0.191	0.152
LHS (mean)	0.045	0.301	0.482	0.217
RHS (mean)	0.070	0.070	0.070	0.070

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black Americans compared to white Americans, calculated according to equation 1.1. Column (2) reports the results of estimating equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white Americans compared to Black Americans. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black Americans compared to white Americans, and less than 1.5 among white Americans compared to Black Americans. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 1.31: Frequency-based Approach, Split by Below/Above Journal Commercialization Impact Factor (JCIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample				
Black-sounding name	0.091**	-0.059***	0.034*	0.024
	(0.038)	(0.020)	(0.020)	(0.016)
Observations	60,958	60,958	60,958	60,958
R-squared	0.019	0.023	0.141	0.102
LHS (mean)	-0.091	0.341	0.499	0.161
RHS (mean)	0.067	0.067	0.067	0.067
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample				
Black-sounding name	0.148***	-0.037	-0.036*	0.073***
	(0.048)	(0.023)	(0.020)	(0.019)
Observations	77,699	77,699	77,699	77,699
R-squared	0.042	0.011	0.184	0.178
LHS (mean)	0.002	0.323	0.479	0.198
RHS (mean)	0.061	0.061	0.061	0.061

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is restricted to those articles linked to a patent (Marx and Fuegi; Marx and Fuegi, 2020; 2022). Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name is taken from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1.1. Column (2) reports the results of estimating Equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 1.32: Frequency-based Approach: Articles Linked to a Patent

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding	0.106*	-0.043	-0.016	0.059***
	(0.059)	(0.030)	(0.027)	(0.022)
Hispanic-sounding	-0.046	0.003	0.010	-0.013
	(0.043)	(0.023)	(0.021)	(0.015)
Asian-sounding	0.004	-0.031***	0.055***	-0.024***
	(0.015)	(0.007)	(0.007)	(0.006)
Am. Native-sounding	-0.563	0.446*	-0.214	-0.233
	(0.403)	(0.261)	(0.232)	(0.170)
Total mortality, log	-0.118***	-0.040***	0.154***	-0.114***
	(0.004)	(0.002)	(0.002)	(0.002)
Observations	41,357	41,357	41,357	41,357
LHS (mean)	-0.070	0.350	0.481	0.169
RHS (mean)	0.058	0.058	0.058	0.058

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. The sample is restricted to those articles linked to a patent (Marx and Fuegi; Marx and Fuegi, 2020; 2022). Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name is taken from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1.1. Column (2) reports the results of estimating Equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals.

Table 1.33: Frequency-based Approach: Threshold Set at 1.3 Higher Mortality

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.047**	0.010	0.036**
	(0.034)	(0.019)	(0.017)	(0.017)
Hispanic-sounding name	0.006	-0.016	0.005	0.011
	(0.024)	(0.014)	(0.012)	(0.012)
Asian-sounding name	-0.006	-0.009*	0.040***	-0.031***
	(0.009)	(0.005)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.082	-0.051	-0.031
	(0.177)	(0.118)	(0.129)	(0.118)
Total mortality, log	-0.108***	0.023***	0.059***	-0.082***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.005	0.035	0.076
LHS (mean)	-0.039	0.416	0.321	0.263
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1.1. Column (2) reports the results of estimating Equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.3 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.3 among Black individuals compared to white individuals, and less than 1.3 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.3 times higher among Black individuals compared to white individuals.

Table 1.34: Frequency-based Approach: Threshold Set at 1.7 Higher Mortality

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Black-sounding name	0.122***	-0.032**	-0.015	0.047***
	(0.034)	(0.016)	(0.017)	(0.013)
Hispanic-sounding name	0.006	-0.003	-0.003	0.006
	(0.024)	(0.013)	(0.013)	(0.009)
Asian-sounding name	-0.006	-0.011**	0.023***	-0.013***
	(0.009)	(0.004)	(0.005)	(0.004)
Am. Native-sounding name	-0.227	0.126	-0.104	-0.022
	(0.177)	(0.113)	(0.124)	(0.085)
Total mortality, log	-0.108***	-0.067***	0.147***	-0.080***
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	138,657	138,657	138,657	138,657
R-squared	0.032	0.052	0.189	0.105
LHS (mean)	-0.039	0.239	0.607	0.153
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015. All columns include year FE. Column (1) reports the result of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1.1. Column (2) reports the results of estimating Equation 1.3 where the dependent variable is a dummy = 1 if relative mortality is at least 1.7 times higher among white individuals compared to Black individuals. Column (3) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is less than 1.7 among Black individuals compared to white individuals, and less than 1.7 among white individuals compared to Black individuals. Column (4) reports the results of estimating equation (1.3) where the dependent variable is a dummy = 1 if relative mortality is at least 1.7 times higher among Black individuals compared to white individuals.

Table 1.35: Frequency-based Approach: Alternative Definitions of Frequency

	(1)	(2)	(3)	(4)	(5)	(6)
	Relative mortality (log)			Mortality, Black population (log)		
Black-sounding name	0.122***	0.116***	0.083***	0.106***	0.100***	0.077***
	(0.034)	(0.034)	(0.026)	(0.033)	(0.033)	(0.026)
Hispanic-sounding name	0.005	0.017	-0.004	-0.002	0.009	-0.006
	(0.025)	(0.024)	(0.019)	(0.024)	(0.024)	(0.018)
Asian-sounding name	-0.006	-0.004	0.000	-0.002	-0.000	0.003
	(0.009)	(0.010)	(0.008)	(0.009)	(0.009)	(0.008)
Am. Native-sounding name	-0.229	-0.135	-0.098	-0.195	-0.109	-0.063
	(0.178)	(0.181)	(0.154)	(0.174)	(0.178)	(0.153)
Mortality, (log)	-0.108***	-0.106***	-0.057***			
	(0.002)	(0.002)	(0.002)			
Mortality, White population (log)				0.818***	0.821***	0.861***
				(0.002)	(0.002)	(0.002)
Year FE	X	X	X	X	X	X
State FE		X			X	
Journal FE			X			X
Observations	138,657	138,171	138,601	138,657	138,171	138,601
R-squared	0.032	0.044	0.293	0.675	0.679	0.755
LHS (mean)	-0.039	-0.039	-0.039	10.961	10.960	10.961
RHS (mean)	0.064	0.064	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, and Asian-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1), (2), and (3) control for the log of total mortality among white individuals + Black individuals over the period 1999 to 2015. All columns include year FE. Columns (1), (2), and (3) report the results of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log of relative mortality among Black individuals compared to white individuals, calculated according to equation 1.1. Column (1) is the baseline. In Column (2), I add FE for the state where the institution of the first author is affiliated is located. In column (3), I add journal FE. Columns (1), (2), and (3) control for the log of total mortality among white individuals + Black individuals over the period 1999 to 2015. Columns (4), (5), and (6) report the results of estimating equation (1.3) where the dependent variable is a continuous variable equal to the log total mortality among Black individuals in 1999 to 2015. Column (4) is the baseline. In Column (5), I add FE for the state where the institution of the first author is affiliated is located. In column (6), I add journal FE. Columns (4), (5), and (6) control for the log of total mortality among white individuals over the period 1999 to 2015.

Table 1.36: Demographics-based Approach, Total Citations

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
				Total Citations		
Black-sounding name	-0.275***	-0.130	-0.277***	-0.199**	-0.123	-0.188**
	(0.040)	(0.136)	(0.042)	(0.079)	(0.156)	(0.091)
MeSH: Black or African American	0.068	-0.363***	0.114**	0.112**	-0.323***	0.230***
	(0.046)	(0.104)	(0.050)	(0.054)	(0.113)	(0.058)
MeSH: Black or A. A. X Black-s. name	0.281	0.069	0.301	0.281	0.145	0.290
	(0.254)	(0.531)	(0.278)	(0.281)	(0.557)	(0.310)
Article on human subjects				X	X	X
Observations	651,253	41,944	609,309	199,455	33,793	165,662
LHS (mean)	71.340	98.402	69.477	66.904	98.406	60.477
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE, and controls for Hispanic-sounding, Asian-sounding, and American Native-sounding last name of the first author, and controls for these variables interacted with "MeSH: Black or African American". The vector of racial frequencies by last name comes from the 2010 U.S. Census. In all columns, the dependent variable is the total number of citations. "MeSH: Black or African American" is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in Column (1), but on the subsample of articles linked to a clinical trial. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (4) to (6) are symmetric to columns (1) to (3), but the model is estimated on the subsample of articles focusing on human subjects. All columns report coefficients from the estimation of a Poisson count model.

Table 1.37: Demographics-based Approach, Citations (Relative Citation Ratio)

	All	Linked to trial	Not linked to trial	All	Linked to trial	Not linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
			Relative Citation Ratio			
Black-sounding name	-0.245***	-0.130	-0.240***	-0.167**	-0.121	-0.149*
	(0.040)	(0.126)	(0.042)	(0.073)	(0.144)	(0.086)
MeSH: Black or African American	0.074*	-0.342***	0.125***	-0.006	-0.302**	0.089*
	(0.042)	(0.123)	(0.045)	(0.048)	(0.136)	(0.051)
MeSH: Black or A.A. × Black-s. name	0.250	-0.085	0.275	0.217	-0.076	0.235
	(0.246)	(0.513)	(0.268)	(0.274)	(0.549)	(0.301)
Article on human subjects				X	X	X
Observations	651,247	41,944	609,303	199,453	33,793	165,660
LHS (mean)	2.100	3.339	2.015	2.380	3.397	2.173
RHS (mean)	0.063	0.072	0.063	0.072	0.072	0.071

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE, and controls for Hispanic-sounding, Asian-sounding, and American Native-sounding last name of the first author, and controls for these variables interacted with "MeSH: Black or African American". The vector of racial frequencies by last name comes from the 2010 U.S. Census. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). "MeSH: Black or African American" is a dummy = 1 if the article has Black or African American among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in Column (1), but on the subsample of articles linked to a clinical trial. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. Columns (4) to (6) are symmetric to columns (1) to (3), but the model is estimated on the subsample of articles focusing on human subjects. All columns report coefficients from the estimation of a Poisson count model.

Table 1.38: Frequency-based Approach, Total Citations

	(1)	(2)	(3)	(4)
	Total Citations			
Black-sounding name	-0.114	-0.070	-0.148*	-0.161*
	(0.071)	(0.087)	(0.089)	(0.082)
Relative mortality (log)	0.424***			
	(0.107)			
Relative mortality (log) X Black-sounding name	0.101			
	(0.072)			
Typically White		0.664***		
		(0.129)		
Typically White X Black-sounding name		-0.177		
		(0.141)		
Similar incidence			-0.625***	
			(0.135)	
Similar incidence X Black-sounding name			0.038	
			(0.139)	
Typically Black				0.825***
				(0.158)
Typically Black X Black-sounding name				0.231
				(0.160)
Observations	138,657	138,657	138,657	138,657
LHS (mean)	80.731	80.731	80.731	80.731
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality (white + Black Americans) over the period 1999 to 2015. All columns include year FE. In all columns, the dependent variable it the total number of citations. In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table 1.39: Frequency-based Approach, Citations (Relative Citation Ratio)

	(1)	(2)	(3)	(4)
	Relative citations ratio (RCR)			
Black-sounding name	-0.111	-0.064	-0.143	-0.157*
	(0.072)	(0.088)	(0.090)	(0.082)
Relative mortality (log)	0.366***			
	(0.109)			
Relative mortality (log) × Black-sounding name	0.096			
	(0.069)			
Typically White		0.635***		
		(0.131)		
Typically White × Black-sounding name		-0.181		
		(0.137)		
Similar incidence			-0.624***	
			(0.141)	
Similar incidence × Black-sounding name			0.038	
			(0.139)	
Typically Black				0.579***
				(0.159)
Typically Black × Black-sounding name				0.220
				(0.168)
Observations	138,656	138,656	138,656	138,656
LHS (mean)	2.366	2.366	2.366	2.366
RHS (mean)	0.064	0.064	0.064	0.064

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015 interacted with Black-sounding name. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table 1.40: Demographics-based Approach, Split by Below/Above median Journal Impact Factor (JIF)

	All	Linked to trial	Not Linked to trial	All	Linked to trial	Not Linked to trial
	(1)	(2)	(3)	(4)	(5)	(6)
	MeSH: Black or African American					
Panel A: Journal Commercialization Impact Factor (JCIF) below median in the sample						
Black-sounding name	0.015***	0.037***	0.012***	0.026***	0.040***	0.023***
	(0.003)	(0.012)	(0.003)	(0.005)	(0.011)	(0.005)
Observations	253,951	22,606	231,345	122,497	18,983	103,514
R-squared	0.001	0.003	0.001	0.001	0.003	0.001
LHS (mean)	0.010	0.010	0.010	0.018	0.010	0.019
RHS (mean)	0.067	0.072	0.067	0.072	0.072	0.072
Panel B: Journal Commercialization Impact Factor (JCIF) above median in the sample						
Black-sounding name	0.007***	0.011	0.007***	0.033***	0.015	0.037***
	(0.002)	(0.008)	(0.002)	(0.007)	(0.010)	(0.008)
Observations	397,302	19,338	377,964	76,958	14,810	62,148
R-squared	0.001	0.002	0.001	0.001	0.002	0.001
LHS (mean)	0.006	0.009	0.005	0.022	0.010	0.025
RHS (mean)	0.061	0.071	0.060	0.070	0.072	0.070

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to the race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Column (1) reports the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has "Black or African American" among its MeSH codes, = 0 otherwise. Column (2) reports the results of estimating the same equation as in column (1), but on the subsample of articles linked to a clinical trial. Column (3) reports the results of estimating the same equation as in column (1), but on the subsample of articles not linked to a clinical trial. In columns (4) through (6), I report the results of columns (1) to (3) estimated on the subsample of articles focusing on human subjects.

Table 1.41: Frequency-based Approach, Split by Below/Above Median Journal Impact Factor (JIF)

	(1) Relative mortality, log	(2) Typically White	(3) Similar incidence	(4) Typically Black
Panel A: Journal Impact Factor (JIF) below median in the sample				
Black-sounding name	0.122***	-0.088***	0.048**	0.040**
	(0.040)	(0.023)	(0.023)	(0.017)
Observations	49,872	49,872	49,872	49,872
R-squared	0.023	0.020	0.145	0.114
LHS (mean)	-0.098	0.363	0.476	0.161
RHS (mean)	0.066	0.066	0.066	0.066
Panel B: Journal Impact Factor (JIF) above median in the sample				
Black-sounding name	0.118***	-0.022	-0.027	0.049***
	(0.045)	(0.022)	(0.020)	(0.018)
Observations	84,170	84,170	84,170	84,170
R-squared	0.040	0.014	0.190	0.178
LHS (mean)	-0.003	0.313	0.496	0.191
RHS (mean)	0.063	0.063	0.063	0.063

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution, and published in a journal in the top 1,000 by Commercialization Impact Factor. Black-sounding name refers to race or ethnicity of the first author listed on the publication. All columns include controls for Hispanic-sounding, Asian-sounding, and American Indian-sounding, and for those variables interacted with Relative mortality (log) (Column (1)), with Typically white (Column (2)), with Similar incidence (Column (4)), with Typically Black individuals (Column (4)). In all columns, white-sounding is the omitted category. All columns control for the log of total mortality among white individuals and Black individuals over the period 1999 to 2015 interacted with Black-sounding name. All columns include year FE. In all columns, the dependent variable (Relative Citation Ratio) is calculated as the citations of a paper normalized to the citations received by publications in the same area of research and year (Hutchins, Yuan, Anderson and Santangelo, 2016). In Column (1), relative mortality (log) is the log of rate among Black individuals compared to white individuals. In Column (2), Typically white is a dummy = 1 if relative mortality is at least 1.5 times higher among white individuals compared to Black individuals. In Column (3), similar incidence is a dummy = 1 if relative mortality is less than 1.5 among Black individuals compared to white individuals, and less than 1.5 among white individuals compared to Black individuals. In Column (4), Typically Black is a dummy = 1 if relative mortality is at least 1.5 times higher among Black individuals compared to white individuals. All columns report coefficients from the estimation of a Poisson count model.

Table 1.42: Sickle Cell Anemia, by MeSH Code

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Anatomy		Organisms		Diseases		Chemicals & Drugs		Diagnostic equip.		Biological sciences		Healthcare	
Black-s.	0.000	0.030	0.002*	0.092***	0.003*	0.091***	0.001	0.085**	0.001	0.078**	0.001	0.095**	0.002*	0.119***
	(0.001)	(0.037)	(0.001)	(0.034)	(0.001)	(0.034)	(0.001)	(0.040)	(0.001)	(0.037)	(0.001)	(0.041)	(0.001)	(0.046)
Hispanic-s.	-0.000	-0.023	-0.000	-0.014	-0.001	-0.015	-0.000	-0.004	-0.000	-0.009	-0.000	-0.014	-0.001	-0.029
	(0.001)	(0.022)	(0.000)	(0.017)	(0.001)	(0.017)	(0.001)	(0.020)	(0.000)	(0.016)	(0.000)	(0.018)	(0.001)	(0.024)
Asian-s.	-0.000**	-0.011	-0.000**	-0.004	-0.000	-0.004	-0.001***	-0.009	-0.000**	-0.005	-0.000***	-0.010	0.000	0.012
	(0.000)	(0.009)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.008)	(0.000)	(0.013)
Am. Native-s.	-0.007***	-0.316*	-0.006***	-0.231**	-0.011***	-0.232**	-0.008***	-0.394**	-0.007***	-0.242**	-0.008***	-0.525**	-0.006**	-0.170**
	(0.002)	(0.183)	(0.002)	(0.099)	(0.003)	(0.100)	(0.002)	(0.162)	(0.002)	(0.097)	(0.002)	(0.262)	(0.003)	(0.072)
Observations	353,238	6,875	599,099	13,268	316,191	13,297	486,374	10,257	435,345	9,536	489,509	9,572	243,224	6,265
R-squared	0.000	0.006	0.000	0.004	0.000	0.004	0.000	0.006	0.000	0.006	0.000	0.005	0.000	0.007
LHS (mean)	0.001	0.058	0.001	0.062	0.003	0.062	0.001	0.061	0.001	0.057	0.001	0.060	0.002	0.075
RHS (mean)	0.060	0.059	0.064	0.063	0.066	0.063	0.061	0.061	0.063	0.063	0.061	0.061	0.069	0.067

Notes. Robust s.e. in parentheses clustered by last name. The unit of observation is a research article in the PubMed database published between 2002 and 2018, with first author affiliated with a US institution. Black-sounding name, Hispanic-sounding name, Asian-sounding name, and American Native-sounding name refer to race or ethnicity of the first author listed on the publication. white-sounding name is the omitted category. All columns include year FE. The vector of racial frequencies by last name comes from the 2010 U.S. Census. Columns (1), (3), (5), (7), (9), (11), (13) report the result of estimating Equation 1.2 where the dependent variable is a dummy = 1 if the article has sickle cell anemia among its MeSH codes, = 0 otherwise. Columns (2), (4), (6), (8), (10), (12), and (14) report the results of estimating the same equation as in column (1), but on the subsample of hemoglobin-related diseases. "Hemoglobin-related" diseases are defined as those articles with neoplasm among their MeSH codes.

1.10 Appendix: Model

In section 1.6 of the main text, I estimate the model of occupational choice built by Hsieh, Hurst, Jones and Klenow (2019). In this model, each individual selects the occupation where they obtain the highest utility given their talents and preferences. In this model of the labor market, there are three forces which cause individuals to choose occupations where they do not have a comparative advantage: i) discrimination in the labor market; ii) barriers to acquiring human capital; iii) group-specific preferences or social norms. I use the data to calibrate the magnitude of these three forces for the period 2001 to 2018. Then I compute a policy counterfactual where barriers to human capital acquisition and in the labor market are lifted for Black individuals.

1.10.1 Workers

The economy is composed by a continuum of individuals j who are either white or Black. Their group is indexed by $g = \{\text{white, Black}\}$. Each individual chooses an occupation j to maximize their lifetime utility. Individuals choose their lifetime occupation and decide how much time to dedicate to schooling before entering the labor market (the pre-period), and live for three periods. Their lifetime utility is equal to:

$$\log U_i = \alpha \sum_{t=1}^{3} \log \mathrm{c}_{it} + \log\,(1 - s_{ij}) + \log\, z_{jg} + \log\, \mu_{ij} \tag{1.12}$$

α represents the tradeoff between utility in the pre-period and utility over the remaining of the lifetime. s is time spent in school to acquire human capital (so that $(1-s)$ is leisure), z is group-specific utility derived from working in occupation j, and μ is individual utility from working in occupation j. The parameter z_{jg} relaxes the assumption that, in the absence of barriers, all groups will select occupation j at the same rate. It can be interpreted as preferences, beliefs, or experience.[41] Individual consumption c_{it} is equal

[41]For example, Hsieh, Hurst, Jones and Klenow (2019) documents a decrease of z in the US between 1960 and 2010 a decrease women in the home sector, which they interpret as changes in social norms for

to:

$$c_{it} = (1 - \tau_{jg}^{w}) w_{jt} \varepsilon_{ij} h_{ij} - (1 + \tau_{j,g}^{h}) e_{ij} \tag{1.13}$$

Where τ_{jg} is the job- and group- specific tax to work in occupation j, w_{jt} is the efficiency wage, ε_{ij} is individual productivity of working in occupation j, (i.e., their "talent" for occupation j), h_{ij} is human capital acquired to work in occupation j. Over the lifetime, individuals repay the loan they got in the first period to acquire education. They repay it equally across all three periods. In every period they have to repay 1/3 $(1 + \tau_{jg}^{h}) e_{ij}$, where τ_{jg}^{h} is the tax on acquisition of human capital.

Occupation-specific human capital acquired in the pre-period equals:

$$h_{ij} = \bar{h}_{jg} \gamma s_{ij}^{\phi_j} e_{ij}^{\eta} \tag{1.14}$$

Where $\bar{h}_{jg}$ are differences in human capital endowment that are specific to a group and a given occupation, γ is the return to experience, ϕ_j is the occupation-specific return to time investments in human capital, and η is the elasticity of human capital with respect to human capital expenditures.

Individuals draw a vector of idiosyncratic talent ϵ_j or preferences μ_j across occupations. When they draw idiosyncratic talent, then preference μ_j is assumed to be the same in all occupations, and $= 1$ in each occupation j.

Talent in occupation j, ϵ_j, is assumed to be distributed according to the multivariate Fréchet distribution:

$$F_g(\epsilon_{i1}, \ldots, \epsilon_{iJ}) = \exp\left[-\sum_{j=1}^{J} \epsilon_{ij}^{-\theta} \right] \tag{1.15}$$

Where J is the total number of occupations j, and θ is the shape parameter that governs the dispersion of talent across occupations, with higher θ corresponding to smaller dispersion. The mean of the distribution is normalized to one in each occupation and each group (white individuals and Black individuals).

women working in the market sector or changing preferences for fertility.

When individuals draw a vector of idiosyncratic preferences, μ_j is assumed to be distributed according to same Fréchet distribution, but with shape parameter $\frac{\theta(1-\eta)}{3\beta}$. In this way, the labor supply elasticity of a given occupation of individuals with heterogeneous preferences is equal to the one with heterogeneous talent. In this case, talent ϵ_i is assumed to be the same in all occupations, and equals $\Gamma^{1-\eta}$, where $\Gamma \equiv \Gamma\left(1-\frac{1}{\theta(1-\eta)}\right)$. In this way, the average talent is the same in the case of heterogeneous preferences, and in the case of heterogeneous talent.

1.10.2 Occupational choice

Solving the worker's problem for a given young cohort c at time t, indirect utility of an individual in occupation j and group g (omitting from now on the individual subscript i) is equal to:

$$U^*_{j,g} = \mu_j \left[\bar{\gamma}\tilde{w}_{j,g}\epsilon_j\right]^{\frac{3\beta}{1-\eta}} \tag{1.16}$$

where:

$$\tilde{w}_{j,g} \equiv w_j s_j^{\phi_j} \left(1-s_j\right)^{\frac{1-\eta}{3\beta}} \cdot \frac{\bar{h}_{j,g}\tilde{z}_{j,g}}{\tau_{j,g}} \tag{1.17}$$

$$\tau_{j,g} \equiv \frac{\left(1+\tau^h_{j,g}\right)^{\eta}}{1-\tau^w_{j,g}} \tag{1.18}$$

$$\tilde{z}_{j,g} \equiv z_{j,g}\frac{1-\eta}{3\beta} \tag{1.19}$$

Utility of working in occupation j increases with the occupation-specific preference (μ_j) and occupation-specific talent (ϵ_j). A higher value of $\tau_{j,g}$ is associated with lower individual utility. A higher value of $\tilde{z}_{j,g}$, which represents the utility of group g from working in occupation j is also associated with higher utility.

Each individual chooses the occupation with the highest U^*. Because the heterogeneity is drawn from an extreme value distribution, the highest utility can also be characterized by an extreme value distribution (McFadden, 1974). The overall occupation share is obtained by aggregating the optimal choice across people. The occupation choice problem is equivalent to selecting occupation j with the highest value of $U^*_{j,g}$.

1.10.3 Workers' equilibrium

The equilibrium is described by the following five propositions.

Proposition 1: Occupational choice

The fraction of individuals from group g (Black, or white) who select into occupation j (i.e., $p_{j,g}$), is equal to:

$$p_{j,g} = \frac{\tilde{w}_{j,g}^{\theta}}{\sum_{s=1}^{J} \tilde{w}_{s,g}^{\theta}} \tag{1.20}$$

where

$$\widetilde{w}_{j,g} \equiv w_j s_j^{\phi_j} \left[1 - s_j\right]^{\frac{1-\eta}{3\beta}} \cdot \frac{\bar{h}_{j,g} \tilde{z}_{j,g}}{\tau_{j,g}} \tag{1.21}$$

Equation (1.20) does not depend on the cohort because the choice is made once, when individuals are young, and remains the same throughout their lifetime.

Proposition 2: Average quality of workers in occupation j

The geometric average of worker quality in each occupation is equal to:

$$\exp\left(\mathbb{E}\log\left[h_{j,g,c,t}\,\epsilon_{j,g,c}\right]\right) = \bar{\Gamma} s_{j,c}^{\phi_{j,t}} \gamma(t-c) \left(\frac{\eta s_{j,c}^{\phi_{j,c}} \bar{\gamma} \bar{h}_{j,g} w_{j,c} \left[1 - \tau_{j,g,c}^{w}\right]}{1 + \tau_{j,g,c}^{h}}\right)^{\frac{\eta}{1-\eta}} \left(\frac{1}{p_{j,g,c}}\right)^{\frac{1-\delta}{\theta(1-\eta)}} \tag{1.22}$$

The average quality of workers in occupation j varies by c, group g, and time t.

Proposition 3: Average wages

The geometric average of earnings in occupation j by cohort c in period t of group g equals:

$$\begin{aligned}\overline{\text{wage}}_{j,g,c,t} &\equiv \left(1-\tau^{w}_{j,g,t}\right) w_{j,t} e^{\mathbb{E}\log[h_{j,g,c,t}\epsilon_{j,g}]} \\ &= \bar{\Gamma}\bar{\eta}\left[p^{\delta}_{j,g,c} m_{g,c}\right]^{\frac{1}{\sigma(1-\eta)}} \tilde{z}^{-\frac{1}{1-\eta}}_{j,g,c} \left[1-s_{j,c}\right]^{-\frac{1}{3\beta}} \times \frac{1-\tau^{w}_{j,g,t}}{1-\tau^{w}_{j,g,c}} \frac{w_{j,t}}{w_{j,c}} \frac{\gamma(t-c)}{\bar{\gamma}} \frac{s^{\phi_{j,t}}_{j,c}}{s^{\phi_{j,c}}_{j,c}}\end{aligned} \tag{1.23}$$

Proposition 4: Relative propensities

The proportion of Black individuals of group g in cohort c employed in occupation j relative to white individuals equals:

$$\frac{p_{j,Black}}{p_{j,white}} = \left(\frac{\tau_{j,Black}}{\tau_{j,white}}\right)^{-\frac{\theta}{1-\delta}} \left(\frac{\bar{h}_{j,Black}}{\bar{h}_{j,white}}\right)^{\frac{\theta}{1-\delta}} \left(\frac{\overline{\text{wage}}_{j,Black}}{\overline{\text{wage}}_{j,white}}\right)^{-\frac{\theta(1-\eta)}{1-\delta}} \tag{1.24}$$

Proposition 5: Relative labor force participation

The proportion of Black individuals in the home occupation relative to white individuals for cohort c is equal to:

$$\frac{1-\text{LFP}_{Black}}{1-\text{LFP}_{white}} = \left(\frac{\overline{\text{wage}}_{j,Black}}{\overline{\text{wage}}_{j,white}}\right)^{-\theta(1-\eta)} \left(\frac{\tilde{z}_{j,Black}}{\tilde{z}_{j,white}}\right)^{-\theta} \left(\frac{p_{j,Black}}{p_{j,white}}\right)^{\delta} \tag{1.25}$$

1.10.4 Firms

In this economy, output is produced by one firm that aggregates labor inputs from J occupations through the production function:

$$Y = \left[\sum_{j=1}^{J} \left(A_j \cdot H_j\right)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \tag{1.26}$$

Where H_j is total efficiency units of labor in each occupation, σ is the elasticity of substitution across occupations, A_j is the exogenously-given productivity of occupation j.

1.10.5 Equilibrium

$H_{j,t}^{demand}$ that satisfies the profit maximization equals:

$$H_{j,t}^{demand} = \left(\frac{A_{j,t}^{\frac{\sigma-1}{\sigma}}}{w_{j,t}} \right) Y_t \tag{1.27}$$

And $w_{j,t}$ clears the labor market in each occupation such that $H_{j,t}^{supply} = H_{j,t}^{demand}$.

1.10.6 Estimation

To estimate the model, I follow the steps of Hsieh, Hurst, Jones and Klenow (2019). I use both internally calibrated parameters (on occupational shares and average wages by group and occupation) from the CPS and externally calibrated moments from Hsieh, Hurst, Jones and Klenow (2019). I use data from the March Annual Demographic Survey files of the Current Population Survey (CPS), and build an occupation that I call "scientists and inventors" combining individuals who report working in "engineering" and "natural scientists".

I start from proposition 4, which governs the propensity of Black individuals compared to white individuals to be employed in occupation j. This equation tells us that the propensity of Black individuals to work in occupation j compared to white individuals depends on relative frictions, on relative talent, and on the average wage Black-white gap.

To estimate the model, I impose a series of additional assumptions following Hsieh, Hurst, Jones and Klenow (2019). First, in the main estimation, I assume that individuals only

select occupations based on talent (not on individual preferences). That is, that $\mu_j = 1$ for each occupation j. Secondly, I assume that talent is distributed equally across groups in a given sector. Third, I assume that white individuals do not face any barriers in the acquisition of human capital or in the labor market. That is, $\tau^h_{white} = 0$ and $\tau^w_{white} = 0$ in all occupations and all periods. Fourth, I normalize preference for the home sector to be equal to 1 for all groups. Fifth, I assume that the return to experience γ is the same for all sectors, groups, and cohorts.

I begin by estimating ϕ_j (return to schooling for occupation j), $z_{j,white}$ (preference of white individuals to work in occupation j with respect to staying at home), and w_j (wage per efficiency unit for occupation j).

The return to schooling for occupation j (ϕ_j) is pinned down by the first order condition for schooling:

$$\phi_j = \frac{1-\eta}{3\beta} \cdot \frac{s_j}{1-s_j} \tag{1.28}$$

To recover s_j, I assume that the pre-market period is 28 years long (although the estimation results are virtually unchanged if I use a lower year threshold).

$$s_j = \frac{\text{Years of education}}{28} \tag{1.29}$$

To estimate the preference of white individuals to work in occupation j with respect to staying at home (z_j) for J occupations, I use data on the earnings of white individuals in the young cohort. Defining $m_{white} = \sum_{j=1}^{J} \widetilde{w}^{\theta}_{j,white}$ and rearranging the equation for average wage:

$$m_{white} = \left[\frac{\overline{\text{wage}}_{j,white} \tilde{z}_{j,white} \left(1-s_j\right)^{\frac{1}{3\beta}}}{\bar{\Gamma}\bar{\eta}} \right]^{\theta(1-\eta)} \tag{1.30}$$

To recover m_{white} (which does not vary by occupation), I plug in values for j = home. By assumption, $\tilde{z}_{home,white}$ is normalized to 1. Since there is no wage data for the home

occupation, I assume that earnings in home occupation are equal to the average earnings in another occupation. I follow Hsieh, Hurst, Jones and Klenow (2019) and use the average wage in occupation "Secretaries". To compute $\tilde{z}_{j,white}$ for all occupations, I plug m_{white} into equation (1.30).

The wage per efficiency unit (w_j) is obtained by rearranging the occupational share equation for white individuals:

$$w_j = \frac{\left[p_{j,white} \cdot m_{white}\right]^{\frac{1}{\theta}}}{\bar{\gamma} \cdot s_j^{\phi_j} \left[\left(1 - s_j\right) z_{j,white}\right]^{\frac{1-\eta}{3\beta}}} \tag{1.31}$$

After making an initial guess about the return to experience γ, all other variables are known, which allows to pin down the value of w_j (which does not vary by group or cohort).

To estimate the return to experience (γ), I use the change in the average wage of cohort c and occupation j over time. The ratio of average wage in occupation j at time t with respect to time c (when that cohort is young) is equal to:

$$\frac{\overline{\text{wage}}_{j,white,c,t}}{\overline{\text{wage}}_{j,white,c,c}} = \frac{w_{j,t} \cdot \gamma(t-c) \cdot s_j^{\phi_{j,t}}}{w_{j,c} \cdot s_j^{\phi_c}} \tag{1.32}$$

Since I have $\overline{\text{wage}}$ from the empirical moments, w_j, s_j, and ϕ_j from the previous steps, I can recover γ. By assumption, it is the same across all occupations j and cohorts c. $\bar{\gamma}$ is then computed as the average across all occupations j and cohorts c. The final step to recover $z_{j,white}$ and w_j is to iterate over equations (1.30) through (1.32).

Now, I turn to estimating τ^h and τ^w. After applying assumptions 1. to 3. and rearranging terms, I obtain the following expression for $\tau_{j,Black,c}$:

$$\tau_{j,Black,c} = \left(\frac{p_{j,Black,c}}{p_{j,white,c}}\right)^{-\frac{1}{\theta}} \left(\frac{\overline{\text{wage}}_{j,Black,c}}{\overline{\text{wage}}_{j,white,c}}\right)^{-(1-\eta)} \tag{1.33}$$

The first term on the right comes from the data: it is the ratio of Black to white individuals in occupation j and cohort c. The second term also comes from the data: it is the ratio between the geometric average of the wage of Black individuals in occupation j and cohort c to white individuals same occupation-cohort. By assumption, $\theta = 2$ and $\eta = 0.103$.

As $\tau_{j,Black,c}$ is estimated from the data only for the young cohort in each period, τ varies at the period- and occupation- level.

Finally, I use the cohort structure to recover the components of composite τ: τ^w and τ^h. For each occupation j and period t, $\tau^w_{j,Black,t}$ represents *labor market* barriers faced by all Black individuals in the labor market at time t. For each occupation j and cohort c, $\tau^h_{j,Black,c}$ represents *human capital accumulation* barriers, faced by all Black individuals who entered the labor market in period $t = c$. By definition, $\tau_{j,Black,c,t}$ is equal to:

$$\tau_{j,Black,c,t} \equiv \frac{\left(1 + \tau^h_{j,Black,c}\right)^\eta}{1 - \tau^w_{j,Black,t}} \tag{1.34}$$

To recover the τ's, I start from equation 1.34 for the young cohort in each period t. In what follows, I will omit subscripts j (for occupation) and Black individuals, but all τs are always specific to an occupation, and specific to Black individuals (relative to white individuals).

1. Recovering $\tau^h_{Cohort=3}$ and $\tau^w_{t=1}$

   (a) Define α as the Cobb-Douglas split of composite τ. Specifically:

   $$\tau^\alpha_{c,t} = \frac{1}{1 - \tau^w_t} \tag{1.35}$$

   and

   $$\tau^{1-\alpha}_{c,t} = (1 + \tau^h_{c-t})^\eta \tag{1.36}$$

   (b) Setting the initial value of α to 0.5 for the young cohort in period 1 (i.e. cohort 3), and plugging in $\tau_{Cohort\ 3,Period\ 1}$ from the data, I compute τ^w for period 1, and τ^h for cohort 3 (i.e. the young in period 1).

2. Recovering $\tau^w_{t=2}$

   (a) When $\delta = 0$, individuals are heterogeneous on talent (and not on preferences), average talent (and, by Proposition 2, also average wages) are inversely related to the share of the group working in occupation j. Rearranging Proposition 3 combined with Proposition 2, wage growth for cohort c in a given occupation by group (white or Black) is equal to:

$$\frac{\overline{\text{wage}}_{j,g,c,t+1}}{\overline{\text{wage}}_{j,g,c,t}} = \frac{1-\tau^w_{j,g,t+1}}{1-\tau^w_{j,g,t}} \cdot \frac{w_{j,t+1}}{w_{j,t}} \cdot \frac{(s_{j,c})^{\phi_{j,t+1}}}{(s_{j,c})^{\phi_{j,t}}} \tag{1.37}$$

   (b) By assumption, both τ^h and τ^w equal zero for white individuals. Therefore, equation 1.37 for white individuals is equal to:

$$\frac{\overline{\text{wage}}_{j,white,c,t+1}}{\overline{\text{wage}}_{j,white,c,t}} = \frac{w_{j,t+1}}{w_{j,t}} \cdot \frac{(s_{j,c})^{\phi_{j,t+1}}}{(s_{j,c})^{\phi_{j,t}}} \tag{1.38}$$

   (c) Dividing equation 1.37 by equation 1.38 and rearranging terms:

$$\frac{\overline{\text{wage}}_{j,Black,c,t+1}}{\overline{\text{wage}}_{j,Black,c,t}} = \frac{1-\tau^w_{j,Black,t+1}}{1-\tau^w_{j,Black,t}} \cdot \frac{\overline{\text{wage}}_{j,white,c,t+1}}{\overline{\text{wage}}_{j,white,c,t}} \tag{1.39}$$

   For a given occupation-cohort, the wage growth for Black individuals is equal to the wage growth for white individuals, times the growth rate of labor market barriers faced by Black individuals.

   (d) Wage growth comes from the data, and $\tau^w_{j,Black,t}$ is known from above. Therefore, this equation allows to compute $\tau^w_{j,Black,t=2}$.

3. Recovering $\tau^h_{j,Cohort=2}$

   To compute the barriers to human capital accumulation faced by the cohort who enters the market in period 2, I plug $\tau^w_{j,t=2}$ into the definition of composite τ for the young cohort in period 2. In the benchmark estimation, the minimum value of τ^h is set to -0.80.

4. Recovering $\tau^w_{j,t=2}$

To recover labor market barriers for period 2, I follow the steps outlined in point 2. plugging in $t = 2$, $t + 1 = 3$, and $c = 2$.

5. Recovering $\tau^{h}_{j,Cohort=1}$

   To compute human capital barriers faced by the cohort who enters the labor market in period 2, I plug $\tau^{w}_{j,t=2}$ into the definition of composite τ (equation 1.34) for the young cohort in period 2.

Finally, since in the model occupations are chosen when young, labor force participation remains the same when middle-aged and old. The wage moments need to be corrected for the fact that the composition of workers changes as individuals move in and out of the labor force. Hsieh, Hurst, Jones and Klenow (2019) apply a common adjustment across all occupations to obtain the wage growth estimate. As the elasticity of labor force participation with respect to wage growth is $\frac{\theta(1-\eta)}{1-\delta}$, the common adjustment is:

$$\left(\frac{\text{wagegrowth}_{\ j,Black}}{\text{wagegrowth}_{\ j,white}}\right)^{\text{Estimation}} = \left(\frac{\text{wagegrowth}_{\ j,Black}}{\text{wagegrowth}_{\ j,white}}\right)^{\text{Data}} \left(\frac{\text{LFPgrowth}_{\ j,Black}}{\text{LFPgrowth}_{\ j,white}}\right)^{\frac{1-\delta}{(1-\eta)}} \tag{1.40}$$

Chapter 2

Dealing With Adversity: Religiosity or Science? Evidence From the Great Influenza Pandemic

2.1 Introduction

Throughout history, the occurrence of adverse events—such as natural disasters and pandemics—has posed challenges to societies worldwide and continues to do so today. Understanding how individuals cope with adverse events has key social, economic, and political implications and has been the focus of a vast literature in economics and other social sciences. Specifically, a strand of research documents that negative shocks lead to an increase in religiosity (Bentzen, 2019). Another strand finds that economies react by boosting scientific efforts (Miao and Popp; Moscona, 2014; 2022).[1]

In this paper, we show that these two responses can occur *simultaneously*, making societies both more religious *and* more science-oriented—a finding at odds with the existing

[1]For example, Bentzen (2019) documents that, across countries and within regions, individuals become more religious when hit by earthquakes. Moscona (2022) finds an increase in innovation efforts towards technologies that mitigate environmental distress in U.S. counties more exposed to the Dust Bowl during the 1930s.

evidence documenting a negative relationship between religiosity and science (Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni; Lecce, Ogliari and Squicciarini, 2015; 2022; 2021). To investigate the possible mechanism behind this pattern, we study how individuals *within* society react to an adverse shock. We uncover heterogeneous responses, with religion and science acting as substitute ways through which different individuals react to adversity. These individual-level findings help reconcile our aggregate results with the existing literature.

The setting of our study is the Great Influenza Pandemic (1918–1919) in the United States. Historical records document that many people turned to or strengthened their religious faith to cope with the pandemic. At the same time, the period following the pandemic saw increased scientific progress and fundamental medical advances.[2] To conduct our empirical analysis, we construct a novel data-driven measure of religiosity at a geographically disaggregated level. This measure is based on naming patterns of babies born between 1900 and 1930 from the historical full-count censuses. Complementing this dataset with information from the Census of Religious Bodies, we empirically identify religious names and construct a measure of "revealed religiosity." The underlying idea is that the first name given to a child conveys information on the religiosity of their parents. Our metrics of scientific progress are the share of people in STEM occupations and the universe of geo-coded patents granted in the U.S.[3]

Using a difference-in-differences framework, we first show that counties hit harder by the shock experienced an increase in religiosity. A one-standard-deviation increase in excess deaths—our main measure for the intensity of the influenza shock—led to a 0.16 standard deviation increase in overall religiosity. We further document that these same counties also experienced an increase in scientific progress. A one-standard-deviation increase in excess deaths led to a 0.17 standard deviation increase in overall patenting activity. In addition, we find that employment in scientific occupations grew in counties hit harder

[2]An increase in religiosity and scientific progress has also been documented after the COVID-19 outbreak. Bentzen (2021), using Google search data, finds a sharp increase in the intensity of prayers during the early days of the pandemic. Agarwal and Gaule (2022) show that the COVID-19 pandemic catalyzed R&D expenditure on pharmaceuticals and digital technologies.

[3]We refer to science and scientific progress interchangeably, and we use two main proxies mentioned in the text.

by the pandemic. This effect is mainly due to the occupational choices of young cohorts. Event-study analyses illustrate the absence of pretrends, providing further support for the validity of the research design. Interestingly, as a result of the contemporaneous increase of religiosity and science, their relationship turned from negative before the pandemic to positive afterward. This is especially puzzling because it contrasts with the existing evidence documenting a negative relationship between the two (Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni; Lecce, Ogliari and Squicciarini, 2015; 2022; 2021).

What is the mechanism behind the contemporaneous increase in religiosity and science? To answer this question, in the second part of the analysis, we move to a within-county analysis and study individual-level reactions to the pandemic. We obtain three main results.

First, we find that individuals from more religious backgrounds were more likely to turn to religion in the aftermath of the pandemic, while those from less religious backgrounds were more likely to select a scientific occupation.[4] This pattern suggests that individuals coped with negative shocks in heterogeneous ways: some turned to religion, while others turned to science. Second, we show that science-oriented individuals became less religious than the rest of the population after the shock. Third, we document that the pandemic widened preexisting differences in religious sentiment. Individuals from more (less) religious backgrounds became even more (less) religious. Consequently, the distribution of religiosity in counties more exposed to the pandemic became more polarized. Importantly, the individual-level analysis reconciles the county-level findings with the existing literature. While a county may have become more religious and innovative, individuals seemed to react differently to the same shock—based, for instance, on their religious background or pre-pandemic scientific orientation. Religiosity and science appear to have been alternative ways of reacting to the pandemic, with individuals becoming even more distant in terms of their religious sentiment than before the shock.

We perform several checks to gauge the robustness of our findings. For both religiosity

[4]We measure religious background using individuals' names (as opposed to their children's), aiming to capture the religious upbringing of a person instead of their current faith.

and scientific progress, we show that our results are robust to weighting regressions by county population and when running our analysis at the city level. These exercises suggest that our findings are not driven by small counties or by individuals residing in rural areas. Second, we perform a series of checks on our religiosity measure. In particular, we validate it internally across several dimensions (e.g., by computing our indicator excluding firstborn babies and accounting for potential heterogeneity in fertility patterns). In addition, we run a few robustness checks to ensure that the increase in religiosity is not driven by migration.[5] Next, we externally validate our data-driven measure of religiosity by using alternative indicators. In particular, we show that results are robust when using the share of biblical and saints' names and the share of people affiliated with a religious denomination. We also perform a series of robustness checks on our measure of innovative activity (e.g., we show that the increase in patenting was not driven by low-quality innovations). Finally, we address the concern that other factors may be related to the pandemic and may have contemporaneously affected the evolution of religiosity and science, confounding our results. To do so, we start by documenting that initial religiosity and scientific progress are not related to the intensity of the shock. Then, using an event-study design, we show that religiosity and innovation were on a similar path across treated and control groups before the pandemic. Additionally, we rule out that a separate yet overlapping shock—World War I—may partly explain our findings. Taken together, our empirical results, supported by historical records, provide evidence that the influenza pandemic was conceivably the main driver behind the aggregate increases in both religiosity and scientific progress.

Concerning our within-county results, one key question is why some individuals became more religious while others selected a scientific occupation. Our findings on religiosity are in line with the religious coping hypothesis, which posits that religious faith can represent a coping device to deal with personal distress following a negative shock.[6] What

[5]In particular, we first run a placebo exercise where we test for the impact of the pandemic on the names of adults. The results show no impact of the shock on adults' names, which we interpret as evidence that the observed increase in religiosity was not driven by ex-ante more religious people moving to areas hit harder by the shock. Then, we show that our results hold when excluding from the estimation sample all those who, in the 1930 census, reside in a state different from the one where they were born.

[6]An alternative explanation could be that individuals turn to religion as an insurance mechanism against the negative economic effects of the pandemic. While we cannot fully exclude this channel, we

motivated people to turn to science is less obvious. We propose a broad interpretation of "scientific coping," with individuals turning to science either to deal with their psychological distress—as in the case of religious coping—or to try to actively mitigate the negative (e.g., health- and economic-related) effects of the pandemic.[7] While our findings cannot directly uncover the individual-level motivations behind these different behaviors—this would go beyond the scope of this paper—they show that people from different backgrounds may have reacted in different ways to the same shock and that this may have increased the polarization of religiosity within society.

Related Literature. This paper is closely related to the literature studying how societies react to negative shocks. Previous work has shown that, in accordance with the religious coping hypothesis (Pargament; Ano and Vasconcelles; Norenzayan and Hansen, 2001; 2005; 2006), natural disasters are associated with an increase in religiosity, both historically (Belloc, Drago and Galbiati; Bentzen, 2016; 2019) and in contemporary scenarios (Sibley and Bulbulia; Bentzen, 2012; 2021).[8] Another set of studies documents that economic crises (Babina, Bernstein and Mezzanotti, Forthcoming), wars (Gross and Sampat, 2021), climate change (Miao and Popp; Clemens and Rogers; Moscona, 2014; 2020; 2022), and pandemics (Gross and Sampat; Agarwal and Gaule, 2021; 2022) all shape innovation activity. To the best of our knowledge, this is the first study to provide evidence that natural disasters may foster a contemporaneous increase in religiosity *and* innovation, and also the first to document the ensuing polarization of religiosity within society.[9]

Additionally, we inform the broad literature on the economics of religion, pioneered by Weber (1905). In particular, we contribute to those studies that analyze the linkage be-

believe it is unlikely (as discussed in Section 2.6).

[7]Another possibility is that individuals turned to science because of increased labor demand in STEM occupations. However, the heterogeneity by religious background suggests that, beyond market forces, individual preexisting religiosity played a key role in their decision to turn to science.

[8]The religious coping hypothesis, first developed in the psychology literature, posits that people who are subject to economic and social shocks turn to religious faith as a coping device to deal with personal distress.

[9]Many studies have looked at the impact of natural disasters on, among others, social norms (Posch, 2022), migration (Boustan, Kahn and Rhode, 2012), and economic activity (Boustan, Kahn, Rhode and Yanguas, 2020).

tween religiosity and science.[10] While most papers adopt a historical (Deming; Mokyr, 2010; 2011), theoretical (Bénabou, Ticchi and Vindigni, 2022), or cross-sectional perspective (Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni, 2015; 2022), to our knowledge, we are the first to study the interaction between religion and science in a panel setting and to uncover the individual-level dynamics behind their coevolution.[11]

Finally, we contribute to a growing literature that exploits the informational content of names to capture individuals' characteristics. Names have been used, for example, to measure race and ethnicity (Abramitzky, Boustan and Eriksson; Fouka, 2016; 2019), individualism (Bazzi, Fiszbein and Gebresilasse, 2020), socioeconomic background (Biavaschi, Giulietti and Siddique; Olivetti, Paserman, Salisbury and Weber, 2017; 2020), nationalism (Jurajda and Kovač, 2021), and religiosity (Abramitzky, Boustan and Eriksson; Andersen and Bentzen, 2016; 2022). While all of these papers assume a preexisting rule to classify names (e.g., whether one has a biblical or saint name or a name shared by a major religious figure), to the best of our knowledge, we are the first to identify the religiosity of names *directly from the data*.[12]

Outline of the Paper. The paper is structured as follows. Section 2.2 describes the Great Influenza Pandemic in the United States and discusses the historical evidence on its effects on religiosity and innovation. In Section 2.3, we present the data and our indicator of religiosity. In Section 2.4, we explain the empirical strategy and the core results. In Section 2.5, we explore the possible mechanisms underlying our findings. Section 2.6 discusses the results and the limitations of the analysis. Section 2.7 concludes.

[10]Other studies analyze the relationship between religion and accumulation of human capital, more broadly (Becker and Woessmann; Botticini and Eckstein; Squicciarini, 2009; 2012; 2020). For an overview of the literature on the economics of religion, see Iannaccone (1998), Iyer (2016), and Becker, Rubin and Woessmann (2021).

[11]Lecce, Ogliari and Squicciarini (2021) study how religiosity impacts the birth and migration of scientists in 19th-century French cantons, and Andersen and Bentzen (2022) show that individuals with religious names are less likely to become engineers, scientists or doctors and that cities with more religious individuals grew slower. However, none of these studies analyze how an adverse shock affects society's heterogeneous response in terms of religion and science and the underlying individual-level dynamics.

[12]For details on how we construct our religiosity measure, see Section 2.3.

2.2 Historical Background

This section provides an overview of the Great Influenza Pandemic in the United States and how it influenced religion and innovation.

2.2.1 The Great Influenza Pandemic

Between 1918 and 1919, the Great Influenza Pandemic—also known as the "Spanish Flu"[13]—killed an estimated 40 million people worldwide (approximately 1 in 30 people); it was one of the deadliest natural disasters in modern times (Barro, Ursúa and Weng, 2020). In the United States, the pandemic started in the spring of 1918 with sporadic outbreaks. Then, a second, more severe wave began in September 1918. The final wave started in January 1919, ending that spring. In total, it killed about 500,000 Americans, corresponding to 0.7% of the U.S. population (Crosby, 1989).[14]

Historical and modern accounts suggest that the pandemic hit across the U.S. quasi-randomly. The National Research Council stated that neither demographic characteristics, such as the ethnic composition of the population, nor geographic factors seemed to explain the difference in the intensity of the pandemic across the country. Crosby (1989) writes that the states with the highest mortality displayed diverse geographical, climatic, and demographic characteristics. The pandemic hit with varying intensity within states as well. For example, in Minnesota, the death rate in Saint Paul was about 70% higher than in Minneapolis, despite the two cities being just 8 miles apart. In Ohio, Dayton experienced an 80% higher mortality rate than Columbus, even though the two cities had similar demographic characteristics (Huntington; Almond, 1923; 2006).

The infection was caused by strains of the A/H1N1 influenza virus, whose origin is still

[13]The Great Influenza Pandemic is popularly known as "Spanish Flu" because media in Spain—which was neutral during World War I (WWI)—were free to report news on this disease. Conversely, countries involved in WWI imposed press censorship on the topic. This gave the (incorrect) impression that Spain was either more severely hit by the disease or that the pandemic had originated in Spain.

[14]By comparison, COVID-19 caused 1.13 million deaths in the United States, approximately 0.3% of the U.S. population, between March 2020 and February 2023 (https://covid.cdc.gov/covid-data-tracker/#datatracker-home; accessed February 12, 2023).

unknown. Neither antiviral drugs to treat the primary disease nor antibiotics to cure secondary bacterial infections were available. Doctors had to rely on an array of mostly ineffective–sometimes fatal–medicines such as aspirin and quinine (Spinney, 2018). It is debated whether nonpharmaceutical interventions (NPIs)–such as using masks, canceling public events, closing schools, and implementing isolation measures and quarantines—were effective in limiting the spread of the disease.[15]

2.2.2 The Pandemic and Religion

A large literature documents that individuals become more religious in response to adverse events. One explanation comes from the "religious coping hypothesis," which posits that individuals turn to religious beliefs or practices as a way to cope with sudden dramatic circumstances (Pargament, 2001).[16]

The influenza pandemic inflicted substantial emotional and socioeconomic distress and could have acted as a powerful amplifier of religious sentiments (Phillips, 2020). Historical records document that spiritualism gained momentum in the aftermath of the pandemic. Not all confessions reacted in the same way. In the United States, modern evangelism benefited from the pandemic, as evidenced by a sharp rise in the circulation of evangelical magazines (Frost, 2020). Membership in Christian Science also soared during these years, reaching an all-time peak in the 1930s.[17] Catholics and Orthodox Jews identified the influenza as a manifestation of divine anger, the expiation of which called for prayers. On the other hand, some groups of progressive Protestants called for a more scientific

[15]Some authors assert that NPIs were effective in reducing mortality (Markel, Lipman, Navarro, Sloan, Michalsen, Stern and Cetron; Berkes, Deschenes, Gaetani, Lin and Severen, 2007; Forthcoming), while others show that the effect of NPIs on overall deaths was small and statistically insignificant (Barro, 2022).

[16]For example, Bentzen (2019) documents that individuals become more religious when hit by earthquakes. Religion may also represent an insurance mechanism when negative shocks occur: Ager, Hansen and Lønstrup (2016) shows that after the 1927 Great Mississippi Flood, demand for social insurance led to higher churchgoing, while Ager and Ciccone (2018) document that in the 19th-century United States, a larger share of the population was organized in religious communities in counties that were exposed to higher common agricultural risk.

[17]Christian Science, founded in 1879, is part of the religious movements belonging to the metaphysical family. It seeks to restore the healing and thaumaturgic virtues of primitive Christianity and has been associated with avoidance of mainstream medicine (Stark, 1998).

interpretation of the pandemic (Phillips, 2020).[18]

2.2.3 The Pandemic and Science

Historical evidence suggests that the period after 1918 was one of sharp intellectual and scientific progress and that the Great Influenza Pandemic was particularly influential in shaping the development of medical sciences (Barry, 2020). Despite being ineffective during the pandemic, medicine evolved enormously in subsequent years. In 1928, Alexander Fleming discovered the medical use of penicillin in treating bacterial infections. By the 1930s, virology had become an established branch of medicine, and the first influenza vaccines were being developed (Spinney, 2018). During this time, medicine became more "scientific" and, hence, effective (Barry, 2020).

These advancements in medicine went hand in hand with increased trust in scientific progress. For instance, in her personal journal, Canadian author L. M. Montgomery wrote, "[...] the Spirit of God no longer works through the church for humanity. It did once but it has worn out its instrument and dropped it. Today it is working through Science" (Montgomery, 1924). Barry (2020) argues that the pandemic was the key driver behind this paradigm shift because it fostered scientific thinking in the face of such a sudden and dramatic shock.

This overview suggests that the 1918–1919 pandemic fostered both scientific progress and religiosity—a result that might seem at odds with theoretical and empirical evidence, which depicts religion and science as opposing forces (Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni; Lecce, Ogliari and Squicciarini, 2015; 2022; 2021). In this paper, we provide causal evidence that the influenza shock led to a simultaneous increase in religiosity and scientific progress at the aggregate level. We then reconcile this apparent puzzle by showing that it induced polarization within society, with some people turning to religion and others turning to science.

[18]There were also conservative Protestant churches, such as those in the Bible Belt—i.e., the region chiefly comprising Alabama, Arkansas, Georgia, Kentucky, Louisiana, Mississippi, Missouri, North Carolina, Oklahoma, South Carolina, Tennessee, and large parts of Florida and Texas—refractory to scientific and medical advancements.

2.3 Data

To conduct our analysis, we construct a new dataset that combines information on religiosity, scientific progress, and the incidence of the Great Influenza Pandemic. This section describes the outcome variables and the main explanatory variables. Appendix 2.10 describes the data in detail. In the first part of the analysis, counties are the geographical unit of observation.[19] In the second part of the analysis, we use individual-level data. Table 2.6 provides descriptive statistics for the main variables.

2.3.1 Religiosity Measure

The key challenge when studying religiosity is that it is not easy to measure it, today as well as in the past. It is especially challenging to find an indicator of religiosity that combines geographical granularity and high-frequency time variation.[20]

In our analysis, we propose a novel measure of revealed religiosity based on the naming patterns of newborn babies. The motivating argument is that parents who give comparatively more religious names are more likely to be religious themselves. Therefore, naming patterns provide a measure of "revealed religiosity" of parents, rather than of the children themselves.[21]

We now describe how we compute the religiosity score associated with first names. The key advantage of this approach is that it allows us to obtain a disaggregated yearly measure of religiosity and to study its changes in the short-to-medium term. The metric

[19]To address concerns related to counties changing their boundaries over time, we use 1920 counties as our geography of reference.

[20]This lack of data is clear in historical settings—Squicciarini (2020), for instance, uses different measures of religiosity, available for only a few points in time—but it poses substantial limitations to contemporary studies as well. Recent papers leverage information from surveys such as the World Value Survey to measure religiosity (Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni, 2015; 2022). Yet, because waves are typically years apart, survey-based measures are not useful for studying the dynamics of religiosity at high time frequency.

[21]A natural corollary is that names carry informational content on the religiosity of an individual's background: while we cannot infer that an individual called "Paul" is comparatively more religious than one named "Harold," we assume that the parents of "Paul" are likely to be more religious than those of "Harold."

we define is conceptually similar to that developed by Abramitzky, Boustan and Eriksson (2016) and Andersen and Bentzen (2022), who measure religiosity, depending on whether children have a biblical/saint names or names of a major religious figure, respectively. Our approach differs from theirs: we *empirically* identify our religious names, using data on the entire population of newborns and existing indicators of religiosity.

Estimating Religiosity Scores for First Names

We use two main sources to compute religiosity scores. First, we construct naming patterns at the county-cohort level from the full-count U.S. censuses between 1900 and 1930 (Ruggles, Flood, Foster, Goeken, Pacas, Schouweiler and Sobek, 2021). More precisely, we take the first name of all children born in the United States between 1896 and 1930 and collapse them at the name-county-cohort level, thus obtaining a panel of name-county pairs at a yearly frequency.[22] Next, we use county-level data from the Census of Religious Bodies. This census—taken once every ten years between 1906 and 1936—allows us to construct, for every county and census decade, the share of people affiliated with any religious denomination, as well as the share of people affiliated with a Catholic or Protestant one.[23]

To obtain the religiosity scores, we first compute the frequency of names. We denote with N_{cd} the total number of individuals born in county c in decade d and with $\text{Name Frequency}_{cd}^{k}$ the number of children with name k born over the years $[d-10, d)$ in county c. Then, we estimate the following model:

$$y_{cd} = \alpha_c + \alpha_d + \beta \times \log(N_{cd}) + \sum_{k=1}^{K} \gamma^k \times \log(1 + \text{Name Frequency}_{cd}^{k}) + \varepsilon_{cd}, \qquad (2.1)$$

[22] A cohort is defined as all babies born in a given year. The first cohort in our sample comprises all babies born in 1896. The underlying rationale is that the first Census of Religious Bodies was published in 1906, and we consider the ten cohorts preceding that year.

[23] To gather information on the number of religious members in each county, a report was obtained directly from local churches and congregations. The shares are computed as the number of people affiliated with these groups, normalized by the population of each county. Our analysis focuses on Catholics and Protestants, as they jointly account for more than 90% of the people enumerated by the census.

where y denotes either the share of people affiliated to any denomination, or the share of Catholics, or the share of Protestants in county c in decade d; d corresponds to the two prepandemic decades of the religious censuses (1906 and 1916); α_c and α_d are, respectively, county and decade fixed effects.[24] The term K is the total number of names that occur in at least 0.3% of the overall sample.[25] To measure name frequencies, we include all babies born within ten years before each pre-pandemic census. Hence, we restrict the sample to cohorts between 1896 and 1916. Then, we aggregate these frequencies from cohorts to decades to estimate equation (2.1).

We label the coefficient (γ^k) as the *religiosity score* associated with name k; we interpret names with larger estimated religiosity scores ($\hat{\gamma}^k$) as conveying a more intense religious sentiment. Because model (2.1) includes county fixed effects, larger religiosity scores are attached to names that become comparatively more frequent in counties that experienced larger increases in religiosity. In Figure 2.1, we report the estimated religiosity scores from model (2.1), where the outcome variable is the share of people affiliated with any religious organization. The figure shows that typically religious-sounding names, such as "Joseph," "Paul," and "Elizabeth," all feature positive and large estimated religiosity scores. Because our estimation method seeks to isolate *distinctively* religious names, relatively common ones such as "Mary" or "John" are associated with negative scores. In the case of "Mary," for instance, its popularity during this period was such that religious and non-religious people alike used it, thus preventing it from being associated with distinctively religious people. Moreover, names with little connection to saints or biblical episodes are associated with negative religiosity scores. This is the case for Germanic names, such as "Bertha," "George," and "Harold," and other non-religious names, such as "Pearl." By considering the shares of people affiliated with Catholicism

[24] In one of our robustness checks, we compute an alternative measure of religiosity that does not include any fixed effect. The results hold. Moreover, the results are robust to including the raw name frequencies in equation (2.1) or, alternatively, apply an inverse hyperbolic sine transformation and add .01 to the frequency inside the log.

[25] We follow Fouka (2020) and restrict the number of names included in model (2.1) primarily to avoid overfitting. Fouka (2020) uses a threshold of 1,000 for a name to be included in the analysis. In our preferred specification, we instead consider all names whose share in our overall sample is at least 0.3% and run checks around this threshold to assess the robustness of our results. We include name frequencies in log to reduce their skewness and add one to avoid dropping all counties where there is at least one name with no newborn children (approximately 40% of the sample).

or Protestantism, we can also obtain religiosity scores for both religious denominations separately. Figure 2.6 reports the results.

A Yearly County-Level Measure of Religiosity

From model (2.1), we obtain a set of estimated religiosity scores $\{\hat{\gamma}^k\}_{k=1}^K$, which we use to construct a *yearly* indicator of religiosity at the county level. More specifically, our synthetic measure of religiosity is defined as the predicted values of model (2.1):

$$\hat{y}_{ct} = \sum_{k=1}^{K} \hat{\gamma}^k \times \log(1 + \text{Name Frequency}_{ct}^k), \tag{2.2}$$

where t denotes a cohort between 1900 and 1930. In addition, by considering religiosity scores associated with different denominations, we can construct a synthetic series for Catholic and Protestant religiosity separately.

A concern about our religiosity indicator is how much variation in county-religiosity names explain, net of that captured by fixed effects. In Appendix 2.11, we provide several robustness and validation exercises for our synthetic measure. First, Figure 2.7 provides county-binned scatters of synthetic and measured religiosity by denomination. The figure summarizes the results from two distinct exercises. Plots in the left column show in-sample correlations, thus comparing Census-measured and predicted religiosity in 1906 and 1916. Plots in the right column compare synthetic and measured religiosity in 1926 instead.[26] We refer to this as an "out-of-sample" correlation, as data from the Censuses of Religious Bodies carried out after the pandemic are not used to estimate religiosity scores. All graphs show a positive correlation between actual and predicted religiosity across all denominations. This exercise provides reassuring evidence that naming patterns capture meaningful variation in overall religiosity and further validates our measure.

Next, following Abramitzky, Boustan and Eriksson (2016), we use biblical and saint names as an alternative name-based measure of religiosity. Finally, as additional indicator

[26]Our results do not change if we include data from the 1936 Census of Religious Bodies. However, growing discontent resulted in substantially lower reporting rates in this last Census for some religious groups. Following Stark (1992), we consider it less reliable and exclude it from our analysis.

of religious sentiment, we use the county-level share of the population with a religious affiliation (for all affiliations, and separately for Catholics and Protestants) recorded by the Census of Religious Bodies for 1906, 1916, and 1926.

2.3.2 Measuring Scientific Progress

We measure scientific progress at the local level using the share of individuals employed in STEM occupations. The rationale for this measure is that a STEM occupation requires that an individual receive a science-oriented education. In turn, receiving a science-based education plausibly correlates with more favorable attitudes toward science and scientific progress at the local level (Bianchi and Giorcelli; Biasi, Deming, Moser and Dillon, 2020; 2022). For each county and census year (1900 to 1930), we compute the share of individuals employed in a STEM occupation relative to (i) the entire population; (ii) the number of people employed in high-skilled occupations.[27] We also use these two classifications into STEM and non-STEM occupations when performing the individual-level analysis.

We complement our main measure of scientific progress by using patent data from the Comprehensive Universe of U.S. Patents (Berkes, 2018). The CUSP contains information about the universe of U.S. patents issued between 1836 and 2015. The data for the time period considered in our paper (1900–1930) are extracted from digitized patent documents obtained from the U.S. Patent and Trademark Office. For the purpose of our analysis, we first assign each patent to a county, based on the residence of its inventor, and a year, based on the year in which the patent was filed. When a patent lists multiple inventors, we give equal weights to the location of each inventor. From the CUSP, we also collect the technology classes associated with each grant (according to the U.S. Patent Classification system) and assign them to technology groupings following the crosswalk provided by the National Bureau of Economic Research (Hall, Jaffe and Trajtenberg,

[27]This second measure increases the comparability of the control group with STEM individuals. Table 2.7 lists the set of occupations that we label as STEM (Panel A) and the occupations that we classify as high-skilled (Panel B). By construction, STEM occupations are also high-skilled. Individuals in STEM occupations represent approximately 6% of those employed in skilled professions in the 1930 census.

2001).[28] Importantly, while the use of patents as a proxy for scientific progress may be subject to debate (Moser, 2005), the main advantage of these data is that they are available at a high-frequency. We will especially use them to quantify the dynamic treatment effect of the influenza.

2.3.3 Exposure to the Great Influenza Pandemic

To measure the incidence of the Great Influenza Pandemic across U.S. counties, we use mortality statistics assembled by the U.S. Department of Commerce. These were first collected in 1915, and throughout the 1915–1918 period, they covered 1,274 counties (40% of the total), accounting for more than 60% of the U.S. population. We follow the methodology developed by Beach, Clay and Saavedra (2020) and measure mortality caused by the flu as average deaths during the flu period (1918–1919) relative to the three years before the pandemic (1915–1917). Formally, excess mortality in county c is defined as

$$\text{Excess Deaths}_c = \frac{\frac{1}{2}\sum_{t=1918}^{1919} \text{Deaths}_{ct}}{\frac{1}{3}\sum_{t'=1915}^{1917} \text{Deaths}_{ct'}}. \tag{2.3}$$

This measure represents our baseline treatment. We also report results from a categorical treatment variable equal to one if the baseline treatment (Excess Deaths_c) is above its median and zero otherwise. Figure 2.2 displays the geographical variation in the intensity of the pandemic in terms of excess deaths. We find that the severity of the pandemic varies substantially across counties, even geographically close ones. The rationale behind our excess-mortality measure is that—all else being equal—deaths during the pandemic that exceed those before the pandemic are likely due to the pandemic itself. A possible threat to this argument might be the U.S. involvement in WWI and the fact that WWI deaths are confounding our results. However, there does not appear to be a significant correlation between WWI morality and the pandemic, as displayed in Figure 2.9. In

[28] Whenever a patent is assigned to more than one field, we split it with equal weights across fields. We conflate the "chemical" and "drugs" NBER classes into a single class, which we label "pharmaceuticals." We follow this approach because of the high collinearity between the number of chemical and drug patents at the county level, which would make it difficult to study them separately. All results for the pharmaceutical class also hold if we consider drug and chemical patents separately. An example of a pharmaceutical patent is shown in Figure 2.8. For historical consistency, we relabel the "computer and communication" class as simply "communication."

Section 2.4, we show that our results are robust to controlling for a post-1918 time indicator interacted with WWI-related deaths.

2.4 Main Results: County-Level Analysis

In this section, we present the baseline empirical strategy and we show that exposure to the influenza pandemic led to an increase in both religiosity and scientific progress across counties. Then, we explore the mechanism behind the aggregate patterns and provide evidence of heterogeneous responses to the pandemic *within* counties.

2.4.1 Empirical Strategy

In the first part of the analysis, we study the impact of the pandemic separately on religiosity and scientific progress at the county level. Our sample consists of a panel of U.S. counties observed over the 1900–1929 period at a decade or yearly frequency. In particular, we leverage quasi-random variation in exposure to the pandemic across U.S. counties in a difference-in-differences (DiD) setting and estimate regression models of the form:

$$y_{ct} = \alpha_c + \alpha_{s(c)\times t} + \boldsymbol{x}'_{ct}\boldsymbol{\beta} + \delta \times (\text{Post}_t \times \text{Excess Deaths}_c) + \varepsilon_{ct}, \tag{2.4}$$

where the subscripts c and t denote county and time (decade or year), respectively; y_{ct} measures either religiosity or scientific progress; α_c and $\alpha_{s(c)\times t}$ are county and state-by-time fixed effects; Post_t is an indicator variable equal to one if $t \geq 1918$ and zero otherwise; Excess Deaths_c measures the intensity of the pandemic in terms of excess deaths, as explained in Section 2.3.3; and ε_{ct} is the error term. In addition, in all regressions, we control for an interaction term between 1910-population and the post indicator $\boldsymbol{x}'_{ct}$. Standard errors are clustered at the county level. Our coefficient of interest, δ, captures the impact of the pandemic on religiosity or scientific progress. To investigate possible heterogeneity of treatment effects over time, we also estimate a more flexible model that, rather than interacting Excess Deaths with the Post indicator, interacts Excess Deaths

with biennial time dummies:[29]

$$y_{ct} = \alpha_c + \alpha_{s(c)\times t} + \boldsymbol{x}'_{ct}\boldsymbol{\beta} + \sum_{\substack{k=1909 \\ k\neq 1917}}^{1928} \delta^k \times I\{t = k \,|\, t = k+1\} \times \text{Excess Deaths}_c + \varepsilon_{ct}, \quad (2.5)$$

where $I\{t = k \,|\, t = k+1\}$ is an indicator variable that takes value one if t is in the two-year window indexed by k, and zero otherwise.

Did the influenza spread randomly? We perform three main exercises to test this in the data. First, in Table 2.8, we report the correlation between the intensity of the pandemic and religiosity, scientific progress, and a set of county covariates, all measured in 1910, the last census before the pandemic (or in the decade 1901–1910 in the case of yearly variables) accounting for population and state-level fixed effects.[30] Counties more exposed to the pandemic are observationally equivalent with respect to all variables except the share of foreigners. This aligns with the pandemic being comparatively more severe in urban areas, where immigrants were clustered.

Additionally, to rule out that these differences confound our analysis, we check whether control and treatment counties were on different trends before the shock by estimating event studies. Formally, in Equation (2.5), this implies that the estimates of δ^k would not be statistically different from zero before the pandemic hit, i.e., for all $k < 1917$.[31] Our estimates support the parallel-trends assumption. However, this approach could still be invalid in the presence of shocks correlated with the intensity of the pandemic that positively affected both science and religiosity but that were *not* caused by the pandemic itself. A plausible candidate is the number of soldiers that counties lost in WWI: their deaths might have driven either the religiosity of their families or the ability

[29] In the dynamic DiD specifications, we code periods over two-year windows to reduce noisy fluctuations in estimated treatment effects and to improve efficiency by pooling observations. We consider a 10-year period before and after 1917-18.

[30] State fixed effects control for the fact that the pandemic spread from East to West between August 1918 and November 1918.

[31] Since the setting is not staggered—because the pandemic hit each county in the same period—equations (2.4) and (2.5) can be estimated through standard two-way fixed effects (Callaway and Sant'Anna; Sun and Abraham, 2021; 2021). Callaway, Goodman-Bacon and Sant'Anna (2021), however, caution against using continuous treatments. We code a binary indicator equal to one for counties with above-median excess deaths. Throughout the paper, we show that the continuous and binary treatments yield qualitatively similar results.

(or motivation) of a county to produce innovation (or both). To test for this possibility, in Tables 2.10 and 2.18, we control for the number of deaths in WWI in our regression equation and show that the results remain robust.

2.4.2 The Effect of the Influenza Pandemic on Religiosity

Table 2.1 displays the DiD estimates focusing on religiosity. In columns (1–3), the dependent variable is the share of individuals affiliated with any religious denomination (column 1), with a Catholic religious denomination (column 2), or with a Protestant religious denomination (column 3), as enumerated in the Census of Religious Bodies. The estimates suggest that counties comparatively more exposed to the pandemic experienced an increase in religiosity, with no significant differences between Catholics and Protestants.

This measure of religiosity has the advantage of including the U.S. population across different age groups. However, it has two main caveats: (i) census-based religiosity is available only at three points in time (1906, 1916, and 1926) and thus does not allow us to study high-frequency variation in religiosity; (ii) the choice to join a religious denomination could be more likely to be affected by social insurance considerations, rather than by religious reasons, thus inducing an upward bias in our results (Ager and Ciccone, 2018). Thus, in columns (4–6), we use our main indicator of religiosity, the name-based measure described in Section 2.3.1. This measure allows us to observe counties every year between 1900 and 1929. We find that a one-standard-deviation increase in excess deaths led to a 0.16 standard deviation increase in name-based religiosity at the county level (column 4). Similarly, moving from a county at the 25th percentile of the excess mortality distribution to one at the 75th percentile led to an increase in religiosity of 10%.

In Figure 2.3, we report the coefficients of the interactions between the treatment variable and biennial dummies using the name-based measure of overall religiosity as the dependent variable. The event study estimates support the patterns observed in the DiD

analysis and confirm the absence of pre-trends. In addition, we observe that the increase in religiosity appears to persist over the decade after the pandemic. These findings are in line with the literature documenting a substantial persistence of religiosity (Squicciarini, 2020).

We now perform a series of additional exercises to gauge the robustness of these findings. First, one may be worried that our results were driven by small counties where the variation in the share of individuals affiliated with religious denominations and in naming patterns may have been more substantial. Table 2.9 replicates the specifications of Table 2.1, weighting regressions by county population in 1910. The baseline findings are confirmed.

Next, we focus on our name-based measure of religiosity and perform three sets of exercises. First, we show that our results hold when changing the specification or the sample. In column (2) of Table 2.10, we code the treatment as a binary variable equal to one if the baseline treatment is above its median and zero otherwise. Then, column (3) controls for mortality due to WW1. In addition, one concern related to our religiosity measure could be that firstborns are often named after a deceased grandparent. Thus, their names would reflect the higher religiosity of previous generations rather than their parents' religious attitudes. If, due to higher mortality, households in areas more affected by the pandemic were also more likely to have recently lost a grandparent, then our results might reflect a mechanical effect. Column (4) reports estimates when dropping firstborn children in every household. Another concern is that numerous households may display different naming behaviors for later-born children. In column (5), we drop children beyond the fourth. Then, if religious families displayed higher fertility rates, one may worry that our results are driven by an increase in the number of religious names due to the higher fertility of already religious households. In column (6), we compute within-household average religiosity to check whether our findings are driven by larger households and differential fertility. All results hold through these alternative specifications. Another concern could be that comparatively more religious people moved into counties where the pandemic had been more severe, perhaps motivated by slacker labor markets. If that were the case, our estimated pandemic effect on name-based religiosity would reflect movers' religiosity

and fertility. To address this concern, we compute a county-decade measure of religiosity based on the names of the adult population only. The in-migration mechanism would predict a positive impact of the pandemic on this variable. Estimates reported in column (7) show no evidence of any such effect, thus ruling out this potential alternative interpretation.

Table 2.11 addresses the possibility that immigration confounds our results. In particular, one may be worried that immigration into more exposed areas by selectively more religious individuals may drive our estimates. We thus exclude all those who, in the 1930 census, are recorded residing in a state that is different from the one where they were born. The results remain unchanged. Finally, we explore the effect of the pandemic within urban areas using the city-level sample constructed by Clay, Lewis and Severnini (2019) and discussed in Appendix 2.10.8. In columns (1–3) of Table 2.12, we estimate equation (2.5) using a balanced panel of cities observed over the 1900–1929 period. The name-based religiosity measure is computed leveraging variation in naming patterns of children born in each city. Results are consistent with the baseline county-level analysis: Religiosity increased in cities more severely affected by the pandemic. This exercise ensures that our results are not driven by individuals residing in rural areas. Moreover, the city-level sample includes several cities in Southern states, which were plausibly more religious.[32] We thus view the city-level exercise as shedding additional internal validity to the county-level analysis.

In the second set of exercises, we test whether the results are robust to alternative ways of constructing our religiosity measure. First, in Table 2.13, we report the baseline result, but using religiosity scores estimated through equation (2.1) *without* county fixed effects. These scores are thus obtained using the "stock" of religiosity in a given county instead of its deviations from the mean. The results from this alternative strategy are consistent with our baseline estimates. Second, we test the robustness of our results to the number of names included in the sample. In our baseline analysis, we exclude names appearing in less than 0.3% of the overall population. Table 2.14 shows that our findings are qualitatively unchanged under different frequency thresholds. Next, a possible concern

[32]Figure 2.12 reports the number of cities included in the sample by state and their location.

could be that the results capture a "fashion" effect, whereby more religious names became more fashionable after the pandemic. If this were the case, even though the initial increase in religious names would suggest a positive shift in religiosity, the effect for the following periods would be biased upwards and driven by this fashion effect. In Table 2.15, we regress a set of indices reflecting the concentration of the name distribution against our baseline treatment and find no evidence of such a mechanism.

Finally, we perform our analysis using an alternative indicator of religiosity. In particular, in Table 2.16, we use biblical and saint names as an alternative name-based measure of religiosity, following Abramitzky, Boustan and Eriksson (2016). We find that the number of children named after biblical or saint names increased in areas more exposed to the pandemic after 1918. In addition, in Figure 2.10, we show that the county-level share of Biblical and Saint names, computed using data from Abramitzky, Boustan and Eriksson (2016), is strongly and positively correlated with the religiosity measure constructed using our data-driven approach.

Overall, we document that the pandemic positively affected religiosity through different specifications, samples, and indicators. This finding is consistent with the religious coping hypothesis, which posits that religion may serve as a coping device to deal with mental and psychological distress (Pargament; Bentzen; Bentzen, 2001; 2019; 2021).

2.4.3 The Effect of the Influenza Pandemic on Scientific Progress

We now turn to study how the influenza pandemic impacted scientific progress. We show that the pandemic positively impacted the share of people in STEM occupations and patenting activities.

Table 2.2 (column 1) runs the specification of equation (2.4) using as the dependent variable the share of individuals employed in STEM relative to the overall population. We perform the analysis at the decade level since this measure is taken from population censuses (1900–1930). We observe an increase in the share of workers in STEM occupations in counties more severely hit by the pandemic. A one-standard-deviation

increase in excess deaths is associated with a 0.985 standard deviation increase in the share of individuals in scientific occupations. Equivalently, moving from the 25th to the 75th percentile of the excess mortality distribution leads to a 31% increase in the share of individuals in STEM. Column (2) replicates this specification, focusing on the skilled sub-sample of the population, and provides similar results.

Next, we use individual-level data on occupations to better understand what drives the change in occupational shares. Specifically, we test whether individuals who were young at the time of the shock, i.e., between 18 and 25 years old, were more likely to be employed in a STEM occupation in 1930 compared to older cohorts in areas that were comparatively more exposed to the pandemic.[33] We estimate the following linear probability model, where we define the treated individuals as those aged between 18 and 25 years old in 1918:

$$\text{STEM}_i = \alpha_{c(i)} + \alpha_{t(i)} + \boldsymbol{x}_i'\boldsymbol{\beta} + \delta \times \left(\text{Excess Deaths}_{c(i)} \times \text{Young}_i\right) + \varepsilon_i, \tag{2.6}$$

where i denotes an individual living in county $c(i)$ and born in year $t(i)$. The terms $\alpha_{c(i)}$ and $\alpha_{t(i)}$ respectively denote county and cohort fixed effects, STEM_i is a dummy variable equal to one if i is employed in a STEM occupation, and zero otherwise; $\boldsymbol{x}_i$ includes urban status and race. The categorical variable Young_i equals one if individual i is between 18 and 25 in 1918, and zero otherwise. Our coefficient of interest is δ, which measures the causal effect of the pandemic on the probability of being employed in a STEM occupation.

Table 2.3 reports the results. In counties more exposed to the pandemic, young individuals were significantly more likely to sort into STEM occupations. Why did young cohorts respond disproportionately more to the shock? We have two potential explanations for this finding. The first is mechanical: the pandemic may have affected everyone similarly, but young cohorts were the only ones choosing an occupation. The second one is that the pandemic may have specifically affected the attitudes and preferences of individuals in their *impressionable years* (i.e., the young cohorts). Thus, the differential occupation

[33]To construct the sample, we use the cross-section of all individuals in the 1930 full-count census. We exclude all individuals born after 1900, as they may have been too young to select an occupation, and we restrict the sample to the working population. We drop individuals with no valid occupational response, and we exclude farmers because they display disproportionately high intergenerational occupational persistence (Long and Ferrie, 2013).

choice reflects a change in attitudes occurring only for these cohorts.[34]

Then, we use our second proxy for scientific progress and focus on patents. Even if patents may be an imperfect measure for innovation and scientific attitudes (Moser, 2005), the main advantage of the patents data is that they allow us to construct a high-frequency indicator of scientific progress. Column (3) of Table 2.2 runs the specification of equation (2.4) and reports the estimated impact of the influenza shock on the total volume of innovation—measured as the log(1 + number of patents) in a given county-year. We find that a one standard deviation increase in excess deaths led to a 0.17 standard deviation increase in the number of patents. Similarly, moving from a county at the 25th percentile to one at the 75th percentile of the excess-deaths distribution leads, on average, to an increase of 7% in the number of patents granted by county year.

Figure 2.4 shows the results in an event-study framework. Each dot in the plot reports the dynamic treatment effect of the pandemic on innovation in the indicated two-year window, as specified in equation (2.5). The coefficients suggest that the number of patents granted after the pandemic increased significantly more in more exposed counties, implying that the pandemic induced a sizable increase in innovative activities that persisted for at least one decade after the shock.

We then investigate the heterogeneous effects of the pandemic across technology classes. Specifically, we ask whether the influenza shock affected not only the volume but also the *direction* of innovation. To do so, we study the effect of the shock on the number of patents in each sector, controlling for the total number of patents. Columns (4)–(8) of Table 2.2 show the results of this exercise. For each field, we report the estimated DiD coefficients. We find that the influenza shock has a positive and statistically significant effect only on pharmaceutical patents. Keeping the total number of patents constant, a county at the 75th percentile of the excess-deaths distribution saw an average increase of 6% in pharmaceutical patents, compared to one at the 25th percentile.

[34] According to the "impressionable years" hypothesis—which represents a long-standing argument in psychology—economic, social, and cultural attitudes and beliefs are formed during early adulthood, approximately between the ages of 18 and 25, and change slowly thereafter (Giuliano and Spilimbergo, 2023). Another explanation could be that there is a higher demand for STEM jobs.

We now run a series of robustness checks. First, as in the case of religiosity, one concern is that our results were driven by small counties where scientific progress was comparatively low during the pandemic. In Table 2.17, we replicate the specification of Table 2.2 weighting regressions by their 1910 population. The results hold.

Next, we focus on the patent data. Table 2.18 uses as a dependent variable the total number of patents irrespective of their field (columns 1–4) and the total number of patents in pharmaceuticals (columns 5–9). In columns (2) and (7), we restrict the sample to an unbalanced county-year panel that includes only county-years with at least one filed patent. Columns (3) and (8) report results coding the treatment as a binary variable. Columns (4) and (9) control for WWI deaths interacted with the post-treatment indicator and confirm that WWI-related mortality is not driving our result. Column (6) omits the total number of patents as a control, thus reporting the impact of the pandemic on the volume of pharmaceutical patents. The corresponding coefficient should be interpreted as the impact of the pandemic on the total number of pharmaceutical patents. The estimated DiD coefficients remain positive and statistically significant throughout.

In the baseline specifications, we take the logarithm of the number of patents and add one to avoid dropping zeros. In Tables 2.19 and 2.20, we use alternative transformations of the dependent variable—e.g., the share of patents in pharmaceuticals—and obtain quantitatively similar findings. Additionally, we estimate the baseline model as a Poisson Quasi-Maximum Likelihood regression. The results hold.

A plausible concern is that our results may be driven by "low-quality" innovation. Newspapers of the day often advocated remedies for influenza that were not science- or evidence-based, some of which may have been granted a patent. To address this concern, we use the text-based measures of "importance" developed by Kelly, Papanikolaou, Seru and Taddy (2021).[35] Table 2.21 shows the results. In particular, we assign to every patent a dummy equal to one if the patent's importance is in the top 20% of the

[35]As discussed by Berkes (2018) and Andrews (2021), citation-based quality measures during this period are noisy and mostly uninformative due to the lack of a mandatory reference section until 1947. The measure built by Kelly, Papanikolaou, Seru and Taddy (2021) identifies important patents based on the textual similarity of a given patent to previous and subsequent work. Important patents are those that are distinct from previous work but are similar to subsequent innovations.

distribution and zero otherwise. The number and share of these "breakthrough patents" substantially increase in counties hit harder by the pandemic, both in all sectors (columns 1–2) and in pharmaceuticals (columns 3–4). In addition, in column (5), we show that the number of breakthrough pharmaceutical patents grows more than the average number in other sectors.

Table 2.22 deals with the potentially confounding role of immigration. In particular, we exclude from the estimation sample all those who, in the 1930 census, are recorded residing in a state that is different from the one where they were born. The results remain unchanged. Finally, while most innovation activity clusters in urban areas, we perform our baseline analysis at the level of counties. To ensure that the results do not conflate rural-urban disparities, we estimate the effect of the pandemic on innovation at the city level. Table 2.23 reports the estimates of equation (2.1) for the panel of cities described in Section 2.10.8. The results confirm the county-level evidence: despite the smaller sample size, we estimate the pandemic's positive and statistically significant effect on innovation, especially in pharmaceuticals.

2.4.4 Joint Dynamics of Religiosity and Innovation

After studying the impact of the pandemic separately on religiosity and scientific progress, we now look at their joint evolution. Specifically, we test whether the *same* counties were affected along both dimensions or whether some counties saw an increase in religiosity while others saw an increase in scientific progress.

We estimate the following equation:

$$\begin{aligned} y_{ct} =& \alpha_c + \alpha_{s(c)\times t} + \boldsymbol{x}'_{ct}\boldsymbol{\beta} + \delta_1 \times (\text{Excess Deaths}_c \times \text{Post}_t) + \delta_2 \times \text{Religiosity}_{ct} + \\ &+ \delta_3 \times (\text{Religiosity}_{ct} \times \text{Post}_t) + \delta_4 \times (\text{Religiosity}_{ct} \times \text{Excess Deaths}_c) + \\ &+ \delta_5 \times (\text{Excess Deaths}_c \times \text{Post}_t \times \text{Religiosity}_{ct}) + \varepsilon_{ct}, \end{aligned} \tag{2.7}$$

where y_{ct} is the number of patents normalized by county population in 1910, follow-

ing Bénabou, Ticchi and Vindigni (2022), and Religiosity_{ct} is the religiosity measure described in Section 2.3.1. The coefficient δ_1 measures the impact of the pandemic on scientific progress, δ_2 captures the correlation between the outcome and religiosity before the pandemic, and the term δ_5 captures how the correlation between the outcome and religiosity changes after 1918 as a function of exposure to the pandemic. As before, the vector $\boldsymbol{x}_{ct}$ includes an interaction term between the county population in 1910 and a post-treatment indicator.

In Table 2.24, we report the estimates of equation (2.7). The results suggest that counties comparatively more affected by the pandemic experienced a joint increase in religiosity and innovation. Columns (1) and (2) report the correlation between innovation and religiosity before and after the pandemic, respectively. Interestingly, this relationship shifts from negative to positive–as shown in Figure 2.11. Indeed, in the period before the shock, there was a negative correlation between scientific progress and religiosity at the county level. This pattern aligns with contemporary evidence reported by Bénabou, Ticchi and Vindigni (2015). After the pandemic, however, religiosity and science became positively correlated. This pattern is confirmed in column (3), which pools together observations before and after the pandemic. In column (4), we then estimate regression (2.7) and find that this shift in the correlation between religiosity and innovation co-moved with county-level exposure to the pandemic. In Section 2.5, we use individual-level data to uncover the possible mechanisms underlying these results.

2.5 Mechanisms: Individual-Level Analysis

Two questions naturally arise after observing a contemporaneous increase in religiosity and scientific progress. Within counties, who turns to religion, and who turns to science? Are these the same or different individuals? In this section, we leverage individual-level data to answer these questions. In particular, we focus on individuals who were heads of household in the 1930 census.[36]

[36]The "head of household" variable is provided by the census. During this period, the father and/or husband were usually the head of the household whenever present.

First, we show that the pandemic led to an increase in the religiosity of individuals from initially more religious backgrounds, while individuals from less religious backgrounds were more likely to select STEM occupations. Second, we show that STEM individuals became less religious than the rest of the population. Third, we document that the pandemic led to the polarization of religiosity.

Taken together, these three results suggest that the pandemic shock led to different reactions within society—based, for instance, on individuals' religious background or initial scientific orientation—with people becoming even more distant in terms of their religiosity than they were before the pandemic. This within-county analysis reveals important heterogeneity in how individuals react to a negative shock, and it helps reconcile our aggregate findings with the existing literature on the negative relationship between religion and scientific progress.

2.5.1 Turning to Religion or Turning to Science

We first study whether preexisting differences in individuals' religious backgrounds could have led to a heterogeneous response to the influenza shock. The full-count census data, in addition to covering the universe of the U.S. population, has the advantage of being deanonymized. These data allow us to construct two measures of religiosity for each individual: one is their revealed religiosity, based on the names individuals gave to their children; the other is their religious background, based on their own name. Specifically, we interpret an individual's own name as conveying information about the religiosity of their parents and, thus, the religious background of that individual.

Combining these measures, we first study how an individual's religious background shaped their response to the pandemic in terms of religiosity. Next, we explore whether, following the pandemic, an individual's religious background may have also shaped their propensity to work in a scientific occupation. To measure this, we use an indicator equal to one if they were employed in a STEM occupation and zero otherwise.[37]

[37]A natural way to construct a measure of scientific background, symmetric to the religiosity one, would be to look at whether individuals had a parent working in a scientific occupation. Unfortunately,

We estimate two triple-difference specifications, one for religiosity and one for the likelihood of selecting a STEM occupation. In the first case, we observe each household multiple times, once per child, and estimate the following regression:

$$\begin{aligned}\text{Religiosity}_{it} = \alpha_{c(i)\times t} + \alpha_{c(i)\times B(i)} + \alpha_{B(i)\times t} + \boldsymbol{x}_i'\boldsymbol{\beta} + \\ + \delta_1 \times \left(\text{Excess Deaths}_{c(i)} \times \text{Post}_{it} \times \text{High Religious Background}_i\right) + \varepsilon_{it},\end{aligned} \tag{2.8}$$

where Religiosity_{it} denotes the religiosity score of a child born in year t in household i, living in county $c(i)$. The term ($\text{High Religious Background}_i$) is an indicator variable returning a value of one if the average religiosity of the adults in the household is in the top 50% of the overall distribution of the religiosity background and zero otherwise. The term Post_t is a categorical variable equal to one for children born after 1918 and zero otherwise. We estimate model (2.8) on the sample of children born between 1900 and 1929, and each child is weighted by the inverse of the number of children in each household.

To investigate the heterogeneous responses of occupational choice, we estimate the following regression:

$$\begin{aligned}\text{STEM}_i = \alpha_{c(i)\times t} + \alpha_{c(i)\times B(i)} + \alpha_{B(i)\times t(i)} + \boldsymbol{x}_i'\boldsymbol{\beta} + \\ + \delta_2 \times \left(\text{Excess Deaths}_{c(i)} \times \text{Young}_i \times \text{High Religious Background}_i\right) + \varepsilon_i,\end{aligned} \tag{2.9}$$

where i denotes an adult individual born in year $t(i)$. In this case, each adult is observed once, and the term Young_i is an indicator equal to one for individuals between 18 and 25 when the pandemic hit. The background religiosity term ($\text{High Religious Background}_i$) denotes an indicator returning a value of one if the religiosity score of the name of individual i is in the top 50% of the overall distribution and zero otherwise. The dependent variable is an indicator variable returning the value one if the head of household i is employed in a STEM occupation in 1930 and zero otherwise.

this is not possible due to data limitations, as this exercise would require tracking individuals across several census waves, thus greatly reducing our sample size. The advantage of our measure of religious background is that it can be constructed for every individual without requiring direct information on or linking to their parents.

In both equations, the terms $\alpha_{c\times t}$, $\alpha_{c\times B}$, and $\alpha_{B\times t}$ denote, respectively, county-by-year, religious-background-by-county, and religious-background-by-year fixed effects, and $\boldsymbol{x}_i$ includes urban status and race of the household head. The coefficients δ_1 and δ_2 quantify the effect of county-level exposure to the pandemic, comparing individuals in the top quintile of the background religiosity distribution with the rest of the population on, respectively, religiosity and STEM employment. Table 2.4 presents the results of the analysis. In columns (1)–(3), the dependent variable is revealed religiosity. Our variable of interest is the interaction between the excess-deaths measure, the "Post" dummy, and the religious background of the household head. In columns (4)–(6), the outcome variable is a dummy variable for STEM occupations, and the main variable of interest is the interaction between the excess-deaths measure, a "Young" dummy, and their religious background.

We find that individuals from more religious backgrounds, who were already more religious before the influenza shock, became even more religious afterward in more exposed counties (columns 1–3).[38] By contrast, individuals who were young during the pandemic and came from less religious backgrounds were more likely to choose a scientific occupation (columns 4–6). These findings suggest that an individual's religious background affects their way of dealing with negative shocks. In particular, those who were raised by religious parents were more likely to resort to religion to deal with adversity. On the other hand, growing up in a less religious household made individuals more likely to approach science, possibly as a coping device in the face of the negative shock.

2.5.2 Science-Oriented Individuals Became Less Religious

In this part of the analysis, we focus on science-oriented individuals and study whether their religiosity changed after the pandemic compared to the rest of the population.

[38]The correlation between revealed religiosity and background religiosity is equal to 0.13 and highly statistically significant ($p< .001$), in line with a large literature on cultural transmission (Bisin and Verdier, 2001).

We estimate the following triple-differences model:

$$\begin{aligned}\text{Religiosity}_{it} = \alpha_{c(i)\times\text{STEM}_i} + \alpha_{t\times\text{STEM}_i} + \alpha_{c(i)\times t} + \boldsymbol{x}_i'\boldsymbol{\beta} + \\ + \delta \times \left(\text{Excess Deaths}_{c(i)} \times \text{STEM}_i \times \text{Post}_t\right) + \varepsilon_{it},\end{aligned} \tag{2.10}$$

where the dependent variable is the religiosity of a child born in household i living in county $c(i)$ in year t. The term Post_t is a dummy variable taking the value one if the child is born after 1918, and zero otherwise; STEM_i is an indicator variable that takes the value one if at least one member of the household is employed in a STEM occupation; and $\boldsymbol{x}_i$ includes urban status and race of the household head. The coefficient δ compares STEM and non-STEM households, before and after the pandemic, by county-level exposure to the pandemic. The sample is composed of all children born between 1900 and 1929. Children are weighted by the inverse of the number of children in each household. Table 2.5 shows the results. In columns (1)–(3), the comparison group is the entire population, while in columns (4)–(6), we focus on high-skilled workers. We find that, for both comparison groups, individuals in STEM occupations became less religious than non-STEM ones in counties more exposed to the influenza shock (columns 1 and 4). This pattern is stronger for Protestants (columns 3 and 6) than Catholics.

These findings further show that different groups within society reacted differently to an adverse shock. In particular, STEM individuals appeared to turn further away from religion than their non-STEM counterparts.

2.5.3 Polarization of Religious Beliefs

We conclude the individual-level analysis by studying the impact of the influenza pandemic on the distribution of religiosity within counties. In particular, we estimate the heterogeneous treatment responses to the pandemic across the initial distribution of background religiosity.

To study this question, we first discretize the distribution of background religiosity into

quintiles, which we label Q^{BR}. Then, we estimate the following model:

$$\begin{aligned}\text{Religiosity}_{it} = \alpha_{c(j)\times t} + \alpha_{c(j)\times Q(i)} + \alpha_{Q(i)\times t} + \\ + \sum_{\substack{k=1 \\ k\neq 3}}^{5} \delta^k \times \left[\text{Excess Deaths}_{c(i)} \times \text{Post}_{it} \times I\left(Q_i^{\text{BR}} = k\right)\right] + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_{it},\end{aligned} \tag{2.11}$$

where the outcome variable measures the religiosity score associated with the name of a child born in household i living in county $c(i)$ in year t. As in the previous analyses, the term Post_{it} is an indicator variable equal to one for children born after the pandemic and zero otherwise. Equation (2.11) includes county-by-time, county-by-background, and background-by-time fixed effects, and the term $\boldsymbol{x}_i$ includes urban status and race of the household head. The term $I\left(Q_i^{\text{BR}} = k\right)$ is a dummy variable that takes the value one if the household average background religiosity is in the k-th quintile and zero otherwise. If the pandemic caused an increase in polarization of religiosity, the set of coefficients $\{\delta^k\}_{k=1}^5$ in equation (2.11) would be monotonically increasing in k. On the other hand, a decreasing sequence of coefficients would provide evidence that the pandemic led to a convergence of religiosity. In model (2.11), the sample comprises all children born between 1900 and 1929. Children are weighted by the inverse of the number of children in each household.

In Figure 2.5, we report the set of $\{\delta^k\}$ coefficients by religious denominations. We normalize the third quintile as the baseline category. The figure provides evidence of an increase in polarization: for individuals with below-median religious backgrounds, the coefficients on exposure to the pandemic are negative, while they are positive for those with above-median religious backgrounds. This pattern suggests that, within the same county, individuals from different religious backgrounds became even more distant in terms of their religiosity, increasing the polarization of religiosity within society.

These three individual-level exercises help us understand the contemporaneous increase in religiosity and science at the county level. They suggest that, within counties, individuals reacted differently to the same shock, based, for instance, on their religious background or their pre-pandemic scientific orientation. Thus, while a county may have become more

religious *and* more innovative, individuals seemed to turn either to religion *or* to science, leading to within-county polarization of religiosity.

2.6 Discussion: Interpretation and Limitations of the Results

Our analysis shows two clear patterns: (i) the 1918–1919 influenza pandemic led to an increase in religiosity and scientific progress across U.S. counties and, as a result of the shock, the same counties became both more religious and more scientific; (ii) *within* counties, there was a heterogeneous response to the same shock, with some individuals turning to religion and others turning to science.

One concern is that other factors related to the pandemic may have affected the evolution of religiosity and science, confounding our results. To address this concern, we proceed in three steps. First, we document that neither initial religiosity nor scientific progress is related to the intensity of the shock. Second, our event-study analysis shows the absence of pretrends, suggesting that religiosity and science were on a similar path in treated and control groups before the shock. Third, we account for other potentially confounding characteristics, such as differential fertility, WWI deaths, and migration patterns. Our results are robust in all these cases. The empirical evidence, supported by historical records, makes it hard to imagine that the pandemic did not trigger an increase in both religiosity and scientific progress.

A second concern regards our main measures of religiosity and scientific progress. First, does our name-based indicator indeed capture religiosity at the local level? We show that our results are robust to alternative ways of constructing our naming measure and using alternative classifications of religious names. In addition, we show that in counties hit harder by influenza, the share of people affiliated with a religious denomination increases, providing further evidence that the pandemic led to an increase in local religiosity. Similarly, we use two measures for scientific progress: the share of individuals in scientific

occupations and patents. While each may have some caveats, they provide consistent and robust results.

One puzzle emerging when looking at the aggregate patterns is whether these results are driven by the same individuals becoming more religious and innovative or by different individuals reacting differently to the same shock. Our findings suggest that the second mechanism is at play. Individuals from more religious backgrounds embrace religion, while those from less religious backgrounds are more likely to choose a scientific occupation. This finding suggests that a group of individuals within society used religion as a coping device, while a separate group turned to science. In addition, we show that the shock widened the distance in religiosity between science-oriented individuals and the rest of the population, as people in scientific occupations moved away from religion. Finally, the pandemic increased the polarization of religiosity in society: individuals from more (less) religious backgrounds became even more (less) religious.

One key question regarding our individual-level results is, what explains the increase in religiosity or the choice of a scientific occupation? The findings on religiosity are in line with the religious coping hypothesis, which suggests that religious faith can represent a coping device to deal with personal distress following a negative shock. An alternative explanation for why individuals may turn to religion is for social insurance. While we cannot fully rule this out (and it goes beyond the scope of our paper), we read our evidence as favoring the religious coping hypothesis. First, this is in line with the literature showing that intrinsic religiosity (rather than churchgoing) responds to unexpected events, as noted by Bentzen (2019). Second, as the increase in religiosity persists up to ten years after the shock, it is more likely to be related to a change in beliefs rather than a temporary increase in the need for social insurance.

What motivates people to turn to science is less obvious. Individuals may turn to science to deal with their psychological distress, similarly to religious coping, or in an attempt to actively mitigate the negative (e.g., health-related or economic) effects of the pandemic. Another possibility could be that individuals turn to science because of increased labor demand in STEM occupations, but our results suggest that, beyond market forces, the

individual's religious background plays a key role in the decision to turn to science. While our findings cannot directly speak to the individual-level motivations behind these different behaviors, they provide evidence of a heterogeneous response to the same adverse event.

One further limitation of our individual-level analysis is that, while we can construct the religious background for every individual, we cannot directly measure their scientific one. This is due to our measure of scientific orientation based on occupational choice, which—contrary to our measure of religious background—does not allow us to know an individual's occupation and the parents' occupation from the same census.

Taken together, we interpret our results as suggestive evidence that, while individuals from religious backgrounds turned to religion as a coping device in the aftermath of the pandemic, those from a scientific background turned to science.

2.7 Conclusions

In this paper, we provide new evidence on how societies react to adversities, studying an exemplary historical episode: the Great Influenza Pandemic of 1918–1919.

First, we show that society reacted to the pandemic by becoming both more religious and more scientific. Second, using individual-level data from full-count censuses, we suggest that religiosity and science are substitute ways of reacting to the shock. When facing adversity, individuals from more religious backgrounds turned to religion, while those from less religious backgrounds turned to science. Third, we show that the pandemic shock widened the distance in religiosity between scientific-oriented individuals and the rest of the population and that it increased preexisting differences in religious sentiment. As a consequence, the distribution of religiosity in counties more exposed to the pandemic became more polarized.

Our paper sheds new light on the relationship between religiosity and science. Throughout history, science and religion have often been in conflict, and recent evidence by

Bénabou, Ticchi and Vindigni; Bénabou, Ticchi and Vindigni (2015; 2022) shows that the two are negatively correlated, both across countries and across U.S. states. We provide novel evidence that—at the individual level—the two are substitute ways of dealing with adversity.

Our analysis also helps shed light on modern events such as the reaction of society to the COVID-19 pandemic. Even though the modern context differs in many ways from the period of influenza pandemic, including the medical advancements of the past century, the reaction of today's society seems in line with what we document for the 1918–1919 pandemic.[39] In particular, our findings can help explain the opposing views that have emerged since the COVID-19 pandemic on science-based responses to the shock, such as the opposing attitudes toward vaccines.

Finally, our results suggest that, in the aftermath of a negative shock, societies may become more polarized in their religiosity. Because religion has become an increasingly polarizing element in the current political debate, facing adversity may strongly affect not only religious polarization but also the polarization of political views and, more broadly, the polarization of society itself.

[39]One key difference between the two pandemics is that no medical remedy or vaccine became available until many years after the earlier pandemic ended.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2016. "Cultural Assimilation During the Age of Mass Migration." *NBER Working Paper*.

Agarwal, Ruchir, and Patrick Gaule. 2022. "What Drives Innovation? Lessons from COVID-19 R&D." *Journal of Health Economics*, 82: 102591.

Ager, Philipp, and Antonio Ciccone. 2018. "Agricultural Risk and the Spread of Religious Communities." *Journal of the European Economic Association*, 16(4): 1021–1068.

Ager, Philipp, Casper Worm Hansen, and Lars Lønstrup. 2016. "Church Membership and Social Insurance: Evidence from the Great Mississippi Flood of 1927." *Southern Denmark University Discussion Papers on Business and Economics*.

Almond, Douglas. 2006. "Is the 1918 Influenza Pandemic Over? Long-term Effects of in Utero Influenza Exposure in the Post-1940 US Population." *Journal of Political Economy*, 114(4): 672–712.

Andersen, Lars H., and J. S. Bentzen. 2022. "In the Name of God! Religiosity and the Transition to Modern Economic Growth." *CEPR Discussion Paper*.

Andrews, Michael. 2021. "Historical Patent Data: A Practitioner's Guide." *Journal of Economics and Management Strategy*, 30(2): 368–397.

Ano, Gene G., and Erin B. Vasconcelles. 2005. "Religious Coping and Psychological Adjustment to Stress: A Meta-analysis." *Journal of Clinical Psychology*, 61(4): 461–480.

Babina, Tania, Asaf Bernstein, and Filippo Mezzanotti. Forthcoming. "Financial Disruptions and the Organization of Innovation: Evidence from the Great Depression." *Review of Financial Studies*.

Barro, Robert J. 2022. "Non-pharmaceutical Interventions and Mortality in US Cities During the Great Influenza Pandemic, 1918–1919." *Research in Economics*, 76(2): 93–106.

Barro, Robert J., José F. Ursúa, and Joanna Weng. 2020. "The Coronavirus and the Great Influenza Pandemic: Lessons from the "Spanish Flu" for the Coronavirus's Potential Effects on Mortality and Economic Activity." *NBER Working Paper*.

Barry, John M. 2020. *The Great Influenza: The Story of the Deadliest Pandemic in History.* London (UK): Penguin.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States." *Econometrica*, 88(6): 2329–2368.

Beach, Brian, Karen Clay, and Martin H Saavedra. 2020. "The 1918 Influenza Pandemic and its Lessons for COVID-19." *NBER Working Paper*.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *The Quarterly Journal of Economics*, 124(2): 531–596.

Becker, Sascha O, Jared Rubin, and Ludger Woessmann. 2021. "Religion in Economic History: A Survey." *The Handbook of Historical Economics*, 585–639.

Belloc, Marianna, Francesco Drago, and Roberto Galbiati. 2016. "Earthquakes, Religion, and Transition to Self-Government in Italian Cities." *The Quarterly Journal of Economics*, 131(4): 1875–1926.

Bénabou, Roland, Davide Ticchi, and Andrea Vindigni. 2015. "Religion and Innovation." *American Economic Review: Papers & Proceedings*, 105(5): 346–351.

Bénabou, Roland, Davide Ticchi, and Andrea Vindigni. 2022. "Forbidden Fruits: The Political Economy of Science, Religion and Growth." *The Review of Economic Studies*, 89(4): 1785–1832.

Bentzen, Jeanet. 2019. "Acts of God? Religiosity and Natural Disasters Across Subnational World Districts." *The Economic Journal*, 126(622): 2295–2321.

Bentzen, Jeanet Sinding. 2021. "In Crisis, We Pray: Religiosity and the COVID-19 Pandemic." *Journal of Economic Behavior & Organization*, 192: 541–583.

Berkes, Enrico. 2018. "Comprehensive Universe of US Patents (CUSP): Data and Facts." *Working Paper.*

Berkes, Enrico, Olivier Deschenes, Ruben Gaetani, Jeffrey Lin, and Christopher Severen. Forthcoming. "Lockdowns and Innovation: Evidence from the 1918 Flu Pandemic." *Review of Economics and Statistics.*

Bianchi, Nicola, and Michela Giorcelli. 2020. "Scientific Education and Innovation: From Technical Diplomas to University STEM Degrees." *Journal of the European Economic Association*, 18(5): 2608–2646.

Biasi, Barbara, David Deming, Petra Moser, and Eleanor Wiske Dillon. 2022. "Education and Innovation." *The Role of Innovation and Entrepreneurship in Economic Growth.* University of Chicago Press.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique. 2017. "The Economic Payoff of Name Americanization." *Journal of Labor Economics*, 35(4): 1089–1116.

Bisin, Alberto, and Thierry Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences." *Journal of Economic Theory*, 97(2): 298–319.

Botticini, Maristella, and Zvi Eckstein. 2012. *The Chosen Few: How Education Shaped Jewish History, 70-1492.* Princeton University Press.

Boustan, Leah Platt, Matthew E Kahn, and Paul W Rhode. 2012. "Moving to Higher Ground: Migration Response to Natural Disasters in the Early Twentieth Century." *American Economic Review*, 102(3): 238–244.

Boustan, Leah Platt, Matthew E Kahn, Paul W Rhode, and Maria Lucia Yanguas. 2020. "The Effect of Natural Disasters on Economic Activity in US Counties: A Century of Data." *Journal of Urban Economics*, 118: 103257.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with a Continuous Treatment." *Working Paper*.

Clay, Karen, Joshua Lewis, and Edson Severnini. 2019. "What Explains Cross-City Variation in Mortality During the 1918 Influenza Pandemic? Evidence from 438 US Cities." *Economics & Human Biology*, 35: 42–50.

Clemens, Jeffrey, and Parker Rogers. 2020. "Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents." *NBER Working Paper*.

Crosby, Alfred W. 1989. *America's Forgotten Pandemic: The Influenza of 1918.* Cambridge: Cambridge University Press.

Deming, David. 2010. *Science and Technology in World History: Early Christianity, the Rise of Islam and the Middle Ages: 2.* Jefferson, NC: McFarland.

Eckert, Fabian, Adres Gvirtz, J Liang, and M Peters. 2018. "A Consistent County-Level Crosswalk for US Spatial Data since 1790." *NBER Working Paper*.

Ferrara, Andreas, and Price Fishback. 2020. "Discrimination, Migration, and Economic Outcomes: Evidence from World War I." *The Review of Economics and Statistics*, 1–44.

Fouka, Vasiliki. 2019. "How do Immigrants Respond to Discrimination? The Case of Germans in the US during World War I." *American Political Science Review*, 113(2): 405–422.

Fouka, Vasiliki. 2020. "Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I." *The Review of Economic Studies*, 87(1): 204–239.

Frost, Peter. 2020. "An Accelerant of Social Change? The Spanish Flu of 1918-19." *International Political Anthropology Journal*, 13(2): 123–133.

Giuliano, Paola, and Antonio Spilimbergo. 2023. "Recessions, Lifetime Experiences and the Formation of Political Beliefs." *Working paper*.

Gross, Daniel P, and Bhaven N Sampat. 2021. "The Economics of Crisis Innovation Policy: A Historical Perspective." *American Economic Association: Papers and Proceedings*, 111: 346–50.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools."

Huntington, Ellsworth. 1923. "Causes of Geographical Variation in the Influenza Epidemic in the Cities of the United States." *Bull. Nat. Res. Council*, 6: 1–36.

Iannaccone, Laurence R. 1998. "Introduction to the Economics of Religion." *Journal of Economic Literature*, 36(3): 1465–1496.

Iyer, Sriya. 2016. "The New Economics of Religion." *Journal of Economic Literature*, 54(2): 395–441.

Jurajda, Štěpán, and Dejan Kovač. 2021. "Names and Behavior in a War." *Journal of Population Economics*, 34(1): 1–33.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring Technological Innovation Over the Long Run." *American Economic Review: Insights*, 3(3): 303–20.

Lecce, Giampaolo, Laura Ogliari, and Mara P Squicciarini. 2021. "Birth and Migration of Scientists: Does Religiosity Matter? Evidence from 19th-Century France." *Journal of Economic Behavior & Organization*, 187: 274–289.

Long, Jason, and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850." *American Economic Review*, 103(4): 1109–1137.

Markel, Howard, Harvey B Lipman, J Alexander Navarro, Alexandra Sloan, Joseph R Michalsen, Alexandra Minna Stern, and Martin S Cetron. 2007. "Nonpharmaceutical Interventions Implemented by US Cities During the 1918-1919 Influenza Pandemic." *Journal of the American Medical Association*, 298(6): 644–654.

Miao, Qing, and David Popp. 2014. "Necessity as the Mother of Invention: Innovative Responses to Natural Disasters." *Journal of Environmental Economics and Management*, 68(2): 280–295.

Mokyr, Joel. 2011. "The Economics of Being Jewish." *Critical Review*, 23(1-2): 195–206.

Montgomery, Lucy Maud. 1924. In *The Selected Journals of L.M. Montgomery, Volume III: 1921-1929*. , ed. Mary Rubio and Elizabeth Waterston. Toronto: Oxford University Press.

Moscona, Jacob. 2022. "Environmental Catastrophe and the Direction of Invention: Evidence from the American Dust Bowl." *Working Paper*.

Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *American Economic Review*, 95(4): 1214–1236.

Norenzayan, Ara, and Ian G. Hansen. 2006. "Belief in Supernatural Agents in the Face of Death." *Personality and Social Psychology Bulletin*, 32(2): 174–187.

Olivetti, Claudia, M Daniele Paserman, Laura Salisbury, and E Anna Weber. 2020. "Who Married, (To) Whom, and Where? Trends in Marriage in the United States, 1850-1940." *NBER Working Paper 28033*.

Pargament, Kenneth I. 2001. *The Psychology of Religion and Coping: Theory, Research, Practice.* New York:Guilford Press.

Phillips, Howard. 2020. "'17, '18, '19: Religion and Science in three Pandemics, 1817, 1918, and 2019." *Journal of Global History*, 15(3): 434–443.

Posch, Max. 2022. "Do Disasters Affect the Tightness of Social Norms?" *Working Paper.*

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS USA: Version 11.0 [dataset]."

Sibley, Chris G., and Joseph Bulbulia. 2012. "Faith After an Earthquake: A Longitudinal Study of Religion and Perceived Health Before and After the 2011 Christchurch New Zealand Earthquake." *PloS One*, 7(12): e49648.

Spinney, Laura. 2018. *Pale Rider: the Spanish Flu of 1918 and How It Changed the World.* London: Vintage.

Squicciarini, Mara P. 2020. "Devotion and Development: Religiosity, Education, and Economic Progress in 19th-Century France." *American Economic Review.*

Stark, Rodney. 1992. "The Reliability of Historical United States Census Data on Religion." *Sociological Analysis*, 53(1): 91–95.

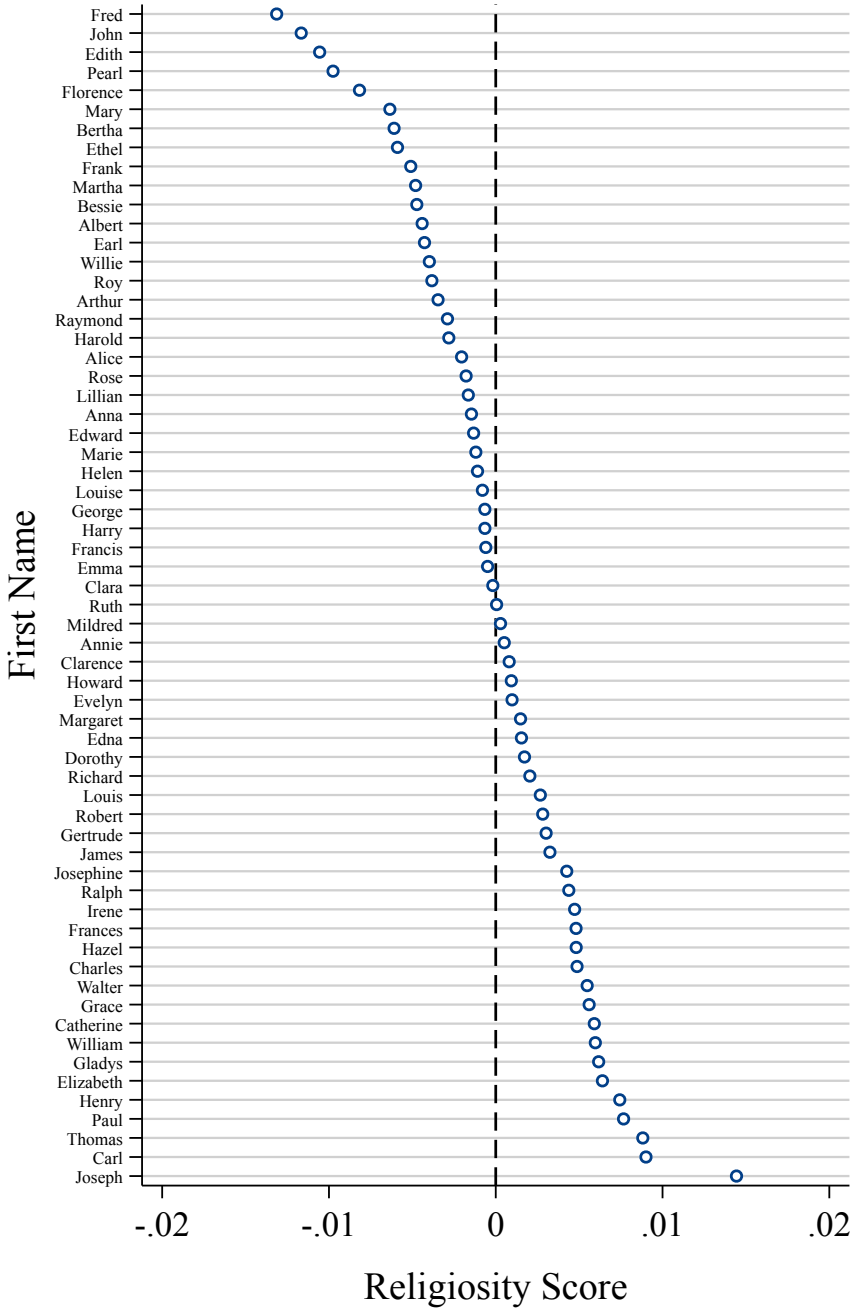
Stark, Rodney. 1998. "The Rise and Fall of Christian Science." *Journal of Contemporary Religion*, 13(2): 189–214.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199.

Weber, Max. 1905. *The Protestant Ethic and the Spirit of Capitalism.* Routledge.

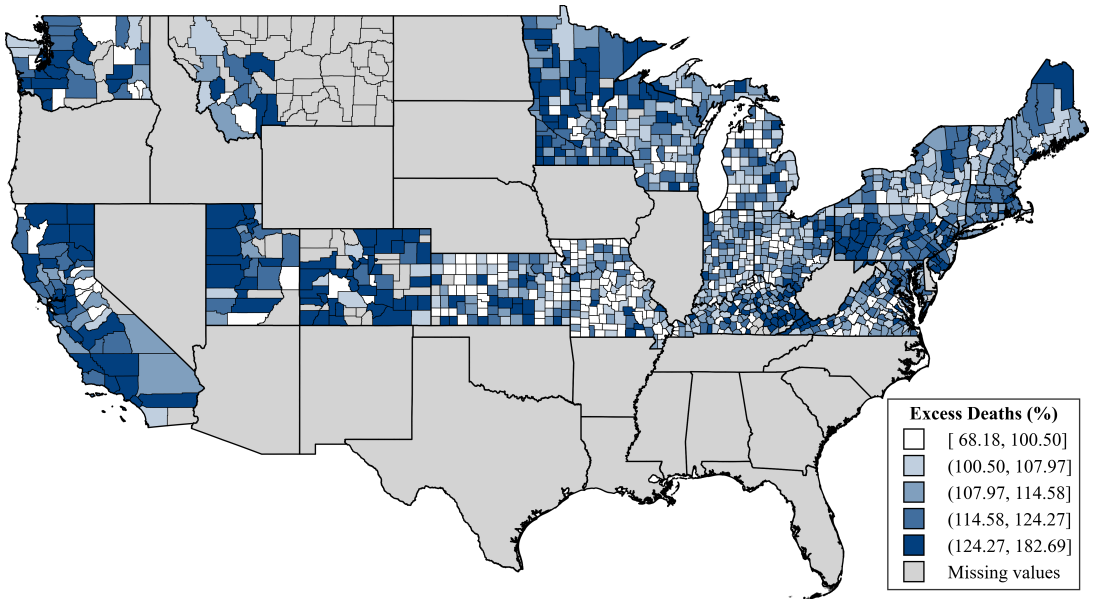
2.8 Figures

Figure 2.1: Estimated Religiosity Scores by Name



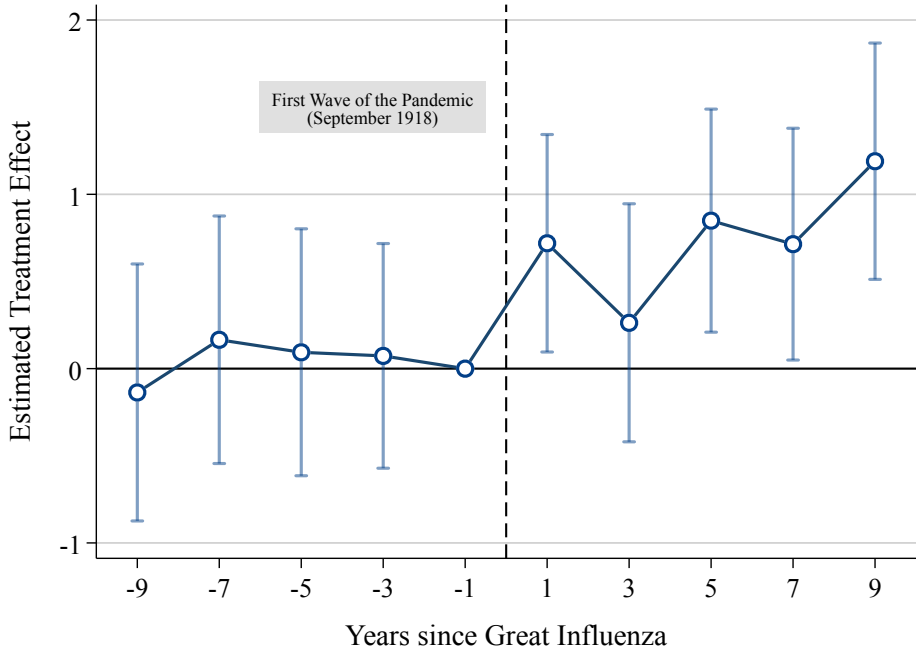
Notes: This figure displays the religiosity scores estimated from model (2.1). Regressions are based on data from the 1906–1916 Censuses of Religious Bodies; they include individuals born between 1896 and 1916. We estimate religiosity scores for names appearing in at least 0.3% of the overall sample. Coefficients are reported in increasing order.

Figure 2.2: Spatial Distribution of Excess Mortality During the Great Influenza Pandemic



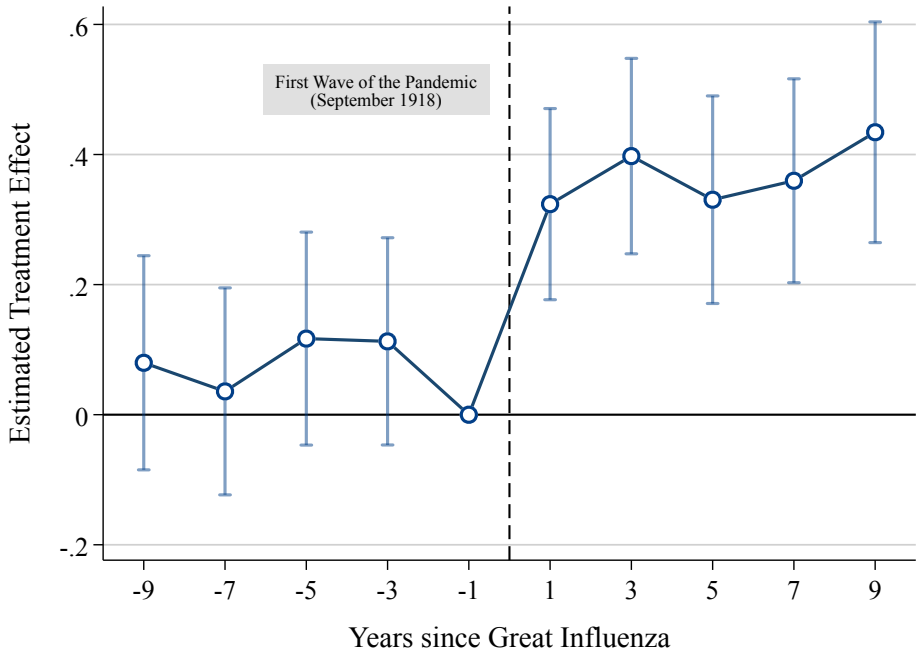
Notes: This figure displays geographic variation in excess deaths, defined in Equation (2.3). Excess mortality is the ratio between the average number of deaths during the pandemic years (1918–1919) and the average number of deaths in the three years before the pandemic (1915–1917). Mortality statistics before 1915 are not available. Excess mortality is displayed in percentage terms. Lighter to darker blue indicates increasing excess deaths. Counties are displayed at their 1920 borders. Morality data are not available for states displayed in gray. Counties displayed in gray are excluded from the analysis sample.

Figure 2.3: Impact of the Influenza on Religiosity



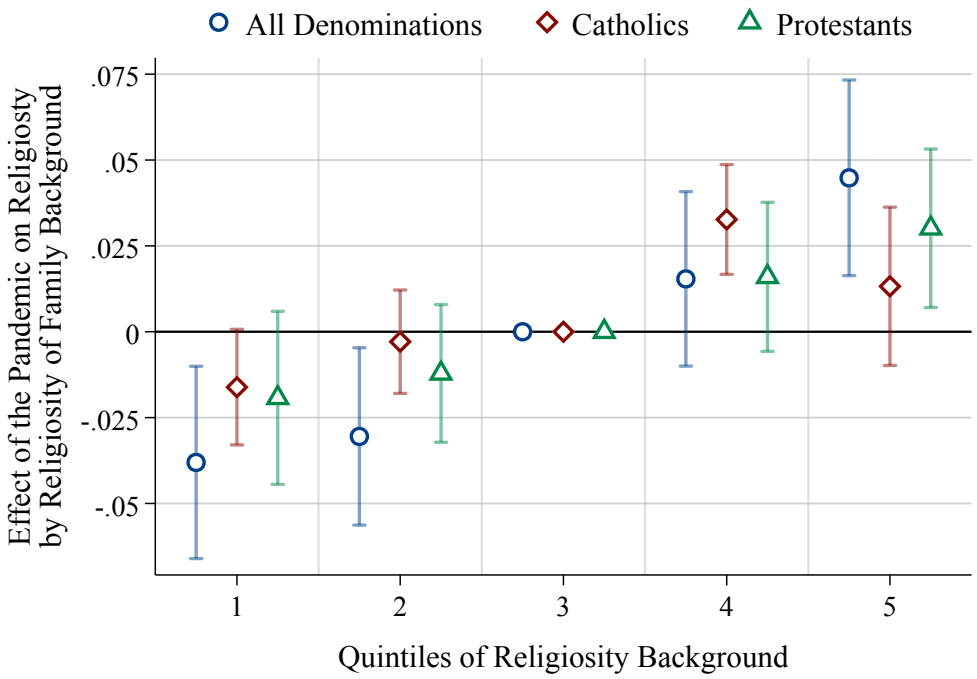
Notes: This figure displays the dynamic treatment effects of the pandemic on overall religiosity. The unit of observation is a county observed at a biennial frequency. Each dot reports the coefficient of an interaction between the baseline measure of excess deaths, defined in Equation (2.3), and a biennial time dummy. The coefficient for the biennial 1917–1918 serves as the baseline category. The model includes county and state-by-year-fixed effects and further controls for an interaction term between the population in 1910 and a post-treatment indicator. Bands report 95% confidence intervals. Standard errors are clustered at the county level. The dashed vertical line indicates the timing of the first wave of the pandemic (October 1918).

Figure 2.4: Impact of the Influenza on Innovation



Notes. The Figure reports the dynamic treatment effects of the pandemic on innovation. The dependent variable is the (log 1+) total number of patents filed in a given year. The unit of observation is a county observed at a biennial frequency. Each dot reports the coefficient of an interaction between the baseline measure of excess deaths and a biennial time dummy. The coefficient for the biennial 1917–1918 serves as the baseline. The model includes county and state-by-year-fixed effects and controls for an interaction term between the population in 1910 and a post-treatment indicator. Bands report 95% confidence intervals. Standard errors are clustered at the county level. The dashed vertical line indicates the timing of the first wave of the pandemic (October 1918).

Figure 2.5: Impact of the Influenza on the Polarization of Religious Beliefs



Notes: This figure reports the estimated impact of the pandemic on the polarization of religious beliefs by religious denominations. Each dot reports the coefficient of an interaction between the baseline measure of excess deaths, a posttreatment indicator, and an indicator for the quintile of background religiosity. The unit of observation is a child born between 1900 and 1930. Treated children are those born after the influenza, i.e., after 1918. The dependent variable is the religiosity score associated with the child's name. Background religiosity is measured as the religiosity score of the child's head of household. Results are reported by confession, and the third quintile serves as the baseline. Regression models include fixed effects for county by cohort, county by quintile of religious background, and cohort by quintile of religious background. Standard errors are clustered at the county level, and the bands report the 95% confidence interval for each coefficient.

2.9 Tables

Table 2.1: The Impact of the Influenza on Religiosity

	Share of Affiliated			Name-Based Religiosity		
	(1) All	(2) Catholics	(3) Protestants	(4) All	(5) Catholics	(6) Protestants
Post × Excess Deaths	0.220***	0.078***	0.094***	0.863***	0.269**	0.688***
	(0.025)	(0.013)	(0.017)	(0.199)	(0.111)	(0.171)
County FE	Yes	Yes	Yes	Yes	Yes	Yes
State-Decade FE	Yes	Yes	Yes	–	–	–
State-Year FE	–	–	–	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1274	1274	1274
Observations	3825	3825	3825	38220	38220	38220
R^2	0.851	0.904	0.929	0.648	0.522	0.678
Std. Beta Coef.	0.678	0.327	0.335	0.157	0.112	0.159

Notes: This table displays the impact of exposure to the Great Influenza Pandemic on religiosity. The unit of observation is a county observed at a decade frequency between 1906 and 1926 (in columns 1–3) and yearly frequency between 1910 and 1929 (in columns 4–6). "Post" is a categorical variable equal to one during and after the pandemic—i.e., over 1918 to 1929—or zero otherwise. The baseline treatment "Excess Deaths" is defined in Equation (2.3). In columns (1–3), the dependent variable is the number of individuals affiliated with religious denominations enumerated in the Census of Religious Bodies, normalized by county population in 1900; in columns (4–6), the dependent variable is the name-based religiosity measure described in the main text. Columns (1) and (4) report the effect of the influenza on overall religiosity, whereas columns (2) and (5)—resp. (3) and (6)—display it on the intensity of Catholicism—resp. Protestantism. Regressions include county and state-by-time (decades in columns 1–3 and years in columns 4–6) fixed effects and control for an interaction term between population in 1910 and a post-treatment indicator. Standard errors, clustered at the county level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.2: The Impact of the Influenza on Innovation

	STEM Employment Share		log(1 + Number of Patents)					
	(1) Whole Population	(2) Skilled Population	(3) All Patents	(4) Pharmaceuticals	(5) Communication	(6) Electrical	(7) Mechanical	(8) Other
Post × Excess Deaths	0.008***	0.103***	0.370***	0.115***	0.015	0.037	0.026	-0.002
	(0.001)	(0.017)	(0.056)	(0.030)	(0.019)	(0.025)	(0.019)	(0.020)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Decade FE	Yes	Yes	–	–	–	–	–	–
State-Year FE	–	–	Yes	Yes	Yes	Yes	Yes	Yes
All Patents	–	–	No	Yes	Yes	Yes	Yes	Yes
Number of Counties	1274	1274	1275	1275	1275	1275	1275	1275
Observations	3822	3822	38250	38250	38250	38250	38250	38250
R^2	0.765	0.625	0.858	0.830	0.715	0.815	0.923	0.934
Std. Beta Coef.	0.985	1.105	0.171	0.085	0.020	0.030	0.014	-0.001

Notes: This table displays the impact of exposure to the Great Influenza Pandemic on innovation. The unit of observation is a county, observed at a decade frequency between 1900 and 1930 in columns (1–2) and at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic—i.e., over 1918 to 1929—or zero otherwise. The baseline treatment "Excess Deaths" is defined in Equation (2.3). In column (1), the dependent variable is the share of people employed in STEM occupations within the population; in column (2), we restrict the denominator to include only those employed in skilled occupations. The dependent variable in columns (3–8) is the (log 1+) number of patent grants. We use this specification of the dependent variable to ensure that we do not drop counties without patents. In columns (4–8), we also control for the overall (log 1+) number of granted patents. Column (3) estimates the impact of the pandemic on the level of innovation, while columns (4)–(8) display this on the direction of innovation. All regressions include county-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Regressions (1–2) include state-by-decade-fixed effects, while regressions (3–8) include state-by-year-fixed effects. Standard errors, clustered at the county level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.3: Impact of the Influenza on Occupational Choice

	Entire Population		Skilled Population	
	(1) No Controls	(2) Controls	(3) No Controls	(4) Controls
Young × Excess Deaths	0.005**	0.005**	0.013*	0.012*
	(0.003)	(0.003)	(0.007)	(0.007)
County FE	Yes	Yes	Yes	Yes
State-Cohort FE	Yes	Yes	Yes	Yes
Household Controls	No	Yes	No	Yes
Number of Counties	1275	1275	1275	1275
Observations	13983936	13983936	5676407	5676407
R^2	0.003	0.004	0.006	0.007
Std. Beta Coef.	0.017	0.016	0.025	0.025

Notes: This table displays the effect of the pandemic on occupational choice. The unit of observation is an individual, observed once in the 1930 population census. Young is a dummy equal to one for all individuals aged between 18 and 25 at the time of the inception of the pandemic. The baseline treatment "Excess Deaths" is defined in Equation (3). The dependent variable is a dummy variable equal to one if the person is employed in a STEM occupation and zero otherwise. The sample includes the entire population in columns (1–2) and only individuals employed in skilled occupations in columns (3–4). In columns (2) and (4), we control for race and urban status of the head of household. Regressions include county and state-by-cohort fixed effects. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.4: Religious Background, Religiosity, and STEM Occupations

	Religiosity			STEM Occupation		
	(1) All	(2) Catholics	(3) Protestants	(4) All	(5) Catholics	(6) Protestants
Excess Deaths × Post × High Religious Background	0.048***	0.033***	0.033***			
	(0.010)	(0.006)	(0.008)			
Excess Deaths × Young × High Religious Background				-0.709**	-0.311	-0.551*
				(0.336)	(0.321)	(0.333)
County-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
County-Background FE	Yes	Yes	Yes	Yes	Yes	Yes
Background-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Household Controls	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1275	1275	1275
Observations	5857923	5857923	5857923	5347751	5347751	5347751
R^2	0.035	0.044	0.028	0.014	0.014	0.014
Std. Beta Coef.	0.047	0.053	0.040	-0.016	-0.007	-0.012

Notes: This table displays the impact of exposure to the pandemic on religiosity—columns (1)–(3)—and occupational choice—columns (4)–(6)—by individual-level background religiosity. The unit of observation in columns (1)–(3) is a head of household, who is observed once for each child born between 1900 and 1930 in the household. In columns (4)–(6), the unit of observation is an adult. Religiosity is defined as the religiosity score associated with the child's name. "Post" is a categorical variable equal to zero for children born during and after the pandemic—i.e., over 1918–1929—or zero for those born before the pandemic—i.e., before 1918. The baseline treatment "Excess Deaths" is defined in Equation (2.3). "STEM" is an indicator variable returning value one if an individual is employed in a STEM occupation—as defined in Table 2.7—or zero otherwise. "Young" is an indicator variable equal to one if an individual is aged between 18 and 25 in 1918 or zero otherwise. "High Background Religiosity" is an indicator variable returning the value one if the religiosity score of the name of the head of the household is in the top 50% of the overall distribution, or zero otherwise. The table displays the coefficient of the interaction between these terms. Each regression includes county-by-cohort, county-by-background, and background-by-cohort fixed effects. We include race and urban status as further household-level controls in each regression. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.5: Effect of the Influenza on Individual Religiosity: STEM and Non-STEM

	Entire Population			Skilled Population		
	(1) All	(2) Catholics	(3) Protestants	(4) All	(5) Catholics	(6) Protestants
Excess Deaths × Post × STEM	-0.061***	0.009	-0.047**	-0.059***	0.011	-0.046**
	(0.021)	(0.013)	(0.019)	(0.022)	(0.014)	(0.018)
STEM-County FE	Yes	Yes	Yes	Yes	Yes	Yes
STEM-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
County-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Household Controls	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1274	1274	1274
Observations	10616377	10616377	10616377	3859347	3859347	3859347
R^2	0.006	0.014	0.007	0.012	0.018	0.013
Std. Beta Coef.	-0.061	0.009	-0.047	-0.059	0.011	-0.046

Notes: This table displays the impact of exposure to the pandemic on STEM and non-STEM individuals' religiosity. The unit of observation is a child born between 1900 and 1930. Religiosity is defined as the religiosity score associated with the child's name. "Post" is a categorical variable equal to zero for children born before the pandemic—i.e., before 1918—or one for those born after the pandemic—i.e., after 1918. The baseline treatment "Excess Deaths" is defined in Equation (2.3). "STEM" is an indicator variable returning a value of one if one parent of the child is employed in a STEM profession or zero otherwise. The table displays the coefficient of the interaction between these terms. This estimates the causal effect of the influenza shock on the religiosity of heads of households employed in STEM occupations compared to non-STEM occupations, leveraging variation in county-level exposure to the influenza. All models include STEM-by-county, STEM-by-cohort, and county-by-cohort fixed effects. The sample includes the entire population in columns (1–3) and only individuals employed in skilled occupations in columns (4–6). Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

2.10 Appendix: Data

This section lists the data sources and describes how we construct the variables used in the analysis.

2.10.1 Names

Data on names are from the individual full-count US 1930 population census (Ruggles, Flood, Foster, Goeken, Pacas, Schouweiler and Sobek, 2021). Since first name records in the census conflate first and middle names, we consider only the first word that appears in each string. The baseline analysis includes all names that appear in at least 0.3% of the population of American-born children between 1900 and 1930. We perform a robustness exercise and show that the baseline results are not sensible to alternative frequency thresholds. We exclude first-generation immigrants because their names would not reflect exposure to the pandemic as they were not born in the United States. Moreover, the frequency threshold we apply implies that only English names are included in the sample.

2.10.2 Religious Affiliations

Data on religious affiliations are supplied by NHGIS and are originally from the Census of Religious Bodies, which took place at decade frequency between 1906 and 1936. We discard the 1936 census because previous research shows that the uptake was low and unequal across counties (Stark, 1992). Census enumerators asked churches, congregations, and other local organizations to report the number of affiliates. The data was then aggregated at the county level. In our analysis, "Total Religiosity" is computed as the simple sum of religious members across all possible denominations; "Catholics" are enumerated as such. We collectively refer to "Protestants" as a set of denominations that we manually mapped to some branch of Protestantism (including, e.g., the Methodist, Evangelical, and–various–Baptist churches).

We use data on the names of saints and biblical characters from Abramitzky, Boustan and Eriksson (2016) to develop an additional indicator of county-level religiosity. To construct this additional measure, we employ the baseline equation (2.2) and substitute the estimated $\hat{\gamma}^k$ coefficients with three indicator variables equal to one if the given name is that of a saint, a biblical character, or both.

2.10.3 Patents

Patent Data Patent data are from Berkes (2018), who performed optical character recognition (OCR) on original patent documents issued by the United States Patents and Trademark Office between 1836 and 2010. Information includes the filing and issue year, author name, latitude and longitude of the inventor(s), and inferred USPC technology class. The data contain a set of additional variables, including the complete text of the patent document and the issue year of the patent, not used in our analysis. We geo-code each patent to its 1920 county using boundary shapefiles supplied by NHGIS. When we collapse by county year, we weigh each patent by the inverse of the number of technology classes and by the inverse of the number of authors. Hence, a patent with two authors and two technological classes appears four times in the original patent-level dataset, and each instance is assigned a .25 weight when aggregating at the county level. We code USPC classes to the NBER classification (Hall, Jaffe and Trajtenberg, 2001). We modify the canonical NBER classification and conflate the "Chemical" and "Drugs" categories into a single "Pharmaceuticals" class. Since multiple USPC codes are typically assigned to a single patent, most patents that fall under "Drugs" would also appear as "Chemical." To avoid this, we recast them into one category. It is worth noting that all the results we present regarding pharmaceutical patents also hold if we keep the "Chemical" and "Drugs" classes separate.

Importance of Patents We measure patent 'importance' using the measure developed by Kelly, Papanikolaou, Seru and Taddy (2021). From their data, we derive two metrics. One is the number of "Breakthrough" innovations, which are defined as any patent whose

importance is in the top 20% of the overall quality distribution. The second variable is the share of breakthrough patents relative to the overall number of issued patents. Both measures are net of grant-year fixed effects. We take forward and backward similarity within a 5-year window around the issue year of the patent.

2.10.4 Occupational Structure

Individual-level data on occupations is extracted from the 1930 individual-level population census. More precisely, we use the 1950 harmonized occupation classification. We then manually map occupational codes to STEM occupations as described in Table 2.7.

2.10.5 Controls & Mortality Statistics

We extract individual-level information on race and urban-rural status from the IPUMS full-count data.

County-level covariates are from NHGIS. This aggregates individual-level data from population censuses and reports data from manufacturing and agricultural censuses. All data come at historical county borders.

Mortality statistics are likewise provided by NHGIS. For the period we are interested in, 1915-1919, they were collected for about 1,200 counties, covering approximately 60% of the US population. We measure Influenza-related mortality as the ratio between deaths during the pandemic and deaths in the three years that preceded the Influenza.[40]

[40]The original documents report, for major cities, deaths broken down by (alleged) cause. We do not use this data for two main reasons. First, they are incomplete and are only available for cities. Second, Beach, Clay and Saavedra (2020) criticizes the methodology adopted to impute the cause of deaths.

2.10.6 Other Data

In several robustness regressions, we control for WW1 mortality. The underlying data were kindly shared by Ferrara and Fishback (2020).

2.10.7 Boundary Harmonization

County-level data from NHGIS and other sources are typically provided at historical borders. To ensure comparability and consistency, we adopt the method developed by Eckert, Gvirtz, Liang and Peters (2018) to compute geographical crosswalks between US counties over time. In a nutshell, their methodology is as follows. Suppose we know the distribution of a given variable y across counties at decade frequency between 1900 and 1930. To harmonize borders to one year, Eckert, Gvirtz, Liang and Peters (2018) overlay the shapefile of counties in a given year, say, 1900, to that in the reference year, say, 1920. They then compute the percentage of land that a given county shares with itself between the two years and that assigned to other counties. To construct the harmonized variable, one multiplies these overlapping area weights by the variable recorded in 1900 and aggregates up by 1920 counties. The underlying assumption is that y is evenly distributed over the county territory. While this may seem untenable in most cases, departures from this assumption are plausibly innocuous in our setting. County borders had undergone major consolidations before 1900 and remained stable thereafter. Moreover, mortality data are mostly available for the Northwest and Midwest areas. Boundary changes in these regions were rare and minor after the 1890s. In our application, we map all county-level variables to 1920-borders.

2.10.8 Details on Sample Construction

In this paragraph, we provide additional technical details on how we construct the estimation samples. The main sample restriction that we impose descends from the fact that we observe mortality for 1,302 out of 3,100 counties. We then discard 27 counties

with values of excess mortality below or above, respectively, the bottom 1% (85% of the pre-pandemic mortality) and top 99% (180% of the pre-pandemic level) of the excess-mortality distribution. Because such figures are due to scarcely-inhabited areas, these 27 counties account for less than 0.1% of the population in the 1302 counties sample. We are left with a set of 1275 counties. In the rest of the paragraph, we explain why we may not always be able to leverage all 1275 for the estimation.

County-Level Religiosity The county-level religiosity estimation sample is a balanced panel dataset where each county is observed at a yearly frequency between 1900 and 1929. This implies that the number of counties in this balanced panel may not be 1275 as long as at least one county is not observed at least once between 1900 and 1929. This happens because, especially in scarcely-inhabited areas, our name-frequency threshold may imply that we cannot match any newborn in a given cohort. If that is the case, the county's religiosity will not be observed every year of the sample, and the county will subsequently be dropped from the estimation sample. This is the case for one county, so the estimation sample, in this case, consists of 1274 counties accounting for 99.5% of the population in the 1302-counties sample.

In one robustness check shown in column (7) of Table 2.10, counties are observed at decade frequency instead. In this case, the sample is constructed from adults observed once per census decade between 1900 and 1930, and the post-treatment indicator returns a value of one for decades 1920 and 1930 and zero otherwise.

County-Level Innovation The county-level innovation sample is a balanced panel dataset where each county is observed at a yearly frequency between 1900 and 1929. Thus, an observation in the dataset can either be a number above zero (if one or more patents are observed in that county year) or zero (if no patents are observed). The estimation sample in this case thus encompasses all 1220 counties for which we observe mortality. In columns (2) and (7) of Table 2.18 we don't include county-year when no patents are observed. This results in an unbalanced panel dataset where a county may not be observed yearly over the estimation period.

Other County-Level Samples In Table 2.15, we use various measures of name concentration as dependent variables. These are the Hirschman-Herfindahl index, the Comprehensive Concentration index, the Rosenbluth index, and concentration ratios equal to the share of children born with the most common k names, for various levels of the threshold k.

Individual-Level We construct two individual-level datasets. The first sample ("adult sample") comprises all those born before 1900. Each individual is observed once in the 1930 census. Because the adult sample is used to study the evolution of the occupational structure, we discard (i) individuals with no valid occupational response, (ii) farmers, who have been shown to display disproportionate occupational persistence (Long and Ferrie, 2013), and (iii) women, since the female labor force participation was extremely low. The second sample ("children sample") comprises all those born between 1900 and 1930. Each child is considered a realization of their household. Each household is thus observed once per child. In the children sample, we exclude households when (i) we do not observe a valid occupational response, and (ii) there is no male head, since here, too, the labor force participation of women was very low. We identify a household as "STEM" if at least one of its members is employed in a STEM occupation. Consistently with the county-level analysis, we do not assign a religiosity score to first-generation immigrants.

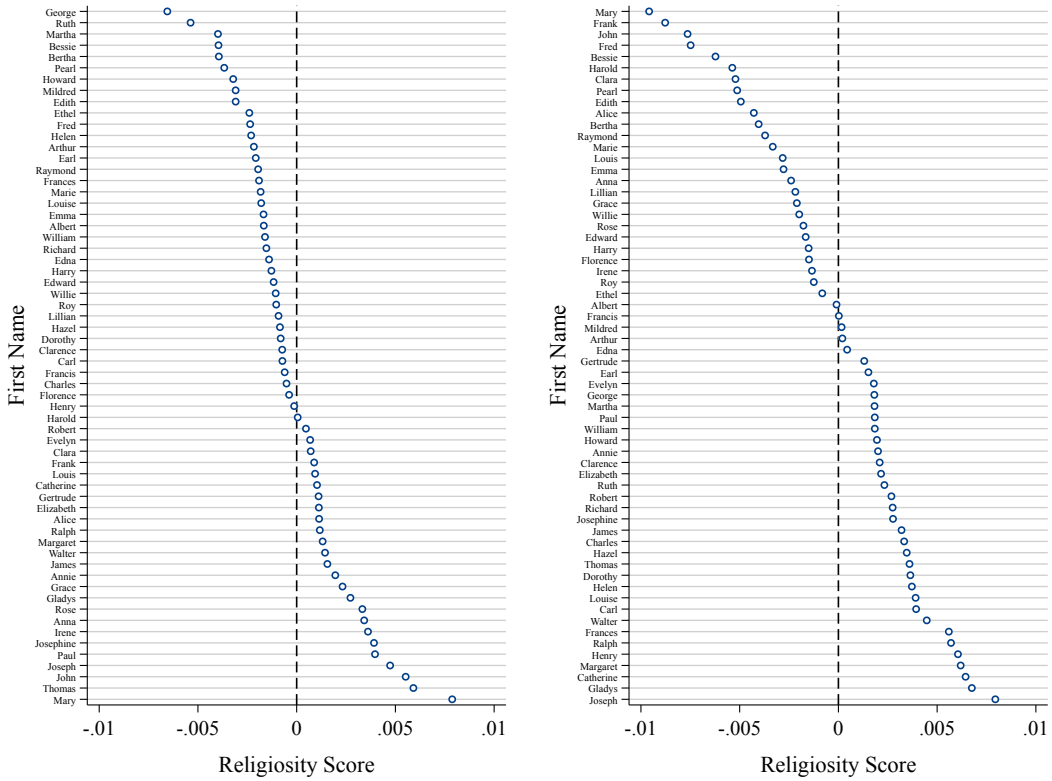
City-Level To build the city-level sample, we construct the baseline excess deaths treatment variable from data by Clay, Lewis and Severnini (2019). The dataset contains mortality information on 976 cities. For 444, however, we observe the number of deaths in one year only, and we do not observe 48 other cities continuously between 1915 and 1919. Moreover, we do not observe population data in 1900 for 41 additional cities. The final sample consists of 443 cities. In Figure 2.12 we report the location of each city and the number of cities included in the sample, by state. We geo-code patents to historical city borders and construct the name-based religiosity measure from the individuals recorded living in each city in the 1930 census. The city-level sample is used in the regressions displayed in Tables 2.12 and 2.23.

2.11 Appendix: Figures and Tables

2.11.1 Figures

Figure 2.6: Estimated Names Religiosity Scores, by Confession

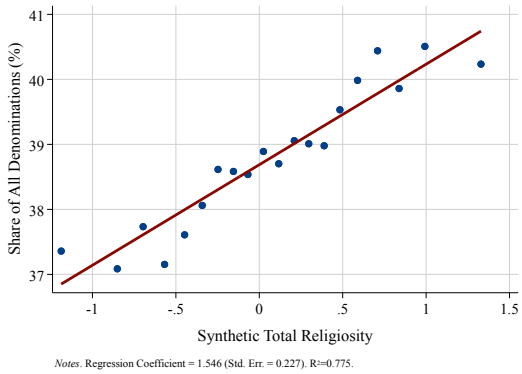
((a)) Catholics ((b)) Protestants



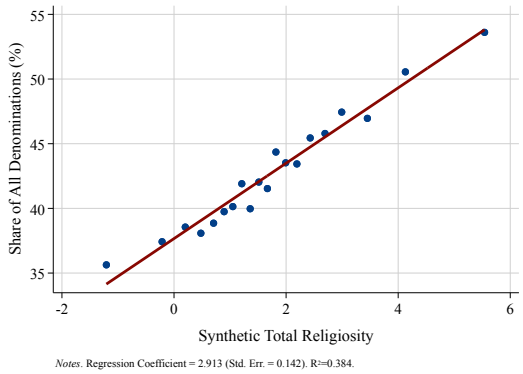
Notes: The Figures display the religiosity scores estimated from model (2.1). Bars report the point estimate of each coefficient. Regressions are based on data from the 1906-1916 Censuses of Religious Bodies and include individuals born between 1896 and 1916. We estimate religiosity scores for names appearing in at least 0.3% of the overall sample. Panel 2.6(a) reports scores for Catholicism; Panel 2.6(b) reports scores for Protestantism.

Figure 2.7: In-sample and Out-of-sample Fit of the Religiosity Measure

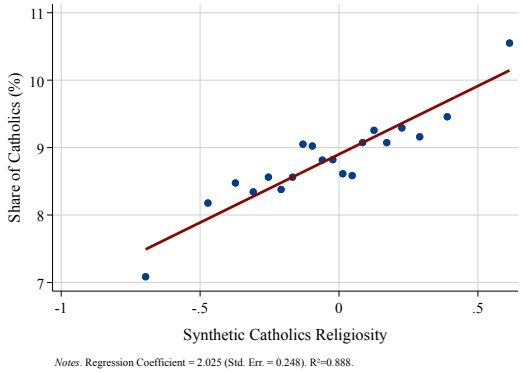
((a)) In-sample: All Denominations



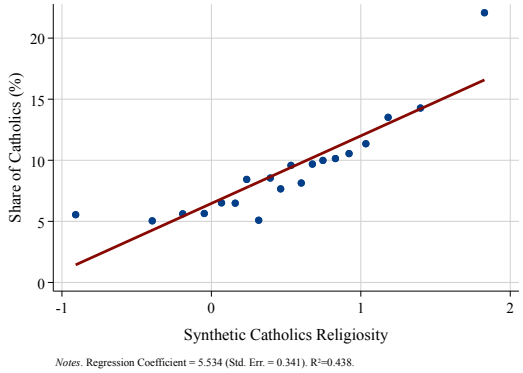
((b)) Out-of-sample: All Denominations



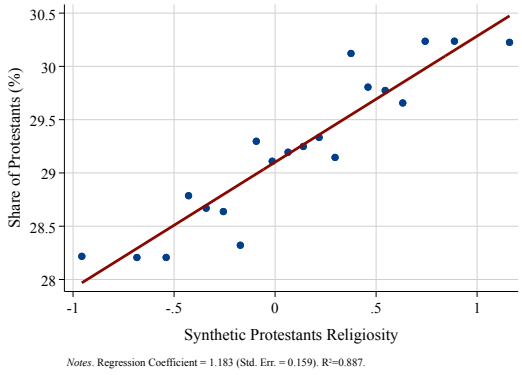
((c)) In-sample: Catholics



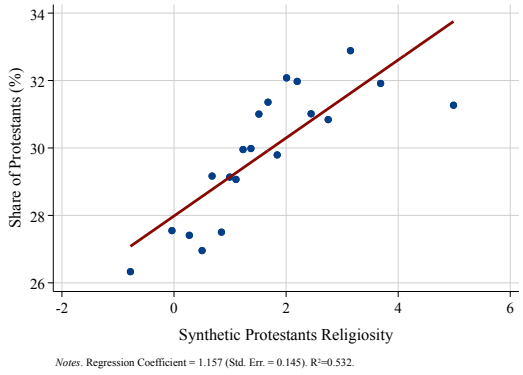
((d)) Out-of-sample: Catholics



((e)) In-sample: Protestants

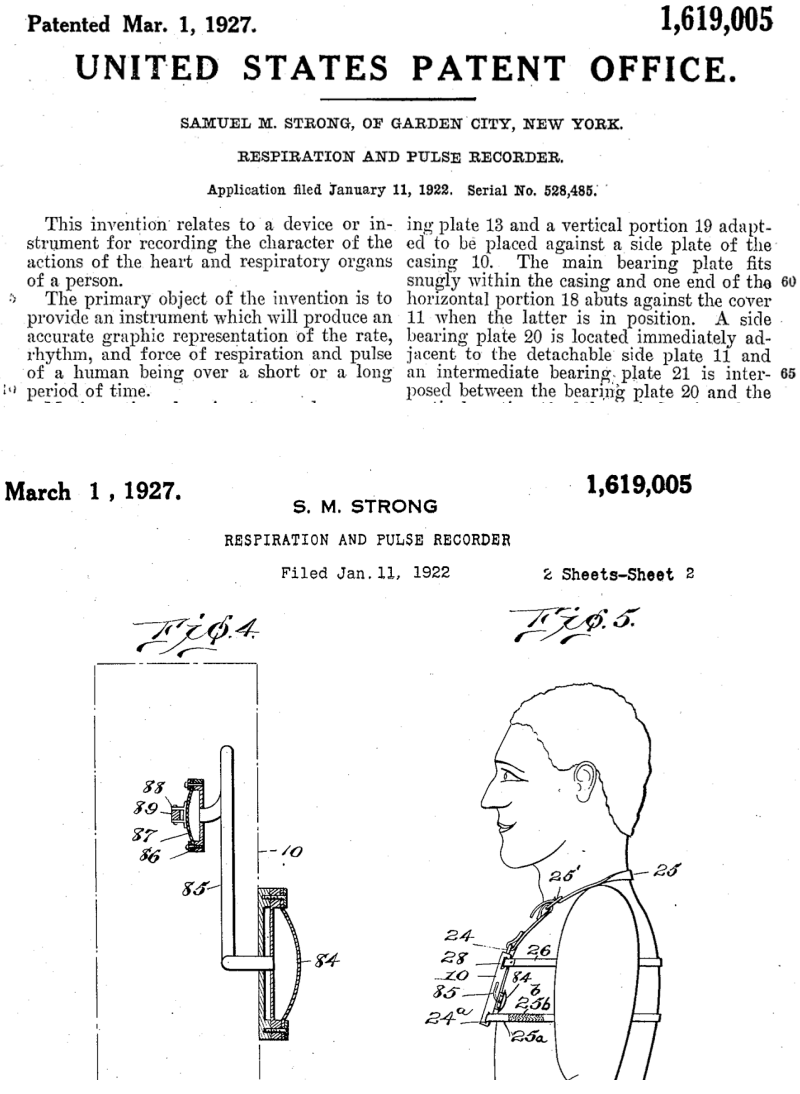


((f)) Out-of-sample: Protestants



Notes: These figures are county-level binned scatter plots reporting the correlation between our religiosity measure and the number of affiliated members to all denominations (2.7(a)-2.7(b)), Catholicism (2.7(c)-2.7(d)) and Protestantism (2.7(e)-2.7(f)) normalized by population in 1900. In-sample figures report data for the 1906 and 1916 censuses of religious affiliations. Out-of-sample figures instead report data for 1926. In-sample regressions control for county-fixed effects; out-of-sample regressions include state-fixed effects. In the note, we report the regression coefficients and the associated R^2.

Figure 2.8: Example of Pharmaceutical Patent



Patented Mar. 1, 1927. **1,619,005**

UNITED STATES PATENT OFFICE.

SAMUEL M. STRONG, OF GARDEN CITY, NEW YORK.

RESPIRATION AND PULSE RECORDER.

Application filed January 11, 1922. Serial No. 528,485.

This invention relates to a device or instrument for recording the character of the actions of the heart and respiratory organs of a person.

The primary object of the invention is to provide an instrument which will produce an accurate graphic representation of the rate, rhythm, and force of respiration and pulse of a human being over a short or a long period of time.

ing plate 13 and a vertical portion 19 adapted to be placed against a side plate of the casing 10. The main bearing plate fits snugly within the casing and one end of the horizontal portion 18 abuts against the cover 11 when the latter is in position. A side bearing plate 20 is located immediately adjacent to the detachable side plate 11 and an intermediate bearing plate 21 is interposed between the bearing plate 20 and the

March 1, 1927. **1,619,005**

S. M. STRONG

RESPIRATION AND PULSE RECORDER

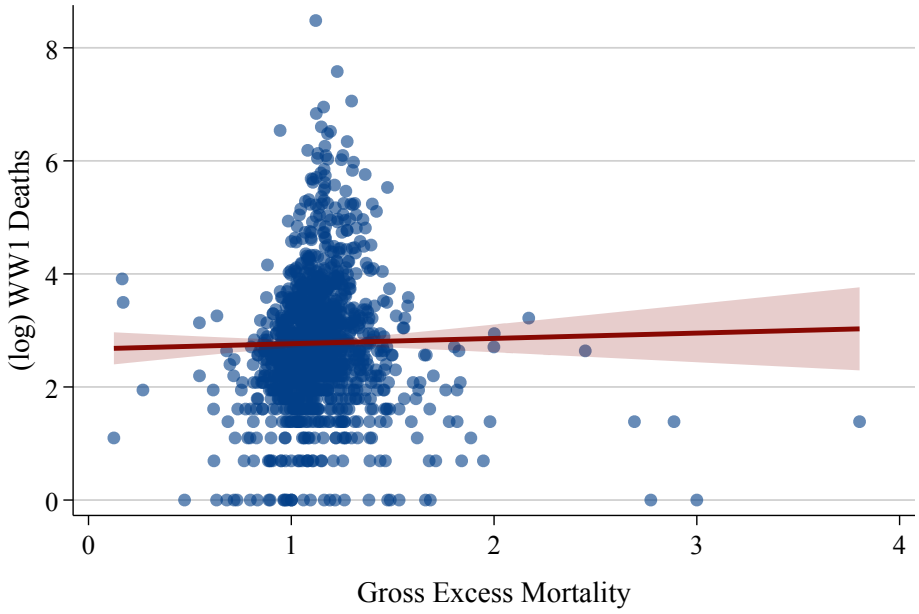
Filed Jan. 11, 1922 2 Sheets-Sheet 2

FIG. 4.

FIG. 5.

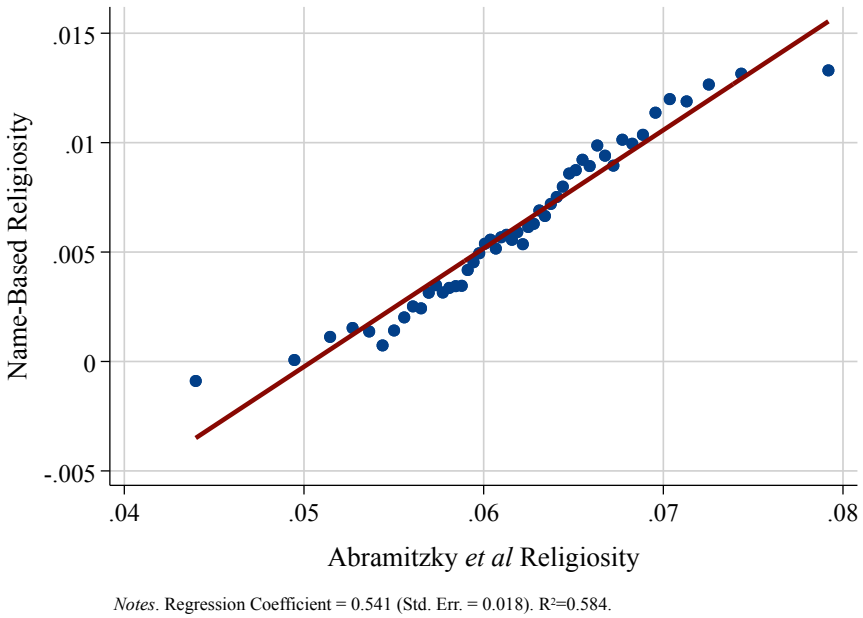
Notes: This Figure displays the text and figures of one sample patent that our classification algorithm assigns to the pharmaceutical class.

Figure 2.9: Correlation Between WW1 Deaths and Excess Deaths



Notes: This figure displays the correlation between WW1 deaths and excess deaths. Gross Excess Mortality is the baseline treatment. WW1 deaths are taken as logs. In the note, we report the regression coefficient between the two variables and the R^2 of the model. Data on WW1 deaths are from Ferrara and Fishback (2020).

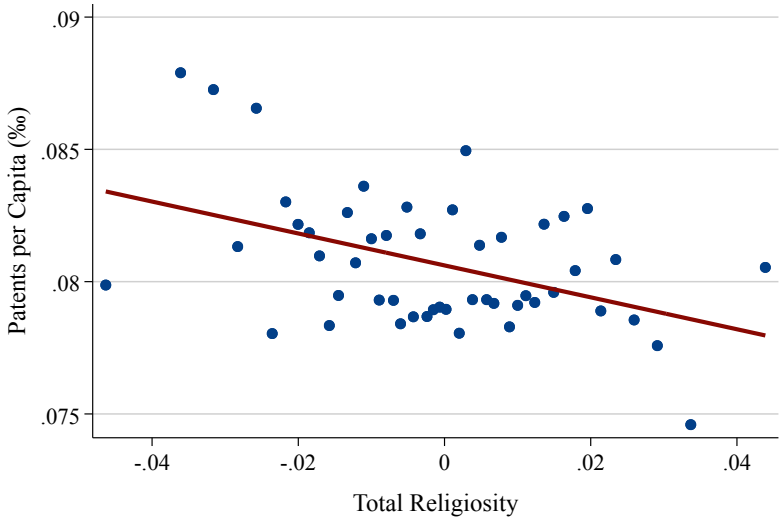
Figure 2.10: Correlation Between Abramitzky, Boustan and Eriksson (2016) Religiosity and Baseline Religiosity



Notes: This figure reports the correlation between our baseline religiosity measure (multiplied by 100) and the share of biblical and saint names, as defined in Abramitzky, Boustan and Eriksson (2016). The unit of observation is a county observed at a yearly frequency between 1900 and 1930. The graph partials out county and year fixed effects. We report in note the regression coefficient and the associated standard error, clustered at the county level, and R^2 coefficient.

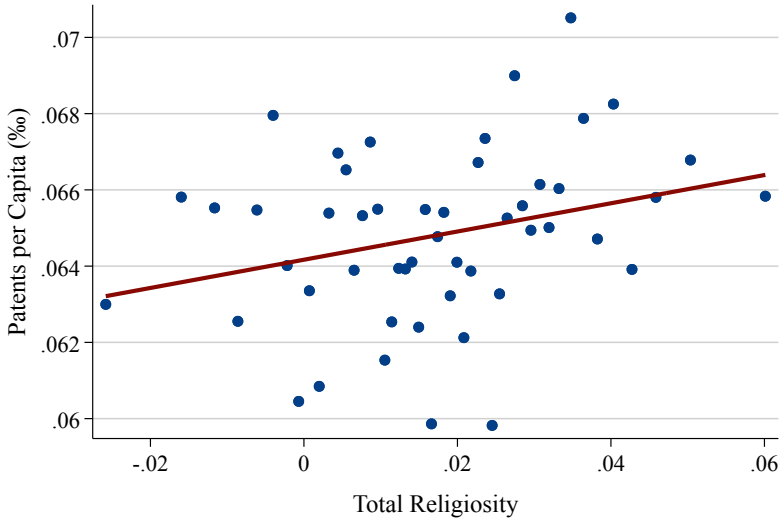
Figure 2.11: Correlation Between Religiosity and Science

((a)) Before the Great Influenza Pandemic (1910–1917)



Notes. Regression Coefficient = -0.060 (Std. Err. = 0.020). R^2=0.492.

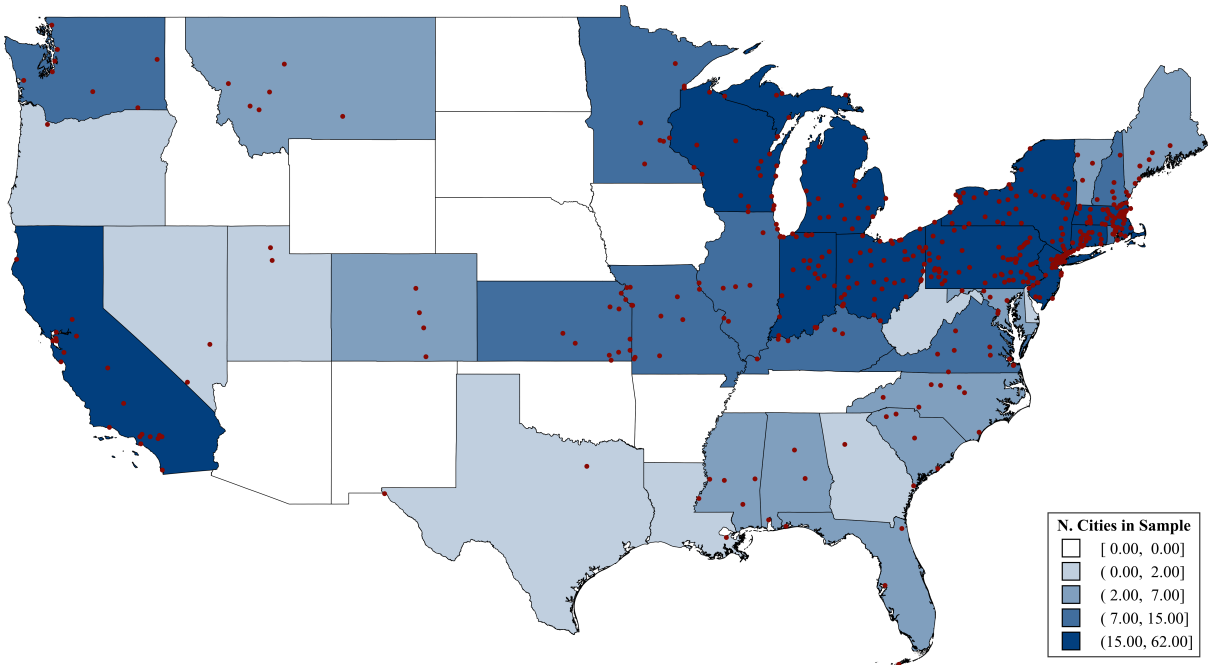
((b)) After the Great Influenza Pandemic (1918–1929)



Notes. Regression Coefficient = 0.037 (Std. Err. = 0.020). R^2=0.576.

Notes: These figures display county-level binned scatter plots reporting the correlation between science–measured as patenting activity normalized by population–and religiosity. The unit of observation is a county observed at a yearly frequency. Religiosity is defined as described in section 2.3.1 and refers to overall religiosity. Graphs absorb for county and year-fixed effects. We report the regression coefficients and associated R^2 separately in each graph.

Figure 2.12: Distribution of Cities in the Alternative Sample



Notes: This figure reports the spatial distribution of the cities in the city-level sample used in Tables 2.12–2.23. We use data from Clay, Lewis and Severnini (2019), which contains information on 483 large cities. The red dots report the coordinates of the 478 cities for which we can construct the excess mortality treatment measure. Lighter to darker shades of blue indicate the state-level number of cities included in the final sample.

2.11.2 Tables

Table 2.6: Summary Statistics

	(1) Mean	(2) Std. Dev	(3) Min	(4) Max	(5) Counties
Panel A. Mortality					
Flu Excess Deaths (%)	1.127	0.163	0.682	1.827	1275
WW1 Deaths	39.834	167.083	0.000	4828.000	1275
Panel B. Religion					
All Denominations - Census Based	40.525	14.099	0.000	94.264	1275
Catholics - Census Based	11.024	12.007	0.000	65.853	1275
Protestants - Census Based	27.448	13.807	0.000	94.264	1275
All Denominations - Name Based	0.016	0.021	-0.039	0.103	1274
Catholics - Name Based	0.001	0.008	-0.030	0.045	1274
Protestants - Name Based	0.014	0.017	-0.027	0.084	1274
Panel C. Patents					
Total	278.976	1095.265	0.000	15000.000	1275
Pharmaceutical	49.756	200.266	0.000	2708.315	1275
Communication	11.166	54.019	0.000	1110.340	1275
Electrical	40.552	196.034	0.000	4005.376	1275
Mechanical	136.779	533.283	0.000	7834.571	1275
Other	149.124	568.327	0.000	7290.446	1275
Share of STEM	0.480	0.313	0.054	2.334	1275
Panel D. Income and Demographics					
Population	130.432	329.177	1.061	4819.392	1275
Area	216.975	280.512	0.277	5205.795	1275
Occupational Score per Capita	835.998	74.425	311.939	2302.648	1275
Share of Men	52.305	4.856	19.177	131.582	1275
Share of Illiterates	72.620	9.326	24.633	196.336	1275
Share of Young	33.241	5.981	12.927	86.595	1275
Share of Whites	94.634	11.952	26.163	100.000	1275
Share of African Americans	5.366	11.952	0.000	73.837	1275
Share of Foreign Born	10.680	9.928	0.008	44.052	1275

Notes. This table displays mean, standard deviation, minimum, maximum, and total number of counties for the main variables in the analysis. Data are measured at the county level. Influenza mortality in Panel A is constructed from the mortality statistics; WW1 deaths are from Ferrara and Fishback (2020). Data in Panel B are from either the Census of Religious Bodies at the decade level or are constructed from name frequencies at the year level. Data in Panel C are from Berkes (2018) and are aggregated at the decade level. The share of STEM is computed from the census and is in 1,000 units. Panel D reports data from the 1910 census. County demographics are measured through the IPUMS full-count census (Ruggles, Flood, Foster, Goeken, Pacas, Schouweiler and Sobek, 2021). Panel B data and Panel D population are expressed in thousand units. All variables are cross-walked to 1920 borders.

Table 2.7: STEM Professions

Code	Occupation Label	Code	Label
(1)	(2)	(3)	(4)
Panel A. STEM Occupations			
7	Chemists	58	Nurses, Professional
12	Agricultural Sciences	59	Nurses, Student Professional
13	Biological Sciences	61	Agricultural Scientists
14	Chemistry	62	Biological Scientists
16	Engineering	63	Geologists and Geophysicists
17	Geology and Geophysics	67	Mathematicians
18	Mathematics	68	Physicists
19	Medical Sciences	69	Miscellaneous Natural Scientists
23	Physics	70	Optometrists
25	Statistics	71	Osteopaths
26	Natural Sciences (n.e.c.)	73	Pharmacists
32	Dentists	75	Physicians and Surgeons
34	Dietitians and Nutritionists	83	Statisticians and Actuaries
41	Engineering, Aeronautical	92	Surveyors
42	Engineering, Chemical	98	Veterinarians
43	Engineering, Civil	240	Officers, Pilots, Purses, and Engineers, Ships
44	Engineering, Electrical	541	Locomotive Engineers
45	Engineering, Industrial	563	Opticians and lens grinders and polishers
46	Engineering, Mechanical	583	Stationery Engineers
47	Engineering, Metallurgical, Metallurgists	772	Midwives
48	Engineering, Mining	781	Practical Nurses
49	Engineering (n.e.c.)		
Panel B. Skilled Occupations (Includes STEM)			
$1 \leq \cdot \leq 99$	Liberal Professions	$200 \leq \cdot \leq 299$	Managers
$500 \leq \cdot \leq 595$	Craftsmen		

Notes: Panel A displays the occupations that we classify as Science, Technology, Engineering, and Mathematics (STEM). Panel B displays the occupations that we classify as skilled: these include all STEM occupations and those listed. Occupation codes and labels are from the IPUMS harmonized 1950 occupation taxonomy (variable "OCC1950").

Table 2.8: Balance Checks Regressions

	(1) Coefficient	(2) Standard Error	(3) 95% C. I.
Panel A. Religion			
All Denominations (Name Based)	0.211	(0.224)	$[-0.229, 0.651]$
Catholics (Name Based)	-0.079	(0.196)	$[-0.464, 0.307]$
Protestants (Name Based)	0.350^{*}	(0.191)	$[-0.025, 0.725]$
All Denominations (Census Based)	0.057	(0.176)	$[-0.288, 0.403]$
Catholics (Census Based)	0.168	(0.147)	$[-0.121, 0.457]$
Protestants (Census Based)	-0.094	(0.149)	$[-0.386, 0.198]$
Panel B. Patents and Science			
Total	0.063	(0.055)	$[-0.045, 0.172]$
Pharmaceutical	0.063	(0.059)	$[-0.053, 0.180]$
Communication	0.019	(0.050)	$[-0.078, 0.116]$
Electrical	0.056	(0.066)	$[-0.073, 0.185]$
Mechanical	0.073	(0.057)	$[-0.038, 0.185]$
Other	0.059	(0.052)	$[-0.043, 0.161]$
STEM Employment Share	0.389	(0.281)	$[-0.162, 0.939]$
Panel C. Income and Demographics			
Population Density	-0.009	(0.206)	$[-0.413, 0.395]$
Occupational Score per Capita	0.100	(0.075)	$[-0.047, 0.246]$
Share of Men	-0.092^{*}	(0.054)	$[-0.197, 0.013]$
Share of Illiterates	-0.147	(0.138)	$[-0.417, 0.123]$
Share of Young	0.097	(0.146)	$[-0.190, 0.384]$
Panel D. Ethnic Composition			
Share of Whites	0.096	(0.092)	$[-0.084, 0.276]$
Share of African Americans	-0.096	(0.092)	$[-0.276, 0.084]$
Share of Foreign Population	0.310^{***}	(0.110)	$[0.095, 0.525]$
Immigrants from:			
Italy	0.191	(0.138)	$[-0.079, 0.461]$
Ireland	-0.141	(0.123)	$[-0.381, 0.100]$
Austria	0.275^{**}	(0.116)	$[0.046, 0.503]$
France	-0.044	(0.074)	$[-0.189, 0.101]$
Spain	-0.007	(0.055)	$[-0.114, 0.100]$
Portugal	-0.063	(0.157)	$[-0.372, 0.245]$

Notes: This table displays the correlation between the Excess Death (defined in (2.3)) and a set of covariates in 1910, i.e., the last census year before the pandemic. Column (1) reports the standardized coefficient of a regression between the row variable and our measure of excess deaths; column (2) reports the associated standard error in round brackets; column (3) reports the confidence interval of the point estimate at the 95% confidence level in square brackets. All variables are expressed as shares of the total population, except for population density. Regressions control for county population and include state-fixed effects. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.9: Impact of the Influenza on Religiosity: Weighted Regressions

	Share of Affiliated			Name-Based Religiosity		
	(1) All	(2) Catholics	(3) Protestants	(4) All	(5) Catholics	(6) Protestants
Excess Deaths × Post	0.264**	0.119**	0.137***	1.052***	0.298	1.084***
	(0.135)	(0.056)	(0.029)	(0.274)	(0.247)	(0.229)
County FE	Yes	Yes	Yes	Yes	Yes	Yes
State-Decade FE	Yes	Yes	Yes	–	–	–
State-Year FE	–	–	–	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1274	1274	1274
Observations	3825	3825	3825	38220	38220	38220
R^2	0.775	0.888	0.925	0.843	0.768	0.844
Std. Beta Coef.	0.696	0.398	0.626	0.165	0.108	0.228

Notes: This table displays the impact of exposure to the Great Influenza Pandemic on religiosity. The unit of observation is a county observed at a decade frequency between 1906 and 1926 (in columns 1–3) and yearly frequency between 1900 and 1929 (in columns 4–6). "Post" is a categorical variable equal to one during and after the pandemic—i.e., over 1918 to 1929—or zero otherwise. The baseline treatment "Excess Deaths" is defined in Equation (2.3). In columns (1–3), the dependent variable is the number of individuals affiliated with religious denominations enumerated in the Census of Religious Bodies, normalized by county population in 1910; in columns (4–6), the dependent variable is the name-based religiosity measure described in the main text. Columns (1) and (4) report the effect of the influenza on overall religiosity, whereas columns (2) and (5)—resp. (3) and (6)—display it on the intensity of Catholicism—resp. Protestantism. Regressions include county and state-by-time (decades in columns 1–3 and years in columns 4–6), fixed effects, and control for an interaction term between the population in 1910 and a post-treatment indicator. Counties are weighted by population in 1910. Standard errors, clustered at the county level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.10: Impact of the Influenza on Religiosity

	Baseline Sample			Family Size Cuts		Household	Adults
	(1) Cont. Treat.	(2) Disc. Treat.	(3) WW1	(4) No firstborn	(5) ¡ 5 Kids	(6)	(7)
Excess Deaths × Post	0.827***		0.824***	1.182***	0.921***	0.123***	0.001
	(0.199)		(0.199)	(0.234)	(0.191)	(0.026)	(0.002)
Excess Deaths Dummy × Post		0.252***					
		(0.060)					
WW1 Deaths × Post			-0.000				
			(0.001)				
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Baseline	Baseline	Baseline	No Firstborn	¡ 5 Kids	Household	Adults
Number of Counties	1274	1274	1274	1274	1274	1274	1267
Observations	38220	38220	38220	38220	38220	38220	5068
R^2	0.648	0.648	0.648	0.623	0.631	0.754	0.782
Std. Beta Coef.	0.151	0.033	0.150	0.224	0.174	0.162	0.028

Notes: This table displays the impact of exposure to the Influenza on overall religiosity. The unit of observation is a county, observed at a yearly frequency between 1900 and 1929 in columns (1)-(6) and at a decade frequency in column (7). "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over the years 1918-1929 in columns (1)-(6) and the decades 1920-1930 in column (7)–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). The dependent variable is the name-based measure of aggregate religiosity described in the main text. Column (1) displays the baseline results. Column (2) reports the results coding the treatment as a binary variable returning value one if the continuous treatment is above its median and zero otherwise. In column (3), we control for WW1-related deaths. Column (4) drops first-born children in every household. In column (5), we compute religiosity, dropping all children beyond the fourth in each household. In column (6) we first compute within-household average religiosity and then aggregate the resulting religiosity series at the county-year level. Column (7) reports results measuring county religiosity using the names stock of adults–which serves as a placebo check. All regressions in columns (1)-(6) include county and state-by-year-fixed effects; the regression in column (7) includes county and decade-fixed effects. Additionally, each regression controls for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.11: Impact of the Influenza on Religiosity: Accounting for Migration

	Religiosity Excluding Internal Migrants		
	(1) All	(2) Catholics	(3) Protestants
Excess Deaths × Post	0.876***	-0.038	0.740***
	(0.199)	(0.103)	(0.172)
County FE	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes
Number of Counties	1274	1274	1274
Observations	38220	38220	38220
R^2	0.646	0.525	0.668
Std. Beta Coef.	0.161	-0.016	0.174

Notes: This table displays the impact of exposure to the Influenza on religiosity. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Religiosity is measured using religiosity scores obtained by estimating equation (2.1). Differently from the main text, we exclude from the sample all those who, in the 1930 census, are recorded residing in a state that is different from the one where they were born. Regressions include county and state-by-year-fixed effects and controls for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.12: Impact of the Influenza on Religiosity: City-Level Analysis

	Dep. Var.: Religiosity		
	(1) All	(2) Catholics	(3) Protestants
Excess Deaths × Post	0.012***	-0.000	0.010***
	(0.003)	(0.002)	(0.003)
City FE	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes
Number of Cities	439	439	439
Observations	13170	13170	13170
R^2	0.640	0.788	0.666
Std. Beta Coef.	0.012	-0.000	0.010

Notes: This table displays the city-level effect of exposure to the Influenza on religiosity. The unit of observation is a city observed at a yearly frequency between 1900 and 1929. We report the location of each city in the sample in figure 2.12. The baseline sample is from Clay, Lewis and Severnini (2019). We include only cities where we can construct the baseline excess mortality measure. The dependent variable is the name-based religiosity measure constructed on the universe of children born between 1900 and 1929 and residing in each city in the 1930 census. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929– and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Each regression includes city and state-by-year-fixed effects. Standard errors, clustered at the city level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.13: Impact of the Influenza on Religiosity: Names Scores without Fixed Effects

	Unweighted			Weighted		
	(1) All	(2) Catholics	(3) Protestants	(4) All	(5) Catholics	(6) Protestants
Excess Deaths × Post	2.508***	3.003***	-0.595	3.572**	4.368**	-0.986
	(0.878)	(0.717)	(0.515)	(1.772)	(1.872)	(0.973)
County FE	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	1274	1274	1274	1274	1274	1274
Observations	38220	38220	38220	38220	38220	38220
R^2	0.910	0.902	0.563	0.977	0.965	0.841
Std. Beta Coef.	0.068	0.091	-0.040	0.058	0.093	-0.049

Notes: This table displays the impact of exposure to the Influenza on religiosity. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Religiosity is measured using religiosity scores obtained by estimating equation (2.1), except that we do not include the fixed effects in the regression specification. In columns (4)–(6), counties are weighted by their population in 1900. Columns (1) and (4) report the results for total religiosity; columns (2) and (5) refer to Catholics; columns (3) and (6) refer to Protestants. Regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.14: Impact of the Influenza on Religiosity: Alternative Thresholds

	All			Catholics			Protestants		
	($\tau = 1$)	($\tau = 2$)	($\tau = 4$)	($\tau = 1$)	($\tau = 2$)	($\tau = 4$)	($\tau = 1$)	($\tau = 2$)	($\tau = 4$)
Excess Deaths × Post	0.790***	0.879***	0.407**	0.300***	0.260**	0.022	0.811***	1.061***	0.219
	(0.244)	(0.238)	(0.168)	(0.116)	(0.121)	(0.108)	(0.210)	(0.214)	(0.140)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	1274	1274	1274	1274	1274	1274	1274	1274	1274
Observations	38220	38220	38220	38220	38220	38220	38220	38220	38220
R^2	0.396	0.421	0.692	0.494	0.467	0.529	0.577	0.614	0.778
Std. Beta Coef.	0.143	0.171	0.079	0.099	0.097	0.010	0.165	0.226	0.049

Notes: This table displays the impact of exposure to the Influenza on religiosity. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Religiosity is measured using religiosity scores obtained by estimating equation (2.1). The term τ denotes the frequency threshold a name must exceed to be included in our sample, in ‰ terms. For instance, $\tau = 2$ implies that at least 2‰ children in our sample must be called with a given name, for that name to be included in the sub-sample of names used to compute the religiosity score. In the various columns, we report the estimated coefficients for different frequency threshold values. As τ decreases, the number of names for which we compute a religiosity score increases. Regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.15: Impact of the Influenza on the Concentration of Names

	HHI	CCI	Rosenbluth	C-5	C-6	C-7	C-8
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Excess Deaths × Post	0.168*	0.002	0.160*	0.004	0.006	0.007	0.009*
	(0.095)	(0.002)	(0.092)	(0.004)	(0.004)	(0.005)	(0.005)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1275	1275	1275	1275
Observations	38250	38250	38250	38250	38250	38250	38250
R^2	0.696	0.823	0.708	0.820	0.833	0.841	0.846
Std. Beta Coef.	0.062	0.039	0.058	0.049	0.064	0.075	0.084

Notes: This table displays the impact of exposure to the Influenza on name concentration. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). The dependent variables measure the concentration of names and are: in column (1) the Herfindahl-Hirschman (HHI) index; in column (2) the Comprehensive Concentration index (CCI), which relative to the HHI assigns more weight to relatively uncommon names; in column (3) the Rosenbluth index (RI), which further refines the CCI because it is more sensitive to the number of uncommon names. In columns (4)–(7), the dependent variable is the k-concentration ratio, *i.e.* the share of children called with the k most common names. More formally, let s_n denote the share of kids with name n, and let N be the total number of names. Suppose that shares are ranked in increasing order, meaning that $\text{rank}(n) \leq \text{rank}(n')$ if and only if $s_n \geq s_{n'}$, and $\text{rank}(n) < \text{rank}(n')$ if and only if $s_n > s_{n'}$ for all n, n'. Then, $HHI \equiv \sum_{n=1}^{N} s_n^2$; $CCI \equiv s_1 + \sum_{n=2}^{N} s_n^2(2 - s_n)$, $RI \equiv \frac{1}{2\sum_{n=1}^{N} n s_n - 1}$; $C_K \equiv \sum_{n=1}^{K} s_n$. Regressions include county and state-by-year fixed effects and the interaction between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.16: Impact of the Influenza on Religiosity measured with Saint and Biblical Names

	Abramitzky et al's Religiosity		
	(1) Saints/Biblical	(2) Saints	(3) Biblical
Excess Deaths × Post	3.446**	3.076**	0.865**
	(1.452)	(1.357)	(0.359)
County FE	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes
Number of Counties	1274	1274	1274
Observations	38220	38220	38220
R^2	0.991	0.991	0.984
Std. Beta Coef.	0.032	0.030	0.029

Notes: This table displays the impact of exposure to the Influenza on religiosity. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). In column (1), the dependent variable is the log number of children by cohort whose name either appears in the bible or is carried by a saint; in column (2), the dependent variable only includes biblical names; in column (3), it only includes names of saints. Biblical and saints' names are from Abramitzky, Boustan and Eriksson (2016). Regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.17: Impact of the Influenza on Innovation: Weighted Regressions

	STEM Employment Share		log(1 + Number of Patents)					
	(1) Whole Population	(2) Skilled Population	(3) All Patents	(4) Pharmaceuticals	(5) Communication	(6) Electrical	(7) Mechanical	(8) Other
Post × Excess Deaths	0.010***	0.106***	0.794***	0.320***	0.167	0.075	0.031	0.003
	(0.003)	(0.021)	(0.134)	(0.081)	(0.140)	(0.080)	(0.039)	(0.038)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Decade FE	Yes	Yes	–	–	–	–	–	–
State-Year FE	–	–	Yes	Yes	Yes	Yes	Yes	Yes
All Patents	–	–	No	Yes	Yes	Yes	Yes	Yes
Number of Counties	1274	1274	1275	1275	1275	1275	1275	1275
Observations	3822	3822	38250	38250	38250	38250	38250	38250
R^2	0.811	0.771	0.958	0.949	0.883	0.941	0.978	0.983
Std. Beta Coef.	1.152	1.436	0.231	0.119	0.089	0.028	0.010	0.001

Notes: This table displays the impact of exposure to the Great Influenza Pandemic on innovation. The unit of observation is a county, observed at a decade frequency between 1900 and 1930 in columns (1–2) and at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic—i.e., over 1918 to 1929—or zero otherwise. The baseline treatment "Excess Deaths" is defined in Equation (2.3). In column (1), the dependent variable is the share of people employed in STEM occupations within the population; in column (2), we restrict the denominator to include only those employed in skilled occupations. The dependent variable in columns (3–8) is the (log) number of patent grants. We use this specification of the dependent variable to ensure that we do not drop counties without patents. In columns (4–8), we also control for the overall (log) number of granted patents. Column (3) estimates the impact of the pandemic on the level of innovation, while columns (4)–(8) display this on the direction of innovation. All regressions include county-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Regressions (1–2) include state-by-decade-fixed effects, while regressions (3–8) include state-by-year-fixed effects. Standard errors, clustered at the county level, are reported in parentheses. Counties are weighted by population in 1910. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.18: Impact of the Influenza on Innovation: Robustness Regressions

	All Patents				Pharmaceutical Patents				
	(1) Baseline	(2) Unbalanced	(3) Disc. Treat	(4) WW1 Deaths	(5) Baseline	(6) No All Patents	(7) Unbalanced	(8) Dummy	(9) WW1 Deaths
Excess Deaths × Post	0.336***	0.416***		0.336***	0.099***	0.209***	0.183***		0.099***
	(0.056)	(0.077)		(0.056)	(0.029)	(0.040)	(0.054)		(0.029)
Excess Deaths Dummy × Post			0.080***					0.035***	
			(0.019)					(0.010)	
WW1 Deaths × Post				3.326					2.136
				(6.144)					(2.115)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
All Patents	No	No	No	No	Yes	No	Yes	Yes	Yes
Number of Counties	1275	1227	1275	1275	1275	1275	1227	1275	1275
Observations	38250	23689	38250	38250	38250	38250	23689	38250	38250
R^2	0.858	0.888	0.858	0.858	0.831	0.792	0.813	0.831	0.831
Std. Beta Coef.	0.336	0.416	0.080	0.336	0.099	0.209	0.183	0.035	0.099

Notes: This table displays the impact of exposure to the Influenza on innovation. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. In columns (1)–(4), the dependent variable is the number of patents across all fields; in columns (5)–(9), it is the number of patents in chemical and drug fields, according to the NBER standard classification. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Columns (1) and (5) display the baseline results. Columns (2) and (7) report results for the unbalanced panel of counties (*i.e.*, the subsample of county-year observations for which we observe at least one filed patent). Columns (3) and (8) report the results when the treatment is coded as a binary variable equal to one if the continuous variable is above its median and zero otherwise. Columns (4) and (9) further control for WW1 deaths interacted with the post-treatment indicator. In column (6), we report the estimated effect without controlling for the total number of patents. All regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Columns (5,7-9) further control for the total number of patents. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.19: Impact of the Influenza on Innovation: Alternative Measures of Overall Innovation

	f (All Patents)			
	(1) $\ln(1+\cdot)$	(2) Count	(3) $\text{arcsinh}(\cdot)$	(4) Poisson
Excess Deaths × Post	0.336***	3.736**	0.410***	1.260***
	(0.056)	(1.790)	(0.069)	(0.226)
County FE	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes
Number of Counties	1275	1275	1275	1256
Observations	38250	38250	38250	37680
R^2	0.858	0.911	0.840	0.907
Std. Beta Coef.	0.336	3.736	0.410	3.526

Notes: This table displays the effect of the Influenza on overall innovation. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. In column (1), the dependent variable is the log number of patents, to which we add one to avoid dropping zeros. The dependent variable in column (2) is the raw patent count. In column (3), the dependent variable is the inverse hyperbolic sine of the raw count of patents. In column (4), the model is estimated as a Poisson regression, and the dependent variable is the raw patent count. Each regression includes county and state-by-year-fixed effects and controls for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.20: Impact of the Influenza on Innovation: Alternative Measures of Pharmaceutical Innovation

	f (Pharmaceutical Patents)								
	(1) $\ln(1+\cdot)$	(2) $\ln(1+\cdot)$	(3) Count	(4) Count	(5) arcsinh$(\cdot)$	(6) arcsinh$(\cdot)$	(7) Share	(8) $\ln(1+\text{Share})$	(9) Poisson
Excess Deaths × Post	0.099***	0.209***	0.629***	1.375***	0.125***	0.256***	0.037***	0.031***	1.953***
	(0.029)	(0.040)	(0.202)	(0.443)	(0.037)	(0.050)	(0.010)	(0.009)	(0.306)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Total Patents	Yes	No	Yes	No	Yes	Yes	No	No	No
Number of Counties	1275	1275	1275	1275	1275	1275	1275	1275	1125
Observations	38250	38250	38250	38250	38250	38250	38250	38250	33736
R^2	0.831	0.792	0.958	0.860	0.815	0.777	0.208	0.231	0.795
Std. Beta Coef.	0.327	0.209	0.629	1.375	0.318	0.256	0.037	0.031	7.049

Notes: This table displays the effect of the Influenza on innovation in pharmaceuticals. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. In columns (1) and (2), the dependent variable is the log number of patents, to which we add one to avoid dropping zeros. The dependent variable is the raw patent count in columns (3) and (4). In columns (5) and (6), the dependent variable is the inverse hyperbolic sine of the raw count of pharmaceutical patents, with and without controlling for the inverse hyperbolic sine of the total number of patents. In column (7), the outcome is the number of pharmaceutical patents relative to patents in all other fields. In column (8), this is taken in logs. Each regression includes county and state-by-year-fixed effects and controls for an interaction term between the population in 1910 and a post-treatment indicator. In column (9), the model is estimated as a Poisson regression, and the dependent variable is the raw patent count. In columns (1), (3), and (6), we further control the total number of patents by county year, transformed according to the column-specific labeled function. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.21: Impact of the Influenza on Patent Importance

	All Patents		Pharmaceuticals		
	(1) Breakthrough	(2) Share Breakthrough	(3) Breakthrought	(4) Breakthrought	(5) Share Breakthrough
Excess Deaths × Post	0.166***	0.016**	0.185***	0.160***	0.042***
	(0.043)	(0.008)	(0.039)	(0.035)	(0.010)
County FE	Yes	Yes	Yes	Yes	Yes
State-Year FE	Yes	Yes	Yes	Yes	Yes
Total Patents	No	No	No	Yes	No
Number of Counties	1275	1275	1275	1275	1275
Observations	38250	38250	38250	38250	38250
R^2	0.797	0.265	0.704	0.731	0.340
Mean Dep. Var.	638.000	638.000	638.000	638.000	638.000
Std. Beta Coef.	0.166	0.016	0.185	0.007	0.042

Notes: This table displays the impact of the Influenza on patent importance. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Importance measures are from Kelly, Papanikolaou, Seru and Taddy (2021). They measure the "importance" of a patent based on the textual similarity between that patent and previous and future works and flag it as important if it is different from previous work but similar to subsequent ones. In columns (1) and (3–4), the dependent variable is the share of breakthrough patents, defined as those in the top 5% of the quality distribution. In columns (2) and (5), the dependent variable is the share of breakthrough patents. Regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. In column (4), we further control for the total number of patents. Standard errors are clustered at the county level and displayed in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.22: Impact of the Influenza on Innovation: Accounting for Migration

	No Internal Migrants	
	(1) Full Sample	(2) High-Skilled
Excess Deaths × Post	0.013***	0.293***
	(0.004)	(0.076)
County FE	Yes	Yes
State-Decade FE	Yes	Yes
Number of Counties	1275	1274
Observations	38230	38220
Sample	Full	Skilled
R^2	0.825	0.739
Std. Beta Coef.	0.031	0.063

Notes: This table displays the effect of the pandemic on the probability of being employed in a STEM occupation. The observation unit is a county at a decade frequency between 1900 and 1930. The dependent variable is the share of individuals employed in STEM occupations relative to the overall population (column 1) or the number of people employed in skilled occupations (column 2). We exclude from the sample internal migrants, defined as those who were born in a different state relative to where they are recorded in the 1930 census. STEM and skilled occupations are enumerated in Table 2.7. The baseline treatment "Excess Deaths" is defined in equation (2.3) and is interacted with a post-Flu indicator. All regressions include county and state-by-decade fixed effects and further control for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.23: Impact of the Influenza on Innovation: City-Level Analysis

	Dep. Var.: Number of Patents	
	(1) All Patents	(2) Pharmaceutical
Excess Deaths × Post	0.554**	0.743**
	(0.253)	(0.327)
City FE	Yes	Yes
State-Year FE	Yes	Yes
Number of Cities	476	474
Observations	14280	14206
R^2	0.949	0.851
Std. Beta Coef.	1.740	2.101

Notes: This table displays the city-level effect of exposure to the Influenza on innovation. The unit of observation is a city observed at a yearly frequency between 1900 and 1929. We report the location of each city in the sample in figure 2.12. The baseline sample is from Clay, Lewis and Severnini (2019). We include only cities where we can construct the baseline excess mortality measure. The dependent variable is the number of patents (column 1) and pharmaceutical patents (column 2). "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). Each regression includes city and state-by-year-fixed effects. Standard errors, clustered at the city level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 2.24: Religiosity and the Intensity of Innovation by Exposure to the Influenza

	Dep. Var.: Patents per Capita			
	(1) Pre Flu	(2) Post Flu	(3) Pooled	(4) DiD
Religiosity	-0.059*	0.084**	-0.207***	-0.649***
	(0.035)	(0.041)	(0.042)	(0.241)
Religiosity × Post			0.659***	
			(0.084)	
Religiosity × Excess Deaths				0.404*
				(0.211)
Excess Deaths × Post				0.053***
				(0.014)
Excess Deaths × Religiosity × Post				0.556***
				(0.072)
County FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Number of Counties	1274	1274	1274	1274
Observations	22932	15288	38220	38220
R^2	0.548	0.734	0.578	0.580

Notes: This table displays the correlation between innovation and religiosity by exposure to the pandemic. The dependent variable is the number of patents normalized by county population in 1910, expressed in 1,000 units. The unit of observation is a county observed at a yearly frequency between 1900 and 1929. "Post" is a categorical variable equal to one during and after the pandemic–*i.e.* over 1918-1929–and zero otherwise. The baseline treatment "Excess Deaths" is defined in equation (2.3). In column (1), we display the correlation between religiosity and innovation before the Flu (before 1918); in column (2), we replicate this exercise for the post-Flu years; in column (3), we pool the years together, and interact religiosity with a post-Flu indicator; finally, column (4) reports the differential effect of the pandemic by religiosity. Regressions include county and state-by-year-fixed effects and control for an interaction term between the population in 1910 and a post-treatment indicator. Standard errors are clustered at the county level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Chapter 3

Return Innovation: The Knowledge Spillovers of the British Migration to the United States, 1870-1940

3.1 Introduction

Technological progress and, thus, economic growth hinge on the diffusion of knowledge across countries (Griffith, Harrison and Van Reenen; Comin and Hobijn, 2006; 2011). Eaton and Kortum (1999), for instance, estimate that in the 1980s, approximately 70% of productivity growth in advanced European countries relied on technology developed in the United States and Japan.[1] Recent models emphasize that exposure to foreign technology is crucial for the cross-country diffusion of innovation (Alvarez, Buera and Lucas; Buera and Oberfield, 2013; 2020). Empirically, however, estimating the impact of exposure to foreign technology on domestic innovation is challenging because it requires observing joint variation in the intensity and composition of exposure across observation units—e.g., firms or regions—and technologies.

[1]Economic historians have long argued that the diffusion of knowledge is a key driver of productivity growth and catching up (Gerschenkron; Rosenberg, 1962; 1982). However, endogenous growth models featuring cross-country diffusion dynamics have emerged only recently (Benhabib, Perla and Tonetti; Perla, Tonetti and Waugh; Van Patten, 2021; 2021; 2023).

In this paper, we overcome this challenge by studying out-migration as a novel source of knowledge diffusion from the country of destination to the country of origin of migrants. Drawing on the British mass migration to the United States (1870–1940), we observe that, within Britain, migration ties impact exposure to US technology along two margins. First, exposure in Britain is more intense in regions with higher US emigration rates. Second, emigrants are exposed to different technologies depending on where they settle across the US. By combining these two components, out-migration offers an ideal test to estimate the impact of exposure to foreign knowledge on innovation.

Leveraging this insight, we present novel, causal evidence that exposure to foreign technology through out-migration linkages contributes to the diffusion of innovation to emigration countries.[2] To the best of our knowledge, this is the first paper to document this phenomenon, which we label the "return innovation" effect. Compared to the influential "brain drain" hypothesis (Docquier and Rapoport, 2012), this paper thus introduces a new and competing perspective on the effects of out-migration on the economic development of emigration countries. We find that physical return migration is an important driver of the return innovation effect. However, we present evidence that the interactions between the emigrants and their origin communities—families and former neighbors—further promote technology diffusion in the absence of physical return migration. Moreover, migration ties foster cross-country market integration, which facilitates innovation diffusion into the emigration country.

The impact of emigration on innovation is ambiguous. Traditional "brain drain" arguments suggest that emigration countries suffer from a loss of human capital Gibson and McKenzie (2011). Growth theory, in turn, predicts that this depletion would negatively hamper their ability to innovate Jones (1995). On the other hand, recent scholarship suggests that exposure to innovation is crucial for innovation activity (Akcigit, Caicedo,

[2]A vast scholarship documents that immigrants actively contribute to several dimensions of economic development in their destination countries spanning entrepreneurship (Kerr and Kerr; Azoulay, Jones, Kim and Miranda, 2020; 2022), innovation (Ganguli; Bahar, Hauptmann, Özgüzel and Rapoport; Burchardi, Chaney, Hassan, Tarquinio and Terry; Bernstein, Diamond, Jiranaphawiboon, McQuade and Pousada, 2015; 2019; 2020; 2022) and science (Moser, Voena and Waldinger; Moser, Parsa and San, 2014; 2020), local specialization (Ottinger, 2020), and the formation of political preferences (Giuliano and Tabellini, 2020). As pointed out by Clemens (2011), emigration has generally generated far less attention than immigration.

Miguelez, Stantcheva and Sterzi; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2018; 2019). Therefore, we argue that as migrants are exposed to innovation in the areas where they settle, they promote knowledge flows between those areas and their origin country, as documented qualitatively by Saxenian (2006). Since this "return innovation" effect and the "brain drain" channel operate in opposing directions, empirical evidence is necessary to assess the impact of out-migration on innovation in emigration countries.

We examine this question in the context of the English and Welsh migration to the United States between 1850 and 1940, when approximately 30 million European migrants settled across the Atlantic. Nearly four million came from Britain.[3] Since Rosenberg (1982), economists have interpreted the spread of the Industrial Revolution in terms of waves of technological diffusion originating in Britain. Hence, existing studies document that European immigrants contributed to the diffusion of (mainly) British technology in the US (Jeremy, 1981). This pattern reflects the British technological leadership during the first half of the nineteenth century. Since as early as the 1860s, however, the US approached the technology frontier in many industries, from interchangeable parts and machine tools to engines and agricultural machinery (David; Rosenberg; Rosenberg and Trajtenberg, 1966; 1970; 2004).[4] It is therefore plausible, although unexplored, that migration ties promoted the diffusion of these technologies back to Britain, which, throughout this period, increasingly lagged behind the newly industrialized countries.

Besides its historical importance, this setting allows us to overcome three limitations of contemporary scenarios that hindered previous attempts to study this question. First, our novel individual-level dataset allows us to look at the entire population of transatlantic migrants. Second, we measure international knowledge flows using historical patent data. This approach would be infeasible with contemporary data due to international in-

[3]This figure does not include the Irish. In the paper, we focus on the English and Welsh migration. We use, for the sake of brevity, the terms "British" and "English" as shortcuts to collectively refer to England and Wales, thus excluding Scotland.

[4]By the 1890s, the American technological primacy was well-established. Nelson and Wright (1992) note that starting in the 1880s, American technology saw major advancements in textiles, sewing machines, clocks, firearms, boots and shoes, locomotives, bicycles, and cigarettes. Starting in the 1890s, mass production led to innovations in consumer products (canned goods, dairy, and grain products), light machinery (typewriters, cameras), electrical equipment, and industrial machinery, such as boilers, pumps, and printing presses.

tellectual property protection in force since 1945.[5] Finally, the near-complete absence of migration regulations targeting British migrants ensures that possibly endogenous policy interventions do not confound the analysis.

To estimate the effect of out-migration on innovation, we observe that districts in the UK would be exposed to different technologies depending on the US county where emigrants from those areas would settle.[6] Our research design thus leverages the joint variation in county-level specialization across technology classes and district-county bilateral migration flows. Consider two English districts, say Staffordshire and Camden, and two US counties, say San Diego and Cook. Assume hypothetically that Staffordshire and Camden are observationally identical, but all emigrants from Staffordshire settle in San Diego County, whereas all those originating in Camden move to Cook County. Suppose that San Diego County specializes in shipbuilding, whereas Cook County specializes in chemistry. Then, Staffordshire will be exposed to shipbuilding technology, whereas Camden will be exposed to innovations in chemistry.

We assemble two novel, detailed, general-purpose datasets to overcome the limitations of the existing sources. First, since the available data do not contain information on the origin of British immigrants within the UK, we leverage confidential individual-level UK and US census data to link individual records of British immigrants in the US to the UK census (Schurer and Higgs; Ruggles, Fitch, Goeken, Hacker, Nelson, Roberts, Schouweiler and Sobek, 2020; 2021). The resulting novel dataset allows us to track individual out-migration and return migration between the US and the UK. Second, to reconstruct the geography of innovation in the UK in the second half of the nineteenth century, we digitize the universe of 300,000 original patent documents issued in England and Wales between 1853 and 1899. We thus assemble the first comprehensive dataset covering patented innovation during the Second Industrial Revolution in the United Kingdom.

[5]At the time, the same invention could be patented by different people in the US and in the UK with no legal penalty. Additionally, the British patent office required that at least one applicant—usually a patent agent for foreign inventors—be a permanent resident in the UK.

[6]In most of the analysis, the units of observation are UK registration districts and US counties. In 1901, there were 631 registration districts in England and Wales. Districts were comparable to US counties in terms of population (approximately 40,000). Unlike counties, however, registration districts were statistical entities that did not enjoy political or budgetary autonomy.

The granularity of our data allows us to deal with the potential endogeneity of exposure to US knowledge. The primary reason that would caution against a causal interpretation of the estimates is assortative matching, namely the possibility that British migrants sort across US counties depending on where they came from. Suppose, drawing on the previous example, that Staffordshire specializes in shipbuilding. Then, Staffordshire would be exposed to shipbuilding technology because emigrants from that district settle in San Diego County, which also specializes in shipbuilding, but the coefficient of a naïve regression between knowledge exposure and innovation would conflate pre-existing specialization patterns into the treatment effect of exposure to US technology.[7]

We develop two approaches to deal with this potential source of endogeneity. First, we build a shift-share instrumental variable that exploits conditional variation in the connection timing to the US railway network to construct county-level immigration shocks (Sequeira, Nunn and Qian, 2020). These shocks allow us to randomize British immigration across counties and avoid the assortative matching issue (Borusyak, Hull and Jaravel, 2022). Second, we note that the return innovation effect would imply that shocks to innovation activity in the United States—defined as unusually large deviations from the average yearly number of patents by technology class—would diffuse to UK districts whose emigrants had settled in the areas where these shocks manifest. To test this hypothesis, we implement a triple differences analysis that compares districts and technology classes by exposure to innovation shocks in the US.

Our main result is that innovation in the UK shifts in response to exposure to US innovation through migration linkages. The instrumental variable design confirms the existence of a causal link between exposure to US knowledge and innovation in the UK. The triple differences analysis provides evidence that innovation shocks in the US diffuse into the United Kingdom through migration ties. We estimate that exposure to an innovation shock in the United States—which, on average, is associated with twenty more patents in a given county-technology class pair—results in two more patents produced in the UK. This implies a 10% pass-through rate of US innovation shocks to Britain through migra-

[7]This example serves illustrative purposes, but our baseline research design non-parametrically rules out the possibility that differences in initial specialization across UK districts drive our results.

tion ties. This figure is sizable, as it accounts for approximately one-third of the average annual number of patents by district technology class in Britain. This result, which we label the "return innovation" effect, is larger in industries in which the UK was relatively more specialized than the US. Exposure to foreign knowledge through migration ties thus appears to nurture existing industries rather than creating new ones.

We then ask whether the knowledge flows generated by migration linkages stimulate technology transfer between the US and Britain or if they propel original innovation in the UK. To do so, we adopt a text-based approach that quantifies (i) the similarity between UK patents and previous US patents and (ii) the "originality" of the former with respect to the latter. We find that areas more exposed to US knowledge produce patents that are more similar to those granted in the United States. We also estimate that those areas produce more innovative patents compared to the existing stock of US knowledge. These results are not contradictory. In the immediate aftermath of a US innovation shock, the similarity of newly produced UK patents with previous US patents increases. However, in later periods, original patents take over the bulk of the increased innovation activity. Taken together, these results indicate that return innovation conflates two margins: a technology transfer catching-up effect *à la* Gerschenkron (1962) and a positive spillover channel that stimulates the production of novel knowledge.

In the second part of the paper, we exploit the richness of our data to explore the mechanisms that underlie the return innovation effect. On the one hand, return innovation may require the physical return of migrants. On the other, however, migration ties may promote the diffusion of technology irrespective of physical return. We find that physical return is an important but not the exclusive driver of return innovation. Interactions between emigrants and their communities of origin further promote technology diffusion even absent physical return. Moreover, we provide indirect evidence that migration ties foster cross-border market integration, further facilitating innovation diffusion into the UK.

Since our data allow us to observe return migration at a high level of spatial granularity, we can measure exposure to US technology through return migrant flows and compare

it with the effect of outward emigration flows. We find that return migration accounts for approximately half of the overall return innovation effect. Importantly, however, the effect of exposure to US technology through out-migration ties remains sizable and significant even if we control for return inflows. This result suggests that return migration is an important determinant of return innovation, but it also indicates that migration ties contribute to the diffusion of knowledge even absent physical return.

We first explore the role of interactions between emigrants and their communities of origin as a driver of technology diffusion. We focus on two factors that could promote such interactions: family ties and geographical proximity. Our results connect to a large literature in development economics which links the diffusion of technology to network interactions Bandiera and Rasul; Conley and Udry; Beaman, BenYishay, Magruder and Mobarak (2006; 2010; 2021). Then, we study the channels through which migration ties impact innovation in emigration countries without directly relying on personal relationships between the emigrants and their origin communities.

We find that the family members of US emigrants display increased patenting activity after their relative moves to the US. It takes about ten years for a British emigrant to contribute to innovation activity back home. Despite this delay, the magnitude of the effect is substantial. Importantly, we can distinguish between emigrants who, at some point, return to the UK from those that do not. The impact of return emigrants is considerably larger than that of those who never return. This further confirms that return migration is a major driver of return innovation. At the same time, however, emigrants promote innovation in their families even if they never return. Overall, since return emigrants account for approximately one-third of the entire migrant stock, the magnitude of these effects is, in aggregate, similar.

The geographical proximity between emigrants and their former neighbors can be interpreted as an alternative proxy for local social networks. To estimate its impact on the innovation activity of stayers, we leverage the individual-level nature of our migration and patenting data. Using a linked patent-census sample and geo-coded information on the universe of the UK population, we find that patenting activity increases for non-migrants

after their neighbor(s) migrate(s) to the United States.[8] Moreover, the estimated effect remains positive and significant when restricting the treatment to include only US migrants who never return. These results strongly suggest that cross-country interactions between emigrants and their origin communities are a key driver of return innovation, even absent physical return.

Building on previous literature, we provide indirect evidence that migration ties contribute to cross-border market integration, thus promoting knowledge flows. We leverage the introduction of the first transatlantic telegraph cable connecting the US and the UK in 1866 as a sudden and sizable increase in the integration of the British and the American markets.[9] In a difference-in-differences setting, we show that districts with higher US emigration rates before the introduction of the transatlantic telegraph cable display higher patenting activity after 1866. Moreover, innovation does not increase evenly across technology classes. The gains in patenting activity manifest in those same technologies that districts had been more exposed to through migration ties. This suggests that the increased economic integration generated by the telegraph accrued relatively more to districts that had pre-existing migration ties with the US market.

To provide additional evidence that migration ties foster market integration, we study trade disruptions arising from the Smoot-Hawley Act (1930), which severely increased US import duties. Trade is commonly interpreted as a measure of cross-border market integration. Importantly, the tariff increase was not homogeneous across goods categories. Leveraging this cross-industry variation, we find that patenting in the UK decreases in districts more exposed—through migration ties—to technologies that the Act targeted more heavily. This result suggests that migration ties promote market integration, which facilitates cross-border knowledge diffusion.

Finally, we investigate whether the information flows generated by migration ties are

[8] In the baseline exercise, two individuals are considered as neighbors if they live in the same street. However, in robustness regressions, we define neighborhoods as areas of a 100-meter radius centered around each individual in the sample.

[9] The telegraph represented a fundamental development in information and communication technology. Steinwender (2018) documents that the transatlantic cable allowed information to flow more rapidly and efficiently across the Atlantic Ocean, thus enabling trade and reducing international arbitrage opportunities.

restricted to innovation or if they encompass a broader set of subjects. We collect data on the coverage of US-related news from a comprehensive repository of historical British newspapers. We find that newspapers in areas with more US emigrants are relatively more likely to cover US-related news. Newspaper coverage of a given state (resp. county) is broader in districts with more emigrants to that given state (resp. county). This exercise suggests that the scope of information flows generated by migration ties is not limited to innovation and encompasses a broader set of topics.

This paper provides new evidence on how knowledge diffuses across countries. More specifically, we find that exposure to foreign technology through migration ties contributes to the diffusion of innovation from the country of destination to the country of origin of migrants. Our results imply that out-migration can promote innovation and, thus, long-term growth by fostering the diffusion of knowledge into emigration countries. Despite cautions on external validity, ever-increasing international human mobility and advancements in communication technology suggest that our results bear relevant policy implications for economic growth, especially in developed countries.

Related Literature. This paper is related to four streams of literature. First, we contribute to the literature that studies the determinants of the direction of innovation and the allocation of research activity across technological sectors. Pioneering work on directed technical change by Hicks (1932) and Habakkuk (1962) was formalized by Acemoglu; Acemoglu (2002; 2010). More recently, this question has been studied both theoretically (Bryan and Lemus; Hopenhayn and Squintani; Acemoglu, 2017; 2021; 2023) as well as empirically (Hanlon; Aghion, Dechezleprêtre, Hemous, Martin and Van Reenen; Moscona; Moscona and Sastry; Einiö, Feng and Jaravel; Gross and Sampat, 2015; 2016; 2021; 2022; 2023; 2022). We inform this literature by introducing one novel determinant of the direction of innovation, namely, international human mobility, through the return innovation effect.[10]

Second, we contribute to the literature that studies the effects of out-migration on coun-

[10]A related literature highlights that the direction of innovation bears relevant consequences in terms of subsequent technical change because it can lead to technology lock-ins (Dosi; Arthur; Acemoglu and Lensman, 1982; 1989; 2023).

tries sending migrants. Emigration has been shown to impact wages (Dustmann, Frattini and Rosso, 2015), attitudes towards democracy and voting (Spilimbergo; Batista and Vicente; Ottinger and Rosenberger, 2009; 2011; 2023) and political change (Chauvet and Mercier; Kapur; Karadja and Prawitz, 2014; 2014; 2019), technology adoption (Coluccia and Spadavecchia, 2022), entrepreneurship (Anelli, Basso, Ippedico and Peri, 2023), and social norms (Beine, Docquier and Schiff; Bertoli and Marchetta; Tuccio and Wahba, 2013; 2015; 2018). This paper provides new evidence that emigration shapes the dynamics and the direction of innovation because it exposes sending countries to novel knowledge produced abroad. This enriches the traditional narrative that reduces out-migration to a mere depletion of the human capital stock.

By its setting, this paper adds to the literature that studies technical change and diffusion of novel technologies during the Age of Mass Migration. A growing number of papers examines the short-run (Arkolakis, Lee and Peters; Moser, Parsa and San, 2020; 2020) as well as the long-run (Akcigit, Grigsby and Nicholas; Burchardi, Chaney, Hassan, Tarquinio and Terry; Sequeira, Nunn and Qian, 2017; 2020; 2020) implications of immigration on US innovation. Ottinger (2020) shows that European immigration influenced US industry specialization. This paper is closest to Andersson, Karadja and Prawitz (2022). They show that mass out-migration in Sweden triggered labor-saving innovation by increasing the relative cost of labor. Instead, we look at the diffusion of technology from the areas where migrants settle to those they originate from. We are thus able to dissect the impact of out-migration on technology diffusion from the US to Britain. We show that migration ties facilitate the cross-border diffusion of technologies and find that information flows, rather than physical return migration, is the main underlying channel of this "return innovation" effect.

Finally, we relate to the literature studying the dynamics and determinants of knowledge flows and technology diffusion across countries (Jaffe, Trajtenberg and Henderson; Griffith, Harrison and Van Reenen; Bahar, Hausmann and Hidalgo; Pauly and Stipanicic, 1993; 2006; 2014; 2021). Specifically, we contribute to the papers documenting how human mobility fosters the diffusion of novel knowledge (Kerr; Hornung; Bahar, Hauptmann, Özgüzel and Rapoport; Bahar, Choudhury, Sappenfield and Signorelli; Prato,

2008; 2014; 2019; 2022; 2021). We contribute to this literature from several perspectives. First, we enlarge the observation sample to include the universe of emigrants instead of a selected subgroup of highly skilled individuals. Second, we leverage recent insights by Akcigit, Caicedo, Miguelez, Stantcheva and Sterzi (2018) and Bell, Chetty, Jaravel, Petkova and Van Reenen (2019) and show that exposure to foreign technology is a major driver of technology transfers. Third, we emphasize that the return innovation effect does not exclusively hinge on the physical return of emigrants. Finally, our setting allows us to uncover the long-run effects of emigration and the mechanisms through which it affects innovation in the home country of emigrants.

Outline. The rest of the paper is structured as follows. In section 3.2, we describe this study's historical and institutional context. Section 3.3 introduces the novel datasets we construct. We present the empirical research design in section 3.4 and discuss the main findings in section 3.5. Section 3.6 uncovers the possible mechanisms underlying the results and discusses possible alternative interpretations. Section 3.7 concludes.

3.2 Historical and Institutional Background

This section offers a concise overview of the historical and institutional features of our study setting. Throughout it, we highlight key aspects and details that were relevant to the empirical investigation. We conclude by presenting three examples of technology transmission to the UK operated by British immigrants in the US.

3.2.1 The English and Welsh Migration to the United States

Between 1850 and 1940—during the so-called Age of Mass Migration—more than 30 million Europeans migrated to the United States (Abramitzky and Boustan, 2017). Migrants from Great Britain—England and Wales in particular—accounted for approximately 10% of this flow (Willcox, 1928). Emigration rates in Britain were among the highest in Europe, except for the years 1890–1900. They steadily increased throughout the period

(Baines, 2002).[11]

Migration Policy in the United Kingdom and the United States

The virtual absence of legal constraints to human mobility represents a major appealing feature of the Age of Mass Migration for economic research. Until 1917, the US applied minor restrictions on European immigration (Abramitzky and Boustan, 2017).[12] Immigrants mostly originated from Northern Europe, particularly the United Kingdom, Ireland, Germany, Sweden, and Norway. This positive attitude towards immigration ceased as flows from Eastern and Southern Europe increased in the 1890s (Goldin, 1994). The restrictive immigration policies of the 1920s, however, allotted generous quotas to the United Kingdom, which were never filled (Abramitzky and Boustan, 2017).[13]

Like in other European countries, out-migration legislation in the UK sought to help emigrants, if not explicitly to foster emigration Baines (2002). Out-migration was encouraged in two ways: reduced and subsidized ticket fares and allotment of agricultural lands. Policy efforts were directed towards the Empire, particularly Canada and Australia, through the Committee of the Emigrants' Information Office. In general, however, these policies were not successful. Baines (2002) argues that less than 10% emigrants traveled under government assistance during the entire 1814-1918 period, and Leak and Priday (1933) report similar figures for the post-War era. Emigration to the United States was neither subsidized nor discouraged. Attitudes towards out-migration remained positive after the First World War. The perceived slowdown of emigrant flows after the War was viewed with concern by policymakers (Leak and Priday, 1933).

[11]Only Ireland, Italy, and Norway had higher emigration rates, although, in England, massive out-migration spanned longer than in the other countries above.

[12]Immigration from China had been severely restricted since as early as 1882. Restrictions on European immigration before 1917 targeted selected groups, such as convicts and disabled persons. In 1917, Congress passed an act that sanctioned legal immigrants' detention and deportation if they committed a crime within five years of their arrival. The act also imposed literacy tests, which, however, did not significantly impact immigration from European countries (Goldin, 1994).

[13]The 1921 (resp. 1924) Act computed the quota for a given country as 3% (resp. 2%) of the population from that country that was recorded in the US census in 1910 (resp. 1880). This scheme favored first-wave immigration countries, such as the United Kingdom and Germany, at the expense of new ones, as recommended by the Dillingham Commission (Higham, 1955).

This overview suggests that institutional constraints to US out- and immigration were largely absent for English and Welsh migrants throughout the XIX and early XX century. Compared to contemporary scenarios, this historical setting thus allows us to abstract from confounding factors arising from endogenous migration legislation.

English and Welsh Emigrants: The Perspective of Great Britain

Compared to the broader European phenomenon, the British migration to the US presents two main distinctive features.[14] First, unlike continental countries, Britain was already highly urbanized and industrialized at the inception of the Mass Migration. Erickson; Erickson (1957; 1972) and Thomas (1954) highlight the centrality of urban areas which, starting in the 1880s, supplied the majority of overseas migrants. Baines (2002) provides some estimates on the origin of migrants based on birth certificates over the years 1850–1900. Emigration ratios were highest in Northern and South-Western England and lowest in Lancashire and neighboring areas. Second, the selection of British migrants radically differed from that in continental countries (Erickson; Abramitzky, Boustan and Eriksson, 1957; 2020). Compared to the occupational structure of Great Britain, migrants were less likely to be employed in agriculture and more likely to be low and high-skilled industrial workers Baines (2002). Until the 1880s, British emigrants generally came from rural areas and, consequently, the vast majority were farmers. However, as cities and smaller urban centers gained prominence, migrants were increasingly employed in industrial manufacturing occupations (Baines, 2002). At the beginning of the 1860s, when the transatlantic migration was taking off, about 15% emigrants were employed in agriculture, and merely five percent were white-collar workers. In the early 1900s, however, this composition had shifted as agriculture workers accounted for a mere five percent of the overall emigrant stock, while those employed in white-collar occupations were 25%.

Our newly constructed migration database allows us to assess the historical evidence quantitatively. In Appendix Table 3.17, we compare individual-level characteristics of

[14]Throughout the period, the US was the most relevant destination for English and Welsh migrants. Between 1850 and 1930, more than 40% emigrants settled in the US. This compares to 25% in Canada, 20% in Australia, and 15% in other destinations Baines (2002).

emigrants with the staying population. On average, emigrants were more likely to come from North West and South East England. Moreover, they were less likely to be farmers. By contrast, emigrants' share of high and low-skilled manufacturing workers is substantially larger than among stayers. Similar—although less marked—patterns were observed for return migrants. Appendix Figure 3.24 displays the origin of emigrants over time at the district level. The data vividly show that rural areas in central and south-western England, which initially feature the highest emigration rates, were gradually replaced by urban industrial districts in the North and South. Taken together, this evidence confirms the qualitative historical knowledge.

English and Welsh Immigrants: The Perspective of the United States

British immigrants have been central throughout the economic and political history of the United States (Berthoff; Fischer, 1953; 1989). Several features distinguish the English from the continental transatlantic migrations. First, English and Welsh immigrants were, especially after the 1880s, artisans and manufacturing workers, who settled where their skills were in highest demand (Berthoff, 1953).[15] Textile workers from Manchester typically settled in Massachusetts, whereas coal miners from Southern Wales mostly settled in the Midwest and Pennsylvania. In 1890, 63% British-born were employed in manufacturing (Thistlethwaite, 1958). Second, English immigrants—unlike the Welsh—did not form ethnic clusters (Furer, 1972). Instead, they tended to be scattered around settlement areas in highly diverse ethnic communities. Finally, British immigrants were economically successful and assimilated relatively easily with the US-born population (Abramitzky, Boustan and Eriksson, 2020).

We quantitatively evaluate these observations in Table 3.18. First, we compare individual-level characteristics observed in the US census between the US-born and the British immigrants. The analysis suggests that British immigrants were substantially different from the average native. For example, they were richer, more literate, and more likely to

[15]Thistlethwaite (1958) presents one instructive example. The pottery industry, a highly skilled and labor-intensive sector, was concentrated in the Five Towns of Staffordshire. As transatlantic migration ensued, ceramic workers located in just two localities: Trenton, New Jersey, and East Liverpool, Ohio.

live in urban centers. Consequently, they were less likely to be farmers and more likely to be employed in manufacturing occupations with high or low-skill content. In addition, English immigrants were comparatively more concentrated in North Atlantic states and the West and less in Southern states. Similar patterns emerge for return migrants.

These results, coupled with Table 3.17, identify British immigrants in the US as part of an urban industrial class of skilled and semi-skilled workers. This is crucial in our analysis: it would have been much more difficult for illiterate farmers to facilitate knowledge flows across the Atlantic Ocean.

3.2.2 Intellectual Property Protection in the US and the UK

We measure innovation and knowledge flows using patent data. In this section, we briefly present the key features of the British and American patent systems and discuss the state of international intellectual property protection in the XIX and early XX centuries.

National Patent Systems

Britain established the world's oldest continuously operating patent system in 1623-1624. Until 1850, access to intellectual property protection was, however, difficult (Gomme; Bottomley, 1948; 2014). Fees amounted to approximately four times the average income in 1860, and the application process was lengthy and rife with uncertainty (Dutton, 1984). A large literature documents the poor performance of this system during the Industrial Revolution (Macleod; Moser, 1988; 2012). The 1852 Patent Law Amendment Act sought to reform this process. The US system inspired the reform effort, which reduced application fees and attempted to streamline bureaucratic procedures. One subsequent reform in 1883 further reduced fees, allowed applications by mail, designed a litigation system and provided for the employment of professional patent examiners (Nicholas, 2011). A technical examination of novelty was introduced only in 1902. Until 1907 patents were granted conditional on the invention being produced in Britain (Coulter, 1991).

The first article of the United States Constitution establishes that inventors be granted exclusive rights over their discoveries. In 1836 the US Congress passed the Patent Act, which formally instituted the US Patent Office (USPTO). The USPTO has been credited as the first modern patent system in the world (Khan and Sokoloff, 2004). Two features distinguished the American patent system from its European counterparts. First, an examination of novelty was carried out by professional examiners to ascertain the originality of patent applications. Second, low application fees ensured that access to intellectual property protection was widespread (Sokoloff and Khan, 1990). Several scholars documented how effectively the US patent system fostered innovation well into the 20$^{\text{th}}$ century (Khan, 2020).

International Intellectual Property Protection

As national patent systems spread across Europe and the US during the 19$^{\text{th}}$ century, demands for international regulation increased. The Paris Convention—formally, the "Paris Convention for the Protection of Industrial Property"—of 1883 governed international patent protection (Penrose, 1951).

The Paris Convention emerged out of a decade of multilateral confrontations spurred by World Exhibitions in Vienna (1873) and Paris (1878). The Convention introduced two major principles. First, nationals and residents of subscribing countries were guaranteed equality of treatment with nationals. This concept, known as "national treatment", rejects the principle of "reciprocity", which maintains that nationals in subscribing countries would be granted the same protection as their origin country. The United States had vigorously demanded reciprocity (Penrose, 1951). Second, upon applying for a patent in one member country under Article 4, inventors were granted a "right of priority" of six months. Patents filed in foreign countries during the priority period would not invalidate the inventor's claim for protection in other member countries. The provisions contained in Article 4 were central within the broader legal apparatus (Penrose, 1951). However, patents obtained in one member state were *not* automatically recognized by other countries. To effectively claim protection, inventors had to submit different patent

applications. This represented a substantial bureaucratic and financial burden. While the Paris Convention—and its numerous amendments—are still in operation today, international patents were established only in 1970. The UK joined the Convention in 1884, while the US waited until 1887.

The state of international intellectual property protection during our period is a major advantage of this historical setting. Since the UK and the US did not mutually recognize patents, we can use them as an informative proxy of knowledge flows between the two countries. This approach would be impracticable in modern settings.

3.2.3 Anecdotal Evidence of Return Innovation

Who were the immigrants that contributed to the diffusion of US technology in Britain? History is rife with examples of skilled artisans, entrepreneurs, and factory workers who were exposed to some novel technology where they settled and promoted its diffusion, or in some cases appropriated it, in the UK.

In this section, we provide three instructive examples. All three are cases of return migration. Historical records typically focus on successful migrants who, upon returning, bring their technology to their origin areas and promote economic development there. The statistical analysis that we present later, however, suggests that this was only part of the story. In fact, we find that emigrants interacted with their origin communities even without returning.

British Puddlers and the Kelly-Bessemer Process

An 1856 article published in Scientific American described a new patent granted in the UK to Henry Bessemer (Wagner, 2008). Bessemer had discovered a new process, the would-be eponymous Bessemer process, that, for the first time, allowed the production of inexpensive steel from molten pig iron.[16] American inventor William Kelly complained:

[16]The Bessemer process was one of the most transformative technological developments of the nineteenth century (Rosenberg and Trajtenberg, 2004).

> "I have reason to believe my discovery was known in England three or four years ago, as a number of English puddlers visited this place to see my new process. Several of them have since returned to England and may have spoken of my invention there."
> (Wagner, 2008)

The veracity of Kelly's allegations remains unverified. They nonetheless indicate three important elements. First, American inventors knew that British immigrants posed a threat to the secrecy of their inventions. Second, technology transfer did not necessitate the very upper tail of the human capital distribution. Skilled workers, such as puddlers, could be the agents of technology diffusion. Finally, the precise mechanism that emerges is return migration. Kelly expects British puddlers to speak of "his" invention upon returning to England.

Henry Marsden and the Industrialization of Leeds

Henry Rowland Marsden was born in Leeds to poor parents in 1823 (Curtis, 1875). At age twenty-five, he emigrated to the United States, first to New York and then to Connecticut. There, he took on apprenticeships in engineering and metal-working firms. He obtained several engineering patents—chiefly related to steam engines and pumps, including a "stone-crusher" which is still in use today. In 1862, Marsden returned to Leeds, where he set up a flourishing business centered around his newly patented inventions. A wealthy man respected for his philanthropic endeavors, he was elected mayor of Leeds in 1873. He died in 1878 and is credited as one of the most prominent figures in the industrial development of Leeds.

Migrants as Agents of Technology Transfer: Wellstood & Smith Ltd.

The case of Stephen Wellstood and John Smith illustrates how international migration spurs technology transfers across countries. At age 16, James Smith (1811–1886) left Bonnybridge, Scotland, and migrated to the US. There, he established himself selling

cooking stoves and married. However, as his wife got ill, Smith returned to Bonnybridge and started re-selling imported stoves from the US. He soon realized, however, that he could manufacture stoves directly in Britain. He then partnered with his long-time friend Stephen Wellstood and opened a foundry. They patented the exact same cooking stove Smith had been selling in the US and started a business that remained active until 1983.

3.3 Data

This section presents our primary data sources and discusses the key methodology we adopt to assemble the final datasets. We provide a more detailed description of the data in Appendix sections 3.10, 3.11, and 3.12. Table 3.1 lists the main variables and provides descriptive statistics.

3.3.1 Migration Data

To conduct our analysis, we need information on the origin of English and Welsh immigrants in the United States *within* the United Kingdom. Currently available data, however, do not contain this information. Neither the US nor the UK collected disaggregated data on, respectively, the origin of immigrants and the destination of emigrants. We tackle this limitation of the data by developing a new dataset that links British immigrants in the US to the UK census. This allows us to observe an individual in the UK and to track him to his US census record after he emigrated.[17] This is the first dataset that reconstructs migration flows at this granular level of aggregation for a major European country in this period.[18]

To construct our linked dataset, we leverage non-anonymized individual-level data from

[17]Throughout the paper, we use the masculine to refer to individuals in our data because, as we explain in detail later, we can only work with male individuals.

[18]Data assembled by Abramitzky, Boustan and Eriksson (2014) and Andersson, Karadja and Prawitz (2022) serve a similar purpose for, respectively, Norway and Sweden. England and Wales, however, were substantially larger in terms of the overall population and the US immigrant population. The population of Sweden and Norway in 1890 was approximately 4.7 and 2 million. In the same year, the population in England and Wales stood at 27 million.

the population censuses in the United Kingdom (Schurer and Higgs, 2020) and the United States (Ruggles, Fitch, Goeken, Hacker, Nelson, Roberts, Schouweiler and Sobek, 2021). We first extracted the universe of British immigrants from the US censuses in 1900, 1910, 1920, and 1930.[19] These list, among other variables, the name and surname, birth year, and immigration year of each migrant. We then match these records to the closest census when they appear. Hence, for example, we try to link an individual who immigrated to the US in 1905 to the 1901 UK census.[20] The matching variables we consider are the name, surname, and reported birth year. We use state-of-the-art census-linking algorithms adapted from pioneering work by Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021). Appendix 3.12.1 lists in more detail the primary sources and the technical implementation of the algorithm. This class of linking algorithms relies on the observation that a simple exact matching routine would artificially discard many plausible links between the two sources because of minor coding errors by the census enumerators. Since human hand-checking is unfeasible, we implement an algorithm that returns a match whenever the string similarity between the US and the UK records is above a certain threshold, conditional on the birth year.

This approach presents some important caveats (Bailey, Cole, Henderson and Massey, 2020). First, it may deliver spurious links if the matching variables are insufficient to restrict the pool of potential matches. Second, the matching probability may be correlated with individual characteristics. This would be the case if, for instance, the likelihood that names and surnames were correctly enumerated in the censuses correlated with education. We discard the matches that do not attain a high level of string similarity to address the first concern. Moreover, we only keep immigrants matched with up to two records in the UK census. This ensures that we minimize the rate of false positives as much as possible. We provide evidence against the second issue in Table 3.13, which shows that the correlation between the number of matches and individual-level observable

[19]We cannot use information contained in the 1870 and 1880 censuses because the immigration year was not recorded. Individual-level data from the 1890 census have not survived.

[20]Because no census was taken in 1870, we match those who migrated between 1870 and 1881 to the 1860 census. Moreover, since the last available UK census was in 1911, we match all those who emigrated after 1911 to that one. This implies that we have no information on migrants born after 1911. Since the median age of migrants is 30 and less than 10% of the distribution is younger than 19 in the rest of the sample, this bears little quantitative implications for the matching rate in the later part of the sample.

characteristics is seldom significant, and always very small in magnitude.

We perform one additional exercise to assess the plausibility of the linked migrant sample. Following Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021), we construct an intergenerational linked sample that identifies individuals in census decade t in the subsequent census in decade $t + 10$. The underlying rationale is that the matching rate in this intergenerational linked sample should be lower for US emigrants than for the non-migrant population. We discuss this approach in more detail in Appendix section 3.12.2. Figure 3.22 reports the results of this exercise. We link approximately 40% non-migrants to an individual in the following decade. As expected, this figure decreases to 20% for US emigrants. This exercise thus provides reassuring evidence that the UK-US linked sample can confidently identify migrants. Moreover, Figure 3.23 confirms that the migratory flows in the baseline sample are highly consistent with those that are obtained by repeating the UK-US linking but excluding individuals that are matched in the intergenerational linked sample.

Finally, we construct a dataset of return migrants. To assemble it, we apply the exact previous logic, except that migrants are matched to the UK censuses taken in the decades *after* their immigration year. Hence, as an example, someone who migrated to the US in 1895 is matched to censuses in 1901 and 1911. To avoid double counting, if a migrant is matched to more than one census, we keep the match(es) in the first. Data on return migrants are generally scant historically and with modern data (Dustmann and Görlach, 2016). This exercise is thus a valuable feature of our methodology.

In Figure 3.1, we report in gray the number of English and Welsh immigrants in the United States by year of immigration, digitized from official statistics (Willcox, 1928). The blue line on the right y-axis tabulates the number of immigrants in our linked dataset. We attain a matching rate of about 60% after dropping multiple matches and links with below-threshold matching quality. Note that we are forced to discard women whose surname was likely to change after marriage. The matching rate aligns with the literature on census linking (Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez, 2021).[21] More-

[21]In Appendix section 3.12.2, we provide a more detailed discussion of the algorithm's performance.

over, reassuringly, our data co-moves with official statistics data. Figure 3.2 reports the spatial distribution of emigration rates across districts in the final sample and highlights its cross-sectional spatial heterogeneity. In Appendix Figure 3.24, we break down the map by decade and uncover substantial variation in the origin of US emigrants over time.

3.3.2 Patent Data

We measure innovation activity using patents, as is standard in the literature (Griliches, 1998).[22] Patents for the United States have been digitized from original documents by Berkes (2018). The data contain, among others, information on the authors' addresses, the filing date, and the CPC patent classification. We use these to construct a balanced panel dataset at the county-technology class-year level.[23]

Patents for the United Kingdom for the period 1895-1939 are collected from PATSTAT, which in turn provides bulk access to data stored at the European Patent Office. These data contain information on authors and CPC classes but do not report the geographic location of inventors. To retrieve the coordinates of the inventors, we merge them with data by Bergeaud and Verluise (2022) and map them to registration districts at their 1890 borders. Patent data for previous years, unfortunately, are not currently available. To tackle this data limitation, we digitize the universe of patents granted in England and Wales between 1853 and 1895. As a result, we assemble a unique patent-level database that leverages textual information from approximately 300,000 original patent documents.[24] We have information on the title, text, inventors' geo-references addresses, filing and issue date, and other variables not used in this paper. Next, we map patents to districts at 1890

[22]Previous research shows that patents are not a flawless measure of innovation because non-patented innovation represents a non-negligible share of overall technological progress (Moser, 2019). We nonetheless believe that this is a comparatively minor issue for our analysis. As discussed in section 3.2.2, before our study period, the US and the UK had enacted important reforms that decreased the cost of access to patent protection (Gomme, 1948). These drastically increased the number of patents in both countries, thus ensuring that patents convey an informative picture of the state of technology in both countries.

[23]We map patents to counties at 1900 borders using the inventors' coordinates. From the three-digit CPC class, we map patents to a coarser taxonomy of twenty sectors. Appendix 3.10.1 provides additional details.

[24]Appendix section 3.11.1 describes the primary sources and methodology we develop to extract and structure the data from the original documents. In section 3.11.2, we compare our series with two existing series and find that the three are highly consistent for the period of common support.

borders. We then employ a simple machine learning classification algorithm, discussed in Appendix section 3.11.1, to assign technology classes using information contained in the titles.

This newly developed dataset is the first with geographical and textual information on the universe of patents granted in England and Wales during the second half of the nineteenth century. Data by Hanlon (2016), for instance, do not list titles or texts and do not report geographic information. This dataset thus expands previous work by Nuvolari and Tartari (2011) and Nuvolari, Tartari and Tranchero (2021) and provides the first comprehensive assessment of innovation in Britain during the Second Industrial Revolution.

In some empirical applications, we link patent data to the census. This allows us to assign a unique, consistent identifier to single inventors appearing in multiple patents and to observe individual-level characteristics recorded in the census. To perform this linking, we match inventors based on the string similarity between their name and surname and those recorded in the census, conditional on geographic proximity. We describe the precise implementation in Appendix section 3.10.3.

3.3.3 Other Variables

In this section, we provide a brief description of the additional data that we assemble. Appendix section 3.10.1 discusses each more diffusely.

UK Census Data

We assemble district-level statistics from population censuses at a decade frequency between 1851 and 1911. Districts are the level of observation in most of the analysis. This is because they were statistical units with neither budgetary nor administrative authority. The average population was 40,000, which makes them roughly comparable to US counties. Districts undergo minor boundary changes during the analysis period. However, to ensure geographical consistency, we cross-walk all variables to districts in 1890 using

the method described in Eckert, Gvirtz, Liang and Peters (2020). In particular, the census allows reconstructing the employment shares across sectors and other demographic information.

Newspapers

We use newspaper coverage of US-related topics as a measure of attention to the United States in public opinion. We collect the data from the British Newspaper Archive. Beach and Hanlon (2022) discuss this dataset in detail. We run three sets of queries. First, we search for the joint mention of the words "United States"; second, we search for mentions of each US state; third, we search for mentions of each US county, jointly with either the state name or "United States". We collect these data at the newspaper level from 1850–1939. Additionally, we know each newspaper's publishing address, which we geo-reference to 1890-border districts. Ultimately, we assemble three datasets at the district, district-state, and district-county levels, each at decade frequency. Figure 3.10 reports the distribution of newspapers.

Telegraph Network

We reconstruct the English and Welsh telegraph network from *Zeitschrift des Deutsch-Österreichischen Telegraphen-Vereins, Jahrgang*, volume IX, 1862. This directory lists all telegraph stations outside of London in 1862. To the best of our knowledge, it is the most comprehensive list before the establishment of the transatlantic telegraph cable connecting the UK and the US (1866). We geo-reference all the stations and assign them to 1890-border districts. Since, however, the source does not list stations in the London area, in the sample of the telegraph analysis, we conflate London urban districts into a single "London" unit, which we assume to be connected to the telegraph network. Figure 3.11 reports the distribution of the stations.

3.4 Empirical Strategy

This section describes our baseline empirical strategy. We discuss the potential caveats that hinder a causal interpretation of the resulting estimates. Then, we discuss two strategies to address these concerns.

3.4.1 Baseline Methodology

The central hypothesis of this paper is that exposure to foreign—in this case, American—knowledge through migrant linkages shapes the direction of innovation of the country of origin of the emigrants. We thus develop a simple measure of exposure to US knowledge that leverages two sources of variation. First, local specialization across counties measures the knowledge that diffuses from those counties. Second, the number of migrants that leave a given district and settle in a given county measures the intensity of the return knowledge channel. To fix ideas, consider two districts, and call them A and B. The same number of emigrants n leaves each district. Emigrants from A settle in county a, which only produces innovation in sector s_a. Emigrants from B settle in county b, which only innovates in sector s_b. Then, we expect district A (resp. B) to innovate comparatively more in sector s_a (resp. s_b).

To implement this intuition, we define knowledge exposure as follows:

$$\text{Knowledge Exposure}_{ik,t} \equiv \sum_{j \in J} \left(\frac{\text{Patents}_{jk,t}}{\text{Patents}_{j,t}} \times \text{Emigrants}_{i \to j,t} \right) \tag{3.1}$$

where i, j, k, and t denote a (UK) district, a (US) county, a technology class, and a decade, respectively.[25] The set J denotes the universe of counties. The knowledge exposure term thus averages district-level exposure to county-level specialization across technology classes. The first term in the summation captures specialization, while the second term codes district-level exposure. One may argue, however, that the relative share

[25] Throughout the paper, we refer to decade t to mean the ten years before the upper bound t. Hence, the decade indexed by 1890 refers to 1881–1890.

of patents may inflate the influence of specialization in counties with a small number of granted patents. While this is unlikely to significantly bias our results as those countries would likely have low district-level exposure, we code an alternative knowledge exposure variable that measures specialization as the raw count of patents in a given technology class. One further challenge to measure (3.1) is that districts with larger bilateral linkages are probably larger and, hence, selected. To account for district-level time-varying confounding variables, we control non-parametrically for district-by-time fixed effects. However, we also report results for an alternative knowledge exposure that measures exposure through relative emigrant shares. We discuss these alternative definitions in more detail in the Appendix table 3.25.

We estimate variants of the following regression model:

$$\text{Patents}_{ik,t} = \alpha_{i\times t} + \alpha_{i\times k} + \beta \times \text{Knowledge Exposure}_{ik,t} + \varepsilon_{ik,t} \tag{3.2}$$

where the coefficient of interest (β) quantifies the correlation between innovation activity and exposure to foreign knowledge. The term $\alpha_{i\times t}$ denotes district-by-decade fixed effects whose inclusion allows to control non-parametrically for time-varying unobserved heterogeneity at the district level; the term $\alpha_{i\times k}$ denotes district-by-technology fixed effects and excludes variation arising, for example, from the possibility that district-level technology specialization and immigration location decisions may be correlated. We comment more on this second point in the next section. The error term is the $\varepsilon_{ik,t}$. Standard errors in this specification are clustered at the district level. We mainly estimate model (3.2) through ordinary least squares. Since the dependent variable presents a non-negligible share of zeros, we also report the estimates of the Poisson regression associated with the baseline model.[26]

[26]In the innovation literature, it is common practice to apply a log transformation to the dependent variable. We do not follow this practice because Chen and Roth (2022) show that average treatment effects for transformations of the dependent variable defined in zero are arbitrarily scale-dependent. In Appendix section 3.13.4, we present alternative specifications with multiple transformations of the dependent variable.

3.4.2 Threats to Identification

The main factor that cautions against a causal interpretation of the estimates of model (3.2) is assortative matching, meaning that there may be a—possibly unobserved—variable that correlates with the location where emigrants settle in the United States and the composition of patenting activity across technology classes.

In section 3.2.1, we discussed that the historical and quantitative evidence suggests that, over time, emigrants originated from increasingly affluent and urbanized areas. Suppose emigrants also settled in comparatively more urban and affluent counties in the United States, and there was a correlation between patenting activity in specific fields and economic growth. In that case, the selection issue may bias the OLS estimates upward. We note that the bias arises only if (i) the correlation between patenting and the underlying confounding variable is heterogeneous across technology classes and (ii) the correlation is the same in the US and the UK. If (i) does not hold, then the omitted confounding variable would be absorbed by district-by-time fixed effects. If (ii) does not hold, the selection bias would be working against our result.

Assortative matching also arises if pre-existing differences in specialization across technology classes predicted the counties where emigrants chose to settle. For example, suppose that emigrants from a largely textile area, say Lancashire, were comparatively more likely to settle in counties with larger textile sectors. Then, the estimated β of model (3.2) would reflect pre-existing innovation similarities between sending and settling areas rather than capture the effect of return innovation. Evidence by Hanlon (2018) and Ottinger (2020), among others, suggest that non-random location decisions may represent a severe threat in this context. We attempt to quantify this issue in Appendix section 3.13.4. We measure the similarity of innovation portfolios between districts and counties and check whether this measure of specialization proximity correlates with observed bilateral migration flows. Table 3.15 reports the results. We find no significant association between innovation similarity and migration choices. This suggests that assortative matching is a plausibly minor concern for our analysis. Moreover, in the baseline estimation equation (3.2), we include district-by-technology fixed effects. Hence, for assortative

matching to bias our estimates, the underlying confounding variable would need to vary over time across district-technology pairs.

While we present evidence against the presence of assortative matching, we ultimately cannot rule it out. We thus develop two strategies that, we argue, ameliorate residual endogeneity concerns.

3.4.3 Shift-Share Instrumental Variable Strategy

We design a shift-share instrument that leverages recent advancements in the econometric literature to deal with selection and assortative matching. Identification critically hinges on the observation that instrument validity can be obtained from the quasi-random assignment of shocks (Borusyak, Hull and Jaravel, 2022). We construct county-specific immigration shocks by interacting aggregate immigration flows in the US with the gradual expansion of the railway network along the lines of Sequeira, Nunn and Qian (2020). These generate exogenous shocks to county-level immigration in a quasi-experimental shift-share design à la Borusyak, Hull and Jaravel (2022).

To construct the shocks, we predict the county-level immigrant share, which is not specific to British immigrants, from a regression between the actual immigrant shares and an interaction between the timing of connection to the railway network and the aggregate inflow of immigrants. Importantly, we control for county-level unobserved time-invariant heterogeneity and several other potential confounding variables at the county level.[27] In our context, shocks are conditionally exogenous if the settlement decisions of British immigrants did not influence the direction of the enlargement of the US railway network. In other words, instrument validity requires that shocks randomly assign British emigrants across counties. Under this assumption, the instrument breaks concerns of assortative matching. This may fail if, for instance, immigrants settled in counties more similar to their area of origin among the counties connected to the network in a given period. Since county-level shocks yield the overall predicted immigrant shares—and not those of

[27] In Appendix section 3.14.2, we describe in more detail the practical computation of the immigration shocks.

the British only—we believe this is a relatively minor concern to rule out by assumption. Following Borusyak, Hull and Jaravel (2022), we show that shocks are uncorrelated with county-level confounding variables and that the instrument does not systematically predict district-level characteristics. Appendix Figure 3.30 shows that while immigrant shares correlate with district-level observable characteristics (Panel A), predicted immigration shares do not (Panel B). Similarly, in Appendix Figure 3.31, we confirm that while out-migration correlated with most district variables, the instrument displays smaller and insignificant correlations with the same variables. These exercises provide evidence in favor of the validity of our research design.

Let $\omega_{j,t}$ be the immigrant share in county j in decade t, and let $\hat{\omega}_{j,t}$ be its prediction. We thus define the instrument as

$$\widehat{\text{Emigrants}}_{i \to j,t} \equiv \hat{\omega}_{j,t} \times \sum_{j \in J} \left(\hat{\omega}_{j,t} \times \text{Emigrants}_{i \to j,1880} \right) \tag{3.3}$$

where $\text{Emigrants}_{i \to j,1880}$ denotes the number of emigrants leaving district i and settling in county j at the beginning of the sample period. Importantly, this exposure term is allowed to be endogenous by design. Identification stems from the quasi-exogeneity of the shocks $\{\hat{\omega}_{j,t}\}$. Given a predicted set of bilateral flows, we construct the instrument for knowledge exposure as in (3.1), except that the predicted flows replace the observed ones.

Even though we present evidence suggesting the opposite, the conditional exogeneity of the timing of railway connection is ultimately an untestable assumption. To validate the results obtained with the instrument (3.3), we construct an additional series of county-level shocks $\{\hat{\omega}_{j,t}\}$ that leverages a different source of variation. Specifically, we compute "leave-out" predicted county-level immigrant shares by interacting start-of-period immigrant shares with aggregate inflows by nationality. Importantly, we exclude British immigrants when calculating these shocks. This ensures that the "leave-out" shares do not reflect the settling decisions of the British. We describe the procedure in more detail in Appendix 3.14.2. This alternative instrument yields results that are highly consistent with the railway-based approach.

3.4.4 Shock Propagation Difference-in-Differences Strategy

The shift-share instrumental variable relies on identifying variation across counties that become connected to the US railway network. Therefore, the associated estimates deliver a local average treatment effect for a complying group of individuals who settle in counties that become connected to the railway network during this period. The literature suggests, however, that it is plausible that these "frontier" migrants would display a relatively higher probability to undertake innovation activity, perhaps due to entrepreneurial attitudes (Bazzi, Fiszbein and Gebresilasse, 2020). Under this interpretation, the IV estimates would yield an upper bound to the effect of overall out-migration on British innovation. In addition, because they rely on the specific group of counties that become connected to the railway network, they do not reflect the overall composition of US innovation across technology classes.

To provide additional causal evidence and circumvent these limitations, we devise a research design that leverages geographically clustered innovation shocks in the United States in a triple-differences setting. We start by observing a logical corollary of the return innovation argument. Suppose we observe a sudden increase in the number of patents granted in some counties in some technology classes. Then, one would expect that districts whose emigrants had settled more extensively in those counties would display increased innovation activity in those classes. In other words, innovation shocks in the United States should "reverberate" in the United Kingdom through pre-existing migration linkages.

We test this prediction using two sets of innovation shocks. First, as we describe in more detail in Appendix 3.14.3, we construct a set of county-technology class synthetic innovation shocks at yearly frequency. The intuition behind these shocks is that we seek to isolate periods of unusual patenting in a given county-technology class-year, controlling for the average volume of patents produced in that county-class cell. We thus regress the number of patents against fixed effects to obtain the residualized innovation activity. Then, we flag an innovation shock $\xi_{jk,t}$ whenever the residualized number of patents in a given county j, technology class k, and year t is in the top 0.1% of the overall

distribution.[28] Appendix Table 3.21 documents that shocks are relevant, as one such shock is associated with an average of more thirty patents in the given county. Second, we leverage recent evidence by Berkes, Coluccia, Dossi and Squicciarini (2023), who document that the Great Influenza pandemic (1918–1919) significantly and positively affected pharmaceutical innovation in counties that were more exposed to the pandemic. We thus claim that districts that were comparatively more exposed to affected counties should feature increased pharmaceutical innovation. We provide additional details on the construction of county-level exposure to the pandemic in Appendix 3.14.3.[29] We code county-level exposure to the pandemic as a dummy φ_j that returns value one if the ratio between deaths during the pandemic (1918–1919) and deaths in the preceding three years (1915–1917) is in the top 25%, and zero otherwise.

We measure district-level exposure to the county-level shocks in terms of the emigrants that had left the given district to settle in the given county *before* the period of analysis.[30] Formally, we compute exposure to synthetic shocks in technology class k as

$$\text{Synthetic Shock Emigrants}_{ik,t} = \sum_{j \in J} \left(\text{Emigrants}_{i \to j,1900} \times \xi_{jk,t} \right) \tag{3.4}$$

and analogously, we define exposure to counties affected by the pandemic as

$$\text{Influenza Emigrants}_{i} = \sum_{j \in J} \left(\text{Emigrants}_{i \to j,1900} \times \varphi_{j} \right) \tag{3.5}$$

To avoid issues of continuous treatment described by Callaway, Goodman-Bacon and Sant'Anna (2021), we recast each exposure metric in terms of a dummy variable that returns value one if the associated continuous measure is in the top 25%, and zero otherwise.[31]

[28] In Appendix Table 3.34, we show that the results remain consistent when imposing different values to flag innovation shocks.

[29] Since the technology taxonomy used in this paper is different from Berkes, Coluccia, Dossi and Squicciarini (2023), in Appendix Table 3.20 we confirm that their result holds in our data. Figure 3.26 reports the associated flexible triple differences specification. Moreover, in Figure 3.33(a), we confirm that the pandemic affected innovation activity only in the pharmaceutical sector.

[30] This part of the analysis restricts the outcome variable to 1900–1930, so we can leverage migrant flows in the preceding decade (1890–1899) to construct fixed exposure shares.

[31] In Appendix Table 3.34 we consider alternative thresholds to code the exposure variable (3.4). In Appendix Table 3.35, we report the results using the continuous measure (3.5).

To estimate the effect of US synthetic shocks on UK innovation activity, we estimate the following triple differences specification:

$$\text{Patents}_{ik,t} = \alpha_{i\times k} + \alpha_{k\times t} + \alpha_{i\times t} + \sum_{h=-a}^{b} \beta^h \times \text{I}\left[D_{ik,t} = h\right] + \varepsilon_{ik,t} \tag{3.6}$$

where $\alpha_{i\times k}$, $\alpha_{k\times t}$, and $\alpha_{i\times t}$ denote, respectively, district-by-technology class, technology class-by-year, and district-by-year fixed effects.[32] The term

$$\left(D_{ik,t} \equiv t - \text{I}\left[\text{Synthetic Shock Emigrants}_{ik,t}\right]\right)$$

denotes the number of years since the district-technology class ik was exposed to a synthetic innovation shock ξ. The roll-out of the treatment is staggered across units. Different district-class pairs may be exposed to the exposure treatment at different points in time.[33] Goodman-Bacon (2021) shows that the standard two-way fixed effects estimator shown in (3.6) fails to estimate the average treatment effect when treatment effects are heterogeneous, either over time or across groups. Several estimators have been proposed to deal with this difficulty. In the main results, we report estimates obtained using the imputation procedure presented in Borusyak, Jaravel and Spiess (2021). Other estimators yield qualitatively similar results, as shown in Appendix Figure 3.34.

We follow a similar approach to estimate the effect of US exposure to the Great Influenza pandemic on UK innovation. In particular, the model is entirely similar to (3.6), except that the treatment variable is defined as $(D_{ik,t} \equiv t - I\left[\text{Influenza Emigrants}_i\right])$ as it codes the number of years since the influenza, and it is interacted with a dummy variable returning value one for the pharmaceutical technology class, and zero otherwise.[34]

[32] When we estimate regression (3.6) using variation in exposure to the pandemic shock, we normalize the dependent variable by the average number of patents granted before the pandemic to ensure that the estimated coefficients' size are comparable.

[33] Notice that the treatment is also potentially repeated, for the same unit can be treated multiple times. This is, however, not the case in the baseline case, where we define synthetic shocks in the top 0.1% of the overall residualized innovation shock distribution.

[34] This specification focuses on the ATE on pharmaceuticals compared to other technology classes. In Appendix Table 3.35, we report the double differences estimates associated with model (3.6). Then, in Figure 3.33(b), we show that, as in the United States, the influenza had a major effect on pharmaceutical innovation only.

The primary estimation strategy in this setting is thus a triple difference estimator (Olden and Møen, 2022). A causal interpretation of the resulting estimates requires that the difference between the within-group differences are not statistically different from zero before the treatment. Several papers highlight that, compared to the standard difference-in-differences estimator, the parallel trends assumption in this setting is relatively weak because it only requires that no contemporaneous shock affects the relative outcome of the treatment and the control group (Gruber, 1994). Throughout the paper, we present flexible triple difference estimates to provide evidence supporting the parallel trends assumption.

3.5 Empirical Results

In this section, we present the main return innovation result. Then, we document that shocks to US innovation diffuse into the UK through migration ties. We interpret these results as evidence that migration flows contribute to the diffusion of innovative knowledge to countries sending migrants.

3.5.1 Exposure to US Innovation Shapes Innovation in the UK

The primary finding of this paper is that exposure to foreign technology through migration ties shapes the dynamics and direction of innovation in the emigrants' country of origin.[35] We label this novel finding "return innovation". We first estimate regression (3.2) through a simple OLS linear probability model to document it. We report the results in columns (1–3) of panel A of Table 3.2. There is a positive, significant, and quantitatively large correlation between the baseline measure of exposure to foreign knowledge and the number of patents at the district-technology class level. Moreover, the correlation persists over time, as the estimates remain statistically significant after

[35] A recent literature produced compelling evidence that exposure to innovation is a key determinant of subsequent innovation activity (Akcigit, Caicedo, Miguelez, Stantcheva and Sterzi; Bell, Chetty, Jaravel, Petkova and Van Reenen, 2018; 2019). Our results can thus be interpreted as additional new evidence in favor of this thesis.

two decades. In columns (1–3) of panel B we repeat this exercise, but we normalize the number of patents by the district-level population at the beginning of the sample (1880). We confirm the positive association between knowledge exposure and per-capita patents. Importantly, all regressions include district-by-decade fixed effects to account for unobserved time-varying heterogeneity at the district level. Moreover, we control for district-by-technology class fixed effects to partial out spurious variation arising from initial district-level specialization across classes.

As discussed in section 3.4.2, at least two factors hinder a causal interpretation of the estimates presented in panel A. First, out-migration is not random across districts. Second, there may be some latent determinant of the settlement location decisions of emigrants that correlates with innovation activity in their origin areas. To ensure that our estimates do not reflect spurious correlation arising from omitted variable bias issues, we estimate model (3.2) using the instrument (3.36). In columns (4–6) of panels A and B, we report the reduced-form association between the instrument and the dependent variable. Figure 3.3 visually compares the OLS and the reduced-form IV regressions. We confirm the positive and statistically significant effect of knowledge exposure on innovation. The effect persists until one decade, as opposed to two from the OLS estimates.[36] Columns (7–9) report the two-stage least-squares (TSLS) estimation results. First, the instrument is relevant.[37] Second, the TSLS estimates confirm knowledge exposure's positive, large, and statistically significant effect on innovation. The magnitude of the TSLS estimates is roughly similar to the OLS, although the latter appears to be slightly upward biased. The OLS estimates possibly reflect the upward bias introduced by assortative matching across district-county pairs.

The evidence in Table 3.2 is at the district-technology level. To explore the heterogeneity of the return innovation effect across industries, however, we estimate model (3.2) at the

[36]Bilateral migration flows are known to be highly persistent over time—a phenomenon known as "chain migration". This would inflate the OLS association between lagged knowledge exposure and innovation. At the same time, this would explain why the associated TSLS estimates are not significant. The instrument, in fact, effectively breaks the persistence of migratory flows using plausibly exogenous county-level immigration shocks.

[37]We report the complete first-stage estimates in Appendix Table 3.30. The instruments are always relevant and capture a substantial share of the variation of the endogenous variables.

district level separately for each technology class. We report the resulting reduced-form coefficients of the knowledge exposure instrument—one for each regression—in Figure 3.4. We estimate the largest treatment effects for industries like electricity and chemistry that were at the forefront of the Second Industrial Revolution (Mokyr, 1998). We employ the UK-revealed comparative advantage to measure the relative sector-level innovation specialization.[38] We find that the return innovation effect is larger in sectors where the UK retained an advantage at the beginning of the period (the 1880s). Rather than igniting the emergence of entirely new sectors, our results suggest that exposure to US knowledge through migration ties nurtured already-existing industries.

The setting of this study allows for gauging the persistence of the association between exposure to foreign knowledge and innovation.[39] In Appendix Figure B27, we report the coefficients of a regression between the number of patents and an interaction term between knowledge exposure in the period 1900–1930 and biennial time dummies from 1940 to 2014. The estimates suggest that the positive effect of knowledge exposure on innovation persists for almost four decades, albeit the magnitude decreases over time. Starting in the mid-1970s, the association gradually becomes small and statistically insignificant. In Appendix Table 3.22, we repeat the exercise by technology class and find consistent results across sectors. Migration ties thus generate enduring knowledge flows that shape innovation activity over the long run.

The analysis presented thus far focuses on how out-migration shaped the *direction* of innovation.[40] A natural question is, however, whether it also impacted the *volume* of patents. Our data are not especially well-suited to answer this question because we lack disaggregated data on outright emigration. Nevertheless, if emigration to countries other

[38] In the international trade literature, the revealed comparative advantage is a widely-employed metric that hinges on the observation that a country's comparative advantage is revealed by the country's relative exports (Balassa, 1965). In our setting, we define the revealed comparative advantage as

$$\text{RCA}_{ik} \equiv \frac{\text{Patents}_{ik} / \sum_{k' \in \mathcal{K}} \text{Patents}_{ik'}}{\sum_{i' \in \mathcal{I}} \text{Patents}_{i'k} / \sum_{i' \in \mathcal{I}, k' \in \mathcal{K}} \text{Patents}_{i'k'}}$$

where i and k denote countries and sectors within sets $\mathcal{I}$ and $\mathcal{K}$. Specifically, $\mathcal{I} = \{\text{UK}, \text{US}\}$. Then, the UK is relatively more specialized in sectors with $\text{RCA}_{\text{UK},k}$ above one.

[39] We discuss the technical details of the long-run analysis in Appendix section 3.13.3.

[40] Appendix section 3.13.4 explores this aspect in more detail and provides the technical details of the analysis.

than the United States correlated with US emigration, we can present some suggestive evidence. In Table 3.19, we estimate the effect of out-migration on innovation, measured as the number of patents granted. The OLS and TSLS estimates show that out-migration has a negative short-term impact on innovation, but this reverses in the medium run (after one decade). Our findings thus appear to reconcile evidence of "brain drain," which views out-migration as a depletion of human capital that hampers innovation, with "brain gain" arguments suggesting that emigrants may be conducive to economic growth via, for instance, monetary remittances (Docquier and Rapoport, 2012). The results suggest that the former hypothesis is predominant in the short-run, while the brain-gain perspective materializes in the medium-to-long term. The effect of out-migration on the volume of innovation has been the focus of many of the existing studies (Agrawal, Kapur, McHale and Oettl; Andersson, Karadja and Prawitz, 2011; 2022). This paper, instead, provides evidence that emigration is a fundamental driver of the direction of innovation.[41] From this perspective, our results thus inform the recent literature studying the determinants of the direction of innovation (Bell, Chetty, Jaravel, Petkova and Van Reenen; Einiö, Feng and Jaravel, 2019; 2023).

We perform several robustness exercises to gauge the robustness of our results. These are reported in the Appendix and discussed in section 3.14.1. First, we consider alternative dependent variable transformations in Table 3.24. Second, Table 3.25 reports the results using five different definitions of knowledge exposure that hold fixed various margins of variation. The baseline specification of model (3.2) includes district-by-decade and technology class-by-decade fixed effects. In Table 3.26, we show that the results are robust to alternative, demanding specifications. The standard errors are clustered at the district level in the baseline specification. In Figure 3.28, we adopt various estimators and confirm that they all preserve the statistical significance of the main results. Another concern is that the return innovation effect concentrates on low-quality innovation and thus bears little relevance in terms of aggregate productivity growth. In Table 3.27 we thus report the results of the baseline and instrumental variable regressions, accounting

[41] Our results resonate with evidence by Fackler, Giesing and Laurentsyeva (2020). While their study essentially leverages cross-country variation in emigration destinations, our analysis is based on within-country disaggregated data on the origin and destination of migrants. This allows us to credibly estimate the causal effect of out-migration and investigate possible underlying mechanisms.

for patent "quality".[42] The results confirm that the number of high-quality patents increases in districts exposed to US knowledge. In fact, the magnitudes are larger than using the raw patent count. This may indicate that the return innovation effect is more intense for marginally more valuable patents. Analogously, in Table 3.28 we compute knowledge exposure weighting US patents by their quality using different thresholds and definitions. The results qualitatively confirm the baseline estimates. We do not have information on the actual adoption of technology by the firms. However, in Table 3.29, we restrict the outcome variable to include only patents that list at least one firm as an assignee.[43] These plausibly reflect actual economic activity carried out by British firms. Here, too, the results remain qualitatively similar to the baseline estimates. The instrument used in Table 3.2 leverages variation in the connection timing to the railway network to randomize immigration across counties. In Table 3.31, we report the results using an alternative "leave-out" instrument, described in section 3.14.2. Importantly, we can also use both instruments simultaneously and provide over-identification tests. In Table 3.32, we confirm that the leave-out instrument results are robust to various alternative definitions of the county-level shocks.

3.5.2 Innovation Shocks in the US Diffuse to the UK

The return innovation result indicates that migration ties shape the direction of innovation in the origin areas of emigrants. We claim that this finding implies that fluctuations in patenting activity in the United States would reverberate in the United Kingdom through migration linkages. We estimate model (3.6) using two different sources of such fluctuations—which we label innovation shocks—to test this hypothesis.

Table 3.3 reports the results of this exercise. Columns (1–4) refer to the synthetic shocks series we construct by residualizing the observed patenting activity against fixed effects

[42]Following Kelly, Papanikolaou, Seru and Taddy (2021), we define a text-based measure of quality which flags as influential those patents that introduce words that did not appear before they were published, and become used thereafter. Because we have full texts for the period 1880–1899 and abstracts only between 1900 and 1939, in this exercise we restrict the sample period to the latter years.

[43]Unfortunately, data on firm assignees is only available for the sub-sample 1870–1900, when we have the full text of the patent specifications form which this information is extracted.

and flagging large increases in the resulting series as "innovation shocks". As a preliminary robustness test, we report the full-sample estimate in column (1), while columns (2–4) exclude districts in the top three areas in terms of patents granted. We estimate a positive, large, and statistically significant effect of US synthetic innovation shocks on innovation activity in the UK. We estimate an average of 0.4 patents per year in the treated technology class after the shock in exposed districts. This is a quantitatively sizable effect since the average number of patents per district-class pair is 1.3. Moreover, the relative size of the effect remains consistent throughout the regression samples. Next, we explore heterogeneous treatment effects over time in Figure 3.5(a). Reassuringly, the figure provides evidence that supports the parallel trends assumption. The effect of the innovation shock is the largest and most significant after two years since the shock initially manifested in the United States. This time lag seems plausible, especially since our data shows an average of 1.1 years delay between the application and issue date at the UK patent office. The effect persists up until six years following the synthetic shock. We estimate the effect of synthetic shocks sector by sector in the appendix Figure 3.32. As in 3.4, we find the largest treatment effect for electricity.

Next, we investigate how exposure to the Great Influenza pandemic across US counties impacted UK innovation. The logic behind this exercise is that exposure to the pandemic fostered innovation in the pharmaceutical sector (Berkes, Coluccia, Dossi and Squicciarini, 2023). We thus expect districts whose emigrants had settled in counties more exposed to the pandemic to display higher patenting rates in pharmaceuticals. We report our findings in columns (5–8) of Table 3.3. We estimate the pandemic shock's effect on British innovation to be positive and sizable. On average, two patents per year are granted in the pharmaceutical sector in districts more exposed to counties severely affected by the influenza. We estimate the associated dynamic treatment effects in Figure 3.5(b). We find only one marginally significant and very small coefficient in the pre-treatment period. By comparison, the post-treatment coefficients are large and highly significant. The effect of the pandemic materialized six-seven years after the shock in the United States. As noted before, this delay is partly due to the shift between patent application and issue by the patent office, except that we now have to compound delays at the US and UK

offices. Moreover, the effect of the pandemic shock on US innovation in pharmaceuticals was not immediate, as shown in Appendix Figure 3.26. Taken together, it is plausible that the propagation of the innovation shock into the UK is observed with some delay. We estimate statistically significant treatment effect coefficients for more than a decade thereafter.

The pandemic shock only impacted innovation in pharmaceuticals in the US (Figure 3.33(a)). We thus expect to retrieve a similar effect in Britain. Figure 3.33(b) shows that, although the point estimates are not as sharp as in the US case, the pharmaceutical sector is the one that benefits the most from the influenza shock. The point estimate for pharmaceuticals is nearly three times larger than the second-largest estimate. The estimated effect in some sectors may be negative because of crowding-out out of those fields into pharmaceuticals, although we cannot entirely disentangle the underlying reason. We interpret this exercise as a falsification check: Figure 3.33 provides convincing evidence that the pandemic shock affected the same sector in the US and the UK.

We assess the robustness of these results through several robustness checks. First, we consider alternative thresholds to (i) flag synthetic shocks and (ii) flag district exposure to synthetic shocks. In Table 3.34, we estimate larger treatment effects for smaller thresholds. This is reasonable since smaller thresholds impute, on average, larger innovation shocks. The synthetic shock triple differences model is a staggered design since shocks generally occur in different periods across technology classes and districts. The baseline estimates are obtained from the imputation estimator developed by Borusyak, Jaravel and Spiess (2021). In Figure 3.34, the estimated treatment effect remains consistent across various estimators. In particular, the one developed by De Chaisemartin and D'Haultfœuille (2022) allows repeated treatments and yields consistent results. In Table 3.35, we report several specifications to gauge the robustness of the pandemic shock results. First, in columns (1–2), we report the double differences estimates that compare pharmaceutical innovation across districts by exposure to counties affected by the pandemic. Then, in columns (3–7), we report various triple differences specifications that exclude districts in areas with very high patenting activity. The results remain consistent throughout.

3.5.3 Technology Transfer and Spillovers: A Text-Based Approach

It is natural, at this point, to ask to what extent the return innovation effect manifests because the *same* patents that are granted in the United States are, at some point, issued in Britain. In other words, is the return innovation effect about UK areas with higher exposure to US knowledge "copying" US-developed innovation?

To quantify the extent of copying and, on the other hand, the spillovers in terms of "original" innovation that would be generated by migrant linkages, we develop two text-based similarity measures.[44] First, we compute the backward similarity between patents granted in the UK and previous patents granted in the US. Leveraging textual information contained in patent titles, this approach allows us to measure whether patents produced in areas with larger exposure to US knowledge become more similar to US inventions. Second, we compute a measure of patent "originality" by comparing patents granted in the UK to previous and subsequent US patents. Specifically, we deem a given patent as more innovative if it is more similar to subsequent US patents relative to previous ones. This approach mirrors the methodology of Kelly, Papanikolaou, Seru and Taddy (2021). Both indices are computed at the patent level and aggregated up at the district-technology class level at a decade frequency.

In Figure 3.6, we report the effect of exposure to US synthetic innovation shocks on the similarity between patents produced in the UK and those issued in the US. In Panel 3.6(a), we estimate a positive and significant effect of exposure to US knowledge on the backward similarity between UK and US patents. This suggests that, to some extent, knowledge flows generated by migration ties stimulate emulation and technology transfer between the United States and Britain. In Panel 3.6(b), however, we estimate the same model using the "originality" of UK patents as the dependent variable. Here, too, we find that areas with more intense exposure to foreign innovation produced more innovative patents (compared to previous US innovations). These results thus suggest that the

[44] In Appendix section 3.11.3, we present the analytical derivation of the similarity metrics that we abridge from Kelly, Papanikolaou, Seru and Taddy (2021).

return innovation effect conflates two distinct margins through which exposure to foreign knowledge affects the production of innovation. First, migrant linkages fostered the technology transfer of already-existing inventions in the UK. Second, they also propelled the development of novel, original inventions. Interestingly, we estimate that a synthetic shock triggers a sudden increase in the backward similarity of innovation, while the effect on original patents manifests more slowly.

In Appendix Table 3.37, we tabulate the associated estimates and report the results for the Great Influenza pandemic shock as well. We confirm that following an innovation shock in the US—either a synthetic one or the Great Influenza pandemic—the similarity between UK and US patents increases (Panel A). At the same time, areas more exposed to the shock start producing more original patents (Panel B).

In Appendix Table 3.36 we report the OLS and instrumental variable estimates of regression (3.2), which confirm the baseline results obtained using the triple differences estimator. We gauge the robustness of these results using three alternative measures of similarity, all displayed in Appendix Table 3.33. First, we use the "raw" similarity measure between titles, which does not take the log of the cosine similarity between patent titles (columns 1–3). Second, we net out the year and technology fixed effects from the patent-level originality and backward similarity measures to ensure that our estimates do not conflate time-varying terminology and fashion trends (columns 4–6). Finally, while in the baseline analysis, we compute the similarity metrics over a ten-year window around each patent, in columns (7–9), we restrict it to five years. The results remain qualitatively unchanged throughout.

3.6 Potential Mechanisms and Discussion

Several concurrent, not necessarily mutually exclusive mechanisms can explain the return innovation result. In this section, we present our analysis to disentangle some. First, we establish whether return innovation is solely a consequence of return migration. Then, we discuss some complementary and possibly quantitatively more substantial channels.

3.6.1 Is Return Innovation Return Migration?

Return migration is a primary candidate to explain our findings through two channels. First, return migrants may engage in innovation activities in the fields they were exposed to abroad. Second, return migrants may facilitate access to US knowledge without directly undertaking innovation activities. The literature does not offer conclusive evidence on the effect of return migration on innovation. On the one hand, several studies estimate modest effects for recruiting programs of high-skilled nationals working abroad (Ash, Cai, Draka and Liu; Shi, Liu and Wang, 2022; 2023). On the other, Giorcelli (2019) shows, although from a different perspective, that those exposed to (managerial) foreign knowledge change their behavior once back in their origin country.[45] In this section, we quantify the relative importance of return migration in generating return innovation.

The baseline linked sample of British emigrants traces them back to the UK census before they migrated. To measure return migration, we instead link them to UK censuses completed after they had migrated to the US. Then, we aggregate return migration flows at the district-by-county level and at decade frequency and compute a measure of "return knowledge exposure" which is analogous to (3.1):

$$\text{Return Knowledge Exposure}_{ik,t} \equiv \sum_{j \in J} \left(\frac{\text{Patents}_{jk,t}}{\text{Patents}_{j,t}} \times \text{Return Migrants}_{j \to i,t} \right) \tag{3.7}$$

where $\text{Return Migrants}_{j \to i,t}$ is the number of migrants that return from county j to district i in decade t. Because UK censuses are available only until 1911, return migration data span the period 1870–1910.

As a first step, we estimate model (3.2) controlling for return knowledge exposure. Table 3.4 reports the results: in columns (1–3), we present specifications with various levels of fixed effects; columns (4) and (5) display the coefficients of lagged values of the independent variables; in column (6) we report the full lag model. Throughout the specifications, the coefficients of baseline and return knowledge exposure remain comparable in size—

[45] Choudhury (2016) shows that R&D firms with returnee managers are disproportionately more likely to file patents in the United States. Bahar, Hauptmann, Özgüzel and Rapoport (2022) show that return migration can influence trade.

by looking at the respective standardized beta coefficients—and statistically significant. These results suggest that return migration is an important driver of return innovation. In our data, approximately 30% emigrants return, and these account for approximately 50% of the total return innovation effect. At the same time, however, a substantial proportion of return innovation is not explained by return migration.

Absent physical return, in the rest of the paper, we provide evidence of two additional mechanisms that underlie the return innovation effect. First, we focus on interactions between the emigrants and local communities in Britain. Second, we explore how migration ties facilitate cross-country market integration, thereby promoting knowledge flows.

3.6.2 Return Innovation Through Interactions

In this section, we explore if and, in case, how emigrants interact with local communities who remained in Britain. We distinguish between two cases. On the one hand, interactions could require physical return. On the other, emigrants may exchange information while abroad. When studying the interactions between emigrants and stayers, one needs to delimit the set of stayers with whom emigrants could plausibly interact. We focus on two factors that could promote social interactions between the emigrants and their communities of origin: family ties and pre-migration geographical proximity (neighbors).

Interactions Between the Emigrants and their Family

Our data do not contain exhaustive information on the families of emigrants. At best, we know those living in the same household. This would be an exceedingly restrictive definition because, in most cases, it would exclude brothers or parents. We thus adopt a less conservative approach. In particular, we assume that it is likely that individuals with the same surname who live in geographical proximity—in the same UK county—are relatives. This assumption is reasonable as long as surnames are not too common: in this analysis, we thus the top 5% most common surnames. Results are robust to alternative sample cuts.

We implement a triple-differences estimation. The outcome variable is the total number of patents granted to inventors with a given surname who are recorded living in a given county in the UK. The treatment leverages variation in the surname of US emigrants by county. Under the previous assumption, this model quantifies how the emigration of family members impacts the patenting activity of those who remain in the UK. Formally, it is

$$\text{Patents}_{sc,t} = \alpha_{s\times c} + \alpha_{c\times t} + \alpha_{s\times t} + \beta \times \text{US Emigrant}_{sc,t} + \varepsilon_{sc,t} \tag{3.8}$$

where s, c, and t denote, respectively, surnames, counties, and years. The treatment ($\text{US Emigrant}_{sc,t}$) is a variable equal to one after an individual with surname s from county c emigrates to the US, and zero otherwise. Therefore, under the standard parallel trends assumption, β estimates the impact of emigration on patenting activity carried by the relatives of the emigrant. To deal with the sharp left skewness of the outcome variable, for each estimate, we report an analogous model that features as the outcome variable a categorical indicator that returns a value of one if the number of patents is strictly positive and zero otherwise.[46]

Columns (1) and (5) of Table 3.5 report the baseline estimates. Emigration to the US has a positive effect on patenting activity by the relatives of emigrants who remain in the UK. The effect is quantitatively sizable. In Appendix figure 3.35 we report the associated flexible triple differences model. The estimates provide evidence in support of the parallel trends assumption. Moreover, they show that it takes, on average, ten years before a British emigrant to the US contributes to the innovation activity of his family in the UK. This delayed effect is plausible inasmuch as it would take time for emigrants to settle in the US and be exposed to technology that they could diffuse back into the UK.

In the second part of the analysis, we distinguish between emigrants who return from those who do not. Emigrants could interact with their families upon returning, but they could also maintain ties while abroad. In columns (2) and (6), we thus restrict to emigrants that, at some point, return to the UK. In columns (3) and (7), by contrast, the treatment includes only those that never return. The impact of return emigrants is

[46]The results remain qualitatively unchanged estimating model (3.8) as a Poisson regression.

four to five times as large as that of those who never return. This difference suggests that return migration is, as we argued in the previous section, a major driver of return innovation. At the same time, however, emigrants promote innovation by their relatives even if they never return. When we compare the two coefficients in columns (4) and (8), we confirm that both return and non-return emigration contribute to innovation in the UK. The relative magnitudes remain unaltered. Because return migration only accounts for approximately 30% of the overall emigration, this analysis suggests that, in quantitative terms, interactions between stayers and emigrants who never return to the UK account for approximately half of the overall return innovation effect.

Interactions Between the Emigrants and their Neighborhood

In this section, we interpret geographical proximity between emigrants and stayers as an alternative proxy for local social networks. The rationale is that, before leaving the UK, emigrants plausibly maintained social ties to those who lived in their neighborhood. We thus hypothesize that stayers could interact with their former neighbors who migrated to the US. Here, too, we further distinguish emigrants who never return to quantify the relative importance of physical return migration on interactions with origin communities.

We leverage the granular nature of our data and perform an individual-level analysis. First, we extract all men aged between 18 and 50 in 1900 that do *not* emigrate from the 1911 census. We then create a yearly balanced panel dataset that reports the number of patents obtained by each individual between 1900 and 1920. To do so, we leverage the linked inventor-census data described in Appendix 3.10.3. Next, each individual is geo-referenced to precise coordinates as described in Appendix 3.10.2. We complement this with information on the geographical proximity between these "stayers" and migrants. More specifically, we define a dummy variable (Neighborhood $\text{Migrant}^{k}_{p,t}$) that returns value one in all periods after the first time at least one individual living within k meters from individual p migrates to the US, and zero otherwise. In the baseline analysis, we consider $k = 0$, meaning that we only consider emigrants in the same street as the observed individuals, and recast (Neighborhood $\text{Migrant}^{100}_{p,t}$) as simply (Neighborhood $\text{Migrant}_{p,t}$)

for brevity. We label this variable an indicator of "neighborhood migration". To estimate the effect of neighborhood migration on the probability of patenting, we thus estimate the following double difference regression:[47]

$$\text{Patents}_{p,t} = \alpha_p + \alpha_t + \beta \times \text{Neighborhood Migrant}_{p,t} + \varepsilon_{p,t} \tag{3.9}$$

where p and t denote, respectively, individuals and years, and α_p and α_t are the associated fixed effects. The term β yields, under a standard parallel trends assumption, the estimated causal effect of neighborhood migration on innovation.

The logic beneath equation (3.9) builds on Bell, Chetty, Jaravel, Petkova and Van Reenen (2019), who document the importance of geographical proximity to inventors as a driver of subsequent innovation activity. A positive and significant estimate of β would be evidence that emigrants promote the innovation activity of their neighbors. Then, we define a (Non-Return Neighborhood Migrant$^k_{pt}$) dummy entirely analogous to the previous treatment, except that we condition the neighborhood emigrant to not return to the UK. In this case, a positive estimate of β would suggest that neighbors benefit from interactions with the emigrants even if those never return.

We report the estimates of equation (3.9) in Table 3.6. The dependent variable is the yearly number of patents. In columns (1–4), the sample includes individuals from all districts; in columns (5–7), we exclude individuals in the top three-producing patents areas (London, Lancashire, and the South-West). In panel A, the treatment is activated by any US neighborhood emigrant. In panel B, we restrict to neighborhood emigrants that never return in the sample period. We estimate a positive effect of neighborhood emigration on innovation by non-migrants. The effects hold in the baseline specification (columns 1 and 5), as well as including parish-by-time fixed effects (columns 2 and 6)

[47]To avoid an excessive computational burden, we estimate model (3.9) on a 10% random sample of the population. Moreover, the model is a staggered difference-in-differences design with (potentially) repeated treatments. We thus estimate regression (3.9) using the estimator proposed by Borusyak, Jaravel and Spiess (2021). In Appendix Figure 3.36 we show that the estimated coefficient remains stable using several different staggered difference-in-differences estimators. In Appendix Figure 3.38, we show that results hold if the neighborhood-migrant treatment is activated whenever emigrants within 100 meters from the individual in the sample migrate.

and applying coarsened exact matching (CEM, columns 3 and 4).[48] Importantly, the estimated coefficient remains if we restrict the sample to exclude all non-inventors, thus reducing the sample size considerably (columns 4 and 8). Panels A and B show that overall and non-return neighborhood migration has a positive statistically significant effect on the probability of inventing regardless of the dependent variable, the fixed effects, and the matching scheme. In Figure 3.37, we report the associated flexible difference-in-differences estimates, which indicate the absence of statistically significant pre-trends. In Appendix Table 3.23 we explore the heterogeneous response to neighborhood out-migration across the age and occupation of the stayer individual. In particular, we find that the gains are larger for relatively young individuals (column 1) and accrue to those employed in skilled occupations (columns 2–4).

Evidence presented in Table 3.6 provides additional evidence that emigrants promote innovation in the communities they come from. Those who never migrate but who were in close geographical proximity with the emigrants before they left, a proxy for local social networks, benefit from interactions with the emigrants. This channel operates even if the former neighbor never returns. These results highlight the importance of cross-country interactions between the emigrant population and their origin communities. Our findings confirm experimental evidence from developing countries linking technology diffusion with network interactions (Bandiera and Rasul; Conley and Udry; Beaman, BenYishay, Magruder and Mobarak, 2006; 2010; 2021). In the rest of the paper, we investigate more aggregate effects of migration ties on innovation activity.

3.6.3 Return Innovation Through Economic Integration

In this section, we explore whether migration ties promote the cross-border diffusion of innovation irrespective of direct interactions between the emigrants and their origin communities. Building on previous literature, we will argue that migration ties foster the

[48]Parishes are very small geographical units with a population of approximately 2,500. Coarsened exact matching weights are calculated to balance individuals in terms of age, parish of residence, and occupation. Appendix Figure 3.38 reports the correlation between treatment status and pre-treatment individual-level observable characteristics for the baseline sample (panel A) and the CEM weighted sample (panel B).

integration between markets. This, in turn, facilitates the diffusion of information and thus fosters knowledge flows.

The Transatlantic Telegraph Increased Innovation In Emigration Districts

We exploit one historically relevant event to provide evidence that migration ties foster cross-border economic integration: the first transatlantic telegraphic cable that connected the US and UK domestic networks (1866). The telegraph represented a major revolution in communication technology that ushered unprecedented market integration (Steinwender, 2018). Before 1866, steam mail was the cheapest and fastest way to communicate between the UK and the US. It took seven to fifteen days to transmit information in this way. This delay was reduced to one day overnight between June 27 and 28, 1866. The connection timing was unanticipated and exogenous (Steinwender, 2018).[49]

We claim that if migration ties fostered cross-border market integration between the British and the American market, then districts with relatively higher US emigration rates would be more exposed to the telegraph shock. Under this interpretation, we thus expect that districts with higher US emigration rates after 1866 would display (i) increased innovation activity in districts with more US emigrants and (ii) increased innovation activity in the fields emigrants were exposed to in the US. To test these hypotheses, we estimate the following difference-in-differences models:

$$\text{Patents}_{i,t} = \alpha_i + \alpha_t + \sum_{h=-a}^{b} \beta^h \left[\text{US Emigrants}_i \times \text{I}(t - 1866 = h)\right] + \varepsilon_{i,t} \tag{3.10a}$$

$$\text{Patents}_{ik,t} = \alpha_{i\times k} + \alpha_t + \sum_{h=-a}^{b} \beta^h \left[\text{Knowledge Exposure}_{ik} \times \text{I}(t - 1866 = h)\right] + \varepsilon_{ik,t} \tag{3.10b}$$

where i, k, and t denote a district, technology class, and yea, respectively. The term (US Emigrants_i) and ($\text{Knowledge Exposure}_{ik}$) code the number of US emigrants and

[49]The project for a transatlantic telegraphic cable had been underway for a long time before 1866. Previous attempts in 1857, 1858, and 1865 all failed due to logistic and technical challenges. The 1866 attempt was thus one among many, and its success had not been anticipated.

exposure to US knowledge.[50] Lastly, the variable $\mathrm{I}(t - 1866 = h)$ denotes the number of years since the transatlantic cable was laid down. In equation (3.10a), the treatment coefficients $\{\beta^h\}$ quantify the effect of the transatlantic cable by comparing districts by the number of US emigrants; in equation (3.10b), we also leverage variation across sectors and exposure to US innovation.

We report the static versions that conflate pre- and post-treatment years in two periods in columns (1) and (4) of Table 3.7. We estimate a positive and significant effect of the transatlantic telegraph on innovation. To provide more convincing evidence on the plausibility of this result, we expect the transatlantic cable to affect innovation only in districts that were connected to the British domestic network.[51] We thus reconstruct the entire telegraph network before the introduction of the transatlantic cable. The exact location of each station is displayed in Appendix Figure 3.11. We refer to districts with at least one station as "connected". In columns (2) and (5), we show the estimated effect of the telegraph on connected districts. By comparison, columns (3) and (6) report the estimates for non-connected districts. The results of this exercise are sharp. We estimate a positive effect of the transatlantic cable only for districts connected to the domestic UK network, as expected. We fail to detect any significant effect on non-connected districts.

Because the location of telegraph stations was not random, one may argue that this exercise only reflects pre-existing differences between connected and non-connected districts. However, identification in this setting requires that patenting in connected and unconnected districts was on the same trend before the introduction of the telegraph and that it would not have differed had the cable not been laid down. In Figure 3.7, we thus report the flexible double-differences estimates of model (3.10a), which we estimate separately on connected and unconnected districts. We find that connected and unconnected districts were on the same trend before 1866. We estimate positive and significant treatment effects only for the former and after 1866, whereas the patenting in the latter does not

[50] The cable was laid down in 1866. Our migration data started in 1870. To construct district-level emigration, we can only use emigrants from 1870–1875. This would be problematic if the telegraph fostered out-migration, which, by available historical accounts, was not the case.

[51] We do not claim that there were no cross-district spillover effects even if districts were not connected to the domestic UK network. We nonetheless believe the effect on connected districts would arguably be more significant.

respond to the shock. In 1873 and 1874, the second and third cables became operational. Our estimates suggest positive treatment effects for those.

Building on Steinwender (2018), we interpret these results as evidence that the increased economic integration between the UK and the US ushered by the transatlantic telegraph was relatively more intense in districts that had previous migration ties with the US. Thus, migration ties facilitate market integration and, indirectly, the diffusion of knowledge across countries. This is in line with evidence by Aleksynska and Peri (2014), who document that migrants promote trade between their origin and their destination countries. We provide additional evidence in this direction in section 3.6.3 and discuss potential additional mechanisms in section 3.6.4.

Newspaper Mentions of United States Topics in Emigration Districts

Thus far, we have restricted the focus of the analysis to information flows that pertain to innovative knowledge (patents). This section provides evidence that migration ties between the UK and the US generated more general-purpose information flows that did not directly concern innovation. We exploit the vast British Newspaper Archive that contains the digitized contents of thousands of historical British newspapers (for a detailed description of the data, see Appendix section 3.3.3 and Beach and Hanlon (2022)). Ideally, we would like to measure the intensity of US-related information flows into the United Kingdom. We tackle the absence of direct hard data by measuring how frequently US-related news appeared in historical newspapers.

We estimate three sets of regressions:

$$\text{US Mentions}_{i,t} = \alpha_i + \alpha_t + \beta^1 \times \text{US Emigrants}_{i,t} + \varepsilon_{i,t} \tag{3.11a}$$

$$\text{US State Mentions}_{is,t} = \alpha_i + \alpha_{s\times t} + \beta^2 \times \text{US Emigrants}_{i\to s,t} + \varepsilon_{is,t} \tag{3.11b}$$

$$\text{US County Mentions}_{ij,t} = \alpha_i + \alpha_{j\times t} + \beta^3 \times \text{US Emigrants}_{i\to j,t} + \varepsilon_{ij,t} \tag{3.11c}$$

where i, j, s, and t denote a UK district, a US county, a US state, and a decade, respectively. Regression (3.11a) is run at the district level and leverages the variation of

the overall US emigration rate; in regressions (3.11b) (resp. (3.11c)), instead, we look at district-by-state (resp. district-by-county) migration flows. We estimate regressions (3.11) using actual out-migration and the shift-share instrument described in section 3.4.3.

Table 3.8 reports the results. Panels A, B, and C respectively display the estimated β^i coefficients of models (3.11a), (3.11b), and (3.11c). In columns (1–3), we report the correlation between measured out-migration flows and newspaper coverage; columns (4–5) report the OLS reduced-form association with the instrument; columns (7–9) display the two-stage least-square estimates. In columns (3), (6), and (9), we restrict the sample to districts with at least one newspaper. We find a strong and positive effect of out-migration on newspaper coverage of general-interest US-related news. Importantly, we always control for time-varying confounding factors at the level of the receiving place, whether it be the country, single states, or single counties. This ensures that the estimates do not reflect shocks in those areas.

We interpret this result as evidence that out-migration generates general—not only innovation—information flows between the areas where emigrants settle and where they originate. We cannot disentangle—-and this goes beyond the scope of this paper—the precise underlying mechanism. For example, increased coverage of US-related news may be demand-driven because the local population may demand information covering the areas where their loved ones settled. On the other hand, US emigrants could have sponsored local newspapers to cover news of the areas where they had located. In this sense, our estimates may reflect a supply-side factor. What is crucial for this paper is that, notwithstanding the precise underlying mechanism, out-migration ignites cross-country information flows. The return innovation effect is thus one of the possibly many effects of out-migration on countries sending migrants.[52]

[52]A recent literature documents the disparate effects of out-migration on attitudes towards democracy (Spilimbergo, 2009), demand for political change (Karadja and Prawitz, 2019), wages (Dustmann, Frattini and Rosso, 2015), technology adoption and innovation (Andersson, Karadja and Prawitz; Coluccia and Spadavecchia, 2022; 2022), social norms (Tuccio and Wahba, 2018). Our analysis confirms that migration ties nurture the exchange of information. These flows prompt the cross-border diffusion of novel knowledge but their influence extends well beyond.

Trade-Induced Technology Transfer

The telegraph analysis suggests that market integration, fostered by migration ties, is a major driver of the return innovation result. Here, we provide one additional piece of evidence to support this interpretation.[53] The scope of this exercise is to explore the possible proximate determinants of information diffusion. First, it may be that emigrants themselves are exposed to novel knowledge, which they contribute to spreading. Alternatively, migration ties may facilitate the establishment of trade linkages, which in turn foster cross-border knowledge flows (Aleksynska and Peri; Ottaviano, Peri and Wright, 2014; 2018).

To study this second effect, we explore one specific historical example. In 1930, the United States passed a tariff—the Smoot-Hawley Act —which sharply increased import duties and hampered trade (Eichengreen, 1986). We leverage variation in the tariff increase across technology classes in a difference-in-differences setting.[54] We find that patenting decreases in districts more exposed to technologies that the Act more heavily targeted. This result suggests that migration ties may facilitate international trade, thus contributing to market integration and nurturing the diffusion of novel knowledge. At the same time, however, the magnitude of the estimated effect is modest, especially given the large increase in tariff duties sanctioned by the Act. We thus view trade-induced knowledge diffusion as one, but likely not the only, determinant of information flows.

The literature identifies several margins along which trade can impact innovation. Since the tariff reform was one-sided, it is unlikely that depressed import competition or access to intermediate inputs drive this result (Bloom, Draca and Van Reenen; Autor, Dorn, Hanson, Pisano and Shu, 2016; 2020). We are unable to conclusively disentangle the impact of export opportunities (Atkin, Khandelwal and Osman, 2017) from the information access effect of migration ties (Aleksynska and Peri, 2014). The research design, however, leverages cross-district variation in previous US emigration rates. A purely export-driven

[53]We discuss the literature and the technical implementation of the empirical analysis in Appendix section 3.13.1.

[54]We thus exploit between-class variation in exposure to the tariff increase to estimate the effect of trade on knowledge flows. The underlying intuition is that trade in industries that were more heavily targeted by the act suffered relatively more. Aggregate trade volumes support this interpretation.

effect would not reflect variation along this margin and would thus be unlikely to drive the results.

3.6.4 Potential Additional Mechanisms

In this section, we discuss some potential additional mechanisms that may explain the return innovation result. It is worth stressing that these may operate on top, and not instead, of return migration, cross-border interactions, and economic integration.

Temporary Migrations

When disentangling the possible mechanisms behind the return innovation effect, we contrasted those requiring physical return migration with those that do not. We concluded that physical return is an important determinant of return innovation, but we provide evidence that other mechanisms operate on top of it that do not require physical return. It may be possible that (unobserved) short-term temporary migrations influence the dynamics of innovation in the UK. We cannot observe temporary migrants because we construct migration flows from census data. Censuses are, in turn, only administered to the residing population every ten years. Our data would thus fail to reflect such temporary migration movements. For the reasons above, we cannot quantify the importance of industrial espionage. Episodes of industrial espionage were relatively common during the Industrial Revolution (Harris, 1998).

Temporary migrations and industrial espionage would confound our estimates if such migrations were correlated with observed migration patterns. We believe that it is unlikely that this factor bears relevant quantitative implications. First, the notion of a "temporary migrant" in XIX-century transatlantic migration is unclear. Piore (1980) refers to Southern and Eastern European migrants as temporary because they planned to return to their origin countries at some point. This could take, however, decades. For example, a one-way cabin travel ticket from New York to Liverpool, at roughly 100$, would cost as much as 20% of the average annual US income (Dupont, Keeling and Weiss, 2017). This

suggests that the extent of short-term stays must have been relatively limited. Moreover, Piore (1980) notes that "temporary" migrants were relatively low-skilled and, thus, less likely to operate technology transfer. Industrial espionage, in turn, does not appear to be quantitatively sufficient to generate the return innovation effect that we estimate.

Furthermore, our research designs speak against the temporary migration and the industrial espionage mechanisms. First, the instrumental variable research design largely rules these mechanisms out. Suppose that measured out-migration and unobserved temporary migrations or industrial espionage were correlated across origin districts and destination counties. Our pull instrumental variable randomizes county-level immigration shocks leveraging (conditional) variation in the decade counties were connected to the railway network. While we show that the resulting instrument predicts actual out-migration, it is likely that the source of pull variation is not as active for temporary "business" migrants or spies. Second, for temporary migration or industrial espionage to explain the double and triple differences result, one would need such flows to be correlated with the county-level innovation shocks. This channel seems unlikely, although it cannot be directly tested and refuted.

Monetary Remittances

Along with classical "brain drain" arguments, monetary remittances have been a major subject of empirical investigations in the migration literature (Clemens, 2011). Remittances have been found to contribute only modestly to the economic development of emigration countries. This notwithstanding, it is possible that the inflow of capital through remittances may have sustained increased innovation activity, perhaps by relaxing financial constraints or access to credit (Gorodnichenko and Schnitzer, 2013). It would be more difficult, however, that it would have impacted the *direction* of innovation and, most importantly, that this effect would have been correlated with variation in knowledge exposure.

Disaggregated data on financial remittances, unfortunately, do not exist. We thus remain silent on the possibility that the documented positive effect of out-migration on innova-

tion depends on financial remittances. This capital inflow, however, cannot explain why out-migration influences the direction of innovation unless knowledge *and* monetary remittances go hand in hand. This is a possibility that we cannot explore. It nonetheless highlights that financial and innovation remittances shape innovation in a complementary, rather than mutually exclusive, fashion.

3.6.5 Discussion

Our results bear potentially far-reaching implications for policy-makers. We show that emigration does not necessarily further underdevelopment or stagnation, as the "brain drain" literature seems to suggest (Docquier and Rapoport, 2012). Instead, out-migration can foster innovation, technology adoption, and diffusion and thus empower long-run economic growth. Rather than focusing on blocking the emigration of skilled individuals, our central recommendation to policy-makers in emigration countries would be to foster cooperation and exchanges between them and the population remaining in the home country. Our results and more recent albeit narrative evidence by Saxenian; Saxenian (1999; 2006) suggest that this approach can yield important and lasting benefits on the economic development of emigration countries.

Consider, as an example, the case of the ongoing diaspora of the Italian scientific community. Italy is often described as an archetypal instance of brain drain (Anelli, Basso, Ippedico and Peri, 2023). In 2020, fifty-five Italian researchers were awarded a European Research Council (ERC) starting grant, possibly the most prestigious award for early-career scholars working in the European Union. Only nineteen ($\approx 35\%$) of them worked in Italian institutions. Italy is the country with the lowest share of ERC-winning researchers working in home institutions among those for which data are reported. This paper sheds new light on the economic contribution of the remaining thirty-six ($\approx 65\%$) on science, innovation, and, ultimately, growth in Italy.[55] Several other countries, how-

[55]These figures are the result of authors' calculations over data released by the ERC, available at this link. We mention the Italian case because Italy underwent a major loss of skilled human capital in recent years. Between 2008 and 2016, more than 500,000 Italians emigrated. Comparable high-skilled emigration, however, concerns several other developed economies (United Nations and OECD, 2013).

ever, have been witnessing important episodes of out-migration, ranging from European countries to India, China, or Pakistan. In fact, Saxenian (1999) specifically analyses the Indian and Taiwanese emigration to Silicon Valley. More generally, we study how the entire stock of emigrants influences the dynamics of innovation in their sending countries. In doing so, we enrich our understanding of the consequences of emigration compared to studies that focus on sub-samples of super-skilled emigrants Prato (2021).

Concerns over the external validity of these results are natural, given the setting we analyze. We nonetheless think that History can inform the scholarly debate and policy-making for two main reasons. First, as previously mentioned, Saxenian; Saxenian (1999; 2006) qualitatively documents similar return innovation effects with respect to the Taiwanese and Indian emigration to the Silicon Valley area. Second, we provide evidence that the UK emigration to the US in the XIX century largely resembles, *mutatis mutandis*, migration between European countries and the United States during the XXI. Compared to the rest of the English population, migrants were positively selected. They were similarly more likely to be employed in skilled occupations than the average native and to live in urban centers. These patterns suggest that a cautious comparison between historical and contemporary migration episodes can yield important insights for policy-makers and scholars.

3.7 Conclusions

The diffusion of innovation across countries is a major factor shaping long-run economic development. In this paper, we argue that international migrations generate knowledge flows that contribute to the diffusion of innovation into emigration countries. This result—which we label "return innovation"—offers a more nuanced view of the effect of emigration on innovation compared to the traditional "brain drain" hypothesis, which interprets it as a depletion of the human capital of countries sending migrants.

To study this question, this paper explores the English and Welsh mass migration to the United States between 1870 and 1940. To construct a granular measure of transatlantic

migration flows, we link census records of British immigrants in the US to the individual-level UK population census. The resulting dataset allows us to observe the universe of (male) British immigrants after they migrate to the US and before leaving the UK. We complement this with newly digitized patent data covering the universe of patents in England and Wales. On top of these unique, high-quality data, the absence of stringent international intellectual property protection and active migration policies provide two prominent appealing features of this historical setting compared to contemporary scenarios.

We provide novel, causal evidence that exposure to US innovation through migration ties contributes to the diffusion of US technology in Britain. Innovation activity in the UK shifts to sectors that emigrants are most exposed to in the US. To address endogeneity concerns arising from the assortative matching of British immigrants in the US, we develop a new shift-share instrument that exploits the conditional timing of connection to the railway network to randomize emigration across counties. Moreover, we implement a triple-differences research design that leverages variation across counties and technology classes. We can thus document a causal link between exposure to foreign knowledge through migration ties and innovation activity. By looking at the textual similarity between UK and US patents, we document that exposure to US knowledge stimulates cross-border technology transfer while also nurturing original innovation in Britain.

What drives the return innovation effect? We find that the physical return of migrants is a crucial driver as it explains approximately half of the return innovation effect. Exploiting the granular nature of our data, we provide evidence that social interactions between the emigrants and their communities of origin represent another important channel through which technologies diffuse into the emigration country, even if emigrants do not return. Additionally, we document that migration ties promote the diffusion of knowledge by fostering information flows and cross-border market integration. Leveraging a comprehensive repository of historical newspapers, we show that migration linkages further promote general-purpose information flows by increasing attention to US-related news in UK media outlets.

The historical evidence suggests that the British mass migration to the United States may be comparable to present-day cross-border movements between developed countries. As the number of international migrants has been steadily rising over the past decades, the role of human mobility as a driver of knowledge and information diffusion across countries in a globalized world economy bears quantitatively relevant implications. History can thus inform the scholarly literature and policymakers on the complex relationship between out-migration, innovation, and, ultimately, long-run economic growth.

References

Abramitzky, Ran, and Leah Boustan. 2017. "Immigration in American Economic History." *Journal of Economic Literature*, 55(4): 1311–45.

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2020. "Do Immigrants Assimilate More Slowly Today Than in the Past?" *American Economic Review: Insights*, 2(1): 125–41.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature*, 59(3): 865–918.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy*, 122(3): 467–506.

Acemoglu, Daron. 2002. "Directed Technical Change." *The Review of Economic Studies*, 69(4): 781–809.

Acemoglu, Daron. 2010. "When Does Labor Scarcity Encourage Innovation?" *Journal of Political Economy*, 118(6): 1037–1078.

Acemoglu, Daron. 2023. "Distorted Innovation: Does the Market Get the Direction of Technology Right?" Vol. 113, 1–28.

Acemoglu, Daron, and Todd Lensman. 2023. "Technology Paradigms, Lock-in, and Economic Growth." *Mimeo*.

Aghion, Philippe, Antoine Dechezleprêtre, David Hemous, Ralf Martin, and John Van Reenen. 2016. "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry." *Journal of Political Economy*, 124(1): 1–51.

Aghion, Philippe, Antonin Bergeaud, Matthieu Lequien, and Marc J Melitz. 2018. "The Impact of Exports on Innovation: Theory and Evidence." *NBER Working Paper*.

Agrawal, Ajay, Devesh Kapur, John McHale, and Alexander Oettl. 2011. "Brain Drain or Brain Bank? The Impact of Skilled Emigration on Poor-Country Innovation." *Journal of Urban Economics*, 69(1): 43–55.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas. 2017. "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age." *NBER Working Paper*.

Akcigit, Ufuk, Santiago Caicedo, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi. 2018. "Dancing With the Stars: Innovation Through Interactions." *NBER Working Paper*.

Aleksynska, Mariya, and Giovanni Peri. 2014. "Isolating the Network Effect of Immigrants on Trade." *The World Economy*, 37(3): 434–455.

Alvarez, Fernando E, Francisco J Buera, and Robert E Lucas. 2013. "Idea Flows, Economic Growth, and Trade." *NBER Working Paper*.

Andersson, David E, Mounir Karadja, and Erik Prawitz. 2022. "Mass Migration and Technological Change." *Journal of the European Economic Association*.

Anelli, Massimo, Gaetano Basso, Giuseppe Ippedico, and Giovanni Peri. 2023. "Emigration and Entrepreneurial Drain." *American Economic Journal: Applied Economics*, 15(2): 218–252.

Arkolakis, Costas, Sun Kyoung Lee, and Michael Peters. 2020. "European Immigrants and the United States' Rise to the Technological Frontier." *Working Paper*.

Arthur, W Brian. 1989. "Competing Technologies, Increasing Returns, and Lock-in by Historical Events." *The Economic Journal*, 99(394): 116–131.

Ash, Elliott, David Cai, Mirko Draka, and Shaoyu Liu. 2022. "Bootstrapping Science? The Impact of a "Return Human Capital" Programme on Chinese Research Productivity." *Working Paper*.

Atkin, David, Amit K Khandelwal, and Adam Osman. 2017. "Exporting and Firm Performance: Evidence from a Randomized Experiment." *The Quarterly Journal of Economics*, 132(2): 551–615.

Autor, David, David Dorn, Gordon H. Hanson, Gary Pisano, and Pian Shu. 2020. "Foreign Competition and Domestic Innovation: Evidence from US Patents." *American Economic Review: Insights*, 2(3): 357–74.

Azoulay, Pierre, Benjamin F Jones, J Daniel Kim, and Javier Miranda. 2022. "Immigration and Entrepreneurship in the United States." *American Economic Review: Insights*, 4(1): 71–88.

Bahar, Dany, Andreas Hauptmann, Cem Özgüzel, and Hillel Rapoport. 2019. "Migration and Knowledge Diffusion: The Effect of Returning Refugees on Export Performance in the Former Yugoslavia." *The Review of Economics and Statistics*, 1–50.

Bahar, Dany, Andreas Hauptmann, Cem Özgüzel, and Hillel Rapoport. 2022. "Migration and Knowledge Diffusion: The Effect of Returning Refugees on Export Performance in the Former Yugoslavia." *The Review of Economics and Statistics*, 1–50.

Bahar, Dany, Prithwiraj Choudhury, James Sappenfield, and Sara Signorelli. 2022. "Talent Flows and the Geography of Knowledge Production: Causal Evidence from Multinational Firms." *Working Paper*.

Bahar, Dany, Ricardo Hausmann, and Cesar A Hidalgo. 2014. "Neighbors and the Evolution of the Comparative Advantage of Nations: Evidence of International Knowledge Diffusion?" *Journal of International Economics*, 92(1): 111–123.

Bailey, Martha J, Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature*, 58(4): 997–1044.

Baines, Dudley. 2002. *Migration in a Mature Economy: Emigration and Internal Migration in England and Wales 1861-1900.* Cambridge University Press.

Balassa, Bela. 1965. "Trade Liberalisation and "Revealed" Comparative Advantage." *The Manchester School*, 33(2): 99–123.

Bandiera, Oriana, and Imran Rasul. 2006. "Social Networks and Technology Adoption in Northern Mozambique." *The Economic Journal*, 116(514): 869–902.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo. 2013. "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics*, 102: 23–47.

Batista, Catia, and Pedro C Vicente. 2011. "Do Migrants Improve Governance at Home? Evidence from a Voting Experiment." *The World Bank Economic Review*, 25(1): 77–104.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States." *Econometrica*, 88(6): 2329–2368.

Beach, Brian, and W Walker Hanlon. 2022. "Historical Newspaper Data: A Researcher's Guide and Toolkit."

Beaman, Lori, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. 2021. "Can Network Theory-Based Targeting Increase Technology Adoption?" *American Economic Review*, 111(6): 1918–1943.

Beine, Michel, Frédéric Docquier, and Maurice Schiff. 2013. "International Migration, Transfer of Norms and Home Country Fertility." *Canadian Journal of Economics/Revue canadienne d'économique*, 46(4): 1406–1430.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *The Quarterly Journal of Economics*, 134(2): 647–713.

Benhabib, Jess, Jesse Perla, and Christopher Tonetti. 2021. "Reconciling Models of Diffusion and Innovation: A Theory of the Productivity Distribution and Technology Frontier." *Econometrica*, 89(5): 2261–2301.

Bergeaud, Antonin, and Cyril Verluise. 2022. "A New Dataset to Study a Century of Innovation in Europe and in the US." *Working Paper.*

Berkes, Enrico. 2018. "Comprehensive Universe of US patents (CUSP): Data and Facts." *Mimeo.*

Berkes, Enrico, Davide Coluccia, Gaia Dossi, and Mara Squicciarini. 2023. "Dealing with Adversity: Religiosity and Science? Evidence from the Great Influenza Pandemic." *Working paper.*

Bernstein, Shai, Rebecca Diamond, Abhisit Jiranaphawiboon, Timothy McQuade, and Beatriz Pousada. 2022. "The Contribution of High-Skilled Immigrants to Innovation in the United States." *NBER Working Paper.*

Berthoff, Rowland. 1953. *British Immigrants in Industrial America 1790-1950.* Harvard University Press.

Bertoli, Simone, and Francesca Marchetta. 2015. "Bringing It All Back Home–Return Migration and Fertility Choices." *World Development*, 65: 27–40.

Bloom, Nicholas, Mirko Draca, and John Van Reenen. 2016. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity." *The Review of Economic Studies*, 83(1): 87–117.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-Experimental Shift-Share Research Designs." *The Review of Economic Studies*, 89(1): 181–213.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting Event Study Designs: Robust and Efficient Estimation." *Mimeo.*

Bottomley, Sean. 2014. *The British Patent System during the Industrial Revolution 1700–1852: From Privilege to Property.* Cambridge (UK):Cambridge University Press.

Bryan, Kevin A, and Jorge Lemus. 2017. "The Direction of Innovation." *Journal of Economic Theory*, 172: 247–272.

Buera, Francisco J, and Ezra Oberfield. 2020. "The Global Diffusion of Ideas." *Econometrica*, 88(1): 83–114.

Burchardi, Konrad B, Thomas Chaney, Tarek Alexander Hassan, Lisa Tarquinio, and Stephen J Terry. 2020. "Immigration, Innovation, and Growth." *NBER Working Paper.*

Bustos, Paula. 2011. "Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms." *American Economic Review*, 101(1): 304–40.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with a Continuous Treatment."

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics*, 19(1): 22–64.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and David Zentler-Munro. 2022. "Seeing Beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes." *Journal of Labor Economics*, 40(S1): S203–S247.

Chauvet, Lisa, and Marion Mercier. 2014. "Do Return Migrants Transfer Political Norms to Their Origin Country? Evidence from Mali." *Journal of Comparative Economics*, 42(3): 630–651.

Chen, Jiafeng, and Jonathan Roth. 2022. "Log-like? ATEs Defined with Zero Outcomes are (Arbitrarily) Scale-dependent." *Working Paper.*

Choudhury, Prithwiraj. 2016. "Return Migration and Geography of Innovation in MNEs: A Natural Experiment of Knowledge Production by Local Workers Reporting to Return Migrants." *Journal of Economic Geography*, 16(3): 585–610.

Clemens, Michael A. 2011. "Economics and Emigration: Trillion-dollar Bills on the Sidewalk?" *Journal of Economic Perspectives*, 25(3): 83–106.

Coelli, Federica, Andreas Moxnes, and Karen Helene Ulltveit-Moe. 2022. "Better, Faster, Stronger: Global Innovation and Trade Liberalization." *The Review of Economics and Statistics*, 104(2): 205–216.

Coluccia, Davide M, and Lorenzo Spadavecchia. 2022. "Emigration Restrictions and Economic Development: Evidence from the Italian Mass Migration to the United States." *Working Paper*.

Comin, Diego, and Bart Hobijn. 2011. "Technology Diffusion and Postwar Growth." *NBER Macroeconomics Annual*, 25(1): 209–246.

Conley, Timothy G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics*, 92(1): 1–45.

Conley, Timothy G, and Christopher R Udry. 2010. "Learning About a New Technology: Pineapple in Ghana." *American Economic Review*, 100(1): 35–69.

Coulter, Moureen. 1991. *Property in Ideas: The Patent Question in Mid-Victorian England.* Kirksville (MO):Thomas Jefferson Press.

Crucini, Mario J. 1994. "Sources of Variation in Real Tariff Rates: The United States, 1900–1940." *The American Economic Review*, 84(3): 732–743.

Curtis, Sidney J. 1875. *The Story of the Marsden Mayoralty: With Sketch of the Mayor's Life.* Leeds (UK):Express Office.

David, Paul A. 1966. "The Mechanization of Reaping in the Ante-Bellum Midwest." In *Industrialization in Two Systems: Essays in Honor of Alexander Gershenkron.* Harvard University Press Cambridge, MA.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review*, 80(2): 355–361.

Davis, Joseph H. 2004. "An Annual Index of US Industrial Production, 1790–1915." *The Quarterly Journal of Economics*, 119(4): 1177–1215.

De Chaisemartin, Clément, and Xavier D'Haultfœuille. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *NBER Working Paper*.

Docquier, Frédéric, and Hillel Rapoport. 2012. "Globalization, Brain Drain, and Development." *Journal of Economic Literature*, 50(3): 681–730.

Dosi, Giovanni. 1982. "Technological Paradigms and Technological Trajectories: A Suggested Interpretation of the Determinants and Directions of Technical Change." *Research Policy*, 11(3): 147–162.

Dupont, Brandon, Drew Keeling, and Thomas Weiss. 2017. "First Cabin Fares from New York to the British Isles, 1826–1914." In *Research in Economic History*. Emerald Publishing Limited.

Dustmann, Christian, and Joseph-Simon Görlach. 2016. "The Economics of Temporary Migrations." *Journal of Economic Literature*, 54(1): 98–136.

Dustmann, Christian, Tommaso Frattini, and Anna Rosso. 2015. "The Effect of Emigration from Poland on Polish wages." *The Scandinavian Journal of Economics*, 117(2): 522–564.

Dutton, H I. 1984. *The Patent System and Inventive Activity during the Industrial Revolution, 1750-1852.* Manchester (UK):Manchester University Press.

Eaton, Jonathan, and Samuel Kortum. 1999. "International Technology Diffusion: Theory and Measurement." *International Economic Review*, 40(3): 537–570.

Eckert, Fabian, Andrés Gvirtz, Jack Liang, and Michael Peters. 2020. "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790." *NBER Working Paper*.

Eichengreen, Barry. 1986. "The Political Economy of the Smoot-Hawley Tariff." *NBER Working Paper.*

Einiö, Elias, Josh Feng, and Xavier Jaravel. 2023. "Social Push and the Direction of Innovation." *Working Paper.*

Erickson, Charlotte. 1957. *American Industry and the European Immigrant, 1860–1885.* Harvard University Press.

Erickson, Charlotte. 1972. "Who Were the English and Scots Emigrants in the 1880s?" *Population and Social Change.* Arnold.

Fackler, Thomas A, Yvonne Giesing, and Nadzeya Laurentsyeva. 2020. "Knowledge Remittances: Does Emigration Foster Innovation?" *Research Policy*, 49(9): 103863.

Finishing Publications, Ltd. 2018. "Early British Patents. A Cradle of Inventions. British Patents 1617 to 1895." MFIS [data collection].

Fischer, David Hackett. 1989. *Albion's Seed: Four British Folkways in America.* Oxford University Press.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M Shapiro. 2019. "Pre-event Trends in the Panel Event-Study Design." *American Economic Review*, 109(9): 3307–38.

Furer, Howard B. 1972. *The British in America: 1578-1970.* Oceana Publications.

Ganguli, Ina. 2015. "Immigration and Ideas: What Did Russian Scientists "Bring" to the United States?" *Journal of Labor Economics*, 33(S1): S257–S288.

Gerschenkron, Alexander. 1962. *Economic Backwardness in Historical Perspective: A Book of Essays.* Cambridge (MA): Harvard University Press.

Gibson, John, and David McKenzie. 2011. "Eight Questions About Brain Drain." *Journal of Economic Perspectives*, 25(3): 107–28.

Giorcelli, Michela. 2019. "The Long-Term effects of Management and Technology Transfers." *American Economic Review*, 109(1): 121–52.

Giuliano, Paola, and Marco Tabellini. 2020. "The Seeds of Ideology: Historical Immigration and Political Preferences in the United States." *NBER Working Paper*.

Goldin, Claudia. 1994. "The Political Economy of Immigration Restriction in the United States, 1890 to 1921." In *The Regulated Economy: A Historical Approach to Political Economy*. 223–258. University of Chicago Press.

Gomme, Arthur Allan. 1948. *Patents of Invention: Origin and Growth of the Patent System in Britain.* British council.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, 225(2): 254–277.

Gorodnichenko, Yuriy, and Monika Schnitzer. 2013. "Financial Constraints and Innovation: Why Poor Countries Don't Catch Up." *Journal of the European Economic Association*, 11(5): 1115–1152.

Griffith, Rachel, Rupert Harrison, and John Van Reenen. 2006. "How Special is the Special Relationship? Using the Impact of US R&D Spillovers on UK Firms as a Test of Technology Sourcing." *American Economic Review*, 96(5): 1859–1875.

Griliches, Zvi. 1998. "Patent Statistics as Economic Indicators: A Survey." In *R&D and Productivity: The Econometric Evidence*. 287–343. University of Chicago Press.

Gross, Daniel P, and Bhaven N Sampat. 2022. "Crisis Innovation Policy From World War II to COVID-19." *Entrepreneurship and Innovation Policy and the Economy*, 1(1): 135–181.

Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review*, 84(3): 622–641.

Habakkuk, Hrothgar J. 1962. *American and British Technology in the Nineteenth Century: the Search for Labour Saving Inventions.* Cambridge (MA): Cambridge University Press.

Hanlon, Walker W. 2016. "British Patent Technology Classification Database: 1855-1882." Unpublished [data collection].

Hanlon, W Walker. 2015. "Necessity Is the Mother of Invention: Input Supplies and Directed Technical Change." *Econometrica*, 83(1): 67–100.

Hanlon, W Walker. 2018. "Skilled Immigrants and American Industrialization: Lessons from Newport News Shipyard." *Business History Review*, 92(4): 605–632.

Harris, John. 1998. *Industrial Espionage and Technology Transfer: Britain and France in the Eighteenth Century.* Ashgate Publishing Limited, Aldershot.

Hicks, John. 1932. *The Theory of Wages.* Macmillan, London (UK).

Higham, John. 1955. *Strangers in the Land: Patterns of American Nativism, 1860-1925.* Rutgers University Press.

Hopenhayn, Hugo, and Francesco Squintani. 2021. "On the Direction of Innovation." *Journal of Political Economy*, 129(7): 1991–2022.

Hornung, Erik. 2014. "Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia." *American Economic Review*, 104(1): 84–122.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *The Quarterly Journal of Economics*, 108(3): 577–598.

Jeremy, David J. 1981. *Transatlantic Industrial Revolution: The Diffusion of Textile Technologies Between Britain and America, 1790-1830s.* Cambridge (MA): MIT Press.

Jones, Charles I. 1995. "R&D-based Models of Economic Growth." *Journal of Political Economy*, 103(4): 759–784.

Juhász, Réka, and Claudia Steinwender. 2018. "Spinning the Web: The Impact of ICT on Trade in Intermediates and Technology Diffusion." *NBER Working Paper.*

Kapur, Devesh. 2014. "Political Effects of International Migration." *Annual Review of Political Science*, 17: 479–502.

Karadja, Mounir, and Erik Prawitz. 2019. "Exit, Voice, and Political Change: Evidence from Swedish Mass Migration to the United States." *Journal of Political Economy*, 127(4): 1864–1925.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring Technological Innovation Over the Long Run." *American Economic Review: Insights*, 3(3): 303–20.

Kerr, Sari Pekkala, and William Kerr. 2020. "Immigrant Entrepreneurship in America: Evidence from the Survey of Business Owners 2007 & 2012." *Research Policy*, 49(3): 103918.

Kerr, William R. 2008. "Ethnic Scientific Communities and International Technology Diffusion." *The Review of Economics and Statistics*, 90(3): 518–537.

Khan, B Zorina. 2020. *Inventing Ideas: Patents, Prizes, and the Knowledge Economy.* Oxford University Press, USA.

Khan, B Zorina, and Kenneth L Sokoloff. 2004. *Institutions and Technological Innovation During the Early Economic Growth: Evidence from the Great Inventors of the United States, 1790-1930.* National Bureau of Economic Research Cambridge, Mass., USA.

Lan, Tian, and Paul Longley. 2019. "Geo-Referencing and Mapping 1901 Census Addresses for England and Wales." *ISPRS International Journal of Geo-Information*, 8(8): 320.

Leak, H, and T Priday. 1933. "Migration from and to the United Kingdom." *Journal of the Royal Statistical Society*, 96(2): 183–239.

Li, Shaobo, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. "DeepPatent: Patent Classification with Convolutional Neural Networks and Word Embedding." *Scientometrics*, 117(2): 721–744.

Macleod, Christine. 1988. *Inventing the Industrial Revolution.* Cambridge (UK):Cambridge University Press.

Maddison, Angus. 2007. *Contours of the World Economy 1-2030 AD: Essays in Macro-Economic History.* Oxford (UK):Oxford University Press.

Mokyr, Joel. 1998. "The Second Industrial Revolution, 1870-1914." In *Storia dell'Economia Mondiale.* , ed. V Castronovo, 219–245. Rome (Italy):Laterza.

Moscona, Jacob. 2021. "Environmental Catastrophe and the Direction of Invention: Evidence from the American Dust Bowl." *Working Paper.*

Moscona, Jacob, and Karthik A Sastry. 2022. "Does Directed Innovation Mitigate Climate Damage? Evidence from US Agriculture." *The Quarterly Journal of Economics*, Forthcoming.

Moser, Petra. 2012. "Innovation Without Patents: Evidence from World's Fairs." *The Journal of Law and Economics*, 55(1): 43–74.

Moser, Petra. 2019. "Patents and Innovation in Economic History." In *Research Handbook on the Economics of Intellectual Property Law.* 462–481. Edward Elgar Publishing.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish émigrés and US invention." *American Economic Review*, 104(10): 3222–55.

Moser, Petra, Sahar Parsa, and Shmuel San. 2020. "Immigration, Science, and Invention. Evidence from the Quota Acts." *Working Paper.*

Nelson, Richard R, and Gavin Wright. 1992. "The Rise and Fall of American Technological Leadership: The Postwar Era in Historical Perspective." *Journal of Economic Literature*, 30(4): 1931–1964.

Nicholas, Tom. 2011. "Cheaper Patents." *Research Policy*, 40(2): 325–339.

Nicholas, Tom. 2014. "Technology, Innovation, and Economic Growth in Britain since 1870." In *The Cambridge Economic History of Modern Britain.* , ed. Roderick Floud, Jane Humphries and Paul Johnson, 181–204. Cambridge (UK):Cambridge University Press.

Nuvolari, Alessandro, and Valentina Tartari. 2011. "Bennet Woodcroft and the Value of English Patents, 1617–1841." *Explorations in Economic History*, 48(1): 97–115.

Nuvolari, Alessandro, Valentina Tartari, and Matteo Tranchero. 2021. "Patterns of Innovation During the Industrial Revolution: A Reappraisal Using a Composite Indicator of Patent Quality." *Explorations in Economic History*, 82: 101419.

Olden, Andreas, and Jarle Møen. 2022. "The Triple Difference Estimator." *The Econometrics Journal*, 25(3): 531–553.

Olivetti, Claudia, M Daniele Paserman, Laura Salisbury, and E Anna Weber. 2020. "Who Married, (To) Whom, and Where? Trends in Marriage in the United States, 1850-1940." *NBER Working Paper*.

Ottaviano, Gianmarco I P, Giovanni Peri, and Greg C Wright. 2018. "Immigration, Trade and Productivity in Services: Evidence from UK Firms." *Journal of International Economics*, 112: 88–108.

Ottinger, Sebastian. 2020. "Immigrants, Industries, and Path Dependence." *Working Paper*.

Ottinger, Sebastian, and Lukas Rosenberger. 2023. "The American Origin of the French Revolution." *Working Paper*.

Pauly, Stefan, and Fernando Stipanicic. 2021. "The Creation and Diffusion of Knowledge: Evidence from the Jet Age." *Working Paper*.

Penrose, Edith. 1951. *The Economics of the International Patent System.* Baltimore (MD):Johns Hopkins University Press.

Perla, Jesse, Christopher Tonetti, and Michael E Waugh. 2021. "Equilibrium Technology Diffusion, Trade, and Growth." *American Economic Review*, 111(1): 73–128.

Piore, Michael J. 1980. *Birds of Passage: Migrant Labor and Industrial Societies.* Cambridge (UK): Cambridge University Press.

Prato, Marta. 2021. "The Global Race for Talent: Brain Drain, Knowledge Transfer, and Economic Growth." *Working Paper.*

Rosenberg, Nathan. 1970. "Economic Development and the Transfer of Technology: Some Historical Perspectives." *Technology and Culture*, 11(4): 550–575.

Rosenberg, Nathan. 1982. "The International Transfer of Technology: Implications for the Industrialized Countries." In *Inside the Black Box: Technology and Economics.* New York: Cambridge University Press.

Rosenberg, Nathan, and Manuel Trajtenberg. 2004. "A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States." *The Journal of Economic History*, 64(1): 61–99.

Ruggles, Steven, Catherine Fitch, Ronald Goeken, J Hacker, M Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS ancestry full count data: Version 3.0 [dataset]." *Minneapolis, MN: IPUMS.*

Saxenian, AnnaLee. 1999. *Silicon Valley's New Immigrant Entrepreneurs.* Public Policy Institute of California.

Saxenian, AnnaLee. 2006. *The New Argonauts: Regional Advantage in a Global Economy.* Harvard University Press.

Schurer, K, and E Higgs. 2020. "Integrated Census Microdata (I-CeM) Names and Addresses, 1851-1911: Special Licence Access." [data collection] Second Edition, UKDS.

Sequeira, Sandra, Nathan Nunn, and Nancy Qian. 2020. "Immigrants and the Making of America." *The Review of Economic Studies*, 87(1): 382–419.

Shi, Dongbo, Weichen Liu, and Yanbo Wang. 2023. "Has China's Young Thousand Talents Program Been Successful in Recruiting and Nurturing Top-caliber Scientists?" *Science*, 379(6627): 62–65.

Shu, Pian, and Claudia Steinwender. 2019. "The Impact of Trade Liberalization on Firm Productivity and Innovation." *Innovation Policy and the Economy*, 19(1): 39–68.

Sokoloff, Kenneth L, and B Zorina Khan. 1990. "The Democratization of Invention During Early Industrialization: Evidence from the United States, 1790–1846." *The Journal of Economic History*, 50(2): 363–378.

Spilimbergo, Antonio. 2009. "Democracy and Foreign Education." *American Economic Review*, 99(1): 528–43.

Spitzer, Yannay, and Ariell Zimran. 2018. "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925." *Journal of Development Economics*, 134: 226–247.

Steinwender, Claudia. 2018. "Real Effects of Information Frictions: When the States and the Kingdom Became United." *American Economic Review*, 108(3): 657–96.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199.

Tabellini, Marco. 2020. "Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration." *The Review of Economic Studies*, 87(1): 454–486.

Thistlethwaite, Frank. 1958. "The Atlantic Migration of the Pottery Industry." *The Economic History Review*, 11(2): 264–278.

Thomas, B. 1954. *Migration and Economic Growth. A Study of Great Britain and the Atlantic Economy.* Cambridge (MA):NIESR and Cambridge University Press.

Tuccio, Michele, and Jackline Wahba. 2018. "Return Migration and the Transfer of Gender Norms: Evidence from the Middle East." *Journal of Comparative Economics*, 46(4): 1006–1029.

United Nations, and OECD. 2013. "World Migration in Figures. A Joint Contribution by UN-DESA and the OECD to the United Nations High-Level Dialogue on Migration and Development."

Van Patten, Diana. 2023. "International Diffusion of Technology: Accounting for Heterogeneous Learning Abilities." *Working Paper.*

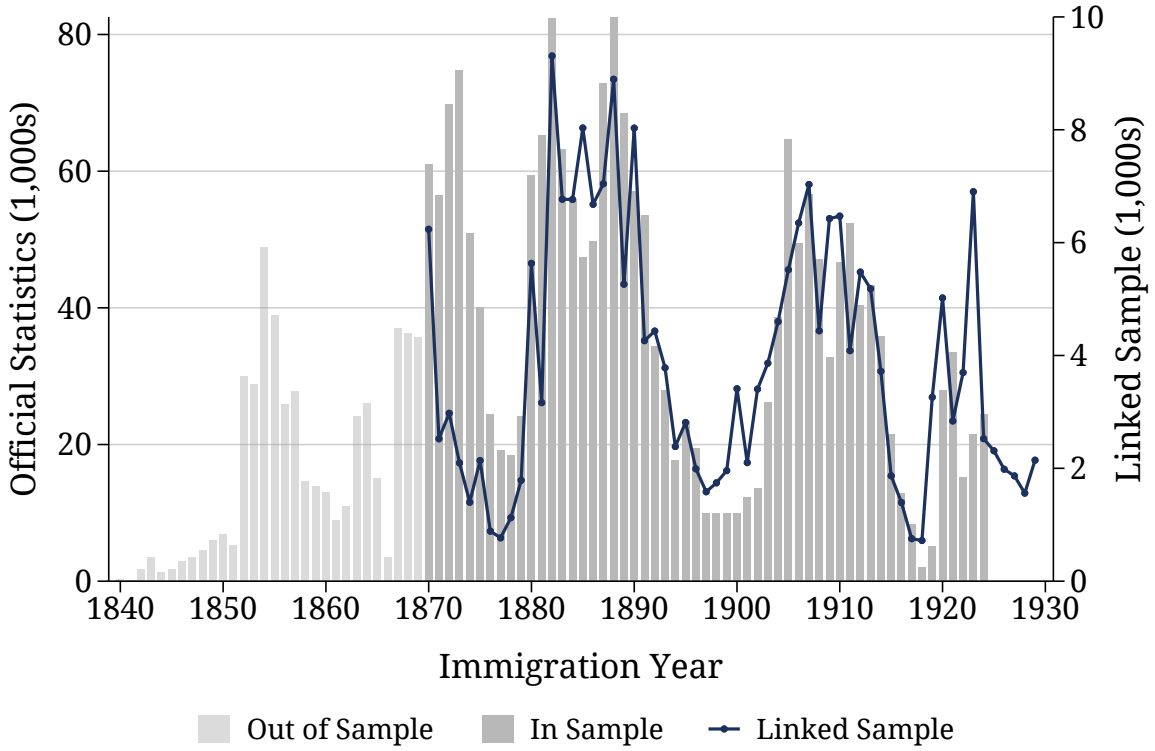
Wagner, Donald. 2008. *Science and Civilisation in China.* Cambridge University Press (UK).

Willcox, William F. 1928. *International Migrations, Volume I: Statistics.* Cambridge (MA):National Bureau of Economic Research.

Xu, Shuo. 2018. "Bayesian Naïve Bayes Classifiers to Text Classification." *Journal of Information Science*, 44(1): 48–59.

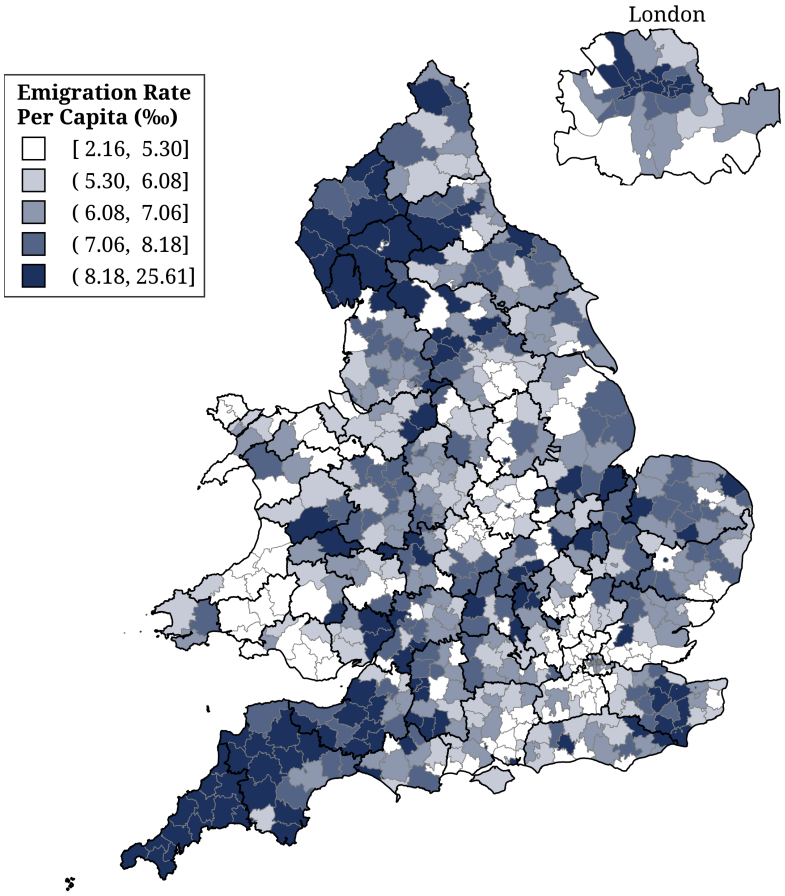
3.8 Figures

Figure 3.1: Number of US Immigrants from the UK and Linked US-UK Migrants, 1840–1930



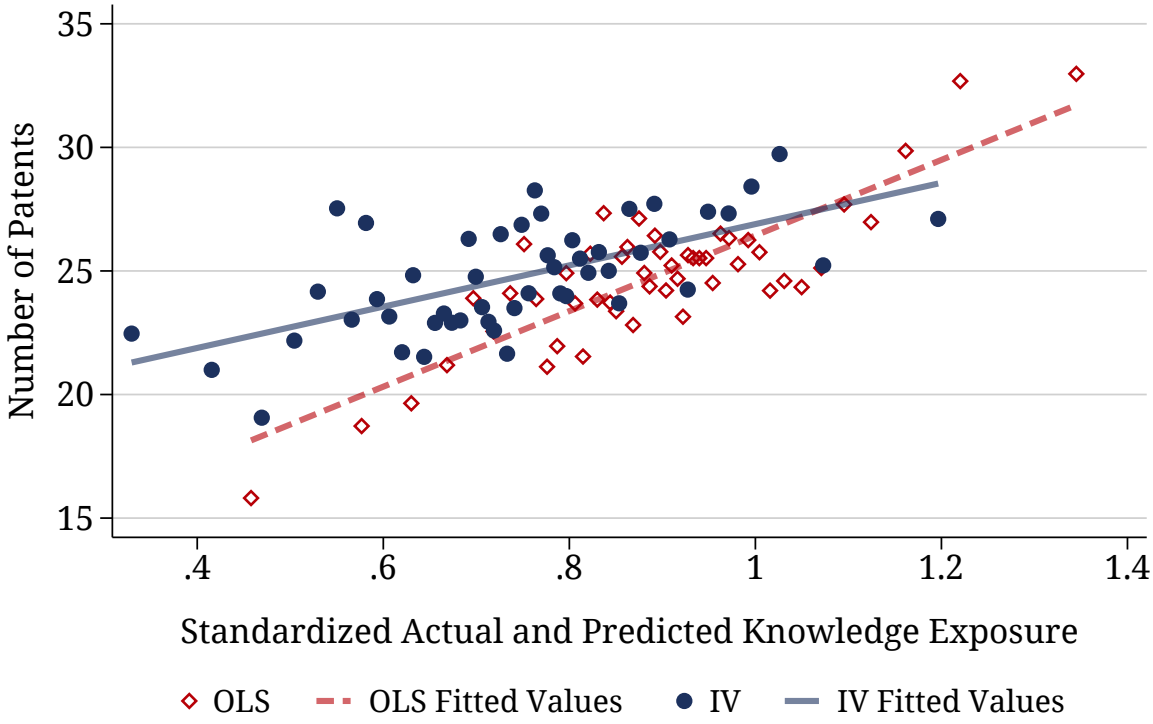
Notes. This figure compares the total number of English and Welsh immigrants in the United States as recorded in official statistics from Willcox (1928) with the linked emigrants' sample developed in this paper. The light gray bars display the total inflow of English and Welsh immigrants in the US over the period 1840–1870, i.e., out of the period we study. The darker gray bars report the same figure for 1870–1924. The blue line, whose values are reported on the right y-axis, reports the total number of English and Welsh immigrants in the US that appear in our matched sample. By construction, we can only match men who appear at least once in one British census. Figures are in thousand units.

Figure 3.2: Spatial Distribution of US Migrants Across British Districts



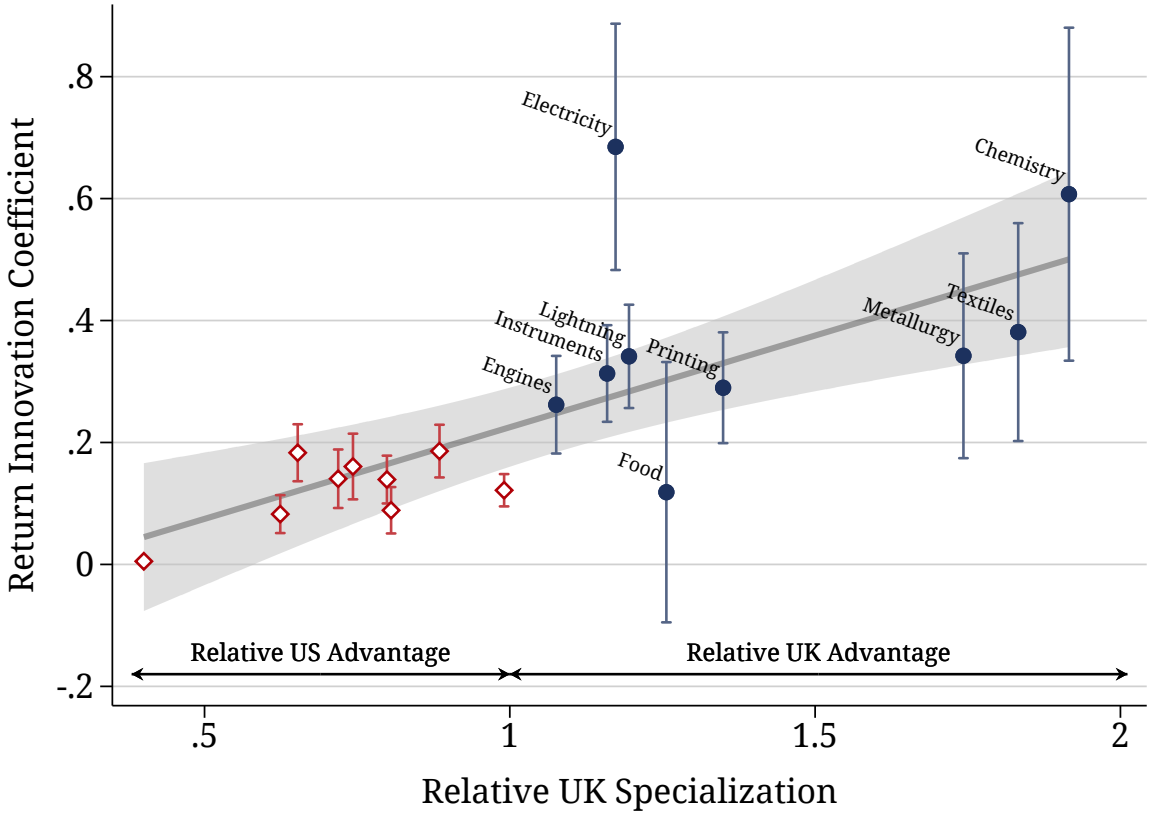
Notes. This figure reports the spatial distribution of emigrants across English and Welsh districts over the period 1870–1930. Data are from the matched emigrants' sample. The total number of emigrants over the period is normalized by district population in 1900 and is reported in ‰ units. Districts are displayed at 1900 historical borders, and the emigrant population is cross-walked to consistent borders as described in 3.10.1. Lighter to darker blues indicate higher emigration rates.

Figure 3.3: OLS and Reduced-Form Association between Knowledge Exposure and Innovation



Notes. This figure is a binned scatter plot of the OLS (in red) and the IV reduced-form (in blue) association between knowledge exposure and innovation. The unit of observation is a district-technology class observed at a yearly frequency between 1880 and 1939. The graph partials out district-by-decade and district-by-technology class fixed effects. The red dots report the correlation between actual knowledge exposure and the number of patents; the blue dots report the reduced-form association between the instrument for knowledge exposure and the number of patents. We report in note the regression coefficient for both the OLS and the IV regression with their standard errors clustered at the district level and the R^2.

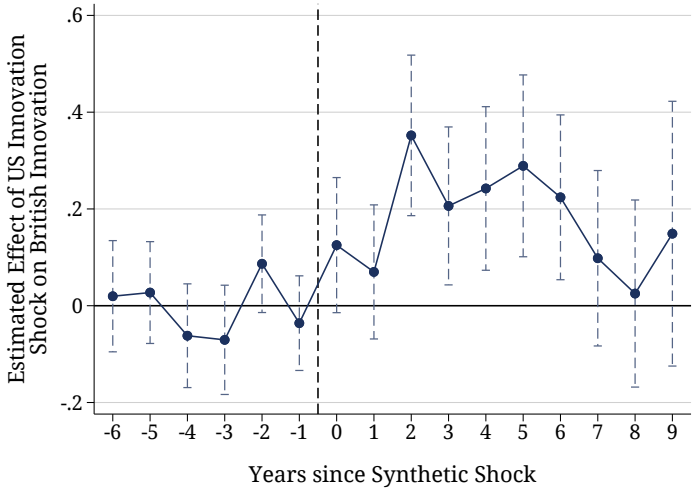
Figure 3.4: Heterogeneous Effects of Return Innovation Across Technology Classes



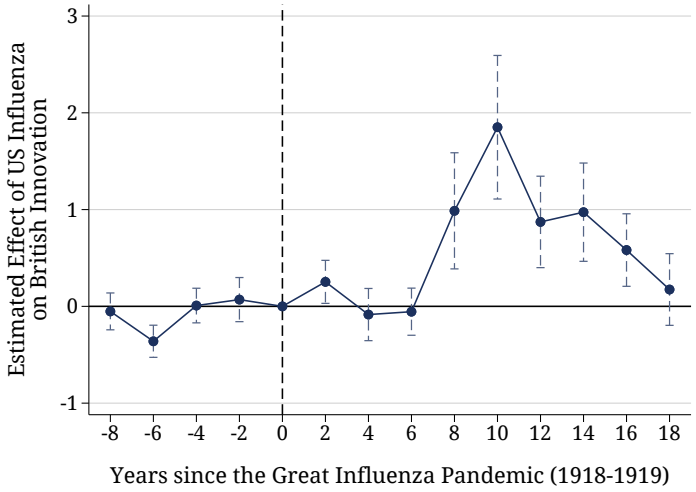
Notes. This figure reports the reduced-form effect of instrumented knowledge exposure on innovation by technology class. Each dot reports the coefficient of a regression between the total number of patents and the instrument for knowledge exposure in a given technology class. The unit of observation in each regression is a district, observed at a decade frequency between 1880 and 1939. Regressions include district and decade fixed effects. Bands report 95% confidence intervals. Standard errors are clustered at the district level. The grey line plots the correlation between the revealed UK specialization on the x-axis and the return innovation coefficient for each sector on the y-axis. Red (resp. blue) dots display the regression coefficients for the US (resp UK) revealed comparative advantage sectors.

Figure 3.5: Effect of Exposure to US Innovation Shocks on UK Innovation

((a)) Synthetic Shocks



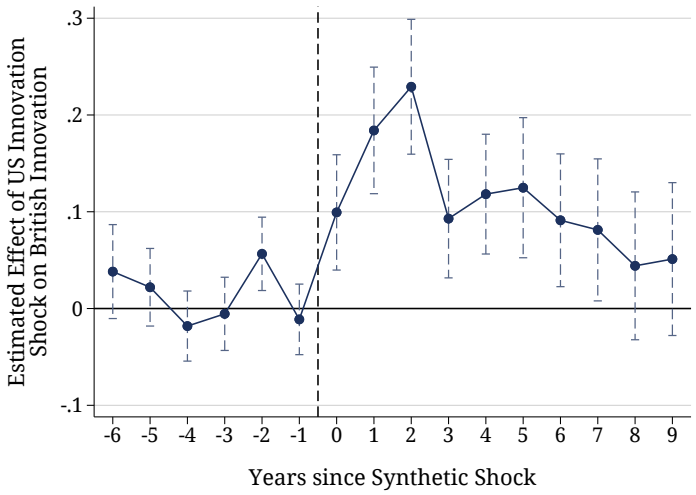
((b)) Great Influenza Pandemic Shock



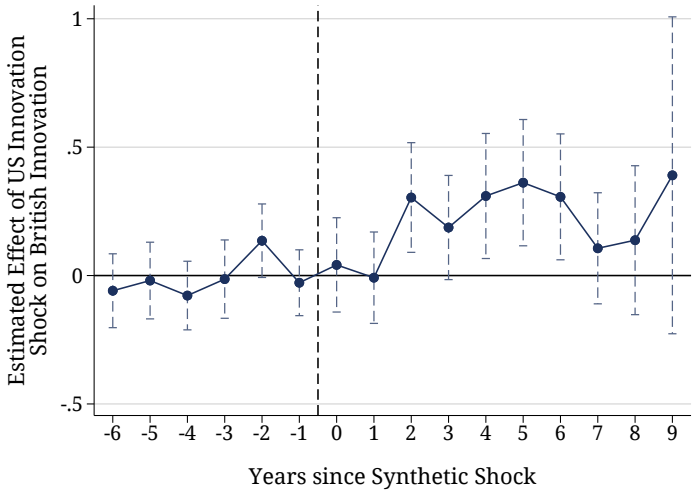
Notes. These figures report the dynamic treatment effects of synthetic shocks (Panel 3.5(a)) and the Great Influenza Pandemic shock (Panel 3.5(b)) on innovation. The units of observation are district-technology class pairs; units are observed at a yearly frequency in Panel 3.5(a) and at a biennial frequency in Panel 3.5(b) between 1900 and 1939. The dependent variable is the number of patents. The treatment is an indicator equal to one if: in Panel 3.5(a), a synthetic shock is observed in a given technology in at least one county where the district has above-median out-migration; in Panel 3.5(b), for pharmaceutical patents, emigration from a given district to counties in the top quartile of the mortality distribution is in the top quartile across districts. The black dashed line indicates the timing of the treatment. Standard errors are two-way clustered by district and technology class; bands report 95% confidence intervals.

Figure 3.6: Effect of Exposure to US Shocks on the UK-US Patent Similarity

((a)) Backward Similarity ("Copying")

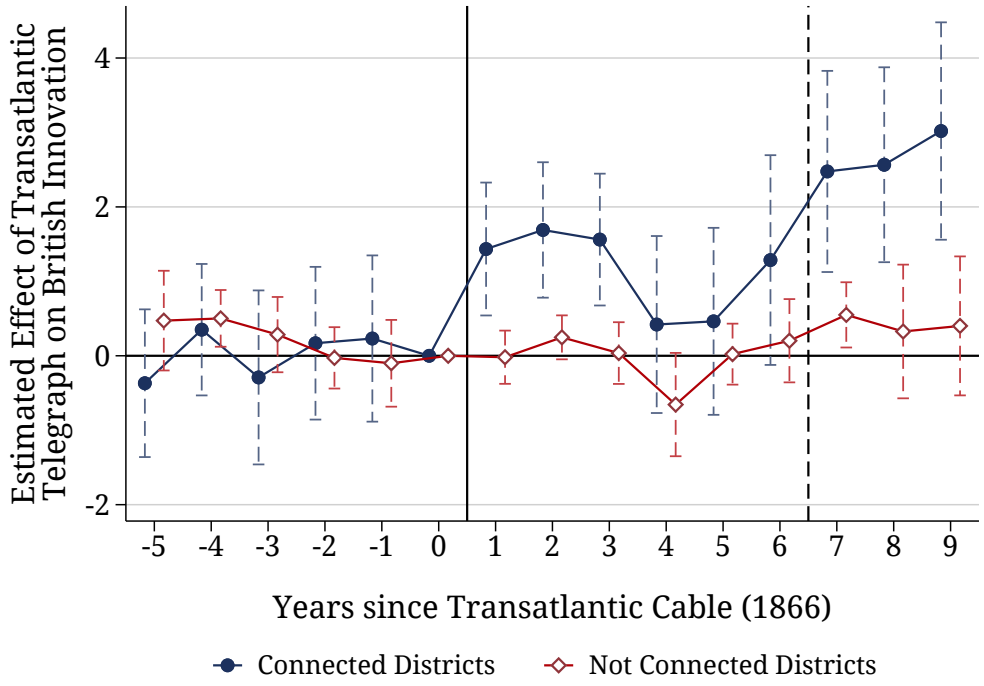


((b)) Excess Forward Similarity ("Originality")



Notes. These figures report the dynamic treatment effects of synthetic shocks on backward (Panel 3.6(a)) and excess forward (Panel 3.6(b)) similarity between UK and US patents. The units of observation are district-technology class pairs observed at a yearly frequency in 1900–1939. In Panel 3.6(a), the dependent variable is the text similarity between UK patents and US patents issued five years before ("copying"); in Panel 3.6(b), the dependent variable is the similarity of UK patents with US patents granted in the subsequent five years, over the similarity of UK patents with US patents granted in the preceding five years ("originality"). The similarity measure is akin to Kelly, Papanikolaou, Seru and Taddy (2021). The treatment is an indicator equal to one if a synthetic shock is observed in a given technology in at least one county where the district has above-median out-migration. Regressions include district-by-year, technology class-by-year, and district-by-technology class fixed effects. The black dashed line indicates the timing of the treatment. Standard errors are two-way clustered by district and technology class; bands report 95% confidence intervals.

Figure 3.7: Effect of the Transatlantic Telegraph Cable on Innovation



Notes. The figure displays the estimated dynamic treatment effect of the connection of the US and UK telegraph lines on innovation in the UK. The units of observation are districts observed at a yearly frequency between 1860 and 1875. The dependent variable is the number of patents. The independent variable is an interaction between the—time-invariant—number of emigrants in the 1870s and a posttreatment indicator that equals one after the transatlantic telegraph cable. Blue dots report the dynamic treatment effects on the sample of districts connected to the domestic UK telegraph network in 1862; red dots report those for the districts not connected to the network. The black solid vertical bar indicates the year the first cable was laid down (1866); the dashed black vertical line flags the year when the second and third cables were laid (1873-1874). Regressions include district and year fixed effects. Standard errors are clustered at the district level; bands report 95% confidence bands.

3.9 Tables

Table 3.1: Descriptive Statistics of Selected Variables

	(1) Observations	(2) Mean	(3) Std. Dev.	(4) Min.	(5) Max.
Panel A. Innovation					
Total Patents	5489	225.826	826.432	1	19789
Electricity Patents	5489	23.565	200.755	0	9430
Instruments Patents	5489	17.69	80.2	0	1850
Personal Articles & Furniture	5489	20.136	73.812	0	1548
Ships & Aeronautics	5489	16.463	55.062	0	1152
Transportation	5489	20.024	74.479	0	1923
Panel B. Emigration					
N. of US Emigrants	3779	61.765	91.36	0.303	1073.998
N. of Return US Emigrants	2494	35.342	52.202	0.064	730
Panel C. Census Tracts					
Population (1,000s)	3773	42.165	54.973	0.092	703.559
Share of Males (%)	3773	47.645	2.586	36.112	62.686
Share of Manufacture Empl. (%)	3773	13.213	6.306	2.569	42.723
Share of Agriculture Empl. (%)	3773	14.43	6.889	1.454	32.914
Share of Transportation Empl. (%)	3773	2.578	1.272	0	13.857
Share of Liberal Professions (%)	3773	1.679	0.65	0.43	6.873
Share of Public Servants (%)	3773	0.897	1.427	0	24.498
Panel D. Individual-Level Panel					
Share of Inventors	471013	0.009	0.094	0	1
N. of Patents	471013	0.018	0.356	0	87
N. of Patents if Inventor	4210	1.993	3.205	1	87
N. of Neighborhood Emigrants	471013	13.62	43.338	0	756
N. of Non-Return Neighborhood Emigrants	471013	12.979	40.888	0	512

Notes. This table displays summary descriptive statistics for a subset of the variables in the dataset. In Panels A, B, and C, variables are observed at the district level and at a decade frequency. In Panel D, the statistics are computed for individuals observed for twenty years around the 1891 and 1911 census years. An individual is labeled an inventor if they obtain at least one patent over this period. Panel A reports statistics for the top five most frequent technological classes. In Panels B and C, the underlying data are cross-walked to 1900 district borders.

Table 3.2: Effect of Exposure to US Technology on Innovation in Great Britain

	Ordinary Least Squares			Reduced Form			Two-Stages Least-Squares		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Dependent variable: Number of patents									
Knowledge Exposure$_t$	1.342***			0.037***			1.224***		
	(0.143)			(0.007)			(0.195)		
Knowledge Exposure$_{t-1}$		0.909***			0.015***			0.488**	
		(0.145)			(0.005)			(0.190)	
Knowledge Exposure$_{t-2}$			0.379***			-0.012			-0.398
			(0.112)			(0.014)			(0.478)
Mean Dep. Var.	10.392	13.345	15.256	8.706	12.045	15.314	8.708	12.049	15.319
Std. Beta Coef.	0.299	0.148	0.050	0.075	0.022	-0.013	0.296	0.088	-0.053
K-P F-stat							109.826	109.826	109.826
Panel B. Dependent Variable: Patents per capita (× 10,000)									
Knowledge Exposure$_t$	0.178***			0.004***			0.146***		
	(0.020)			(0.001)			(0.027)		
Knowledge Exposure$_{t-1}$		0.092***			0.002***			0.078***	
		(0.018)			(0.001)			(0.024)	
Knowledge Exposure$_{t-2}$			0.049***			0.000			0.001
			(0.015)			(0.001)			(0.043)
Mean Dep. Var.	2.066	2.629	2.973	1.748	2.345	2.980	1.747	2.346	2.980
Std. Beta Coef.	0.124	0.054	0.023	0.023	0.011	0.000	0.093	0.046	0.000
K-P F-stat							107.825	107.825	107.825
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	11268	11268	11214	11214	11214	11214	11214	11214
N. of Observations	67549	67549	56295	56070	56070	56070	56047	56047	56047

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. The main explanatory variable is knowledge exposure. In Panel A, the dependent variable is the number of patents; in Panel B, the dependent variable is the number of patents normalized by district-level population in 1880 and multiplied by 10,000 for readability. In columns (1–3), we estimate the OLS correlation with the observed measure of knowledge exposure; in columns (4–6), we estimate the reduced-form association with the railway-based instrument of knowledge exposure through OLS; columns (7–9) report the two-stage least-squares estimate. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.3: Effect of Exposure to US Innovation Shocks on UK Innovation

	Synthetic Shocks				Great Influenza Pandemic Shock			
	(1) Full	(2) No London	(3) No Lancs	(4) No S/W	(5) Full	(6) No London	(7) No Lancs	(8) No S/W
Synth.Shock×Post×Emigr.	0.434***	0.277***	0.578***	0.420***				
	(0.121)	(0.082)	(0.125)	(0.127)				
Pharma×Post×Emigr.					0.613***	0.417***	0.678***	0.461***
					(0.164)	(0.140)	(0.172)	(0.156)
District-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-by-Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Units	10029	9697	8760	9547	10727	10217	9384	10047
Number of Observations	393046	382153	343450	375850	429080	408680	375360	401880
Mean Dep. Var.	1.361	1.029	1.263	1.276	0.725	0.532	0.682	0.668

Notes. This table displays the effect of US innovation shocks on innovation activity in the UK. The unit of observation is a district-technology class pair observed at a yearly frequency between 1900 and 1939. The dependent variable is the number of patents. In columns (1–4), the independent variable is an indicator that, for a given district–technology, returns value one after a synthetic innovation shock in that technology class is observed in at least one county where the district has above-average out-migration. A synthetic innovation shock is observed whenever the residualized number of patents observed in the country is in the top 0.5% of the overall distribution. In columns (5–8), the independent variable is an indicator that returns value one for pharmaceutical patents only and only if emigration from the observed district to counties in the top quartile of the influenza mortality distribution is in the top quartile across districts. Both models should thus be interpreted as triple-difference designs. Since models in columns (1–4) are staggered designs, we estimate them using the imputation estimator developed by Borusyak, Jaravel and Spiess (2021). In columns (2) and (6), we drop districts in the London area; in columns (3) and (7), we exclude districts in the Lancashire area; in columns (4) and (8), we drop districts in the South-West area. Excluded regions are the first three in terms of patents granted. All models include district-by-year, district-by-technology class, and technology class-by-year fixed effects; standard errors, clustered two-way by district and technology class, are shown in parentheses.

*: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.4: Association Between UK Innovation and Exposure to US Technology Through Overall and Return Emigration

	Dep. Var.: Number of Patents					
	(1)	(2)	(3)	(4)	(5)	(6)
Knowledge Exposure$_t$	1.385***	1.347***	1.301***			0.697***
	(0.306)	(0.389)	(0.319)			(0.092)
Return Knowledge Exposure$_t$	0.306***	0.254***	0.311***			0.273**
	(0.075)	(0.084)	(0.084)			(0.108)
Knowledge Exposure$_{t-1}$				1.833***		0.060
				(0.489)		(0.444)
Return Knowledge Exposure$_{t-1}$				0.156***		0.170*
				(0.028)		(0.084)
Knowledge Exposure$_{t-2}$					0.416	0.139
					(0.243)	(0.436)
Return Knowledge Exposure$_{t-2}$					0.218	0.192*
					(0.131)	(0.099)
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes
Technology Class FE	Yes	–	–	Yes	Yes	Yes
Class-Decade FE	No	Yes	Yes	No	No	No
District-Class FE	No	No	Yes	No	No	No
N. of District-Class	11376	11376	11376	11375	11371	11340
N. of Observations	45464	45464	45464	34114	22742	22680
Mean Dep. Var.	9.869	9.869	9.869	12.175	15.871	15.907
Std. Beta (KE)	0.211	0.205	0.198	0.200	0.041	
Std. Beta (Return KE)	0.287	0.238	0.291	0.136	0.054	

Notes. This table reports the association between innovation and the baseline measure of knowledge exposure, accounting for return knowledge exposure. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1920. The dependent variable is the number of patents by district-technology decade. Return knowledge exposure is constructed by interacting county-level specialization with district-county return migration flows analogously to the baseline knowledge exposure measure. In columns (1) and (4–6), we include district-by-decade and technology class fixed effects. In column (2), we add technology-by-decade fixed effects; the specification in column (3) is saturated. Standard errors, clustered at the district level, are displayed in parentheses. The Table reports the standardized beta coefficient of both the baseline knowledge exposure term and the return knowledge exposure term. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.5: Effect of Family Member Emigration on Innovation Produced by Relatives of the Emigrant in the UK

	N. of Patents				I(N. of Patents > 0)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post × Relative US	0.059***				0.035***			
	(0.006)				(0.003)			
Post × Relative Return US		0.335***		0.332***		0.126***		0.124***
		(0.068)		(0.068)		(0.031)		(0.031)
Post × Relative Non-Return US			0.060***	0.059***			0.035***	0.034***
			(0.006)	(0.006)			(0.003)	(0.003)
County-Surname FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Surname-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of Surname-County	416735	416735	416735	416735	416735	416735	416735	416735
N. of Observations	25004100	25004100	25004100	25004100	25004100	25004100	25004100	25004100
Mean Dep. Var.	0.083	0.083	0.083	0.083	0.054	0.054	0.054	0.054

Notes. This table reports the effect of transatlantic emigration on innovation by inventors with the same surname as the emigrant. The unit of observation is a surname-county couple, observed at a year frequency between 1870 and 1929. The dependent variable in columns (1–4) is the number of patents granted to inventors with a given surname in a given county and year; the dependent variable in columns (5–8) is a categorical variable that takes value one if the number of patents is strictly positive. In columns (1) and (5), the treatment takes value one after at least one individual from a given county and with a given surname emigrates to the US, and zero otherwise; in columns (2) and (6), we restrict to emigrants that at some point return; in columns (3) and (7), we restrict to emigrants that never return; in columns (4) and (8) we horse-race the two latter treatments. Each regression includes county-by-surname, surname-by-year, and county-by-year fixed effects. Standard errors, clustered at the surname level, are displayed in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.6: Effect of Emigration on Innovation Produced by Former Neighbors of the Emigrant in the UK

	Baseline Sample				Dropping Individuals in...		
	(1)	(2)	(3)	(4)	(5) London	(6) Lancashire	(7) South-West
Panel A. All Emigrants							
Neighborhood Emigrant × Post	0.167***	0.180***	0.170***	16.208***	0.130**	0.180***	0.198***
	(0.053)	(0.055)	(0.056)	(5.758)	(0.061)	(0.055)	(0.061)
Std. Beta Coef.	0.022	0.024	0.023	0.211	0.018	0.025	0.025
Panel B. Only Non-Return Emigrants							
Non-Return Neighborhood Emigrant × Post	0.165***	0.189***	0.167***	15.749***	0.108*	0.183***	0.211***
	(0.054)	(0.056)	(0.057)	(5.987)	(0.060)	(0.056)	(0.062)
Std. Beta Coef.	0.021	0.024	0.021	0.196	0.014	0.024	0.026
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	–	Yes	Yes	Yes	Yes	Yes
Parish × Year FE	No	Yes	No	No	No	No	No
Matching	No	No	Yes	No	No	No	No
Sample	Full	Full	Full	Inventors	Full	Full	Full
N. of Individuals	473112	473112	469585	4224	410327	422230	352064
N. of Observations	9462240	9419787	9391700	84480	8206540	8444600	7041280
Mean Dep. Var.	0.890	0.892	0.893	99.716	0.794	0.836	0.893
S.D. Dep. Var.	40.291	40.324	40.351	414.695	37.439	39.126	41.333

Notes. This table reports the effect of neighborhood out-migration on innovation. The units of observation are individuals who are observed yearly between 1900 and 1920. In columns (1–3) and (5–7), the sample consists of the universe of males who did not emigrate over the period and that were at least 18 years old in 1900; in columns (4) and (8), we restrict the sample to inventors. The dependent variable is the number of patents obtained annually. In columns (1–4), the sample consists of individuals residing in all England and Wales divisions; in columns (5–7), we exclude the top tree-patents producing areas: London, Lancashire, and the South-West. In Panel A, the independent variable is an indicator that, for a given individual, returns value one after at least one person that was living in the same neighborhood as the individual migrates to the United States; in Panel B, we restrict to emigrants that never return in the period of observation. In this context, "neighborhood" refers to the same street, square, or similar. We explore an alternative distance-based definition in Appendix Table 3.38. Each model includes individual and—at least—year fixed effects; in column (2), we include parish-by-year fixed effects; in column (3), individuals are weighted by their coarsened exact matching weight. The estimates are obtained using the method discussed in Borusyak, Jaravel and Spiess (2021) to account for the staggered roll-out of the treatment across individuals. Standard errors, clustered at the district level, are reported in parentheses.*: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.7: The Transatlantic Telegraph Cable and Innovation in the UK

	Double Differences			Triple Differences		
	(1) All	(2) Connected	(3) Not Connected	(4) All	(5) Connected	(6) Not Connected
Post × Emigrants	1.345***	1.639***	-0.083			
	(0.451)	(0.559)	(0.097)			
Post × Knowledge Exposure				0.027**	0.027**	-0.003
				(0.010)	(0.011)	(0.005)
District FE	Yes	Yes	Yes	–	–	–
Class FE	Yes	Yes	Yes	–	–	–
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
District-Class FE	–	–	–	Yes	Yes	Yes
N. of District-Class	631	463	168	631	463	168
N. of Observations	10096	7408	2688	181728	133344	48384
Mean Dep. Var.	5.610	7.241	1.115	0.312	0.402	0.062
Std. Beta Coef.	0.114	0.125	-0.039	0.035	0.034	-0.007

Notes. This table displays the estimated effect of the connection of the US and UK telegraph lines on innovation in the UK. The units of observation are districts in columns (1–3) and district-technology class pairs in columns (4–6). Units are observed yearly between 1860 and 1875. The dependent variable is the total number of patents granted. In columns (1–3), the independent variable is an interaction between the—time-invariant—number of US emigrants in the 1870s and an indicator variable that returns value one after the transatlantic cable successfully connected the UK and the US in 1866, and zero otherwise; in columns (4–6) the treatment interacts—time-invariant—knowledge exposure in the 1870s with the same posttreatment indicator. In columns (1) and (4), the sample includes all districts; in columns (2) and (5) (resp. 3 and 6), we restrict the estimation to districts that were (resp. were not) connected to the domestic UK telegraph system. Models (3) and (6) should be interpreted as placebo exercises. Regressions include fixed effects for district and year in columns (1–3) and district-by-class and year in columns (4–5). Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.8: Effect of US Emigration on Newspaper Coverage of US-related News

	Dependent Variable: Number of Newspaper Mentions								
	OLS			Reduced Form			2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. US-Wide Coverage									
US Emigrants	6.753***	6.632***	7.207***				24.396***	24.228***	25.061***
	(0.958)	(1.006)	(0.600)				(1.570)	(1.611)	(0.912)
$\widehat{\text{US Emigrants}}$				1.451***	1.440***	1.501***			
				(0.121)	(0.124)	(0.078)			
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Dep. Var.	1017.635	1017.635	2276.989	1017.635	1017.635	2276.989	1017.635	1017.635	2276.989
Panel B. State-Wide Coverage									
US Emigrants	1.050***	1.049***	1.103***				10.060***	10.061***	10.091***
	(0.095)	(0.096)	(0.052)				(0.428)	(0.430)	(0.458)
$\widehat{\text{US Emigrants}}$				0.038***	0.038***	0.039***			
				(0.001)	(0.001)	(0.001)			
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Dep. Var.	57.486	57.486	127.984	64.672	64.672	136.660	64.672	64.672	136.660
Panel C. County-Wide Coverage									
US Emigrants	1.120***	1.120***	1.217***				4.861***	4.863***	5.130***
	(0.148)	(0.148)	(0.079)				(0.471)	(0.460)	(0.291)
$\widehat{\text{US Emigrants}}$				0.055***	0.055***	0.058***			
				(0.006)	(0.005)	(0.004)			
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Dep. Var.	0.003	0.003	0.007	0.004	0.004	0.008	0.004	0.004	0.008
N. of Newspapers	No	Yes	No	No	Yes	No	No	Yes	No
Districts in Sample	All	All	w/News	All	All	w/News	All	All	w/News
N. of Districts	602	602	321	602	602	321	602	602	321

Notes. This table displays the effect of out-migration on newspaper coverage of emigrants' destinations. The observation unit is: in Panel A, a district; in Panel B, a district-US state pair; in Panel C, a district-US county pair. Units are observed at a decade frequency between 1880 and 1930. The dependent variable is the number of articles mentioned: in Panel A, "United States"; in Panel B, US states; in Panel C, US counties. The independent variable is: in Panel A, the number of US emigrants; in Panel B, the district-state emigrants; in Panel C, the district-county emigrants. Models (1–3) estimate the model through OLS; models (4–5) report the reduced-form association between mentions and the out-migration instrument; models (7–9) report the two-stage least squares estimates. Regressions include district fixed effects and: in Panel A, decade fixed effects; in Panel B: state-by-decade fixed effects; in Panel C: county-by-decade fixed effects. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

3.10 Appendix: Data Sources and Methods

This section describes the sources and methods we adopted to assemble and merge the various datasets that underlie the empirical analysis. We defer a more detailed discussion of the novel patent data that we digitize, and the linked international migrants sample to sections 3.11 and 3.12, respectively.

3.10.1 Summary of Data Sources

Patent Data US patent data are from Berkes (2018), who digitizes the universe of patents granted between 1836, when the US patent and trademark office was established, and 2010. In this paper, we are interested in the CPC technology class, the issue year, and the coordinates of residence of each inventor. We then assign each patent to US counties at 1900 borders. Depending on the number of inventors, a single patent may be assigned to multiple counties. In the case of patents with multiple inventors, we weigh each by the inverse of the number of inventors to avoid multiple counting. English and Welsh patents after 1900 are available at the European patent office. To construct our dataset, we leverage bulk access to the PATSTAT dataset. Information contained in PATSTAT includes the CPC class and the issue year. To retrieve the location of each inventor, we merge the PATSTAT data with the PatCity repository, which contains geo-coded information on the universe of English and Welsh patents during this period (Bergeaud and Verluise, 2022). Data before 1900 are not available. In section 3.11, we describe how we digitize the universe of patent documents issued over the period 1853–1900 to fill this substantial gap. Importantly, we map 3-digit CPC classes to a coarser taxonomy of classes. To do that, we reduce them to functional units using the CPC classification scheme. The scheme is publicly available at the following `link`. To accommodate the historical context, we divide the transporting categories into two classes: "Transporting", which includes carriages, railways, and cars, and "Ships and Aeronautics". Moreover, we conflate the "Weapons and Blasting" and the "Mining" classes into the "Metallurgy" category because few patents were observed in those industries. We further augment patent

data by defining a measure of "quality" or "innovativeness" following Kelly, Papanikolaou, Seru and Taddy (2021). This metric flags as influential those patents that introduce terms that were not used before they were granted, and become common thereafter. We evaluate this metric on the abstracts of patents granted after 1900. We apply this sample restriction for consistency: in the period 1880–1899, in fact, we observe the full text of patents instead of their abstract.

Migration Data Disaggregated data on the origin of English and Welsh immigrants—and, more generally, of all other nationalities—do not exist. These we collected neither by receiving US authorities nor by sending UK offices. We thus lack precise information on where British immigrants in the US came from *within* the UK. We fill this gap and link the individual-level UK and US censuses, as described in 3.12. Ideally, we observe the universe of British emigrants to the United States between 1870 and 1930. For those individuals, we know all information contained in the US Census and those detailed in the UK one. Most notably, we know where they came from. As we discuss more in detail later, we also link return migrants. Since the last publicly available UK census dates to 1911, however, we can only construct return migration flows over the period 1870–1910.

Population Census The main data sources we leverage are the individual-level non-anonymized UK and US population censuses. The US census features prominently in the economic history literature as a major source of detailed microdata, and we thus avoid discussing it any further (Ruggles, Fitch, Goeken, Hacker, Nelson, Roberts, Schouweiler and Sobek, 2021). The UK census is relatively less well-known (Schurer and Higgs, 2020). Although not as rich as its US counterpart, the UK population census covers individuals who have resided in the UK since 1850. The first census was run in 1841, but only 1851, 1861, 1881, 1891, 1901, and 1911 are completely digitized.[56] Data in the census include the name and surname, birth year, division, county, district, parish, precise address of residence, the specific occupation detailed through HISCO codes, and other variables that we do not use in the paper, such as the type of dwelling and fertility information.

[56]The 1921 census is currently being digitized and is partially available. We do not use it because its coverage is still not complete and because it is not available in bulk. All censuses after 1921 are subject to privacy restrictions.

We augment these variables by geo-coding the universe of addresses that appear in the census to precise geographical coordinates, as detailed in section 3.10.2.

Newspapers We collect data on newspaper coverage of US-related news from the British Newspaper Archive.[57] Beach and Hanlon (2022) describe this dataset in detail. In this paper, we run a set of three queries. First, we search for the words "United States". Second, we perform fifty searches, one for each state. Finally, we perform approximately three thousand searches, one for each county. Each search spans the period 1850–1939. We collect the information at the article level. For each entry in the database, we know the journal, day, month, and year of publication, whether it is an article or some other type of content—e.g., an obituary–, the page, and the word count. Importantly, we collect information on the universe of newspapers in the archive. Journal-level data contain the publishing address at the city level, the first and last day, month, and year of activity, and the publication frequency—e.g., quarterly, daily. We then geocode each newspaper to the coordinates of the city where it was published and map those to 1891 registration districts. We can thus construct a measure of newspaper coverage at the district-year level.[58] In Table 3.9, we provide a set of summary statistics on the resulting dataset. We collect information for a total of 2022 newspapers: of these, 1459 are based in England, and 93 are published in Wales. We exclude Scottish and Irish newspapers from the analysis. The average life of a newspaper in this period is 40 years. In Panel B, we report district-level statistics by decade. The number of newspapers decreases over time, as noted by Beach and Hanlon (2022), from an average of 2.3 newspapers per district in the 1870s to 0.7 in the 1930s. It is unclear whether this is due to incomplete coverage in the later period. In Panel C, we report the district-level statistics by division and find that, except for the London division, newspapers appear to be quite sparse across the country. Figure 3.10 displays the spatial distribution of the number of newspapers across districts over the period and confirms the impression that newspapers tend to evenly cover a substantial share of districts. London stands as a major outlier: we thus perform

[57] A limited free-tier access to newspaper data is available at the following `link`.

[58] Unfortunately, for newspapers based in London, we only know their city, i.e., London. In the newspaper analysis, we are thus forced to conflate all urban London districts into a single "London" geographical unit.

all exercises dropping London districts and find consistent results.

Miscellaneous To construct the domestic UK telegraph network prior to the first transatlantic UK–US cable (1866), we digitize the list of telegraph stations reported in the *Zeitschrift des Deutsch-Österreichischen Telegraphen-Vereins, Jahrgang*, volume IX, 1862. This directory lists the universe of telegraph stations outside of London in 1862. To the best of our knowledge, it is the most complete directory prior to the introduction of the transatlantic cable. We geo-code each station to precise coordinates. The red dots in Figure 3.11 report each station. We then label each district with at least one telegraph station as "connected" to the domestic network and as "not connected" otherwise.

We construct US county-level exposure to the Great Influenza pandemic using mortality statistics collected by the US Bureau of Census. These data are available for a subset of counties representing approximately 60% of the US population in 1900.

To compute the railway-based instrument, we construct US-county level immigration shocks following the methodology described in Sequeira, Nunn and Qian (2020). We use the same data sources. Hence we defer the interested reader to their paper for a more detailed discussion.

We digitize import and export yearly data from the 1935 edition of the *Statistical Abstract of the United States*.[59] In particular, we collect the yearly tariff rates applied between 1925 and 1929, i.e., before the Smoot-Hawley Act, and between 1930 and 1935, i.e., after the Act. Tariff rates are available by sector. We then map each industry to a technology class, as listed in Table 3.11. In the baseline analysis, we consider an industry protected if its tariff rate increases by more than 50% between 1925-1929 and 1930-1935. We consider alternative thresholds as robustness checks.

GIS Shapefiles & Boundary Harmonization Patents and telegraph stations are mapped to 1900 registration district borders using historical GIS files and their coordinates.[60] However, all data from the population censuses appear at historical borders.

[59]This publication is freely available at the following `link`.

[60]`GIS` data for the US are provided by `NHGIS`, whereas district boundaries have been digitized by the Great Britain Historical `GIS` Project.

Registration districts do not undergo major boundary changes over the period that we study. However, we adapt the method presented by Eckert, Gvirtz, Liang and Peters (2020) to UK districts to ensure that we work with consistent geographical units. To construct geographical crosswalks using their method, one needs to assume that variables are evenly distributed over the area of geographical units. The crosswalk is then obtained by overlapping geographical units over time. Suppose unit x in decade d is split, and 80% of its territory is assigned to itself, while 20% is assigned to another district y. To construct a cross-walk relative to period $d + t_2$ for a generic variable between decades $d - t_1$ and $d + t_2$, for $t_1, t_2 > 0$, one needs to multiply the variable measured in district x in $d - t_1$ by 4/5, and add 1/5 of the variable in x to that measured in y in the same decade. We map registration districts to their boundaries in 1901. Less than 5% of the overall area of England and Wales is re-assigned in this way. We adopt the same methodology to map counties to their 1900 borders.

3.10.2 Geo-referenced Census Records

A notable feature of the UK census is that it contains precise information on the residential address of the universe of British population. This information is extremely valuable because, in principle, it assigns the finest possible location to each individual. In practice, however, it is highly non-standardized and challenging to use. In this section, we discuss the methodology that we apply to assign geographical coordinates to textual addresses. This dataset expands earlier work by Lan and Longley (2019), who adopt a different strategy and only analyze the 1901 census, whereas we geo-reference the entire 1851-1911 censuses. Furthermore, the geo-coded census sample is used in the individual-level analysis only. All other exercises do not rely on these data.

Methodology

There are two ways to geo-reference historical addresses. One approach is to manually digitize historical locations, either streets or enumeration units, from historical maps.

However, this method does not scale up and becomes rapidly unfeasible as the data grows. A second automated approach is to run text-based address matching between historical data sources and address databases that have already been geo-referenced. We follow this latter method since we need to geo-reference 5,464,578 unique addresses.

To implement the latter approach Lan and Longley (2019) exploit open-source address data from `OpenStreetMaps`. In this paper, instead, we take advantage of the commercial geo-referenced database developed by `MapTiler AG`. This has three key benefits compared to `OpenStreetMaps`-powered engines. First, the data has some historical "depth", meaning that historical names of locations are sometimes recorded. Second, `MapTiler AG` provides a flexible address-correction engine that matches the query to the closest address available in their dataset. Finally, this commercial database has better coverage than `OpenStreetMaps` in rural areas.

To perform the actual matching, we first operate a preliminary manual trimming of addresses. First, we remove house numbers because they undergo many changes and re-sequencing over time. Second, we remove uninformative locations, such as "village", "farm", and "rectory". Then, we input the resulting addresses as queries into the geo-referencing engine. Crucially, we discard the match if the resulting coordinates are not within the parish's boundaries where the address is recorded. This consistency check is necessary because homonyms are frequent. Since observing two addresses with the same name within a given parish is extremely rare, this ensures that the algorithm matches are not spurious.

Matching Performance

In Figure 3.9, we report the distribution of the share of geo-referenced addresses by district and census decade. The blue bars refer to the simple matching rate, defined as the share of geo-referenced addresses. The black-contoured bars, instead, adjust for the number of residents recorded in each address. In each figure, we report the average matching rates and their respective standard deviations. The average matching rate ranges between 76% in 1851 and 86% in 1911. All distributions display substantial right-skewness, meaning

there are very few districts with a matching rate lower than 50%. The matching rate increases over time for two reasons. First, the quality of recorded addresses increases in more recent censuses. Second, the urban geography in 1911 is more similar to that in the `MapTiler AG` database than in 1851. This is due to street re-labeling and urban agglomerates' growth and consolidation. Figure 3.8 displays the spatial distribution of the average geo-referencing rates across censuses. Figure 3.8(a) reports the crude rate, whereas Figure 3.8(b), we adjust by address-population. Except for Wales and some rural districts at the center of England, the geo-referencing rates are above 80% everywhere. It is particularly high—above 90%—in North-Western and South-Eastern England. More urbanized areas generally tend to feature larger geo-referencing rates because addresses tend to be more informative. This notwithstanding, differences are quantitatively small as the matching rate is remarkably homogeneous across registration districts. Wales is the single most relevant exception. The geo-referencing rate there is very low because addresses in the census until 1901-1911 tend to be reported in Welsh, especially in Western areas.

Taken together, the results of the geo-referencing algorithms are satisfactory. More than 80% addresses are successfully matched to precise geographical coordinates. This ratio is even higher in areas outside Wales, where innovation and migration activity are more intense.

3.10.3 Linked Inventor Sample

This section presents the methodology we use to link patents to census records. The linked inventors-census sample is used in the individual-level analysis only. All other exercises do not rely on these data.

Methodology

We follow the logic of Berkes (2018), who links patents to census records in the US. We link patents between 1881 and 1899 to the 1891 census and those between 1901 and 1920

to the 1911 census. Relative to our baseline sample, we thus drop patents issued after 1920 because we cannot observe individuals born after 1911. While this is probably a minor issue for patents granted until 1930, it may induce some selection of linked inventors for later patents. Patent data contain the name and surname of inventors, their residence, and the issue year.

Given a patent p, define the set of inventors as $\mathcal{A}_p = \{A_1, \ldots, A_{n_p}\}$. Most patents are solo-authored in this period, meaning $|\mathcal{A}_p| = 1$. Call $\mathcal{L}_p = \{\ell_1, \ldots, \ell_{m_p}\}$ the set of locations patent p is associated to. Each ℓ is a couple of latitude-longitude coordinates. Let $\mathcal{L}_p^{\text{parish}}$ be the set of parishes associated with each coordinate. Analogously, let $\mathcal{L}_p^{\text{district}}$ and $\mathcal{L}_p^{\text{county}}$ be the set of, respectively, districts and counties where each coordinate locates. Notice that these are progressively coarser units: parishes are contained in districts, which form counties. Unfortunately, we do not know the inventor-location pair. To match the generic A_p, we thus perform the following operations:

1. With a slight abuse of notation, let $\mathcal{L}_p^{\text{parish}}$—and, analogously, $\mathcal{L}_p^{\text{district}}$ and $\mathcal{L}_p^{\text{county}}$—denote the set of census records in each parish, district, and county within the respective sets.

2. Take all entries i within the set of parishes $\mathcal{L}_p^{\text{parish}}$ that are at least 18 when the patent p is filed. Let year_i and t_p respectively denote the birth year of i and the issue date:
$$\mathcal{M}_{A_p}^{\text{parish}} = \left\{ i \in \mathcal{L}_p^{\text{parish}} \middle| \, t_p - \text{year}_i \geq 18 \right\} \tag{3.12}$$

3. For each $i \in \mathcal{M}_{A_p}^{\text{parish}}$, compute the distance between the name and surname of i, and that of A_p:
$$\text{Similarity}_i^{A_p} = \alpha \times \text{Name Similarity}_i^{A_p} + (1 - \alpha) \times \text{Surname Similarity}_i^{A_p} \tag{3.13}$$
for some $\alpha \in [0, 1]$. In our baseline setting, we pick $\alpha = .3$ to assign a larger weight to the surname.

4. Define the set of acceptable matches as those with the highest similarity with the

given A_p:

$$\overline{\mathcal{M}}_{A_p}^{\text{parish}} = \left\{ i \in \mathcal{M}_{A_p}^{\text{parish}} \middle| \text{Similarity}_i^{A_p} = \max_{i' \in \mathcal{M}_{A_p}^{\text{parish}}} \text{Similarity}_{i'}^{A_p} \right\} \tag{3.14}$$

and define Similarity^{A_p} as the similarity between all elements in $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$ and A_p. Notice that this is the same across all $i \in \overline{\mathcal{M}}_{A_p}^{\text{parish}}$.

5. Set a threshold τ such that if $\text{Similarity}^{A_p} < \tau$, $\overline{\mathcal{M}}_{A_p}^{\text{parish}} = \emptyset$, otherwise pass.

6. If $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$ is not empty, then inventor A_p is matched to all records in $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$. If it is empty, repeat steps 2–4 conditioning on records in $\mathcal{L}_p^{\text{district}}$. If $\overline{\mathcal{M}}_{A_p}^{\text{district}}$ is empty, repeat steps 2–4 conditioning on records in $\mathcal{L}_p^{\text{county}}$. If $\overline{\mathcal{M}}_{A_p}^{\text{county}}$ is empty, repeat steps 2–4 without imposing geographical conditions on records i. In the baseline setting, we only accept county-level and country-level matches if the name and surname of the match(es) exactly match A_p's.

Patent data have the clear advantage that we have geographical information on the location of inventors. Inventors are mobile, however, and there may be a considerable time between the moment the patent is granted and the 1911 census. For these reasons, we incrementally exploit geographical information on the inventor's location. First, we look for high-quality matches within the same parish where the patent is filed. Parishes are small, as their average population is less than 10,000. When a match at the parish level is feasible, it is usually unique. We then progressively expand the set of records by coarsening their geographic location. Districts are larger than parishes, and counties are, in turn, larger than districts. If we cannot find one match at the county level, we look for one within the entire population of England and Wales. Unlike the migrants sample, we do not have information on the birth year. To ensure that county- and country-level matches are reliable, we require that their name and surname are verbatim those recorded in the patent document.

Matching Statistics

In Figure 3.12, we report the matching rate of this exercise. We focus on two matching rates: the gross rate is the share of inventors that have at least one match, relative to the overall set of inventors; the net rate is the share of inventors with at least one *acceptable* match, relative to the overall group of inventors. In the analysis, a match is acceptable if (i) the similarity between name and surname is above 0.95 and (ii) a given inventor has no more than five matches. Panel 3.12(a) reports both margins over time. The gross matching rate remains consistently above 80% throughout the period. The net matching rate, however, rejects approximately 20% of the matches. This is mainly due to inventors linked to more than five census records. This notwithstanding, the share of acceptable matches is approximately constant and above 60% each year. Our algorithm delivers satisfactory performance compared to standard linking rates in the literature. In panel 3.12(b), we break down the number of matches by the geographical unit where the match is attained. Blue, red, green, and yellow bars report the matching rates at the parish, district, county, and national levels. The share of inventors matched with more than 20 census records is larger at the national level; there, we look for possible matches with no information on the residence. Multiple matches are somewhat common at the parish level as well. This is because we first try to match inventors at the parish level. Hence parish matches represent the large majority of the linked sample, while district-level matches are residual and, thus, more accurate. Figure 3.13 displays the spatial distribution of inventors, who are plotted using the geo-coded census coordinates described in the previous section.

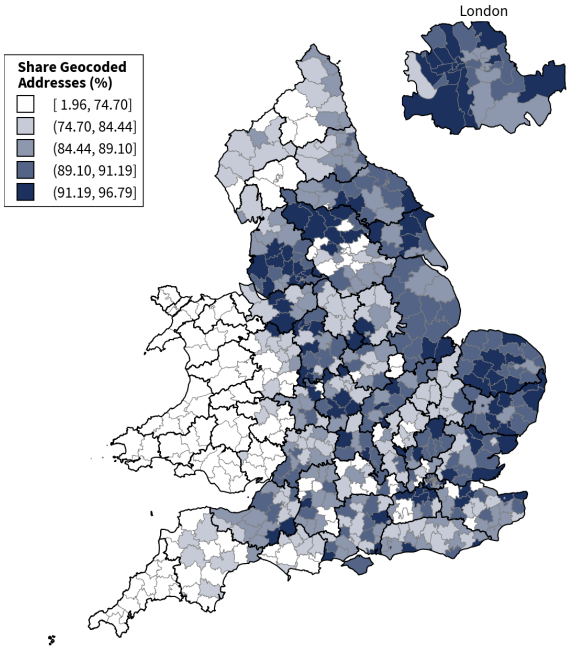
A plausible concern is that the probability of obtaining a link is not random. This may be the case if, for instance, more successful inventors were more educated and, hence, more likely to report their names correctly in the census. On the other hand, if successful inventors were relatively more mobile, we may fail at linking them because we may need to go national to obtain a match, which would most likely be dropped because of the multiple-match issue. While these hypotheses are ultimately challenging to test, in Table 3.10, we compute the correlation between the number of matches in our sample and a

set of individual observed characteristics. In Panel A, we have age; in Panel B, we list the set of occupational categories; in Panel C, we list the residence divisions. We find no clear association between the number of matches and these variables in the overall sample (column 1) and across matches selected by geographical layer (columns 2–5). Overall, we interpret the Table as conveying reassuring evidence that the selection of inventors into the linked sample does not appear to systematically favor particular groups.

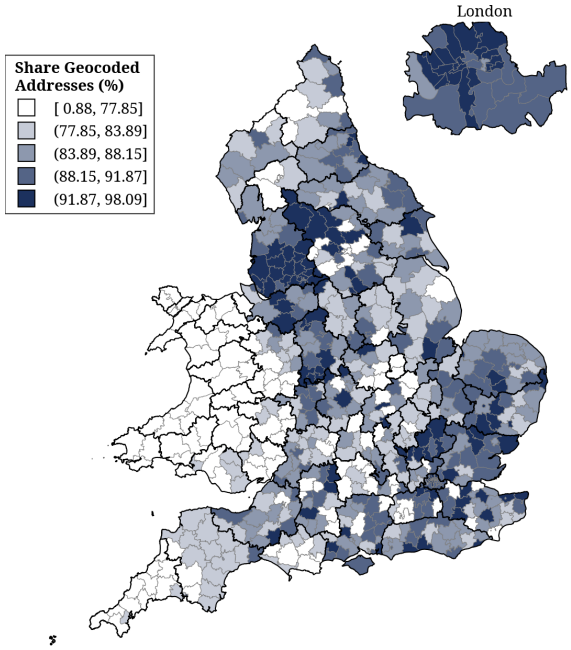
3.10.4 Figures

Figure 3.8: Spatial Distribution of the Share of Geo-coded Addresses in the UK Population Censuses, 1851–1911

((a)) Unweighted Geo-coding Ratio



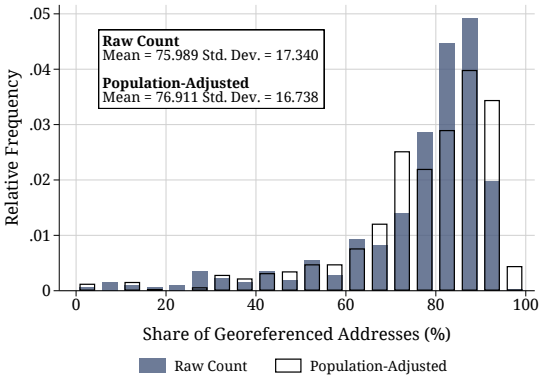
((b)) Population-Weighted Geo-coding Ratio



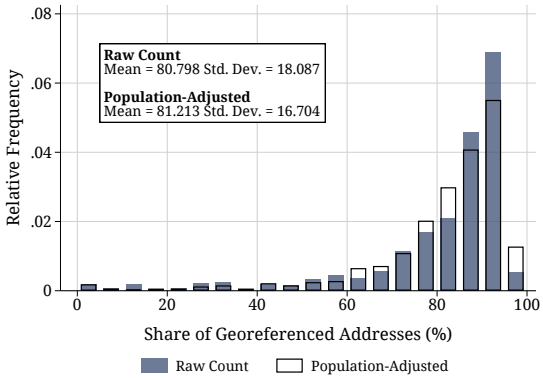
Notes. These figures report the spatial distribution of the share of geo-referenced addresses from the UK censuses, 1851–1911. For each census, we obtain a list of more than five million addresses by fine geographical unit (i.e., parishes). We then geo-reference these addresses to precise geographical coordinates. Panel 3.8(a) reports the district-level share of successfully geocoded addresses. In Panel 3.8(b), we weigh each address by the number of people reported to live in that address. The performance of the geo-referencing algorithm is relatively poor in Wales because addresses there are often reported in Welsh.

Figure 3.9: Distribution of the Share of Geo-coded Addresses by Census

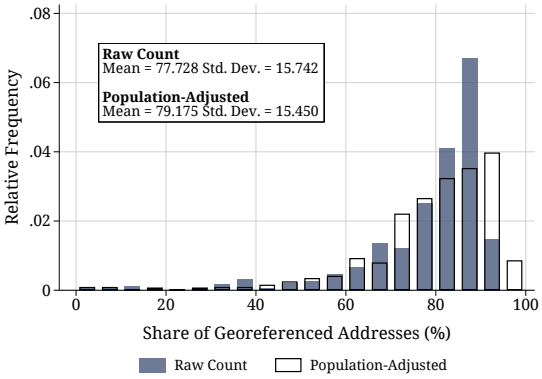
((a)) 1851 Census



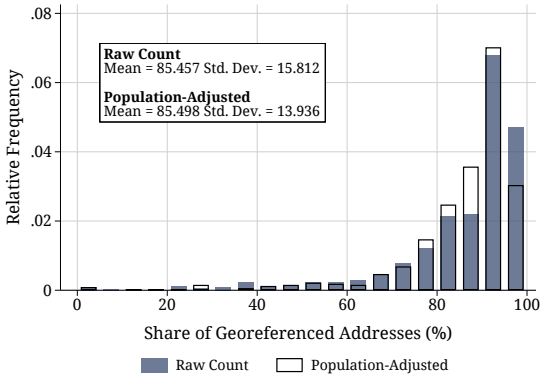
((b)) 1861 Census



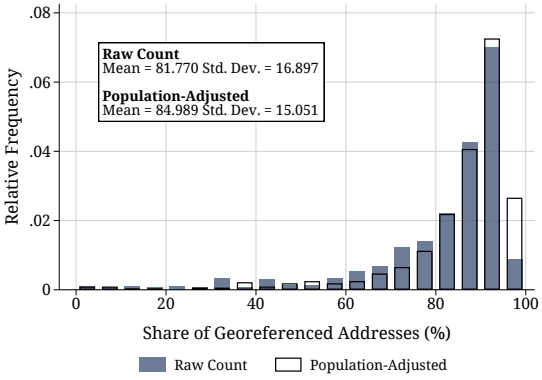
((c)) 1881 Census



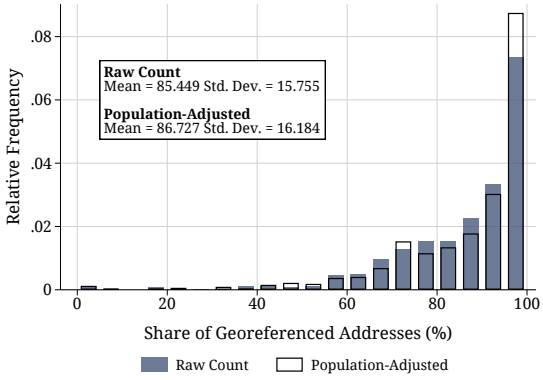
((d)) 1891 Census



((e)) 1901 Census

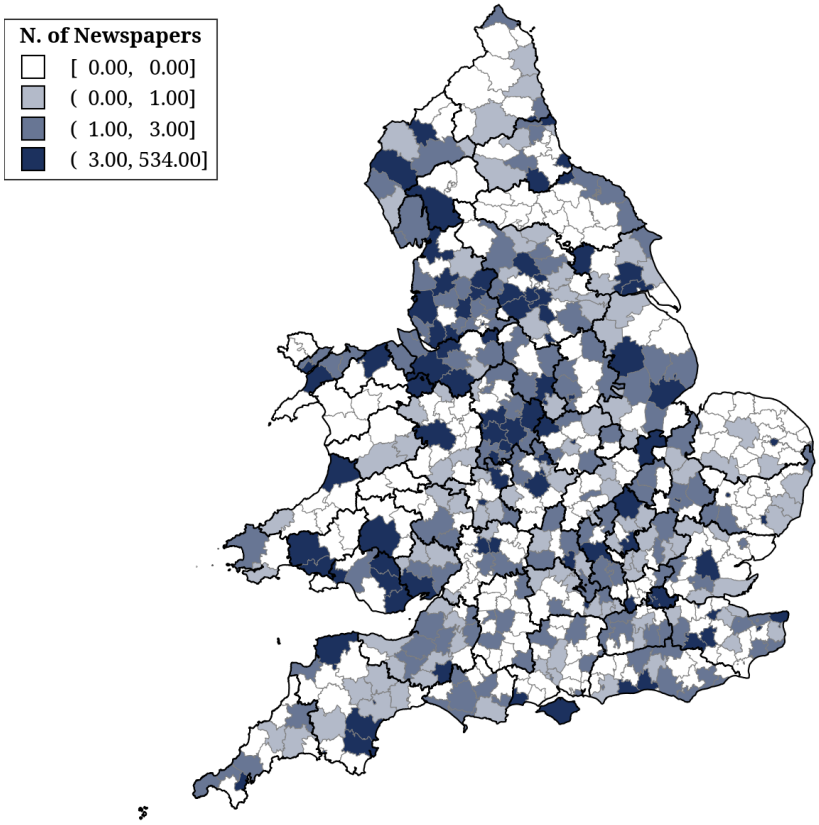


((f)) 1911 Census



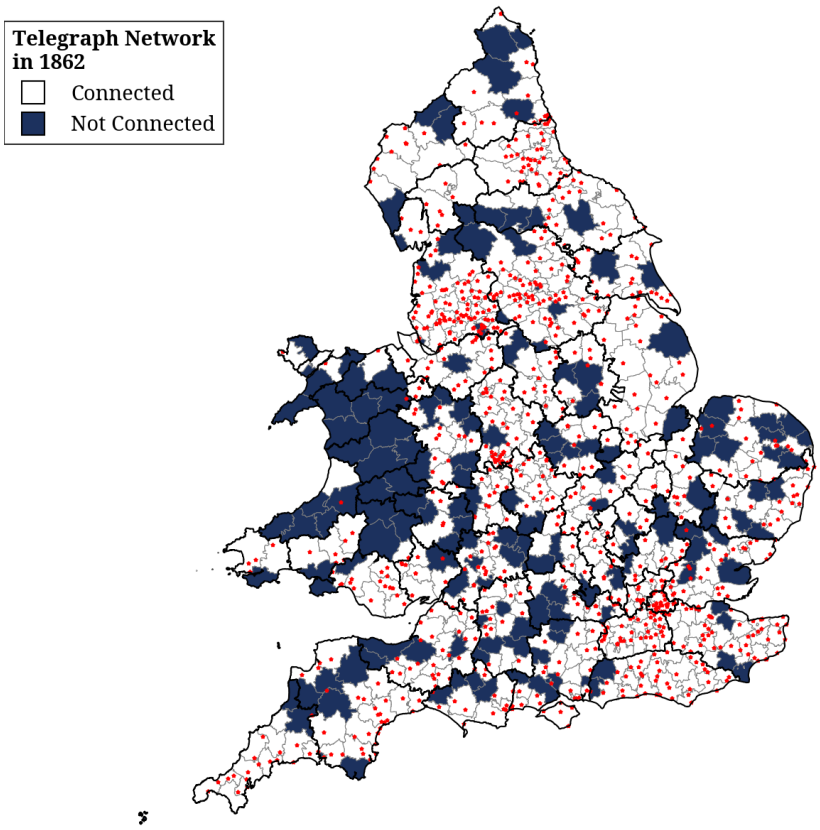
Notes. These figures display the district-level distribution of the share of geo-coded addresses from the UK censuses (1851–1911) by decade. For each census, we obtain a list of more than five million addresses by fine geographical unit (i.e., parishes). We then geo-reference these addresses to precise geographical coordinates. The black-contoured bars report the crude geo-coding rate; the blue bars report the population-adjusted geo-coding rate. Each figure reports the average and standard deviation of the two distributions.

Figure 3.10: Number of Active Newspapers Over the Period 1880–1940, by District



Notes. This figure reports the spatial distribution of the number of active newspapers across districts over the period 1880–1940. To be included in the data, a publication must be active for at least one year between 1880 and 1940. To retrieve the location of each journal, we geo-reference its publishing address and overlay historical district boundaries to assign it to consistent 1900 districts. The publishing address only lists the city. Hence we cannot distinguish across the eleven London urban districts. We consequently dissolve these districts into a single "London" unit.

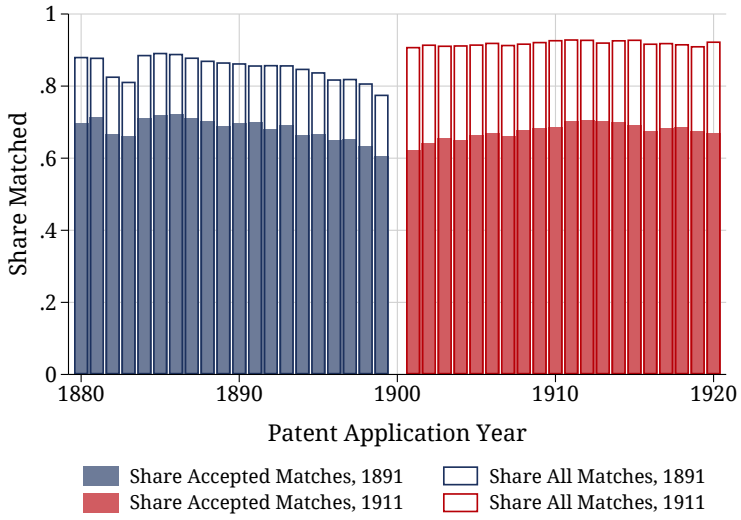
Figure 3.11: Distribution of Districts Connected to the UK Telegraph Network in 1862



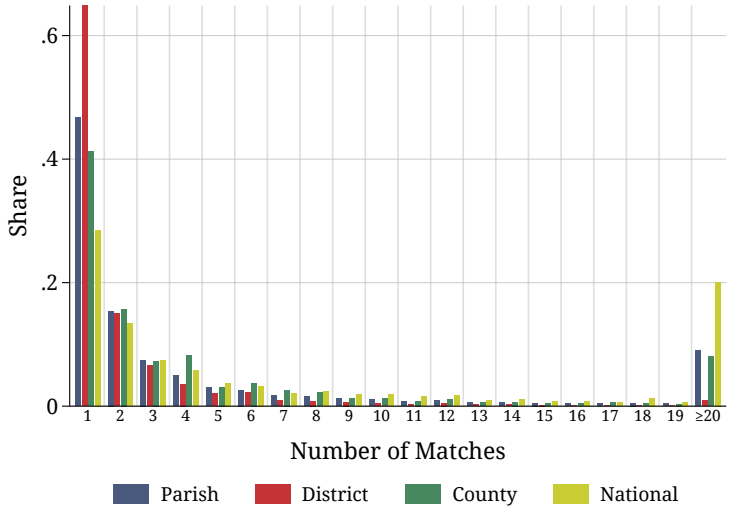
Notes. This figure reports the spatial distribution of telegraph stations across districts in 1862. Red markers display the location of telegraph stations. Districts without any telegraph station are displayed in dark blue. To retrieve the coordinates of each telegraph station, we geo-reference the city where it is located. The list of telegraph stations is taken from the *Zeitschrift des Deutsch-Österreichischen Telegraphen-Vereins, Jahrgang*, volume IX, 1862. This source does not list telegraph stations in London. We thus dissolve urban districts in the London area into a single "London" unit and assume that this unit is connected to the domestic telegraph network.

Figure 3.12: Matching Rate of the Linked Inventors-Census Sample, 1881–1911

((a)) Matching Rate Over Time

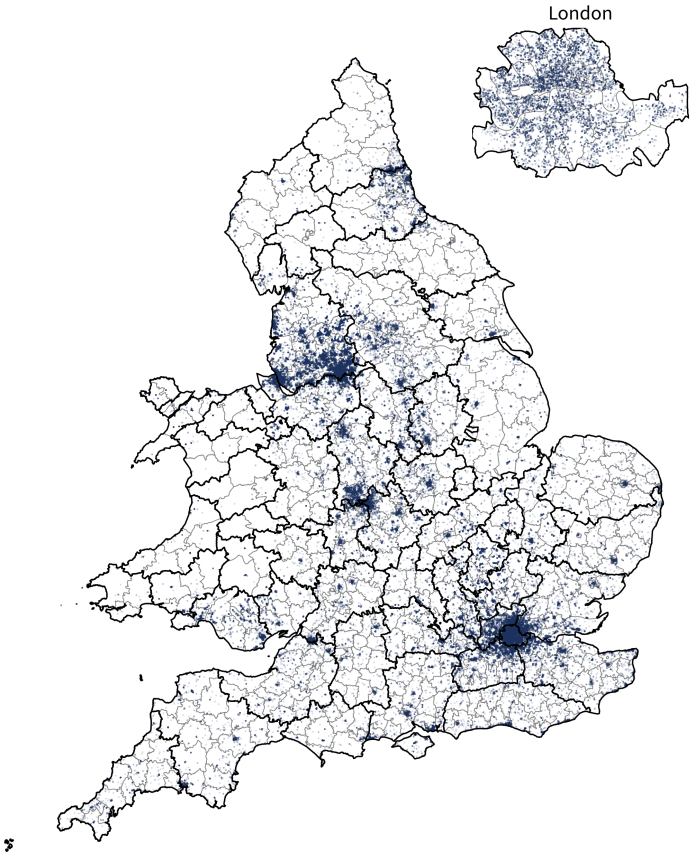


((b)) Matching Rate by Geographic Layer



Notes. These figures report the matching rate for the linked inventor-census sample. Panel 3.12(a) reports the matching rate over time for the 1881–1900 sample (blue bars) and the 1901–1920 sample (red bars). Color-contoured bars report the share of records with at least one match; color-filled bars report the share of acceptable linked matches. A record match is acceptable if it has no more than five multiple matches. Panel 3.12(b) reports the share of matches by the number of matches, broken down by geographical layers. In Panel 3.12(a), we do not show the few matches with quality below .95. In Panel 3.12(b), the sample is restricted to records with at least one match.

Figure 3.13: Distribution of Inventors Across UK Districts, 1881–1911



Notes. This figure displays the spatial distribution of inventors across districts between 1881 and 1911. Each marker reports one inventor, defined as an individual who obtains at least one patent over the sample period. To retrieve the coordinates of the inventors, we first link population censuses, whose entries are, in turn, geo-referenced. The background map displays districts at historical borders in 1900.

3.10.5 Tables

Table 3.9: Descriptive Statistics on Newspapers and Newspaper Coverage in the UK

	(1) Mean	(2) Std. Dev.	(3) Min.	(4) Max.	(5) Observations
Panel A. Journal-Level Statistics					
Number of Issues	2795.843	4959.740	1	46163	2022
First Publication Year	1869.746	44.171	1699	1996	2094
Last Publication Year	1910.692	49.470	1699	2009	2094
Publication Lifespan	40.946	40.490	0	273	2094
Publication Lifespan if English	40.993	41.921	0	273	1459
Publication Lifespan if Welsh	38.161	36.920	0	178	93
Publication Lifespan if Scottish	45.144	41.107	0	251	229
Publication Lifespan if Irish	41.336	34.809	0	170	241
Panel B. District-Level Statistics, by Decade					
1870s	2.309	14.860	0	285	637
1880s	1.885	11.610	0	233	636
1890s	1.494	8.587	0	160	634
1900s	1.166	5.893	0	114	634
1910s	0.942	3.845	0	83	633
1920s	0.809	2.381	0	50	633
1930s	0.714	1.274	0	24	633
Panel C. District-Level Statistics, by Division					
East	1.631	1.272	1	8	111
East Midlands	2.349	2.409	1	14	43
London	18.767	97.312	1	534	30
North East	2.079	1.761	1	8	38
North West	3.600	3.477	1	17	40
South East	1.800	1.271	1	6	100
South West	1.747	1.382	1	8	79
Wales	2.327	2.391	1	10	52
West Midlands	2.342	2.722	1	18	79
Yorkshire	2.186	2.201	1	10	59

Notes. This table reports descriptive statistics on newspapers active in the UK between 1850 and 1940. In Panel A, figures are computed at the newspaper level; Panel B computes district-level statistics on the number of newspapers by decade; Panel C computes district-level statistics on the number of newspapers by division. Panels B and C only restrict the observation sample to English and Welsh districts. Newspapers were geo-coded to their publishing address and assigned to districts based on their borders in 1900.

Table 3.10: Correlation Between Inventor Characteristics and N. Matches

	Overall Sample	Parish Matches	District Matches	County Matches	Nationwide Matches
	(1)	(2)	(3)	(4)	(5)
Panel A. Demographics					
Age	0.005	-0.018	-0.005	-0.014*	0.026***
	(0.010)	(0.021)	(0.012)	(0.007)	(0.005)
Dependent Variable – Dummy = 1 if Matched Inventor is in:					
Panel B. Occupation					
Agriculture	0.123*	0.272**	0.073***	0.000	0.028**
	(0.073)	(0.129)	(0.019)	(0.016)	(0.011)
Chemicals	-0.010***	-0.019***	-0.012*	-0.011***	-0.005***
	(0.004)	(0.005)	(0.006)	(0.004)	(0.001)
Construction	-0.015	-0.032	0.006	0.010	-0.001
	(0.016)	(0.031)	(0.008)	(0.017)	(0.002)
Engineering	-0.018	-0.042*	-0.017**	-0.007	0.004
	(0.012)	(0.025)	(0.007)	(0.007)	(0.003)
Liberal Professions	-0.014***	-0.016***	-0.003	-0.018***	-0.019***
	(0.005)	(0.006)	(0.010)	(0.002)	(0.005)
Metallurgy	-0.020	-0.031	0.018	-0.015	0.004**
	(0.013)	(0.019)	(0.019)	(0.019)	(0.002)
Other Manufacturing	-0.024*	-0.042**	-0.027	0.005	-0.007**
	(0.012)	(0.018)	(0.016)	(0.007)	(0.003)
Public Administration	-0.009	-0.017	-0.014	-0.015	-0.008**
	(0.008)	(0.014)	(0.010)	(0.011)	(0.003)
Textiles	-0.013	-0.042*	0.002	0.058***	0.003
	(0.012)	(0.026)	(0.024)	(0.013)	(0.005)
Trade	-0.031***	-0.044***	-0.038***	-0.021*	-0.025***
	(0.011)	(0.016)	(0.007)	(0.011)	(0.005)
Transport	-0.008	-0.025	0.001	0.006	0.003
	(0.014)	(0.026)	(0.008)	(0.011)	(0.003)
Utilities	-0.013***	-0.020***	-0.031***	-0.016***	-0.007*
	(0.004)	(0.005)	(0.005)	(0.006)	(0.004)
Panel C. Division of Residence					
East	-0.004	-0.059	-0.061	-0.074	-0.010
	(0.015)	(0.067)	(0.065)	(0.089)	(0.011)
East Midlands	0.004	-0.061	0.070	-0.055	0.012
	(0.012)	(0.070)	(0.102)	(0.066)	(0.013)
London	-0.048	-0.039	-0.053	-0.008	-0.025
	(0.069)	(0.173)	(0.053)	(0.079)	(0.022)
North East	0.028	-0.045	-0.041	-0.060	0.016
	(0.031)	(0.052)	(0.049)	(0.072)	(0.016)
North West	-0.057	-0.165	-0.069	0.195***	0.012
	(0.050)	(0.167)	(0.080)	(0.056)	(0.013)
South East	-0.024	-0.050	-0.106	-0.118	-0.028
	(0.030)	(0.057)	(0.106)	(0.136)	(0.027)
South West	0.001	-0.025	-0.049	-0.051	-0.019
	(0.008)	(0.029)	(0.054)	(0.062)	(0.020)
Wales	0.233	0.469**	0.540***	-0.026	0.046
	(0.187)	(0.212)	(0.137)	(0.032)	(0.046)
West Midlands	-0.049	-0.102	-0.104	-0.130	0.004
	(0.050)	(0.112)	(0.103)	(0.148)	(0.007)
Yorkshire	0.005	-0.021	-0.029	0.008	-0.003
	(0.013)	(0.025)	(0.032)	(0.021)	(0.007)
Decade FE	Yes	Yes	Yes	Yes	Yes

Notes. This table reports the correlation between inventor-level variables observed in the UK census and the number of matches in the linked sample. In column (1), the sample is the entire linked dataset. We restrict to matches at the parish (column 2), district (column 3), county (column 4), and national level (column 5). The Table reports standardized beta coefficients for comparability. Regressions include decade fixed effects. Standard errors are clustered at the division level and are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.11: List of Industries By Tariff Rate, 1925–1935

Sector	Technology Class	Tariff Rate Before S-H	Tariff Rate After S-H	Change in Tariff	Treated
(1)	(2)	(3)	(4)	(5)	(6)
Agricultural products and provisions	Agriculture	23.059	40.204	74.352	Yes
Chemicals, oils, and paints	Chemistry	29.577	40.195	35.900	No
Cotton Manufactures	Textiles	34.876	44.764	28.352	No
Earths, earthenware, and glassware	Personal Articles, Furniture	47.321	53.049	12.106	No
Flax, hemp, and jute, and manufacture thereof	Textiles	18.948	26.104	37.766	No
Metals, and manufacture thereof	Metallurgy	34.534	36.803	6.572	No
Pulp, paper, and books	Printing	25.652	25.591	-0.239	No
Silk and silk goods	Textiles	55.768	58.115	4.208	No
Spirits, wines, and other beverages	Food	37.298	59.007	58.226	Yes
Sugar, molasses, and manufactures thereof	Food	68.971	110.022	59.519	Yes
Sundries	Personal Articles, Furniture	38.149	36.587	-4.096	No
Tobacco, and manufactures thereof	Agriculture	58.176	81.636	40.326	No
Wood, and manufactures thereof	Building	23.727	20.672	-12.875	No
Wool, and manufactures thereof	Textiles	49.344	78.255	58.591	Yes

Notes. This table reports the US tariff rate applied to the categories listed in the *Statistical Abstracts of the United States*. Column (1) reports the listed sector; column (2) maps the sector to technology classes in our baseline taxonomy; columns (3) and (4) report the tariff rate applied, respectively, before and after the Smoot-Hawley Act (1930). Tariff rates before the Act are averages in the five years before the reform (1925–1929); tariff rates after the Act are averages in the five years posterior to the reform (1930–1935). Column (5) computes the change in the tariff rates. In column (6), we list the technology classes we considered targeted by the Act, namely, those whose tariff rate increase exceeded 50%. Data are digitized from the 1935 *Statistical Abstracts of the United States*.

3.11 Appendix: Novel Patent Data

3.11.1 Sources and Digitization

This section presents the motivation for developing a new patent dataset for England and Wales that spans the second half of the XIX century. Then, we describe the sources we use and how we structure the textual data they contain into a machine-usable dataset. Finally, we describe two data-augmentation routines that we perform to geocode the patents and assign them a modern technology class.

Motivation

Despite its historical significance, we lack comprehensive patent data for the Second Industrial Revolution period (1850–1900) in the United Kingdom. In particular, it is impossible to reconstruct the geographical distribution of innovation activity during this period. This data limitation sharply contrasts the effort undertaken to document patenting activity since the inception of the English patent law in 1617 up until the end of the First Industrial Revolution in the 1840s (Nuvolari and Tartari; Nuvolari, Tartari and Tranchero, 2011; 2021). We fill this gap by constructing the first dataset of English and Welsh patents that spans the period 1853–1900 and contains detailed information on the text, geographical location, inventors' personal information, and date for the universe of patents.

Data Sources

The UK Intellectual Property Office allowed us access to restricted full-page scans of original patent documents. These are the universe of patents granted in England and Wales between 1617 and 1899. This paper focuses on the period 1853-1899 for two main reasons. First, Nuvolari and Tartari (2011) already digitized patents before 1853 from Bennet Woodcroft's index, although patent documents contain additional information

compared to the index. Second, in 1853 a reform dramatically lowered patent application prices. This makes it challenging to compare patents before and after the reform. Patent documents contain a wealth of unstructured information. We provide two examples in Figure 3.14: in panel 3.14(a) we show the patent granted to Henry Bessemer for the eponymous process to produce steel, and in panel 3.14(b) we display the patent granted to John Starley for the first modern safety bicycle. Both patents are in our dataset. The rectangles identify the location of the textual data that we extract. These comprise (i) a short title, (ii) a long title, (iii) the author(s)'s name(s), (iv) the author(s)'s address(es), (v) the author(s)'s profession(s), (vi) the filing date, (vii) the issue date, (viii) the type of protection, (ix) an indicator of whether the application was filed by an agent on behalf of someone living abroad, and (x) the full text of the patent. Not all (i-x) are available throughout the sample. In particular, (i), (vi), and (viii) are available only until 1873. After that date, a short title is no longer reported, the filing date is reported only sporadically, and the type of protection becomes immaterial, for only granted patents are included in the sample.

Digitization

We perform optical character recognition (OCR) on each patent individually to structure the data in a machine-readable dataset. To ensure state-of-the-art performance, we OCR the first page of each document, where all the (i–ix) variables are located, using Amazon's commercial `textract` engine. To retrieve the rest of the text, which is not used in this paper, we use the open-source engine `tesseract`. An OCR-ed document is a text file. To extract the relevant variables, we implement a script that leverages regular expressions to identify the variables (i–ix). Fortunately, the text of each patent is fairly standardized; hence this routine yields detailed and high-quality results for all variables except (v), which is not used in this paper. This exercise results in a database of approximately 800,000 patents granted between 1853 and 1899.

Geo-Coding

To retrieve each patent's location, we geocode each inventor's listed address using the commercial geocoding engine provided by MapTiler AG. To geocode an address, if a coarse geographical unit is listed on the patent (e.g., the county), we condition the outcome coordinates to lie within that unit. In Figure 3.17, we report the resulting distribution of patents (panel 3.17(a)) and patents per capita (panel 3.17(b)). Reassuringly, these are consistent with underlying population and economic development indicators.

Technology Class Assignment

Naturally, historical patent documents do not list CPC classes. Yet, the technological classification is a key variable in our empirical exercise. To reconstruct the class, we adopt a supervised machine-learning approach. We conjecture, following Xu (2018), that titles are informative of technological classes. We split the PATSTAT data, which covers the years 1900–1939 and for which we observe both titles and classes, in a train and a test set, with a proportion of 4:1. We apply a term frequency-inverse document frequency vectorization algorithm to the titles of both datasets. Then, we estimate a linear support vector machine (LSVC) on the train set. An LSVC is a non-probabilistic classifier that assigns class labels to maximize the width of the gap between classes. Formally, consider a set of points $(\mathbf{x}_i, y_i)_{i=1}^N$ where $\mathbf{x} \in \Re^N$ represent the features—in our case, words—and y is the class. For simplicity, assume $y \in \mathcal{Y} = \{-1, 1\}$. An LSVC solves for the hyperplane $\mathcal{W} = \{\mathbf{w} \in \Re^N \text{ s. t. } \mathbf{w}^\intercal \mathbf{x}_i - \ell = 0\}$ that maximizes the distance between the group i such that $y_i = 1$, and the group where $y_i = -1$. The distance that is most commonly used that allows for non-linearly separable data is the hinge loss, which is defined as $d_i = \max\{0, 1 - y_i(\mathbf{w}^\intercal \mathbf{x}_i - \ell)\}$. In our case, however, we allow for multiple classes, that is, the cardinality of $\mathcal{Y}$ be more than two. We employ an LSVC because the literature notes that it yields particularly robust results. However, the classification outcome would remain fairly unchanged using different algorithms.[61]

[61] In particular, we tested the Naïve Bayes classifier, several Boosting algorithms (e.g., AdaBoost, XGBoost), a random forest classifier, and a simple convolutional neural network. All the above yield similar classification results but slightly lower accuracy than the LSVC. Additionally, we explored alternative

On the training set, the LSVC yields a 95% accuracy, measured as the share of patents with a correctly imputed class relative to the total number of patents. This decreases to 85% on the test set, which is not used to train the algorithm. Given that state-of-the-art models trained on modern US data achieve approximately a test 90% accuracy, we interpret these results as rather encouraging (Li, Hu, Cui and Hu, 2018). We report the confusion matrix on the test set in Figure 3.15. For a given cell, the row label is the true technology class, and the column label is the imputed class. A perfect classifier would thus yield a diagonal confusion matrix. Overall, we find that misclassification errors are evenly distributed, in relative terms, across classes. Hence, even though the classifier is not perfect, there does not seem to be any systematic measurement error in class imputation.

3.11.2 External Validation

To validate our data, we consider the only two series that cover—a portion of—the years 1853–1899. Hanlon (2016) digitized an index of patents issued between 1855 and 1883. His data list, for each patent, the inventor(s) and their profession(s), a technology class, and the issue year. On top of the longer time coverage, our data thus contain several additional information, including the geographical coordinates. The second dataset that we use as a comparison is the "A Cradle of Invention" (COI) series, published by Finishing Publications (2018). These data, too, were digitized from indices and thus only list authors, issue year, and, often, titles. In principle, this series spans the years 1617–1895. However, after 1883 patent applications that were eventually denied protection are also listed. Absent a way to identify granted patents, we do not report figures after 1883 for the COI series.

In Table 3.12, we report the aggregate number of patents issued according to our series (columns 2 and 6), COI (columns 3 and 7), and Hanlon (2016) (columns 4 and 8). Reassuringly, the three series are highly consistent. Our series is closest to Hanlon (2016),

vectorization algorithms using transformers (e.g., BERT and RoBERTa) with no significant performance gains.

but the COI figures are not too far off either. Overall, the Table strongly suggests that our series is as complete as the Hanlon (2016) database. We cannot, however, externally validate it for the later part of the period because there is no data available.

3.11.3 Measuring Pairwise Similarity Between US and UK Patents

In this section, we describe in detail how we construct the patent similarity metric we adopt to measure "copying" and "originality" of UK innovation activity. The approach borrows heavily on Kelly, Papanikolaou, Seru and Taddy (2021). We adapt their methodology to our context by leveraging text information contained in titles only. Even though we do not have access to full US patent texts, the title of a patent is usually very informative about its content. In fact, we previously showed that a title-based machine learning algorithm predicts the technological classification of the patent with nearly 90% accuracy. Titles for UK patents are embedded in the digitized text for the period 1870–1899 and are collected from PATSTAT for the later years; titles for US patents are collected from PATSTAT throughout the sample period.

We start by defining the backward inverse-document frequency associated with each word w. This expresses the inverse frequency with which the word w appears in US patents p issued until year t. Formally, we have

$$\text{BIDF}_{w,t} \equiv \log\left(\frac{\text{Number of Patents Issued Before } t}{1 + \text{Number of Patents Issued Before } t \text{ that contain word } w}\right) \quad (3.15)$$

Then, to each patent-word pair, we associate the term frequency variable that counts the number of instances word w appears in patent p, normalized by the length of the patent. With a slight abuse of notation, let p denote both the index of the patent and the set of words it contains. We shall have

$$\text{TF}_{wp} \equiv \frac{\sum_{c \in p} 1(c = w)}{\sum_{c \in p} 1(c)} \quad (3.16)$$

where the numerator returns how many times word w appears in patent p, and the denumerator is simply the number of words in patent p. Then, we define the TF-BIDF

associated with word w, patent p at time t as the product between these two terms:

$$\text{TF-BIDF}_{wp,t} \equiv \text{TF}_{wp} \times \text{BIDF}_{w,t} \tag{3.17}$$

and, thus, the vector $\text{TF-BIDF}_{p,t}$ collects the term frequency-backward inverse document frequency for all words w in p. For comparability, the vector $\text{TF-BIDF}_{p,t}$ is normalized by its norm to have unit length.

We compute the $\text{TF-BIDF}_{p,t}$ vectors for US and UK patents, but the $\text{BIDF}_{w,t}$ are computed on the corpus of US patents only. Then, we compute the cosine similarity $\rho_{i,j}$ between each UK patent i and each US patent j. This allows us to define two variables. First, we seek to measure the similarity between British innovation and previous American patents. This yields a measure of backward similarity that, for the sake of the narrative of the paper, we define as "copying". Formally we define

$$\text{Backward Similarity}_i^{\tau} \equiv \sum_{j \in \mathcal{F}_i^{-\tau}} \rho_{i,j} \tag{3.18}$$

where the set $\mathcal{F}_i^{-\tau}$ denotes the set of US patents issued within τ years from the issue year of patent i. This measures the degree of similarity between a given patent in the UK and previous patents in the US. Second, we define a measure of "originality" of UK patents compared to previous US patents. This leverages the insight of Kelly, Papanikolaou, Seru and Taddy (2021), who suggest that innovative and influential patents are those that are most dissimilar from existing innovation, while at the same time retaining semantic proximity with subsequent patents. Formally, we have

$$\text{Excess Forward Similarity} \equiv \frac{\sum_{j \in \mathcal{F}_j^{+\tau}} \rho_{i,j}}{\sum_{j \in \mathcal{F}_i^{-\tau}} \rho_{i,j}} \tag{3.19}$$

where $\mathcal{F}_i^{+\tau}$ denotes the set of US patents issued within τ years after the issue year of patent i. In the baseline analysis, we set a symmetric window of $\tau = 5$ years around each patent's issue date. In Table 3.33 we report the results using an alternative threshold of ten years. Moreover, in the same table, we report the results obtained by netting out year and technology class fixed effects at the patent level. As noted by Kelly, Papanikolaou,

Seru and Taddy (2021), this ensures that we do not conflate shifting terminology fashions in the similarity measures.

3.11.4 Summary Statistics and Stylized Facts

We conclude this section by presenting some stylized statistics and facts our new data allow us to uncover. First, as noted in Table 3.12, the number of patents granted generally grows over time, although at a somewhat stagnating path. There is, however, a sizable discontinuity between 1883 and 1884, when the number of patents jumps from 6074 to 9873. In 1883 the Patents Act reduced application fees by 83%, as noted by Nicholas (2014). It seems plausible to attribute the discontinuity to this reform.

Second, in Figure 3.16, we report the composition of patenting activity by technology class. In each year, we compute the share of patents in a given sector with respect to the total number of patents issued that year. We report such shares over time between 1853 and 1939. The composition of innovation exhibits two clear patterns. First, the share of textiles patents, which in the 1850s represented nearly 20% of the total, shrinks considerably, and in 1939 it accounts for less than 5%. This is consistent with the historical preeminence of textiles during the First Industrial Revolution and their subsequent loss of importance. Second, electricity-related innovation grows considerably in the later part of the period. In 1939, it represented more than 20% of the total number of patents issued. Once more, this echoes historical, anecdotal evidence highlighting the centrality of electricity during the later stages of the Second Industrial Revolution and beyond (David; Mokyr, 1990; 1998).

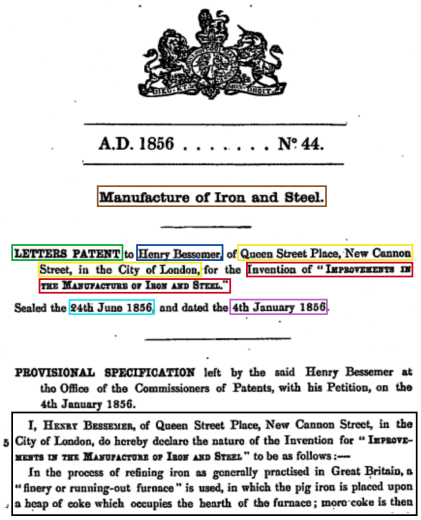
Finally, a crucially novel component of our dataset is that it allows studying the geographical dimension of the innovation process. Thus, in Figure 3.17, we report the spatial distribution of the number of patents in absolute number (panel 3.17(a)) and normalized by population (panel 3.17(b)). These maps attest to the importance of duly considering the geography of innovation. The patenting activity appears to be widely dispersed across England and Wales. Heavily industrial areas, such as Lancashire, the

Midlands, the Tyne, and South Wales, all feature prominently in terms of issued patents. Similarly, the London area is also a major innovation hub. By contrast, Northern Wales, Anglia, Cornwall, and Cumbria perform poorly. In Figure B18, we repeat this exercise, but we break down the number of patents by selected technology classes: chemistry (panel 3.18(a)), electricity (panel 3.18(b)), engineering (panel 3.18(c)), engines and pumps (panel 3.18(d)), metallurgy (panel 3.18(e)), and textiles (panel (panel 3.18(f)). While innovation centers remain roughly similar across sectors, some differences emerge. For example, the metallurgy industry was particularly deep-rooted in the Midlands, where we note the largest concentration of metallurgy patenting. Similarly, textile innovation centers in the Lancashire area, the historic "cotton districts". Our database allows studying a novel, thus far largely unexplored dimension of the innovation and patenting activity. Therefore, the analysis carried out in this paper is one of many that may take advantage of this contribution.

3.11.5 Figures

Figure 3.14: Sample Annotated Patent Documents: the Bessemer Process and the First Modern Safety Bicycle

((a)) Henry Bessemer's 1856 Patent



A.D. 1856 . . . . . Nº 44.

Manufacture of Iron and Steel.

LETTERS PATENT to Henry Bessemer, of Queen Street Place, New Cannon Street, in the City of London, for the Invention of "IMPROVEMENTS IN THE MANUFACTURE OF IRON AND STEEL."

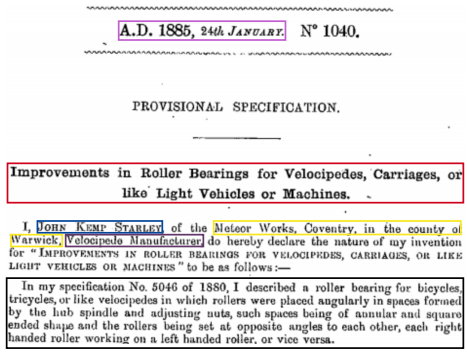
Sealed the 24th June 1856 and dated the 4th January 1856.

PROVISIONAL SPECIFICATION left by the said Henry Bessemer at the Office of the Commissioners of Patents, with his Petition, on the 4th January 1856.

I, HENRY BESSEMER, of Queen Street Place, New Cannon Street, in the City of London, do hereby declare the nature of the Invention for "IMPROVEMENTS IN THE MANUFACTURE OF IRON AND STEEL" to be as follows:—

In the process of refining iron as generally practised in Great Britain, a "finery or running-out furnace" is used, in which the pig iron is placed upon a heap of coke which occupies the hearth of the furnace; more coke is then

((b)) John K. Starley's 1885 Bicycle Patent



A.D. 1885, 24th JANUARY. N° 1040.

PROVISIONAL SPECIFICATION.

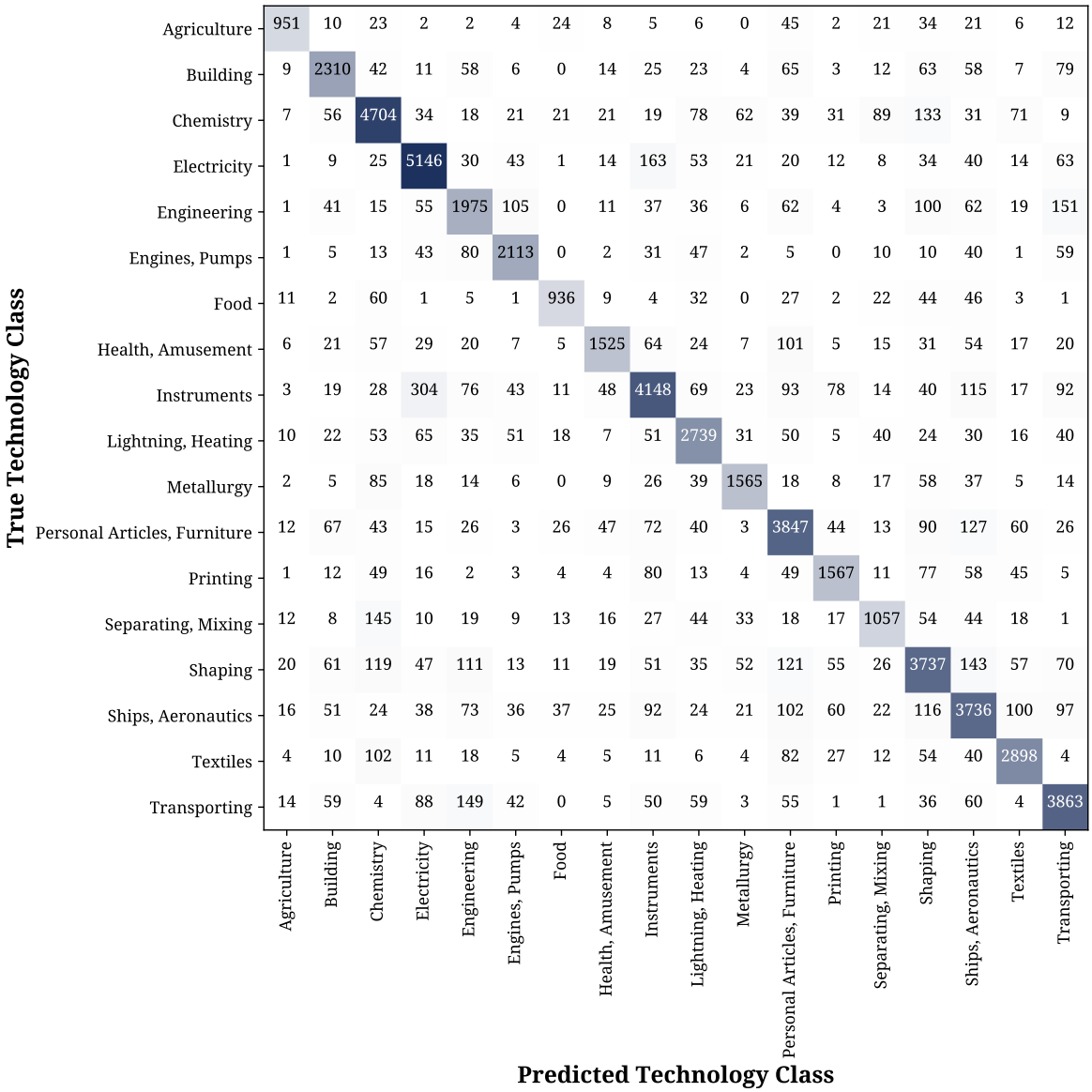
Improvements in Roller Bearings for Velocipedes, Carriages, or like Light Vehicles or Machines.

I, JOHN KEMP STARLEY, of the Meteor Works, Coventry, in the county of Warwick, Velocipede Manufacturer, do hereby declare the nature of my invention for "IMPROVEMENTS IN ROLLER BEARINGS FOR VELOCIPEDES, CARRIAGES, OR LIKE LIGHT VEHICLES OR MACHINES" to be as follows:—

In my specification No. 5046 of 1880, I described a roller bearing for bicycles, tricycles, or like velocipedes in which rollers were placed angularly in spaces formed by the hub spindle and adjusting nuts, such spaces being of annular and square ended shape and the rollers being set at opposite angles to each other, each right handed roller working on a left handed roller, or vice versa.

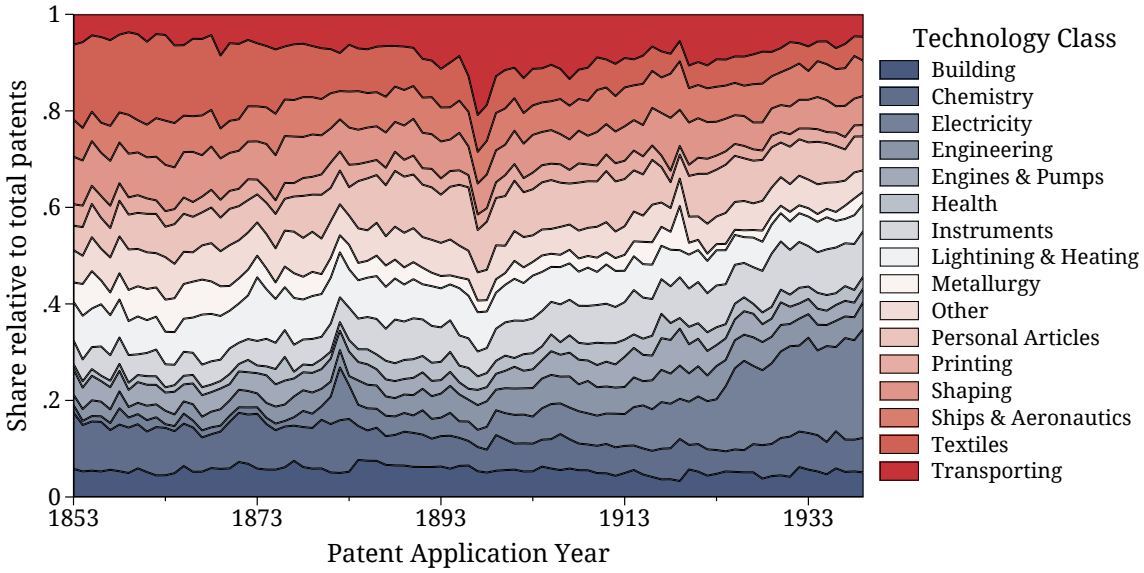
Notes. This figure displays two sample patent documents in our dataset. Panel 3.14(a) was granted to Henry Bessemer in 1856 for the invention of the famous eponymous process for the mass production of steel from the molten pig iron. Panel 3.14(b) was granted to John Starley in 1885 for the invention of the first modern bicycle, which would soon revolutionize mobility in Europe and in the US. Colors mark different variables that we structure in the dataset: (i) in brown, the short title; (ii) in red, the complete title (iii) in green, the type of protection granted; (iv) in blue, the author(s) name(s); (v) in yellow, the author(s)'s address(es); (vi) in light blue, the application date; (vii) in purple, the issue date; (viii) in black, the patent text that continues in the rest of the patent document; (ix) in dark purple, the author(s) profession(s). Not all (i–ix) data are available on every patent and in each year.

Figure 3.15: Confusion Matrix of the Technology Sector Classifier



Notes. This figure displays the confusion matrix of the patent technology classifier. The algorithm assigns to each patent an imputed technology class using information contained in the title. Titles undergo pre-processing and term frequency-inverse document frequency (tf-idf) vectorization. The classifier is trained on an 80% sub-sample of the universe of British patents granted over the period 1900–1940. The figure reports the classifier's performance on the remaining 20% test set, which is not used in training. The y-axis reports the true patent class; the x-axis reports the class imputed by the classifier. A perfect classifier would yield a diagonal confusion matrix. The accuracy in the training (resp. test) set is ≈98% (resp. ≈85%). Lighter to darker blue indicates an increasing number of patents in the cell.

Figure 3.16: Composition of Total Patents Granted in the United Kingdom Across Technology Classes, 1880–1939

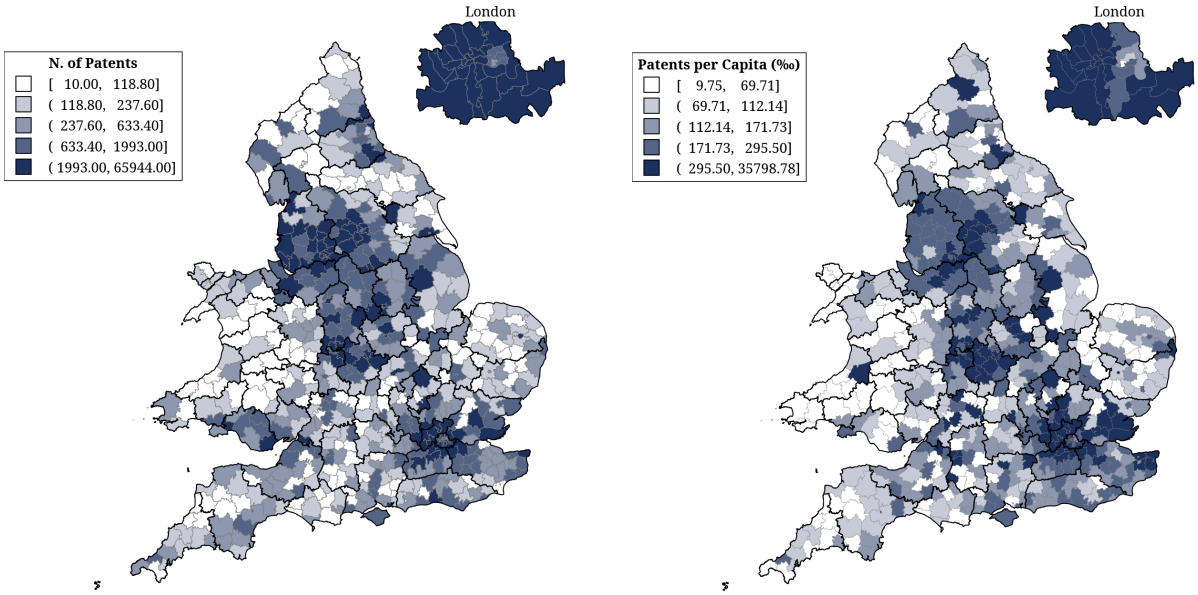


Notes. This figure displays the evolution of innovation in Britain across technology classes from 1853–1939. For each year, we compare the share of patents in each class in our database relative to the total number of patents granted in that year. Data for the period 1853–1899 are from the newly digitized universe of patents; data for the period 1900–1939 are made available by the European Patent Office repository.

Figure 3.17: Distribution of Patents and Patents per Capita Across Districts, 1880–1939

((a)) Number of Patents

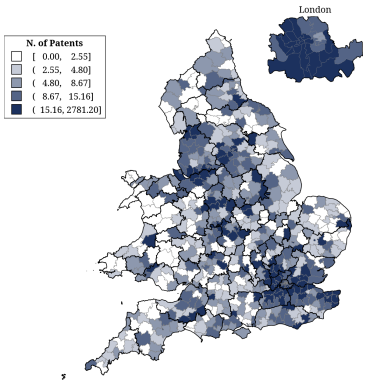
((b)) Patents per Person



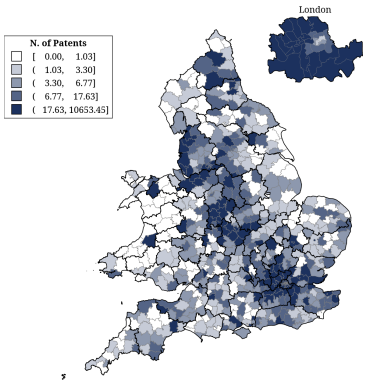
Notes. These figures report the intensity of patenting activity across districts over the period 1880–1939. Panel 3.17(a) reports the total number of patents granted; Panel 3.17(b) normalizes this by district population in 1900 and expresses the resulting rate in ‰ units. Districts are displayed at 1900 borders. To assign patents to districts, we geo-reference the address of each author listed in the patent document and assign districts based on historical district borders.

Figure 3.18: Distribution of the Total Number of Patents Granted Across Districts and Selected Technology Classes, 1880–1939

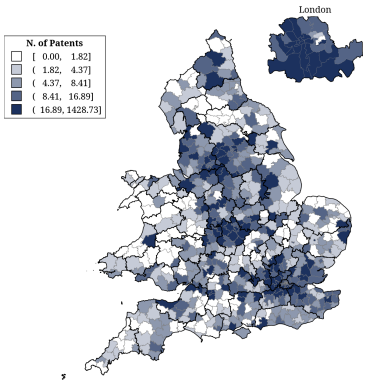
((a)) Chemistry



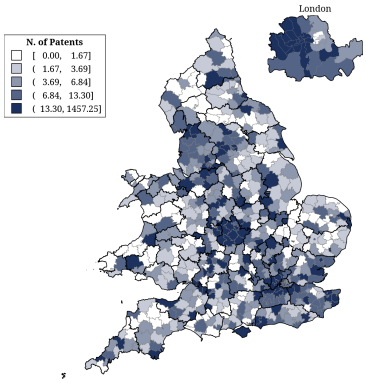
((b)) Electricity



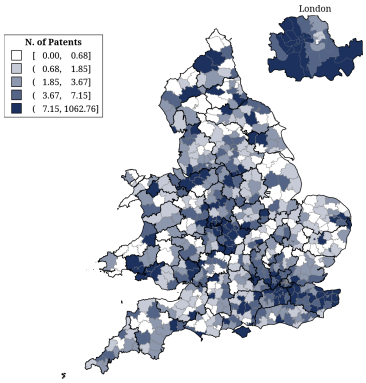
((c)) Engineering



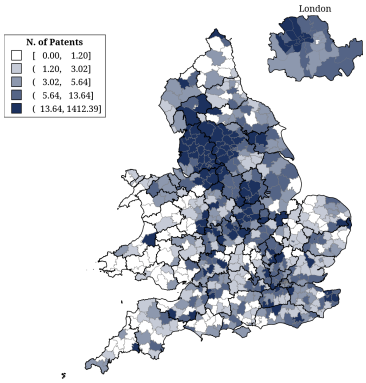
((d)) Engines, Pumps



((e)) Metallurgy



((f)) Textiles



Notes. These figures report the intensity of patenting activity across districts over 1880–1939 for selected technology classes. Districts are displayed at 1900 borders. To assign patents to districts, we geo-reference the address of each author listed in the patent document and assign districts based on historical district borders.

3.11.6 Tables

Table 3.12: Total Number of Patents Granted in the UK: Comparison Across Three Datasets

Years 1853-1876				Years 1877-1899			
(1) Year	(2) Our Series	(3 COI	(4) Hanlon	(5) Year	(6) Our Series	(7) COI	(8) Hanlon
1853	3042	3016		1877	4943	4928	4940
1854	2759	2690		1878	5336	5143	5333
1855	2960	2866	2955	1879	5332	5305	5325
1856	3107	2967	3102	1880	5499	5132	5509
1857	3206	3092	3197	1881	5744	5620	5745
1858	3023	2954	2999	1882	6159	6150	6233
1859	3048	2989	2998	1883	6074	6006	5981
1860	3192	3139	3190	1884	9873		
1861	3261	3269	3272	1885	8783		
1862	3482	3459	3486	1886	8999		
1863	3301	3299	3308	1887	9218		
1864	3256	3225	3257	1888	9331		
1865	3378	3364	3378	1889	10325		
1866	3451	3408	3452	1890	10355		
1867	3724	3692	3720	1891	10686		
1868	4008	3908	3984	1892	11429		
1869	3832	3741	3781	1893	11985		
1870	3407	3288	3405	1894	11648		
1871	3525	3479	3525	1895	12198		
1872	3969	3940	3967	1896	13597		
1873	4276	4281	4282	1897	14249		
1874	4494	4516	4491	1898	13100		
1875	4557	4451	4557	1899	13172		
1876	5049	5012	5064				

Notes. This table reports the total number of patents in England and Wales between 1853 and 1899. Columns (2) and (6) report the series constructed from our novel dataset; columns (3) and (7) tabulate data from *A Cradle of Inventions* (Finishing Publications, 2018); columns (4) and (8) report data from Hanlon (2016). The *A Cradle of Inventions* series potentially stretches until 1899. However, after 1883 there is no way to distinguish between patents granted and applications. Hence we do not report figures for these later years (Nicholas, 2014). Data from Hanlon (2016) only cover the years 1855–1883.

3.12 Appendix: Linked International Migrants Sample

This section discusses our methodology to link English and Welsh immigrants in the US to the UK census and presents key statistics on the resulting dataset.

3.12.1 Sources and Linking Algorithm

We rely on two sources of externally compiled data.[62] For the US, we have access to the IPUMS full-count non-anonymized census (Ruggles, Fitch, Goeken, Hacker, Nelson, Roberts, Schouweiler and Sobek, 2021). A census was taken in the US every ten years starting in 1790, except for 1890. Until 1840, the census was run at the household level. From 1850 on, instead, we have detailed *individual* information on the universe of the US population.[63] For confidentiality, these data are available up until 1940. Our dataset, therefore, contains snapshots of the entire US population at any given decade between 1850 and 1940, although for the sake of this paper, we restrict to the years 1870-1930. Crucially, we have access to the non-anonymized version of the IPUMS data. Hence, besides publicly available information, we also know each individual's recorded name and surname.

In the UK, the I-CeM data mirrors the IPUMS (Schurer and Higgs, 2020) content. More precisely, it contains information on the universe of people living in England, Scotland, and Wales. Similarly to the US—and virtually every other—census, it was run at decade frequency starting in 1851 and until 1911. No census was taken in 1871. As with the IPUMS data, we can access the full-count non-anonymized version of the dataset. Besides publicly available information, this contains full names and addresses of the universe of individuals living in the UK at any given decade.

[62] We are deeply thankful to IPUMS and I-CeM for allowing us access to their confidential data. Without their help, this paper would not have been possible.

[63] By US population, we refer to the universe of individuals who *lived* in the US at a given point in time.

Our methodology relies on Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021). This dataset tackles the problem that neither the US nor the UK—nor any other European countries—recorded where British immigrants came from *within* the UK. We thus try to match British immigrants residing in the US with their entry in the UK census, which records where they come from at a granular geographical level.[64] More precisely, we take the stock of British residing in the US in a given census year—say, 1900—and match them with their entry in the preceding UK census—in this case, 1891.[65] This implies that we measure the *flow* of British immigrants over time rather than their stock.

We use three variables to link individuals: first name, surname, and birth year. The baseline sample we link consists of individuals who report, in the US census, either England or Wales—or analogous denominations, such as Great Britain—as their country of origin. In the 1900 census, we take all those who immigrated between 1870 and 1899. In the subsequent censuses, until 1930, we retrieve stock of those who immigrated in the preceding decade. Then, to match each unit in the sample—call the generic one A—to an entry in the UK census, we perform this sequence of operations:

1. Take the census that precedes the immigration year of A. Hence, for instance, we match all those who immigrated in 1896 to the 1891 census.

2. Select all records in that census with the same reported birth year as A—call the resulting sample $\mathcal{M}^A = \{m_1^A, \ldots, m_N^A\}$.

3. Compute a string-similarity measure between the name and surname of A and that of all elements of $\mathcal{M}^A$. In other words, for every $m_i^A \in \mathcal{M}^A$, compute[66]

$$\text{Similarity}_i^A = \alpha \times \text{Name Similarity}_i^A + (1 - \alpha) \times \text{Surname Similarity}_i^A \qquad (3.20)$$

   for some $\alpha \in [0, 1]$. In our baseline setting, we set $\alpha = 0.3$ to give higher weight to

[64] Since women usually change their name upon marriage, we are unable to match them. This is a common problem in linking algorithms (Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez, 2021).

[65] Since no census was taken in the UK in 1871, we link the 1880 US census to the 1861 UK one. This is not overly problematic because we can still match all those aged ten or older in 1871.

[66] We cannot simply match on exact same name and surname because coding errors are commonplace in historical census data (Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez, 2021).

the surname.

4. The set of matches is defined as

$$\overline{\mathcal{M}}^A = \left\{ m_i^A \in \mathcal{M}^A \,\middle|\, \text{Similarity}_i^A = \max_{m_{i'}^A \in \mathcal{M}^A} \text{Similarity}_{i'}^A \right\} \tag{3.21}$$

which means that we restrict the set of possible matches to include only those whose similarity score with the entry in the US census A is the largest.

5. Finally, for a given threshold $\tau > 0$, we select only the possible matches whose similarity score is above τ. The set of effective matches thus boils down to:

$$\widetilde{\mathcal{M}}_\tau^A = \left\{ m_i^A \in \overline{\mathcal{M}}^A \,\middle|\, \text{Similarity}_i^A \geq \tau \right\} \tag{3.22}$$

Clearly, $\widetilde{\mathcal{M}}^A$ can ideally be empty, meaning that A has no effective matches. It can have one element, in which case we refer to it as a "perfect match," or it can have multiple matches. In our baseline exercise, we set $\tau = 0.7$ as we see a clear elbow in the distribution of similarities there.

We evaluate the distance between two strings i and j in terms of their Jaro-Winkler similarity d_{ij}:

$$d_{ij} \equiv \widehat{d_{ij}} + \ell p(1 - \widehat{d_{ij}}) \tag{3.23}$$

where

$$\widehat{d_{ij}} \equiv \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|i|} + \frac{m}{|j|} + \frac{m-t}{m}\right) & \text{else} \end{cases} \tag{3.24}$$

where m is the number of matching characters, $|i|$ is the length of string i, and t is half the number of transpositions, ℓ is the length of common an eventual common prefix no longer than four characters between i and j, and $p = 0.1$ is a constant scaling factor. Two characters are matching only if they are the same and are not farther than $*\frac{\max(|i|,|j|)}{2} - 1$. Half the number of matching characters in different sequence order is the number of transpositions.[67]

[67]The Jaro-Winkler distance has been recently employed in the economic history literature for inter-

The Jaro-Winker distance has been shown to perform well in linking routines (Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez, 2021). In our particular case, however, this metric outperforms more standard string dissimilarity metrics, such as the cosine or the Levenshtein distances, because the Jaro-Winkler assigns a "bonus" score to strings starting with closer initial substrings. In addition, coding errors are far more frequent at the end of names and surnames than at the beginning. A manual assessment confirmed that the Jaro-Winkler metric outperforms other measures in our setting.

3.12.2 Internal and External Validation

Matching Statistics

We now present key statistics on the dataset that we assemble. In Figure 3.19, we report the matching rate by the number of matches (panel 3.19(a)) and over time (panel 3.19(b)). The matching rate is the ratio between the number of matched individuals and the number of English and Welsh immigrants in the US census. We break down the matching rate by the number of matches every immigrant is associated with. About 40% of the overall immigrant population is matched to one single record in the UK census. Another 10% is matched to two records, and the remaining 50% is matched to three or more records in the UK census. By construction, we can never match someone not appearing in the UK census. This is possible if a child born in, say, 1895 emigrates before 1901, which is the closest subsequent census. In Figure 3.19(a), we report a corrected matching rate whose denominator removes these "unmatchable" observations. Overall, 55% of the total number of English and Welsh immigrants is matched to no more than two records in the UK census. This constitutes the baseline sample that we analyze. A 55% matching rate is consistent with standard historical linking algorithms (Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez, 2021), although a more precise quantitative assessment is complex because the benchmark statistics refer to intergenerational census linking exercises.

generational linking purposes by, among others, Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021)

In panel 3.19(b), we report the matching rate by immigration year. In blue, we report the total number of immigrants; those paired with at least one match are shown in red; the green area reports our baseline sample, which is composed of all those immigrants with no more than two matches. We also impose a quality threshold on names and surnames. Suppose an immigrant is matched to someone born in the same year. In that case, we require both the name and the surname to have a similarity above .85. If an immigrant is matched to someone born either one year before or one year after, we impose a stricter threshold of .9 on both name and surname. We set high thresholds because we are concerned about false positive matches. Following Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021), we are thus willing to give up on power to maximize accuracy. In Figure 3.20, we report the overall distribution of name (panel 3.20(a)) and surname (panel 3.20(b)) match quality. The solid and dashed red lines superimpose the aforementioned coarse and strict thresholds. The quality distribution is substantially skewed to the right: most matches are of excellent quality. Dropping low-quality ones is, therefore, quantitatively second-order.

Since we match immigrants to the UK census before their migration year, the matching rate decreases over a decade. This is clear from the black line in Figure 3.19(b), which jumps up at the turn of each decade until 1911. The matching rate before 1881 is relatively low. This is because no census was taken in the UK in 1871. Therefore, we match all those who migrated to the US between 1870 and 1881 to the 1861 census. This mechanically reduces the matching rate, for we cannot match all those born between 1862 and 1881 who migrate during this period. Similarly, the matching rate decreases after 1911. This is because censuses after 1911 are protected by British privacy law. We thus match all those who migrate after 1911 to that census. However, this implies that we cannot match all those who migrated after 1911 and were born after that year. To ensure that our results are not driven by these asymmetries at the edges of the sample, in robustness analyses, we show that restricting the period to the years 1880-1920 does not affect our main findings.

Number of Matches and Observable Characteristics

One plausible concern is that instances of migrants with multiple matches in the UK census are not randomly distributed. This may be due to various reasons (Bailey, Cole, Henderson and Massey, 2020). First, educated individuals are more likely to report their name and surname in full, with consistent spelling over time. This would generate non-classical measurement error because the matching rate would be higher for a selected sub-sample of the population. This issue does not seem to be relevant in this case, as the matching rate—*i.e.* the share of immigrants that are *eventually* matched, irrespective of the number of matches—approaches the universe of the observations. Second, the number of matches may not be orthogonal to individual characteristics. This may be the case if wealthier individuals give relatively uncommon names, as documented by Olivetti, Paserman, Salisbury and Weber (2020). To assess the severity of this concern, we regress the number of matches on a set of individual-level observable variables observed in the US and UK censuses. Under classical measurement error, we would expect no statistically significant correlation between the number of matches and observable characteristics. Table 3.13 reports the estimates thus obtained. We find minimal and marginally significant correlations between the number of matches and individual-level characteristics observed in the US census. The number of matches correlates positively with agriculture and low-skilled employment. However, these correlations are very small: one more match is associated with a .01% increase in the probability of being employed in agriculture. This association is marginally larger for low-skilled manufacturing employment (0.03%). These very low magnitudes are unlikely to affect the results we document in this paper quantitatively. Moreover, notice that most correlations are not statistically significant. Most importantly, we do not find any significant association between the number of matches and the location of English immigrants. This is reassuring because our identification assumption crucially hinges on the variation arising from settlement decisions. We believe this is solid evidence of our linking algorithm and the novel database we assemble.

Plausibility Checks

Official statistics do not contain disaggregated data on emigration outflows. We thus rely on data compiled by Baines (2002) to attempt a validation of our series. These are not, however, based on official reports. The author tabulates emigration figures estimating the "missing population" from enumeration tables published by the census. This methodology yields necessarily approximate results. Moreover, and more crucially for our analysis, Baines (2002) is only able to construct data by counties, a much coarser level of aggregation than registration districts, and report figures on the overall number of overseas emigrants. These include outflows toward Scotland, Ireland, as well as the US and other overseas destinations. Lastly, the data only cover the last three decades of the nineteenth century. These caveats imply that we do not expect this validation exercise to yield unambiguously conclusive results. This notwithstanding, since the US was a major destination for British emigrants, this comparison is useful to gauge the plausibility of our estimates. Figure 3.21 reports the correlation between the two datasets. We find a positive and statistically significant correlation between overall out-migration and US emigration, both unconditionally (3.21(a)), as well as conditioning on county fixed effects (3.21(b)) and county fixed effects and a time trend (3.21(c)). Overall, this exercise indicates that our linked dataset is consistent with the previous historical literature.

We now describe an exercise to evaluate the plausibility of the linking algorithm. Building on Abramitzky, Boustan, Eriksson, Feigenbaum and Pérez (2021), we construct an intergenerational linked sample of English and Welsh individuals from the population censuses in 1881, 1891, and 1901.[68] The algorithm is very standard: for any given individual in census t, we look at individuals with the same name, surname, and birth year—with a one-year tolerance—who were recorded living in the same parish at year $t + 10$. If at least one record is found, we link that individual to that record(s). Otherwise, we look for potential matches in the same district. If no match is found, we leave that individual unmatched. The idea of the exercise is the probability to link migrants to the census after

[68]We do not use the 1861 because no census was taken in 1881. This would force us to link individuals in the 1861 census to the 1881 one. However, this imbalance may bias the linking rate between 1861 and 1881. We thus prefer to focus on the censuses for which we have the follow-up taken one decade after.

they migrate to the US should be lower than for the rest of the (non-migrant) population.

We compare the linking rate across migrants and non-migrants in Figure 3.22.[69] The blue bars report the linking rate in the intergenerational sample for migrants; the red bars, instead, refer to non-migrants. We find that non-migrants are more than two times more likely than migrants to be linked to the follow-up census. The matching rate of the intergenerational sample is 42% for non-migrants, but it is only 21% for US migrants. The most conservative interpretation of this result is that it provides an upper bound to the share of false positive matches of the international migrant sample. Suppose that *all* matches in the intergenerational linked sample were true positives. Then, the share of false positive links in the migrant sample would be 40%. In other words, even in this "worst-case" scenario, approximately 60% of the linked migrant matches would be true positives. It should be noted, however, that this represents a somewhat unlikely limit case for intergenerational linkage techniques display substantial type-I error rates (Bailey, Cole, Henderson and Massey, 2020). Overall, we view this exercise as evidence in favor of the plausibility of the international migrant sample.

To further assess the robustness of the UK-US linkage, we perform one additional linking exercise that excludes individuals that are matched in the intergenerational linked sample from the pool of entries which we attempt to link US migrants with. In other words, we exclude individuals that we would identify as plausibly living in the UK ten years after a given census is taken. In Figure 3.23 we compare this linked migrant sample with the baseline dataset that does not apply this trimming to the set of potential matches. These two exercises yield extremely consistent migration flows.

3.12.3 Return Migration Data

Following the logic explained in section 3.12.1, we construct a linked sample of return migrants. This identifies English and Welsh immigrants in the US in decade d and looks for possible matches in the UK census in decade $d + 1$, using a minor variation on the

[69]To avoid differential attrition due to mortality across migrants and non-migrants, we restrict the sample to individuals that were no older than 40 in the starting census year.

algorithm described previously. Since the last UK census that we have is the 1911 one, we face a hard upper bound for the coverage of return migration, as we can only construct return migrants linked samples spanning the period 1870–1910.

Previous research suggests that return migration rates during the Age of Mass Migration were substantial (Bandiera, Rasul and Viarengo, 2013), although probably less so in the UK than in second-wave countries such as Italy. Using our linked sample methodology, we find an approximately 30% return migration rate, broadly consistent with previous estimates.

3.12.4 Summary Statistics and Stylized Facts

The newly developed dataset we construct presents some key novelties compared to available data. It is the first dataset that allows retrieving the origin of US immigrants from England and Wales at a fine level of geographical aggregation during a period of massive international migrations (1880–1930).[70] The dataset's granular—individual—structure allows us to observe several individual characteristics of immigrants at home and in the US. This section briefly discusses key stylized facts that our new data allow us to document.

In Figure 3.24, we explore the origin of English and Welsh emigrants to the US over time. Each figure reports the emigration rate normalized by population in 1900, in thousand units. Two patterns emerge. First, substantial cross-sectional heterogeneity exists in the intensity of out-migration across districts throughout the sample period. Second, we find that the intensity of US emigration flows is initially larger in rural districts, especially in the South West and East regions, but this shifts over time toward industrial and urban areas. By the 1910s, the industrialized Lancashire districts featured as a prominent area of emigration. This finding provides a sound quantitative validation of historical—largely anecdotal—evidence (Erickson; Baines, 1972; 2002).

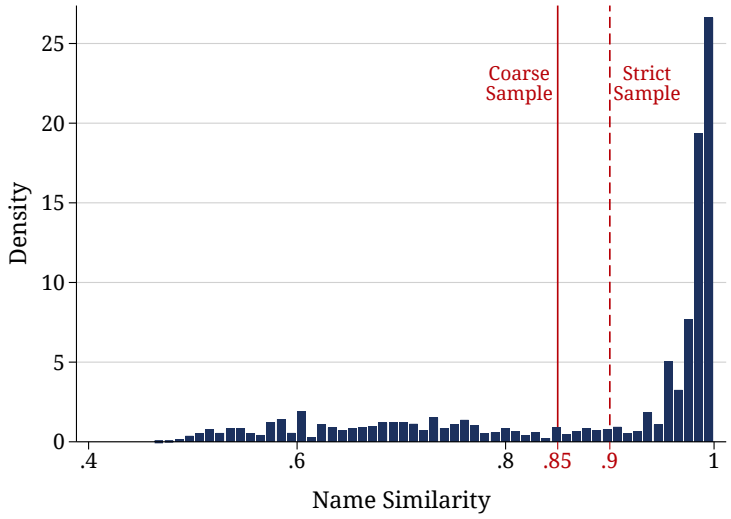
[70]Similar data-sets have been produced for Norwegian Abramitzky, Boustan and Eriksson (2014) and Swedish (Andersson, Karadja and Prawitz, 2022) immigrants. Our is the first such effort for a major European country: in 1890, the population in England and Wales stood at more than 27 million inhabitants. This compares to approximately 2 million Norwegians and 4.7 million Swedes.

Additionally, we can study the selection patterns of English and Welsh emigrants along two margins. Specifically, we can compare them to (i) the native US population in the areas where they settled and (ii) the non-migrant population in England and Wales who lived in their origin areas. These exercises extend seminal historical work by Baines (2002), who performed a similar exercise using incomplete information from the population censuses. We defer a discussion of selection patterns to the main text. Here, we only note that our dataset is well-suited to study the selection of British emigrants because it identifies individuals before they migrate, thus conveying a complete picture of selection issues during the period.

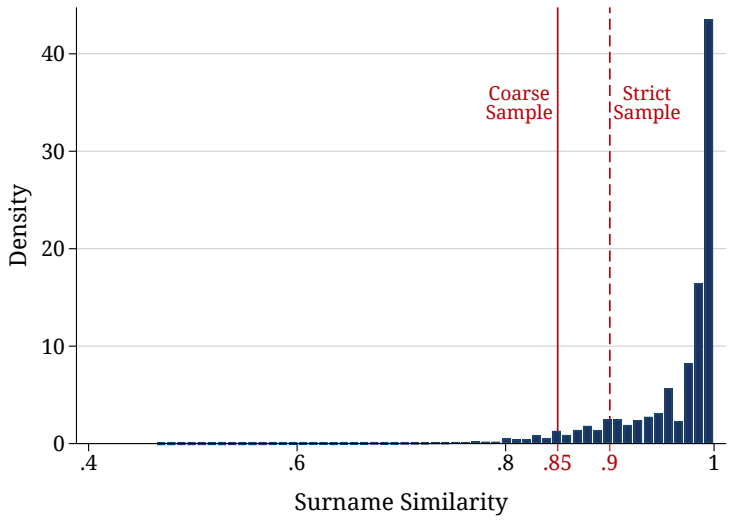
3.12.5 Figures

Figure 3.20: Quality of Matches in the Complete Linked Sample: Names and Surnames

((a)) Name Match Quality



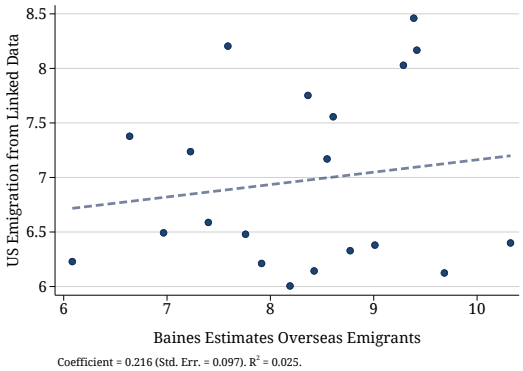
((b)) Surname Match Quality



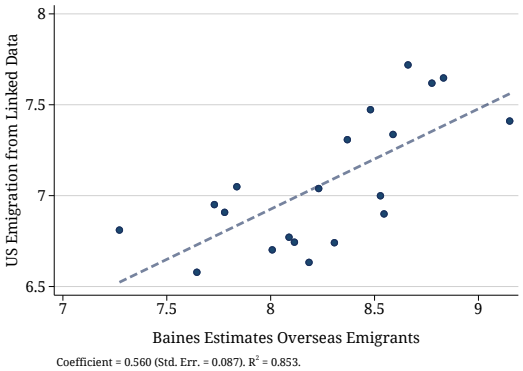
Notes The figures report the distribution of the match quality in terms of name and surname similarity for the set of records with no more than two matches in the baseline sample. The similarity measure we use to construct the links is the Jaro-Winkler. This string metric measures the edit distance between the name and surname of the British immigrant recorded in the US census and their match(es) in the UK census. Panel 3.20(a) reports the distribution of the name similarity; Panel 3.20(b) refers to surnames. The vertical lines mark the quality thresholds we impose for a match to be part of the final linked sample.

Figure 3.21: Comparison Between Linked Data and Estimates from Baines

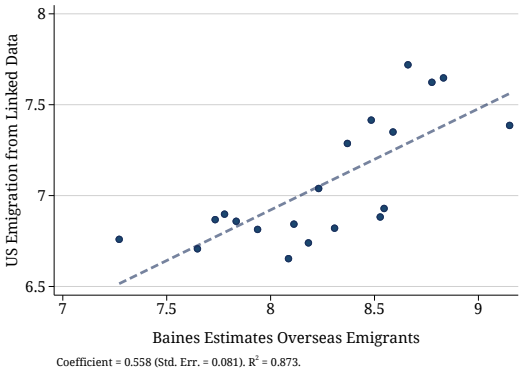
((a)) Unconditional Correlation



((b)) County Fixed Effects



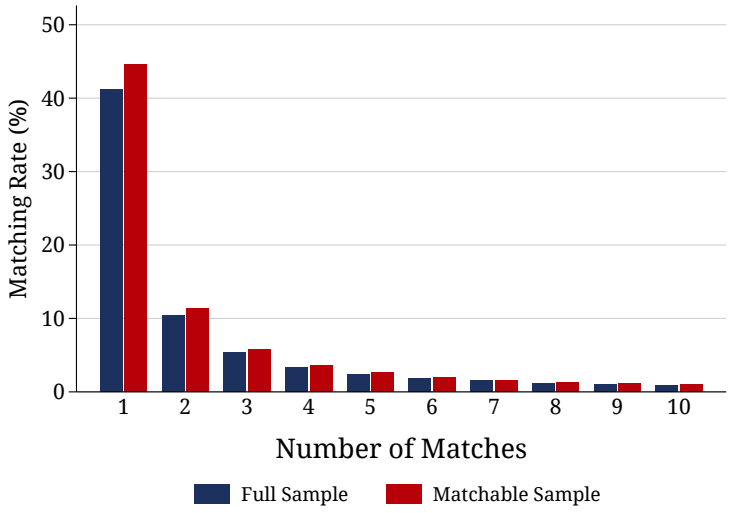
((c)) County Fixed Effects and Decade Trend



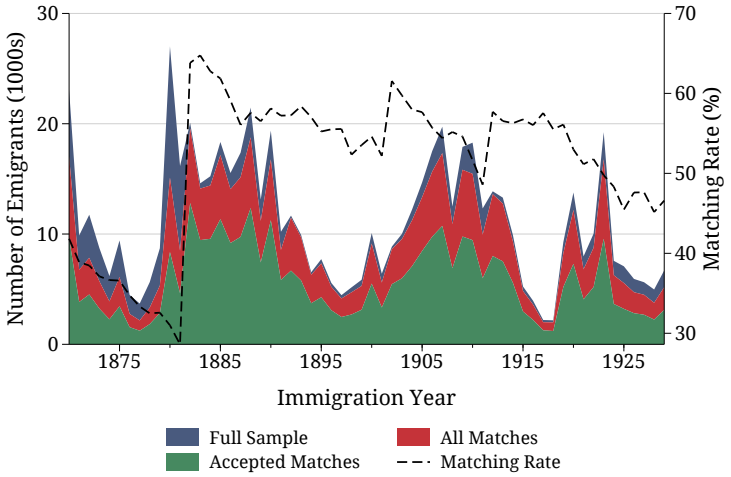
Notes. These figures report the correlation between county-level out-migration measured using our linked emigrant sample and data from (Baines, 2002). The dataset listed by the author is at the county level at a decade time frequency between 1870 and 1900 and reports the overall number of overseas emigrants. Thus, it conflates emigration to Scotland, Ireland, European, and trans-oceanic out-flows. In panel 3.21(a) we correlate the two series; in panels 3.21(b), we control for county fixed effects; in panel 3.21(c), we include a decade time trend. Observations are weighted by county-level population in 1880. Each graph reports in note the regression coefficient, along with its standard error, and the coefficient of determination of each regression.

Figure 3.19: Share of British Immigrants in the US Census Matched to the UK Census

((a)) Matching Rate by Number of Matches

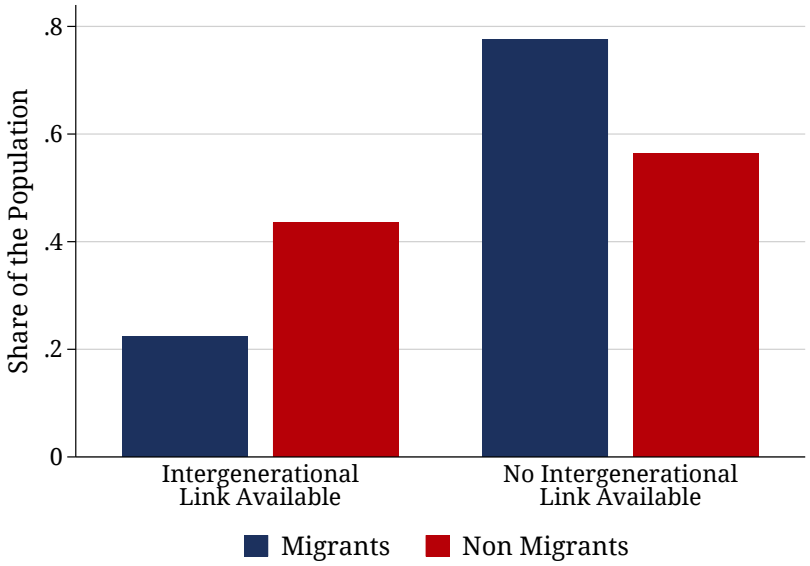


((b)) Matching Rate Over Time



Notes These figures report the share of English and Welsh immigrants recorded in the US census that we match to the UK census. Panel 3.19(a) plots the share of records that we match to the UK census and whose match quality is such that we retain it in the linked sample, broken down by the number of matches. In the baseline sample, we keep records with no more than two matches. Blue bars report ratios relative to the entire number of immigrants, and red bars restrict the set of immigrants to those we can match. Panel 3.19(b) reports the matching rate over time. The blue area reports the total number of US immigrants, the red area reports the entire number of matches we obtain, and the green area reports the matches that eventually enter our baseline linked sample. The black dashed line on the right y-axis is the ratio between the green and the blue areas.

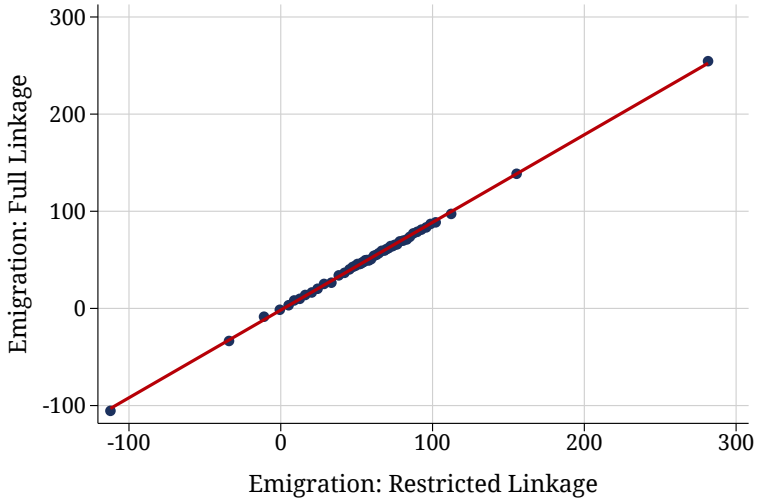
Figure 3.22: Falsification Exercise of the Intergenerational Linked Sample



Notes This figure reports the matching rate in the intergenerational linked sample. The blue bars report the matching rate among individuals that are identified as US migrants in the UK census. The red bars refer to individuals that are not identified as US migrants in the UK census. We exclude the 1861 census because there is no 1871 census in the UK and the intergenerational sample is unbalanced over the years 1861–1881. The intergenerational linked sample includes individuals with one single match obtained at the parish or at the district level. We exclude individuals over forty years old because differential mortality across the life cycle may impact the linking rate of the intergenerational sample.

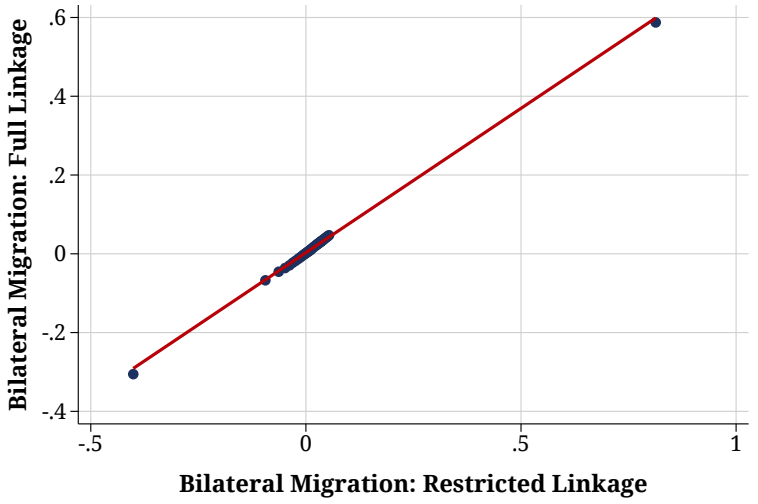
Figure 3.23: Comparison Between Linked Samples: Full and Restricted

((a)) District-Level Migration Flows



Notes. Coefficient = 0.906 (Clust. Std. Err. = 0.011). R^2 = 0.991.

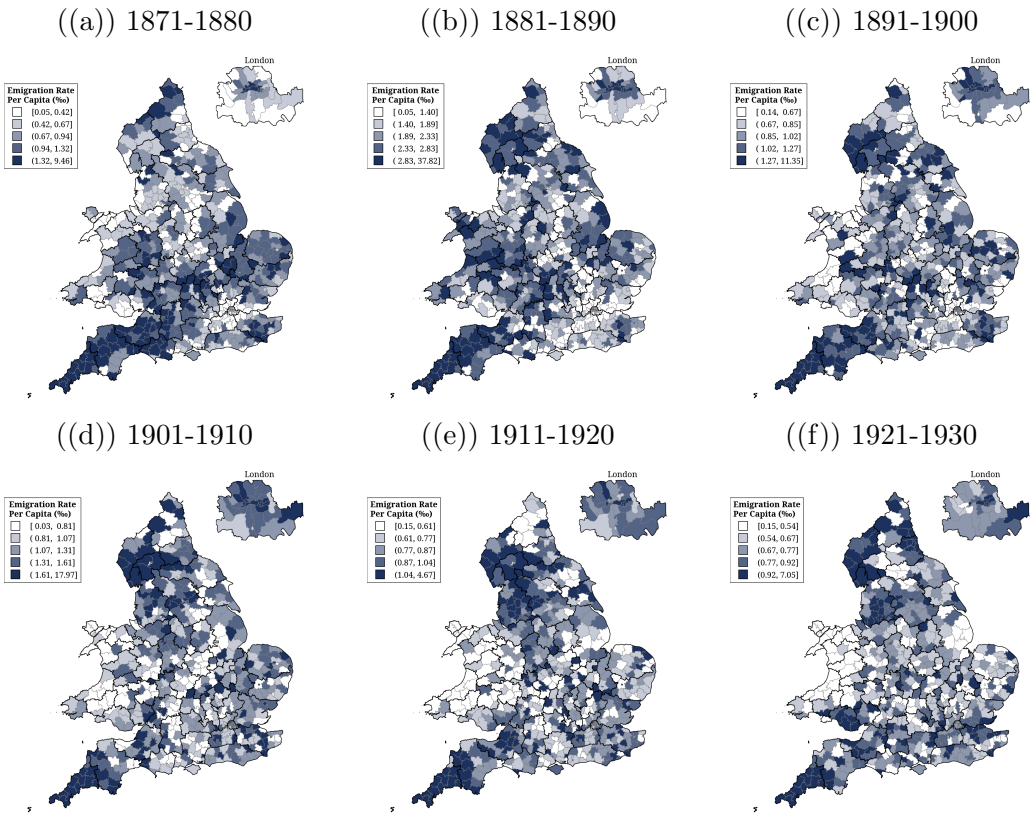
((b)) District-County Migration Flows



Notes. Coefficient = 0.866 (Clust. Std. Err. = 0.007). R^2 = 0.872.

Notes This figure compares the migration flows of the baseline UK-US linked sample, on the y-axis, with those obtained by restricting the pool of potential matches to those entries in the UK census that are not matched to any individual in the following census decade, on the x-axis. Panel 3.23(a) reports the district-level US emigration; the unit of observation is a district, observed at a decade frequency between 1870 and 1900. The figure includes district and decade fixed effects. The reported standard error is clustered at the district level. Panel 3.23(b) reports the district-county-level US migration flows; the unit of observation is a district-county pair, observed at a decade frequency between 1870 and 1900. The figure includes district, county, and decade fixed effects. The reported standard error is clustered at the district-by-county level.

Figure 3.24: Distribution of Emigration Rate Across Districts, 1871–1930



Notes These figures report the distribution of US emigrants across districts in England and Wales over the period 1871–1939 by decade. Data are from the matched emigrants' sample. The number of emigrants in each decade is normalized by population in 1900 and is expressed in ‰ units. Districts are displayed at their 1900 borders. Out-migration is also cross-walked to consistent historical borders. Lighter to darker blues indicate higher emigration rates.

3.12.6 Tables

Table 3.13: Correlation Between Number of Matches and Observable Characteristics

	Dep. Var.: Number of Matches		
	(1)	(2)	(3)
Panel A. Occupations			
Agriculture	0.003	0.004*	0.003**
	(0.001)	(0.001)	(0.001)
Low-Skilled Manufacture	0.013**	0.011**	0.008**
	(0.003)	(0.003)	(0.002)
High-Skilled Manufacture	0.006	0.007	0.008*
	(0.003)	(0.003)	(0.003)
Professionals	–0.001	–0.001	–0.001
	(0.001)	(0.001)	(0.001)
Public Administration	–0.004**	–0.003*	–0.002
	(0.001)	(0.001)	(0.001)
Manager	–0.000	–0.000	0.000
	(0.000)	(0.000)	(0.000)
Service Worker	0.001	0.001	0.002
	(0.001)	(0.001)	(0.001)
Panel B. Origin			
Northeast	–0.004		
	(0.006)		
Midwest	0.004		
	(0.004)		
South	–0.002		
	(0.001)		
West	0.001		
	(0.002)		
State FE	No	Yes	No
County FE	No	No	Yes
Year FE	No	Yes	Yes

Notes. This table reports the correlation between observable characteristics of British immigrants in the US census and the number of matches in the linked sample database. In each row, the table displays the correlation between the number of matches and an indicator equal to one if for immigrants that correspond to the row variable and zero otherwise. The sample is restricted to the set of matches we effectively use in the analysis. Column (1) reports unconditional correlations; column (2) includes state and census decade fixed effects; column (3) adds county fixed effects. In Panel A, the characteristics are the occupations; in Panel B, the variables are the Census Bureau region of residence. Standard errors, clustered at the county level, are shown in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

3.13 Appendix: Additional Results

This section presents in some detail several additional results that are mentioned in passing in the main text.

3.13.1 Trade-Induced Technology Transfer

Our favored explanation of the return innovation result is that migrants facilitate the flow of knowledge between the areas where they settle and those they originate from. We argue that those flows are fostered by the diffusion of information and by market integration. This section presents one more piece of evidence in this direction. We focus on international trade as a measure of bilateral market integration. Previous research documents that trade fosters innovation, either because of increased import competition (Bloom, Draca and Van Reenen; Autor, Dorn, Hanson, Pisano and Shu, 2016; 2020), export opportunities (Bustos; Atkin, Khandelwal and Osman; Aghion, Bergeaud, Lequien and Melitz, 2011; 2017; 2018), access to intermediate inputs (Juhász and Steinwender, 2018), and increased market size (Coelli, Moxnes and Ulltveit-Moe, 2022).[71] In our analysis, we interpret trade as a means of facilitating technology transfer between the UK and the US, following Aleksynska and Peri (2014) and Ottaviano, Peri and Wright (2018).

We consider a major shock to trade flows between the US and the UK: the 1930 Smoot-Hawley Act. The Act was a major trade policy reform enacted in response to the Great Depression (Eichengreen; Crucini, 1986; 1994). Importantly for our setting, the Act did not establish a uniform tariff rate. Instead, as we report in Table 3.11, tariffs vastly differed across industries before and after the shock. We leverage this variation, interacted with the before-Act knowledge exposure in a difference-in-differences setting.[72] The key

[71] Shu and Steinwender (2019) provide a critical assessment of the literature studying the effect of international trade on innovation.

[72] We first map sectors defined in the Act to technology classes. We then assign one class to the treatment group if its average *ad valorem* import duty changes by more than fifty percentage points between 1925–1930 and 1931–1936. Yearly tariff rates have been digitized from the *Statistical Abstract of the United States*.

idea that underlies this approach is that if migration linkages generate return innovation flows through international trade, then an increase in trade costs is expected to reduce patenting in the UK in the sectors that (i) districts were more exposed to, through migrations, and (ii) were targeted by the tariff increase.

We thus estimate the following double differences model separately for protected and non-protected industries:

$$\text{Patents}_{ik,t} = \alpha_{i\times k} + \alpha_t + \sum_{h=-a}^{b} \beta_h \times [1\,(t = 1931 + h) \times \text{Knowledge Exposure}_{ik}] + \varepsilon_{ik,t} \tag{3.25}$$

Where i, k, and t respectively denote a district, technology class, and year, and $\text{Knowledge Exposure}_{ik}$ is the average sector-level knowledge exposure in the decade before the Act (1920–1930). In the baseline analysis, an industry is protected if its tariff rate increases by more than 50 p.p. between 1925–1930 and 1931–1935. Then, we estimate the triple-differences specification that compares treated and non-treated industries:

$$\begin{aligned}\text{Patents}_{ik,t} = {} & \alpha_{i\times k} + \alpha_{i\times t} + \alpha_{k\times t} + \\ & + \sum_{h=-a}^{b} \beta_h \times [\text{Tariff}_k \times 1\,(t = 1931 + h) \times \text{Knowledge Exposure}_{ik}] + \varepsilon_{ik,t}\end{aligned} \tag{3.26}$$

where Tariff_k is an indicator returning value one for protected industries and zero otherwise.

In columns (1–2) of Table 3.16 we report the results of model (3.25). Column (1) presents the estimated coefficient for non-protected industries, while column (2) focuses on protected ones. We find no effect for the former and a negative effect for the latter. This is confirmed when looking at the associated flexible specification, reported in Figure 3.25. This also provides evidence supporting the parallel trends assumption for the two groups of technology classes. In columns (3–5), we report the estimates of the triple differences model (3.26). We consider three possible threshold values of the increase in the tariff rate after the Act to define a protected sector (10%, 30%, and 50%). All yield quantitatively similar estimates. Note, however, that the estimated ATE reassuringly increases

in absolute magnitudes for larger tariff increases.

The analysis suggests that trade—which we interpret as a proxy for market integration—is a relevant channel through which migration ties generate knowledge flows and technology transfer. However, it is worth noting that the magnitude of the estimated treatment effects of the tariff reform on UK innovation is modest, despite the large increase in tariff duties. We thus interpret trade as one additional, although plausibly not the pivotal, factor driving return innovation.

3.13.2 Selection of British Migrants

The historical scholarship argues that the English and Welsh mass migration to the US starkly differed from that of other countries (Berthoff; Baines, 1953; 2002). Unlike other European countries, such as Germany, Sweden, or Italy, UK emigration to the US in the second half of the nineteenth century was not a low-skilled rural phenomenon. Especially after the 1880s, people started to leave urban, industrial areas. Importantly, emigrants did not represent the bottom of the human capital distribution, as was the case in Italy (Spitzer and Zimran, 2018) or Norway (Abramitzky, Boustan and Eriksson, 2014). This is crucial for our analysis, as it is unlikely that illiterate farmers would facilitate the flow of novel knowledge back to their origin areas. Even if this was the case, it would be equally unlikely that those rural areas would have the ability to reproduce US patents. While these considerations are helpful for our analysis, they largely rely on anecdotal evidence or analyses of incomplete census sources. In this section, we present evidence on the selection of English emigrants to the US and their integration into the US. To construct these statistics, we leverage the novel linked sample that allows us to observe individual-level characteristics before emigrants left—in the UK census—and after they settled—in the US census.

Table 3.17 compares UK emigrants with the non-migrant population. Column (1) refers to non-migrants, and columns (2) and (5) refer to emigrants and return migrants, respectively. In columns (3) and (6), we compute the difference between non-migrants and

emigrants and non-migrants and return migrants, respectively. Migrants are less likely to work in agriculture and as professionals. They are, however, more likely to be employed in industrial sectors, such as textiles and metallurgy. This overall confirms the historical analysis of Baines (2002). Emigrants mainly originated from the North West, including Lancashire, and South West, chiefly, Devon and Cornwall. Similar patterns emerge when looking at return migrants, who are even less likely to be employed as agricultural workers. Return rates in high-emigration areas of the South West appear low compared to the rest of the country, while they are very high in the London area.

In Table 3.18, we compare English and Welsh immigrants to the rest of the US population. Column (1) refers to natives, and columns (2) and (5) refer to emigrants and return migrants, respectively. In columns (3) and (6), we compute the difference between natives and emigrants and natives and return migrants, respectively. UK immigrants differ substantially from the rest of the US population: they are less likely to work in agriculture and as civil servants. By comparison, they are more likely to be employed in metallurgy, textiles, and trade. This aligns well with evidence by Erickson (1972), who argues that English immigrants in the US tended to specialize in industries where they had a comparative advantage. Similar patterns emerge for return migrants. Regarding their geographical distribution, UK immigrants settled most commonly in the New England and Mid-Atlantic regions.

3.13.3 Long-Run Effect of Return Innovation

We now investigate the persistence of the effect of exposure to foreign knowledge through migration ties on the direction of patenting activity. While this exercise cannot be tasked with any claim of causality, it nonetheless suggests the possible far-reaching effects of out-migration on innovation.

We estimate the following regression:

$$\text{Patents}_{ik,t} = \alpha_{i\times k} + \alpha_t + \sum_{\tau\in\mathcal{T}} \beta^{\tau} \left[\text{Knowledge Exposure}_{ik} \times 1\left(t = \tau \,|\, t = \tau + 1\right)\right] + \varepsilon_{ik,t} \tag{3.27}$$

where i, k, and t denote a district, technology class, and year, respectively. In this setting, we have $t \in [1940, 2015]$. The term $\text{Knowledge Exposure}_{ik}$ refers to knowledge exposure in the years 1900–1930, i.e., before the sample period. To reduce noise in the estimated β^{τ} coefficients, we conflate years in $\mathcal{T}$ in biennial windows. The estimated set of β^{τ} expresses the conditional correlation between historical exposure to knowledge exposure and innovation activity in the two-year window indexed by τ.

In Figure B27, we report the set of estimated β^{τ} over time. The correlation between historical knowledge exposure and patenting activity remained positive and significant until the early 1980s, although it—reassuringly— decreased over time. We interpret this as evidence that exposure to foreign knowledge through migration ties has a potentially long-lasting effect on the composition of innovation activity over time. In Table 3.22 we re-estimate model (3.27), sector-by-sector, by decade. Compared to (3.27), we can thus only include district and decade fixed effects. Columns report the estimated β^{τ} by decade. The estimated correlation between historical exposure and patenting decreases over time in almost all sectors and only a few display significant coefficients after the 1980s.

3.13.4 Further Additional Results

Out-Migration and the Volume of Innovation

The main analysis concentrates on the effect of knowledge exposure on the direction of innovation. Knowledge exposure leverages variation in specialization across US counties and bilateral flows between UK districts and US counties. In this section, we briefly comment on the effect of out-migration on the *volume* of innovation.

We estimate variations on the following model:

$$\text{Patents}_{i,t} = \alpha_i + \alpha_t + \beta \times \text{US Emigrants}_{i,t} + \varepsilon_{i,t} \tag{3.28}$$

where $\text{US Emigrants}_{i,t}$ is the total number of emigrants from district i in decade t. As in the main text, we instrument total out-migration flows with the shift-share instruments constructed using railway-based and leave-out immigration shocks. Compared to the model estimated in the main text, endogeneity concerns in (3.28) are severe. However, if the instruments are valid, then the estimated β coefficients measured the causal effect of out-migration on patenting. A perhaps more crucial concern in regression (3.28) is that we do not have information on emigration to countries other than the US. Suppose emigration rates to, say, Australia or Canada (the second and third most common destinations) were correlated with US out-migration. In that case, we may fail to single out the effect of out-migration.

With these caveats in mind, in Table 3.19, we report the estimates of regression (3.28). In panel A, columns (1–3), we report the correlation between measured out-migration and patenting, while columns (4–6) and (7–9) display the reduced form association with, respectively, the railway-based and the leave-out instruments. In panel B we report the 2SLS estimates. We find that the contemporaneous effect of out-migration on innovation is negative. This is reasonable given that out-migration entails a loss of human capital, which, in the light of the selection analysis, was probably relatively skilled and is consistent with the "brain drain" literature. Once we lag emigration by one decade, however, we find a positive effect. This sign reversal is robust across the two instruments in the reduced form and the two-stage least-square estimates. It is tempting to interpret it as evidence of "brain gain", that is, that return innovation increases the volume of innovation (Docquier and Rapoport, 2012). While the results are consistent with this interpretation, they are not conclusive because of the caveats that underlie this exercise.

Assortative Matching

In this section, we lay down a simple framework to test whether British immigrants sort into US counties depending on the innovation similarity between the settlement location and their origin district. Let $\mathbf{P}_{j,t} = \{p_{1j,t}, \ldots, p_{Nj,t}\}$ denote the patent portfolio of county j in decade t, whose generic entry p_{kjt} returns the number of patents in technology class k. Analogously, let $\mathbf{P}_{i,t}$ be the portfolio of district i. We define a metric of innovation similarity as follows:

$$\text{Innovation Similarity}_{ij,t} \equiv \frac{\mathbf{P}_{i,t}^{\intercal}\mathbf{P}_{j,t}}{\|\mathbf{P}_{i,t}\| \cdot \|\mathbf{P}_{j,t}\|} = \frac{\sum_k p_{ki,t}p_{kj,t}}{\sqrt{\sum_k p_{ki,t}^2}\sqrt{\sum_k p_{kj,t}^2}} \leq 1 \tag{3.29}$$

which is a simple cosine similarity. The similarity measure returns value one if the patent portfolios of district i and county j are equal, meaning their composition across classes is the same. The index is normalized between zero and one.

We then estimate variations on the following simple linear probability model:

$$\text{Emigrants}_{i \to j,t} = \alpha_{i \times j} + \alpha_t + \beta \times \text{Innovation Similarity}_{ij,t} + X_{ij,t}^{\intercal}\Gamma + \varepsilon_{ij,t} \tag{3.30}$$

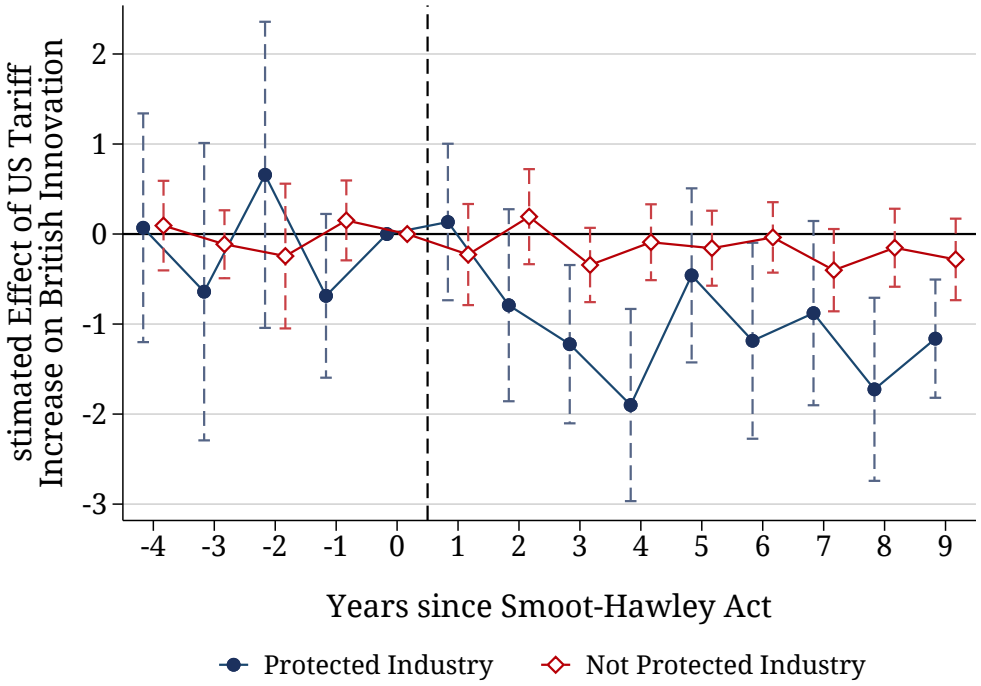
where the dependent variable is the flow of emigrants from district it to county j in decade d, and $\alpha_{i \times j}$ denotes county-by-district fixed effects. The coefficient β thus yields the correlation between the similarity of innovation activity and migration flows. The dependent variable is measured in logs, and standard errors are two-way clustered by district and county. Under sorting, one would expect $\hat{\beta} > 0$.

We test this prediction in Table 3.15. We find no correlation between the similarity of innovation portfolios across county districts and the migration flow between them. This holds irrespective of whether we take the contemporaneous similarity (columns 1–2) or if we lag by one (columns 3–4) or two (columns 5–6) decades. Notably, the standardized beta coefficient of the innovation similarity term is always minimal in magnitude. This suggests that assortative matching based on innovation similarity between origin and destination places is probably not a significant threat to a causal interpretation of our

estimates. This notwithstanding, since the similarity of innovation portfolios is measured with error, we do not claim that we can exclude it *tout court*.

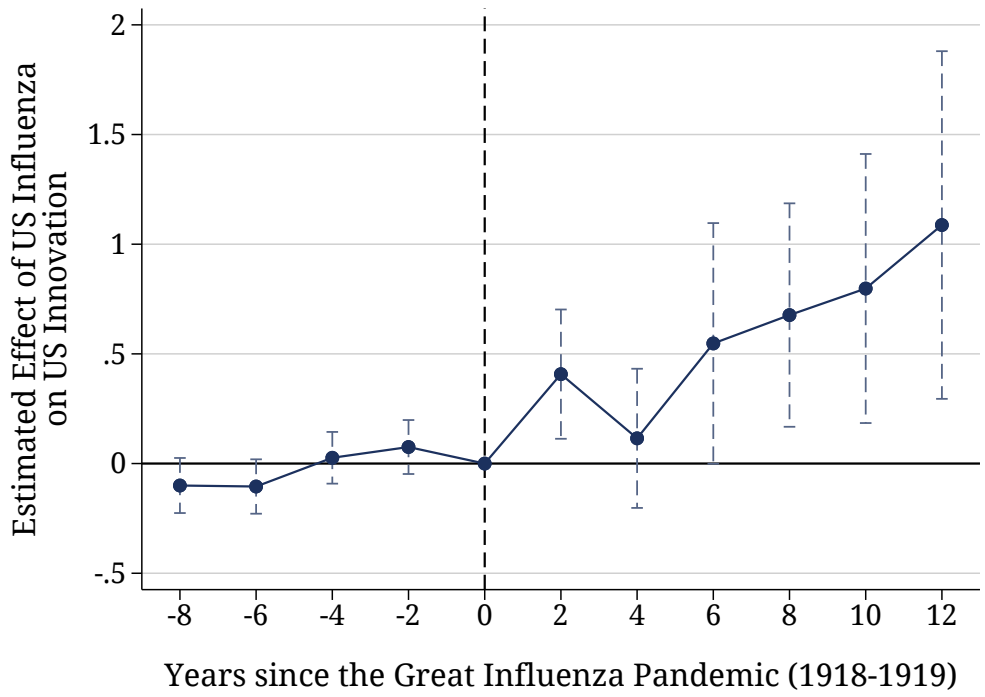
3.13.5 Figures

Figure 3.25: Flexible Difference-in-Differences Estimated Effect of Tariff Reform on Innovation



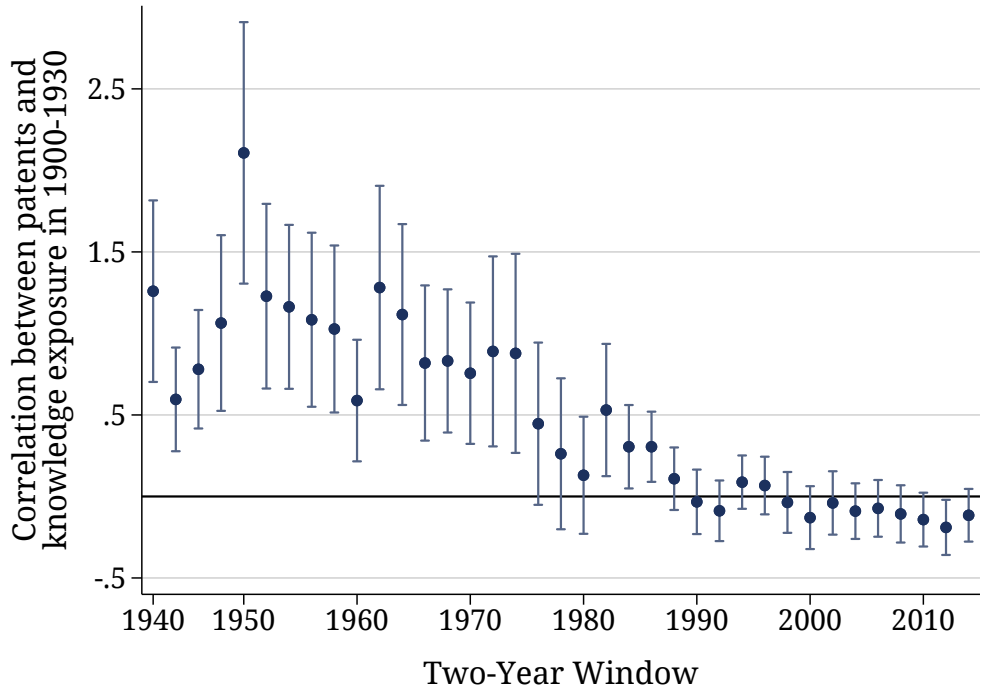
Notes. This figure reports the estimated dynamic treatment effects of increased US tariff rates on innovation in Britain. The unit of observation is a district-technology class pair observed at a yearly frequency between 1920 and 1939. The dependent variable is the number of patents. The independent variable is the interaction between knowledge exposure over 1910–1920 and year dummies. The last year before the Reform, 1929, is the baseline category. The blue dots report the estimated treatment effects for technology classes targeted by the Act; the red dots restrict the sample to non-treated technology classes. We define a class as "targeted" if its average tariff rate increases by more than 50% after the Smoot-Hawley Act. Regressions include district-by-class and year fixed effects. Standard errors, clustered at the district level, are reported in parentheses.

Figure 3.26: Flexible Triple Differences Estimated Effect of the Influenza Pandemic on US Innovation



Notes. These figures report the dynamic treatment effects of exposure to the Great Influenza Pandemic on innovation in the US. The units of observation are county-technology class pairs; units are observed at a yearly frequency between 1900 and 1939. The dependent variable is the number of patents. The treatment is an indicator equal to one for pharmaceutical patents and districts in the top quartile of the excess mortality distribution. The graph displays the interaction coefficients between the treatment and biennial time dummies, where the last dummy before the pandemic—1916–1917—serves as the baseline category. Excess mortality is computed as the average number of deaths during the pandemic over the average number of deaths in the three years before the pandemic. The black dashed line indicates the timing of the treatment. The regression includes county-by-technology class, technology class-by-biennial, and county-by-biennial fixed effects. Standard errors are two-way clustered by district and technology class; bands report 95% confidence intervals.

Figure 3.27: Long-Run Association Between Knowledge Exposure and Subsequent Innovation Activity, 1940–2020



Notes. This figure reports the correlation between knowledge exposure in the period 1900–1930 and subsequent innovation activity. The unit of observation is a district-technology class pair. Units are observed at a biennial frequency between 1940 and 2015. Each dots report the coefficient of an interaction term between—time-invariant—knowledge exposure and biennial time dummies. The last biennial, 2014–2015, serves as the baseline category. The model includes district-by-technology class and decade fixed effects. Standard errors are clustered at the district level. Bands report 95% confidence intervals.

3.13.6 Tables

Table 3.14: Zero-Stage Regressions Between Immigrant Shares and Railway Access

	Baseline	Excluding States in...			
	(1)	(2) Northeast	(3) Midwest	(4) South	(5) West
$I^{\text{Rail}}_{t-1} \times$ Immigrant Flow$_{t-1}$	0.372***	0.399***	0.198**	0.461**	0.252**
	(0.102)	(0.111)	(0.097)	(0.198)	(0.101)
I^{Rail}_{t-1}	0.845	4.014	-4.288	-17.024***	3.374
	(2.765)	(2.775)	(2.798)	(5.918)	(2.775)
County FE	Yes	Yes	Yes	Yes	Yes
Decade FE	Yes	Yes	Yes	Yes	Yes
N. of Counties	2759	2543	1742	1513	2479
N. of Observations	17308	15803	10919	9222	15980
R^2	0.905	0.903	0.921	0.880	0.915
Mean Dep. Var.	79.842	74.284	55.019	132.174	72.101

Notes. This table reports the results of the zero-stage regressions that we estimate to construct the railway-based county-level immigration shocks. This table largely replicates Sequeira, Nunn and Qian (2020). The unit of observation is a county observed at a decade frequency between 1870 and 1930. The dependent variable is the share of the foreign-born population. The main dependent variable is an interaction between the one-decade-lagged national inflow of immigrants and an indicator variable that returns value one if the county was connected to the national railway network in the previous decade and zero otherwise. The regressions also control for the railway indicator, the lagged share of foreign-borns, an interaction between lagged national industrial production and the railway indicator, an interaction between lagged GDP and the railway indicator, population density, the share of the population living in urban centers, and an interaction between the share of the urban population and the national inflow of immigrants. The parameter restriction imposed by the instrument's logic requires that the railway indicator's coefficient be non-positive. In column (1), the sample is the universe of counties; in columns (2), (3), (4), and (5), we drop states in, respectively, the North-East, Midwest, South, and West Census Bureau regions. Each regression includes county and decade fixed effects. Standard errors, clustered at the county level, are displayed in brackets. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.15: British Immigrants Assortative Matching Across US Counties

	Contemporaneous		10 Years Lag		20 Years Lag	
	(1)	(2)	(3)	(4)	(5)	(6)
Innovation Similarity	0.083	201.876				
	(2.847)	(155.968)				
Innovation Similarity$_{t-1}$			0.419	333.940		
			(2.800)	(205.476)		
Innovation Similarity$_{t-2}$					1.370	-172.428
					(2.485)	(218.951)
District-County FE	Yes	Yes	Yes	Yes	Yes	Yes
Decade FE	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Counties	1743283	32176	1665941	31383	1505060	29948
N. of Observations	9029476	88636	7266084	86789	5476833	83220
Sample	All	Non-Zero	All	Non-Zero	All	Non-Zero
R^2	0.473	0.675	0.553	0.675	0.662	0.676
Mean Dep. Var.	0.022	1.617	0.027	1.635	0.034	1.670
Std. Beta Coef.	0.000	0.010	0.000	0.018	0.001	-0.010

Notes. This table reports the association between the similarity of innovation activity and migration flows between US counties-UK districts pairs. The unit of observation is a county-district pair, observed at a decade frequency between 1870 and 1920. The dependent variable is the number of emigrants that leave the given district and settle in the given county. The independent variable is the similarity of the innovation portfolios between the county and the district. The innovation similarity is computed as the cosine distance of the respective patent portfolios over the decade. Columns (1), (3), and (5) report results for the universe of county-district pairs; columns (2), (4), and (6) restrict to pairs with non-zero migration flows. Columns (1) and (2) estimate the contemporaneous correlation; in columns (3) and (4), innovation similarity appears with a one-decade lag; in columns (5) and (6), it is included with a two-decade lag. Regressions include district-by-county and decade fixed effects. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.16: Double and Triple Differences Effect of The Smoot-Hawley Act on Innovation

	Double Differences		Triple Differences		
	(1) Not Protected	(2) Protected	(3) +10%	(4) +30%	(5) +50%
Knowledge Exposure × Post	-0.040	-0.463**			
	(0.063)	(0.219)			
Knowledge Exposure × Post × Protected (+10%)			-0.469***		
			(0.078)		
Knowledge Exposure × Post × Protected (+30%)				-0.478*	
				(0.269)	
Knowledge Exposure × Post × Protected (+50%)					-0.685***
					(0.206)
Year FE	Yes	Yes	–	–	–
District-Year FE	No	No	Yes	Yes	Yes
District-Class FE	Yes	Yes	Yes	Yes	Yes
Class-Year FE	No	No	Yes	Yes	Yes
N. of District-Class	632	632	632	632	632
N. of Observations	63200	37920	101120	101120	101120
R^2	0.653	0.563	0.713	0.713	0.713
Mean Dep. Var.	2.125	1.260	1.801	1.801	1.801
Std. Beta Coef.	0.000	0.000	0.000	0.000	0.000

Notes. This table reports the estimated effect of an increase in the US tariff rate on innovation in Britain. The unit of observation is a district-technology class pair observed at a yearly frequency between 1920 and 1939. The dependent variable is the number of patents by district technology class. In columns (1–2), the independent variable is the interaction between knowledge exposure over 1910–1920 and a post-reform (1930) indicator variable. The regression in column (1) is estimated over technology classes not targeted by the Act; in column (2), we focus on classes that the Act targets. We define a class as "targeted" if its average tariff rate increases by more than 50% after the Smoot-Hawley Act. In columns (3), (4), and (5), the treatment interacts the previous one with an indicator that returns value one for technology classes whose tariff rates increases by more than, respectively, 10%, 30%, and 50% after 1930. Regressions (1–2) are thus double-difference designs; regressions (3–5) are triple-difference designs. Consequently, in columns (1–2), we include district-by-class and year fixed effects, while in columns (3–5), we add district-by-year and technology-by-year fixed effects. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.17: Selection of US Emigrants Compared to the Rest of the British Population

	Non Migrants	Emigrants			Return Migrants		
	(1) Mean	(2) Mean	(3) Difference	(4) Std. Err.	(5) Mean	(6) Difference	(7) Std. Err.
Panel A. Employment (Dependent variable = 1 if individual employed in:)							
Agriculture	0.281	0.271	-0.009***	(0.001)	0.252	-0.028***	(0.002)
Chemicals	0.008	0.008	-0.001**	(0.000)	0.009	0.001**	(0.000)
Construction	0.141	0.142	0.001	(0.001)	0.145	0.004***	(0.002)
Engineering	0.138	0.138	-0.000	(0.001)	0.148	0.010***	(0.002)
Liberal Profession	0.035	0.032	-0.002***	(0.000)	0.036	0.002*	(0.001)
Metallurgy	0.029	0.034	0.005***	(0.001)	0.032	0.003***	(0.001)
Other Manufacturing	0.074	0.074	-0.000	(0.001)	0.074	-0.001	(0.001)
Public Administration	0.030	0.028	-0.001***	(0.000)	0.032	0.002***	(0.001)
Textiles	0.090	0.099	0.009***	(0.001)	0.082	-0.008***	(0.001)
Trade	0.072	0.079	0.007***	(0.001)	0.078	0.007***	(0.001)
Transport	0.097	0.090	-0.007***	(0.001)	0.103	0.006***	(0.001)
Utilities	0.007	0.006	-0.000	(0.000)	0.009	0.002***	(0.000)
Panel B. Region of Residence (Dependent variable = 1 if individual lives in:)							
East	0.102	0.086	-0.016***	(0.001)	0.089	-0.014***	(0.001)
East Midlands	0.065	0.057	-0.007***	(0.000)	0.058	-0.006***	(0.001)
London	0.132	0.129	-0.003***	(0.001)	0.139	0.006***	(0.001)
North East	0.067	0.070	0.003***	(0.000)	0.070	0.003***	(0.001)
North West	0.179	0.194	0.015***	(0.001)	0.199	0.020***	(0.001)
South East	0.120	0.110	-0.009***	(0.001)	0.117	-0.003***	(0.001)
South West	0.063	0.085	0.022***	(0.001)	0.065	0.002***	(0.001)
Wales	0.070	0.064	-0.006***	(0.000)	0.069	-0.001	(0.001)
West Midlands	0.114	0.110	-0.004***	(0.001)	0.108	-0.006***	(0.001)
Yorkshire	0.088	0.094	0.006***	(0.001)	0.087	-0.001*	(0.001)

Notes. This table compares observable individual characteristics of US emigrants with the rest of the British population. In each row, we define a dummy variable equal to one for individuals in the given employed in the given sector (Panel A) or residing in the given division (Panel B) and compute the average for non-migrants (column 1), emigrants (column 2), and return migrants (column 5). Columns (3) and (6) report the difference between columns (1) and, respectively, columns (2) and (5). Robust standard errors are reported in columns (4) and (7). *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.18: Selection of British Immigrants Compared to the Rest of the US Population

	US Population	Immigrants			Return Migrants		
	(1) Mean	(2) Mean	(3) Difference	(4) Std. Err.	(5) Mean	(6) Difference	(7) Std. Err.
Panel A. Employment (Dependent variable = 1 if individual employed in:)							
Agriculture	0.215	0.121	-0.094***	(0.001)	0.129	-0.086***	(0.001)
Chemicals	0.006	0.010	0.004***	(0.000)	0.006	-0.001***	(0.000)
Construction	0.044	0.096	0.052***	(0.001)	0.087	0.042***	(0.001)
Engineering	0.434	0.199	-0.235***	(0.001)	0.262	-0.172***	(0.002)
Liberal Profession	0.042	0.078	0.036***	(0.001)	0.063	0.021***	(0.001)
Other Manufacturing	0.076	0.159	0.083***	(0.001)	0.148	0.072***	(0.001)
Public Administration	0.014	0.009	-0.005***	(0.000)	0.008	-0.006***	(0.000)
Textiles	0.015	0.076	0.061***	(0.001)	0.080	0.066***	(0.001)
Trade	0.069	0.104	0.035***	(0.001)	0.092	0.023***	(0.001)
Transport	0.056	0.087	0.031***	(0.001)	0.085	0.029***	(0.001)
Utilities	0.028	0.059	0.031***	(0.001)	0.041	0.013***	(0.001)
Panel B. Region of Residence (Dependent variable = 1 if individual lives in:)							
East North Central	0.205	0.210	0.005***	(0.001)	0.192	-0.014***	(0.001)
East South Central	0.087	0.008	-0.079***	(0.000)	0.008	-0.078***	(0.000)
Mid Atlantic	0.208	0.350	0.143***	(0.001)	0.365	0.157***	(0.002)
Mountain	0.030	0.058	0.028***	(0.000)	0.062	0.032***	(0.001)
New England	0.068	0.165	0.097***	(0.001)	0.187	0.120***	(0.001)
Pacific	0.054	0.101	0.047***	(0.001)	0.072	0.018***	(0.001)
South Atlantic	0.130	0.026	-0.104***	(0.000)	0.023	-0.107***	(0.000)
West North Central	0.123	0.067	-0.055***	(0.001)	0.077	-0.045***	(0.001)
West South Central	0.095	0.014	-0.081***	(0.000)	0.013	-0.082***	(0.000)

Notes. This table compares observable individual characteristics of British immigrants with the rest of the US population. In each row, we define a dummy variable equal to one for individuals in the given employed in the given sector (Panel A) or residing in the given division (Panel B) and compute the average for non-migrants (column 1), immigrants (column 2), and return migrants (column 5). Columns (3) and (6) report the difference between columns (1) and, respectively, columns (2) and (5). Robust standard errors are reported in columns (4) and (7). *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.19: Association Between Out-Migration and the Volume of Innovation

	Dependent Variable: Number of Patents								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. OLS Estimates									
	Measured US Emigration			Railway Instrument			Leave-Out Instrument		
US Emigrants$_t$	-1.453***								
	(0.303)								
US Emigrants$_{t-1}$		0.951***							
		(0.258)							
US Emigrants$_{t-2}$			-0.702						
			(0.646)						
Railway-Predicted Emigrants$_t$				-0.141***					
				(0.038)					
Railway-Predicted Emigrants$_{t-1}$					0.283***				
					(0.074)				
Railway-Predicted Emigrants$_{t-2}$						0.212**			
						(0.107)			
Leaveout-Predicted Emigrants$_t$							-0.307***		
							(0.093)		
Leaveout-Predicted Emigrants$_{t-1}$								0.473***	
								(0.126)	
Leaveout-Predicted Emigrants$_{t-2}$									0.162
									(0.134)
Std. Beta Coef.	-0.228	0.125	-0.086	-0.185	0.284	0.196	-0.095	0.106	0.034
Panel B. Two-Stage Least-Square Estimates									
	Railway Instrument			Leave-Out Instrument			Overidentified 2SLS		
US Emigrants$_t$	-1.479***			-2.351***			-1.433***		
	(0.372)			(0.396)			(0.389)		
US Emigrants$_{t-1}$		1.670***			1.424***			1.662***	
		(0.458)			(0.366)			(0.455)	
US Emigrants$_{t-2}$			-28.128			59.484			-18.895
			(30.883)			(377.778)			(15.412)
Std. Beta Coef.	-0.232	0.219	-3.441	-0.368	0.187	7.276	-0.224	0.218	-2.311
K-P F-stat	71.165	215.018	0.805	10.054	70.998	0.026	43.918	106.788	0.933
Sargan-Hansen J							3.473	2.735	0.397
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District	620	620	618	618	618	618	618	618	618
N. of Observations	2474	1858	1236	2472	1854	1236	2472	1854	1236
Mean Dep. Var.	179.519	221.154	288.871	179.672	221.617	288.871	179.672	221.617	288.871

Notes. This table reports the association between US out-migration and the number of patents. The unit of observation is a district, at a decade frequency between 1880 and 1939. The dependent variable is the number of patents. In Panel A, we report the association with measures out-migration (columns 1–3), the reduced-form railway instrument (columns 4–6), and the reduced-form leave-out instrument (columns 7–9). In Panel B, we report the two-stage least-square estimates of the railway (columns 1–3), leave-out (columns 4–6), and combined (columns 7–9) instruments. All regressions include district and decade fixed effects; standard errors are clustered at the district level and are displayed in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.20: Estimated Effect of the Influenza Pandemic on US Innovation

	Double Differences		Triple Differences			
	(1) Level	(2) Level	(3) Level	(4) Level	(5) Share	(6) Share
Excess Deaths × Post	3.120***					
	(0.751)					
1(Q. of Excess Deaths ¿ 75) × Post		1.474***				
		(0.504)				
Excess Deaths × Post × Pharma			2.641***		0.063**	
			(0.678)		(0.032)	
1(Q. of Excess Deaths ¿ 75) × Post × Pharma				1.311***		0.042***
				(0.451)		(0.016)
County FE	Yes	Yes	–	–	–	–
Year FE	Yes	Yes	–	–	–	–
County-Year FE	–	–	Yes	Yes	Yes	Yes
County-Class FE	–	–	Yes	Yes	Yes	Yes
Class-Year FE	–	–	Yes	Yes	Yes	Yes
N. of County-Class	1272	1272	21624	21624	21624	21624
N. of Observations	50880	50880	864960	864960	864960	864960
Classes in Sample	Pharma	Pharma	All	All	All	All
R^2	0.405	0.405	0.683	0.683	0.114	0.114
Mean Dep. Var.	0.991	0.991	0.534	0.534	0.077	0.077
Std. Beta Coef.	0.296	0.083	0.191	0.041	0.032	0.009

Notes. This table reports the effect of exposure to the Great Influenza Pandemic (1918–1919) on innovation in the US. The units of observation are counties (columns 1–2) and county-class pairs (columns 3–6). Units are observed at a yearly frequency between 1900 and 1939. In columns (1–4), the dependent variable is the number of patents granted; in columns (5–6), the dependent variable is the number of pharmaceutical patents divided by the total number of patents granted. In column (1), a post-influenza indicator is interacted with a measure of excess mortality, namely, the ratio between the average number of deaths during the pandemic (1918–1919) and the three previous years. In column (2), the treatment interacts the post-influenza indicator with a dummy variable equal to one for counties in the top quartile of the excess deaths distribution. In columns (3) and (5), the treatment interacts the excess deaths measure with a post-influenza dummy and an indicator variable for pharmaceutical patents; in columns (4) and (6), the excess deaths variable is coded as binary, and returns value one for counties in the top quartile of the excess mortality distribution. In columns (1–2), regressions include county and year fixed effects; in columns (3–6), regressions include county-by-year, county-by-technology class, and technology class-by-year fixed effects. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.21: Double and Triple Differences Effect of Synthetic Innovation Shocks on Subsequent US Innovation

	Baseline	Excluding States in ...				Innovation Shock Treshold		
	(1)	(2) Northeast	(3) Midwest	(4) South	(5) West	(6) 0.1%	(7) 1%	(8) 10%
Innovation Shock × Post	32.727***	40.194***	26.937***	32.514***	33.336***	94.663***	18.948***	3.958***
	(2.610)	(4.018)	(2.613)	(2.633)	(2.802)	(8.056)	(1.384)	(0.207)
District-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-by-Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Counties	2101783	1942416	1346809	1177807	1892929	2101824	2101250	2093047
Number of Observations	51263	47376	32849	28727	46169	51264	51250	51097
Mean Dep. Var.	0.772	0.511	0.730	1.241	0.765	0.772	0.752	0.615

Notes. This table reports the effect of synthetic innovation shock on US innovation. These coefficients are not interpreted as causal but as evidence that synthetic shocks capture relevant variation in county-technology-specific patenting activity. The unit of observation is a county-technology class pair observed at a yearly frequency between 1900 and 1939. The baseline treatment is an interaction between an innovation shock and a post-shock indicator. An innovation shock occurs when the residualized patenting activity in a given county technology is in the top 0.5% of the overall distribution of residualized values. Because the setting is staggered, all regressions are estimated using the methodology of Borusyak, Jaravel and Spiess (2021). Column (1) reports the estimate for the entire panel of counties; in columns (2), (3), (4), and (5), we exclude counties in, respectively, the North-East, Midwest, South, and West Census Bureau regions. In columns (6), (7), and (8), instead, we consider different thresholds for the definition of innovation shocks at the top 0.1%, 1%, and 10% of the overall distribution of residualized patents, respectively. All regressions include county-by-year, county-by-technology class, and technology class-by-year fixed effects. Standard errors, clustered at the county level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.22: Long-Run Sector Correlation Between Knowledge Exposure and Innovation

	1940s	1950s	1960s	1970s	1980s	1990s	2000s
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable: Number of Patents in:							
Agriculture	3.183***	2.654***	1.914***	1.909**	1.750**	0.647	0.516
	(0.751)	(0.617)	(0.602)	(0.790)	(0.823)	(0.609)	(0.641)
Building	6.065***	4.748***	5.312***	3.918***	4.540***	3.265***	1.804*
	(1.181)	(1.085)	(1.093)	(0.942)	(1.031)	(1.010)	(1.045)
Chemistry	23.553***	15.352***	23.228***	29.817***	11.953**	5.042	2.337
	(7.640)	(5.371)	(6.691)	(12.031)	(5.354)	(6.310)	(6.625)
Electricity	32.018***	31.842**	22.787***	16.405**	8.087	2.124	2.665
	(8.010)	(13.812)	(7.136)	(8.273)	(6.442)	(7.260)	(7.732)
Engineering	9.870***	7.969***	9.549***	8.374***	4.229***	1.167	0.702
	(1.705)	(1.563)	(1.754)	(2.008)	(1.406)	(1.666)	(1.720)
Engines, Pumps	6.551***	7.387***	5.926***	8.172*	4.480*	0.200	0.784
	(2.090)	(2.559)	(1.960)	(4.461)	(2.299)	(2.228)	(2.263)
Food	10.948***	9.570***	8.089***	10.390***	4.173**	0.495	-0.291
	(1.995)	(2.262)	(2.486)	(3.452)	(1.830)	(2.340)	(2.344)
Health, Amusement	4.430***	4.959***	3.988***	7.074***	6.700***	4.318***	5.210***
	(1.307)	(1.405)	(1.419)	(2.111)	(1.536)	(1.526)	(1.708)
Instruments	14.172***	14.338***	15.127***	14.236***	10.658***	4.819	4.079
	(2.937)	(3.244)	(3.480)	(4.101)	(2.905)	(3.072)	(3.599)
Lightning, Heating	11.553***	8.118***	8.113***	5.359***	3.534***	1.581	0.513
	(2.154)	(1.447)	(1.774)	(1.397)	(1.270)	(1.528)	(1.476)
Metallurgy	18.803***	9.443***	13.905***	10.698***	6.346**	1.834	0.849
	(3.888)	(2.573)	(3.541)	(3.493)	(2.722)	(3.110)	(3.236)
Personal Articles, Furniture	6.810***	6.784***	5.250***	2.813***	1.599**	1.367*	1.674**
	(1.014)	(1.175)	(0.811)	(0.755)	(0.803)	(0.811)	(0.821)
Printing	6.914***	7.830***	8.202***	6.030***	3.245***	1.984*	1.455
	(1.226)	(1.341)	(1.573)	(1.205)	(1.045)	(1.190)	(1.313)
Separating, Mixing	7.892***	7.493***	7.633***	8.032***	5.290***	1.602	1.166
	(1.707)	(1.508)	(1.681)	(1.922)	(1.458)	(1.557)	(1.696)
Shaping	9.833***	7.901***	8.591***	7.795***	3.555***	1.156	0.426
	(1.584)	(1.377)	(1.520)	(1.629)	(1.214)	(1.421)	(1.491)
Ships, Aeronautics	8.433***	8.800***	9.757***	6.624***	3.441***	1.319	0.905
	(1.032)	(1.156)	(1.379)	(1.193)	(0.946)	(1.081)	(1.161)
Textiles	14.865***	12.841***	11.039***	10.100***	3.649***	0.752	0.475
	(2.044)	(1.653)	(1.760)	(2.000)	(1.263)	(1.464)	(1.496)
Transporting	5.102***	4.368***	3.974***	2.988**	1.704*	0.471	0.147
	(1.157)	(1.194)	(1.231)	(1.391)	(0.934)	(1.123)	(1.167)
Number of Districts	632	632	632	632	632	632	632

Notes. This table reports the correlation between knowledge exposure in the years 1900–1930 and subsequent patenting activity by sector. For each class displayed in the rows, we estimate a model that interacts knowledge exposure with decade dummies, and we report the coefficients for each decade in the respective column. The 2010s decade serves as the baseline category. All regressions include district and decade fixed effects. Robust standard errors are displayed in parentheses.= *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.23: Heterogeneity Analysis of the Effect of Neighborhood Out-Migration on Innovation

	By Age	By Occupation						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Emigrant × Post	0.105*	0.348***	0.595***	0.172*	-0.040	-0.209	-0.030	0.069
	(0.057)	(0.134)	(0.222)	(0.103)	(0.110)	(0.202)	(0.212)	(0.051)
Age ∈ [18, 30) × Emigrant × Post	0.195**							
	(0.092)							
Age ∈ [30, 40) × Emigrant × Post	0.006							
	(0.068)							
Age ∈ [50, 60) × Emigrant × Post	-0.033							
	(0.083)							
Age ≥ 60 × Emigrant × Post	0.040							
	(0.182)							
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Full	Engineering	Metallurgy	Construction	Textiles	Trade	Pub. Adm.	Agriculture
N. of Individuals	469250	62716	12875	65013	31144	40576	15420	102463
N. of Observations	13608250	1818764	373375	1885377	903176	1176704	447180	2971427
R^2	0.135	0.120	0.097	0.148	0.080	0.097	0.295	0.103
Mean Dep. Var.	0.616	0.871	0.672	0.564	0.548	0.988	0.745	0.253
Std. Beta Coef.	0.002	0.004	0.010	0.003	-0.001	-0.003	-0.000	0.002

Notes. This table reports some heterogeneity analysis on the individual-level effect of neighborhood migration on patenting activity. The units of observation are individuals who are observed yearly between 1900 and 1920. The baseline treatment is an indicator that, for a given individual, returns value one after at least one person that was living in the same neighborhood as the individual migrates to the United States. In column (1), we interact this treatment with age category dummies and normalize the dummy for the age range 40–50 as the baseline category. In columns (2–8), we estimate the baseline double differences model by recorded occupations. Hence, in column (2), we estimate the model only for individuals employed in engineering occupations. All models include individual and year fixed effects. Standard errors are clustered at the district level and are reported in parentheses.

3.14 Appendix: Robustness Analysis

This section provides details on the technical implementation of the analyses discussed in the main text and briefly describes the exercises we perform to ensure the results' robustness.

3.14.1 Alternative Baseline Specifications

In this section, we list and comment on the alternative specifications of equation (3.2) that we estimate in the main text.

Alternative Dependent Variables

In the principal analysis, we use the raw number of patents at varying levels of aggregation as the dependent variable. We thus follow Chen and Roth (2022), who note that under transformations of the dependent variable defined at zero—as would be our case, to avoid dropping zero-patents observations—, the estimates of the average treatment effect are scale-dependent. Since it is common practice in the innovation literature to take the log-transformation, in Table 3.24, we show that the results are robust using a battery of alternative transformations.

Alternative Definitions of Knowledge Exposure

In Table 3.25, we employ four alternative measures of knowledge exposure. First, we take the log of the baseline. Second, we construct a measure that fixes bilateral emigrant flows:

$$\text{Knowledge Exposure}^{2}_{ik,t} = \sum_{j} \left(\frac{\text{Patents}_{jk,t}}{\text{Patents}_{j,t}} \times \text{Emigrants}_{i \to j,1880} \right) \tag{3.31}$$

which, compared to the main measure, restricts assortative matching to the first decade of the analysis. Third, we define the mirror measure that holds fixed specialization patterns

across counties:

$$\text{Knowledge Exposure}^{3}_{ik,t} = \sum_{j} \left(\frac{\text{Patents}_{jk,1880}}{\text{Patents}_{j,1880}} \times \text{Emigrants}_{i \to j,t} \right) \tag{3.32}$$

Compared to the main measure, this ensures that knowledge exposure does not conflate variation in patenting activity across counties determined or influenced by English immigrants. Finally, we define an alternative measure that leverages the *stock*, instead of the *flow* of patents issued:

$$\text{Knowledge Exposure}^{4}_{ik,t} = \sum_{j} \left[\sum_{\tau \leq t} \left(\frac{\text{Patents}_{jk,\tau}}{\text{Patents}_{j,\tau}} \right) \times \text{Emigrants}_{i \to j,t} \right] \tag{3.33}$$

The idea behind (3.33) is that specialization can be defined in terms of the cumulative number of patents filed before the given period. In Table 3.25, we show that all these measures yield quantitatively similar results.

Alternative Fixed Effects

In the main text, we report the results for a specification that includes district-by-time and technology class-by-time fixed effects. These are intended to capture time-varying unobserved heterogeneity at the district technology levels that we do not observe. In Table 3.26, we show that the—OLS and 2SLS—results are robust when including a wide array of alternative fixed effects. First, in columns (1) and (6), we report the unconditional correlation between innovation and knowledge exposure. This documents that knowledge exposure alone explains a sizable (30%) share of the variation in patenting activity. Then, in columns (2–5) and (7–10), we incrementally include additional fixed effects and show that the significance and magnitude of the coefficients remain very stable. In particular, in columns (5) and (10), we saturate the model with all couples of fixed effects to non-parametrically control for heterogeneity at the district-time, technology-time, and district-technology levels. The results are confirmed even in this demanding specification.

3.14.2 Instrumental Variable Strategy

This section discusses how we construct the county-level shocks necessary to compute the predicted bilateral flows, as described in section 3.4. We first present the strategy to construct the shocks for the main railway-based instrument. Then, we explain how we compute the shocks for the additional, leave-out instrument.

Railway-Based Instrument

The baseline instrument leverages county-level immigration shocks obtained from variations in the conditional timing when each county was connected to the US railway network. This strategy closely mimics the instrument developed by Sequeira, Nunn and Qian (2020) to estimate the long-run effect of immigration in the US.

To construct such shocks, we follow a two-step procedure. We first estimate the following zero-stage equation:

$$\begin{aligned}\text{Immigrant Share}_{j,t} = \alpha_j + \alpha_t + \beta\text{Immigrant Share}_{j,t-1} + \gamma I^{\text{Rail}}_{j,t-1} + \\ + \delta\left(I^{\text{Rail}}_{j,t-1} \times \text{Immigrant Flow}_{t-1}\right) + \zeta\left(\text{Industrialization}_{t-1} \times I^{\text{Rail}}_{j,t-1}\right) + \\ + \eta\left(\text{GDP Growth}_{t-1} \times I^{\text{Rail}}_{j,t-1}\right) + X^{\intercal}_{j,t-1}\Theta + \varepsilon_{j,t}\end{aligned} \tag{3.34}$$

where (Immigrant Share) is the share of foreign-born individuals, $I^{\text{Rail}}_{j,t}$ is a dummy variable returning value one if county j is connected to the railway network in decade t, and zero otherwise, (Immigrant Flow) is the aggregate immigration inflow computed from Willcox (1928), (Industrialization) is an index of industrial production computed by Davis (2004), and annual average GDP growth is obtained from Maddison (2007) data. The other terms control for confounding factors and non-random connections to the railway network. The term X includes log-population density, lagged urbanization, and an interaction between lagged urbanization and lagged aggregate immigrant flow. The core of the identification strategy that we borrow from Sequeira, Nunn and Qian (2020) is to exploit variation generated by the interaction between aggregate immigration inflows and connection to the railway network (δ). The underlying idea is that connection to

the railway only induces a larger immigrant inflow if it occurs during a period of high immigration. If this reasoning holds, the estimate of β should be close to zero, and that of δ should be positive. We confirm these predictions in Appendix Table 3.14.

We construct a synthetic series of county-level time-varying immigration shocks from equation (3.34) as follows:

$$\widehat{\text{Immigrant Share}}_{j,t} = \hat{\delta}\left(I^{\text{Rail}}_{j,t-1} \times \text{Immigrant Flow}_{t-1}\right) \tag{3.35}$$

where $\hat{\delta}$ is simply the OLS estimates from the previous model. We thus generate a set of county-level immigration shocks that are orthogonal to economic development and other characteristics that may induce sorting into the US. Variation, in other words, is solely due to the timing when a county is connected to the railway network.

Alternative Instrumental Variable

As further robustness to the railway instrument, we develop a simple leave-out instrument that borrows heavily on the literature that uses shift-share instruments to estimate the effects of immigration (Card; Tabellini, 2001; 2020). The rationale that underlies this approach is that if assortative matching across counties by British immigrants is the main threat to identification in the baseline regression, then it is possible to leverage the distribution of immigrants from *other* countries to construct county-level immigration shocks that yield consistent estimates because they do not reflect such assortative matching effects.

In practice, let ω^M_j be the share of immigrants from country M that settle in county j in the period 1860-1870, i.e., before the beginning of the analysis years. We then compute the aggregate inflow of immigrants from country M in each subsequent decade and construct the predicted immigrant inflows as

$$\widehat{\text{Immigrant Share}}_{j,t} = \frac{1}{\text{Population}_{j,t}} \sum_{\substack{M \neq \text{UK} \\ M \in \mathcal{M}}} \left(\omega^M_j \times \text{ Immigrant Inflow}^M_t\right) \tag{3.36}$$

where $\mathcal{M}$ is a set of origin countries. Both (3.35) and (3.36) yield a set of county-specific immigration shocks that do not conflate the immigration patterns of the British. They leverage very different sources of variation, though, which enables us to use the resulting instruments jointly and perform over-identification tests.

We allow multiple sets of origin countries $\mathcal{M}$. The baseline exercise, whose first-stage relevance is shown in Table 3.30 and results are displayed in Table 3.31, collates all countries except for the UK.[73] To account for possible correlation between British immigrants and those from other nationalities, however, we vary the set of included countries in Table 3.32. In particular, we drop all countries in Northern Europe (column 3), which may have been more similar to England and Wales. Moreover, in column (6), we only include non-European countries and show that results hold nonetheless. The coefficients remain relatively stable across all specifications, indicating the possibility that assortative matching may be a quantitatively mild issue.

Tests on Instrument Validity

The validity of the shift-share instrument for knowledge exposure that we construct hinges on the exogeneity of the shocks constructed using either (3.35) or (3.36), following Borusyak, Hull and Jaravel (2022). In practice, they advise conducting two types of falsification tests. First, shocks should be orthogonal to observed county-level characteristics. Second, the instrument should not be systematically correlated with district-level observable variables. The first test provides evidence of the exogeneity of the shocks, while the second should support the exclusion restriction that underlies the instrument.

We perform the first exercise in Figure 3.30. Panel (A) displays the correlation of the observed immigration shares with county-level observable characteristics. As expected, immigration is not random as it tends to be concentrated in larger counties, which also display higher patenting activity. In panels (B) and (C), we report the correlation between the predicted immigrant shares using the railway-based and the leave-out

[73] In Figure 3.29 we report binned scatter plots of the association between predicted and actual knowledge exposure using the two instruments.

approaches, respectively. We fail to detect a statistically significant correlation between the so-constructed immigrant shares and the large majority of county-level observable variables.[74] This provides reassuring evidence in favor of the validity of the instruments.

We report the second exercise in Figure 3.31. Each dot displays the correlation between district-level observable variables and actual, railway-based, and leave-out emigration in panels (A), (B), and (C), respectively. Unsurprisingly, districts featuring higher emigration flows are larger, produce more patents, and have a larger share of the population working in agriculture and textiles. On the other hand, synthetic out-migration, whether constructed using the railway or the leave-out shocks, is not correlated with any such variables. Once more, we interpret these results as evidence supporting the validity of the shift-share research design.

3.14.3 Shock Propagation

This section describes the technical definition of the synthetic innovation shocks and exposure to the influenza pandemic, along with two falsification exercises and several sensitivity analyses.

Details on the Construction of the Synthetic Shocks

We define a synthetic innovation shock as an unusual deviation from the number of patents granted in a given county, technology class, and year. Formally, we estimate the following fixed-effects regression:

$$\text{Patents}_{jk,t} = \alpha_{j\times k} + \alpha_{k\times t} + \alpha_{j\times t} + \varepsilon_{jk,t} \tag{3.37}$$

where j, k, and t denote a county, technology class, and year respectively, and α is the associated fixed effect. In particular, we include county-by-year fixed effects to remove fluctuations in patenting activity due to, for instance, economic growth. We remove

[74] Even when the correlation remains significant, the standardized beta coefficient is substantially lower than in the benchmark panel (A).

technology-by-year fixed effects to ensure that the shocks do not reflect aggregate changes in the propensity to patent in a given class. Finally, we average out county-by-class fixed effects to remove asymmetries due to initial specialization. We then construct a series of residualized innovation activity from the residuals of (3.37).

In the baseline analysis, we define an innovation shock as an observed residualized patenting activity in the top 0.1% of the overall distribution. Let $\Gamma(\cdot)$ be the cumulative distribution of the residuals of regression (3.37). Then, the set of shocks $\xi(\tau)$, for $\tau = 0.001$, is given by the set $\xi(\tau) = \{\xi \in \text{supp}(\Gamma) \mid \Gamma(\xi) - \Gamma(\tau) \geq 0\}$. In Table 3.34, we use two other threshold values of τ (1% and 0.5%). We find that the average treatment effect decreases as τ increases. This is compelling since larger τ's flag smaller residualized patenting activity as instances of treatment. In Table 3.21, we show the "effect" of synthetic shocks on US innovation. This is not a causal effect but rather a measure of the relevance of such shocks. There is a strong and positive increase in the number of patents after the shock is observed, and this also holds excluding specific areas (columns 2–5). In columns (6–8), we show that larger levels of τ are associated with a lower increase in patenting.

Details on the Construction of the Influenza Shock

To construct exposure to the influenza across counties, we follow Berkes, Coluccia, Dossi and Squicciarini (2023). From the mortality statistics collected by the Bureau of Census, we define a metric of excess deaths as the ratio between average deaths during the pandemic (1918–1919) relative to the average in the preceding three years.[75] Formally, we have

$$\text{Excess Deaths}_j = \frac{\frac{1}{2}\sum_{t=1918}^{1919} \text{Deaths}_{j,t}}{\frac{1}{3}\sum_{t'=1915}^{1917} \text{Deaths}_{j,t'}} \tag{3.38}$$

We then recast it as a binary variable equal to one if county j is in the top 25% of the excess deaths distribution to avoid issues of continuous treatment (Callaway and Sant'Anna, 2021).

[75]Due to data limitations, this is the pre-pandemic period that maximizes the sample size. Mortality statistics thus allow covering 60% of the US population.

The baseline estimation equation for US counties is then

$$\text{Patents}_{jk,t} = \alpha_{j\times k} + \alpha_{k\times t} + \alpha_{j\times k} + \delta\left(\text{Excess Deaths}_c \times \text{Pharma}_k \times \text{Post}_t\right) + \varepsilon_{jk,t} \quad (3.39)$$

where Pharma_k is an indicator variable returning value one if k is pharmaceutical patents, and zero otherwise, and Post_t is an indicator variable returning value one for years after 1918, and zero otherwise. Figure 3.26 reports the associated flexible triple differences estimates, which, with no evidence of statistically significant pre-treatment coefficients, suggests that the influenza had a strong, positive, and significant effect on pharmaceutical innovation in the US.

Robustness of the Synthetic Shock Analysis

We perform two main exercises to ensure that the results using the synthetic shocks are robust. First, in Figure 3.34, we check that the estimated effect of US innovation shocks on UK innovation remains significant and is quantitatively consistent under different estimators that allow for staggered roll-out of the treatment across units. The estimated ATE remains significant, and its magnitude is preserved under various estimators.

Second, in Table 3.34, we vary two margins along which a district is considered to be treated. First, as previously discussed, we consider different thresholds τ (1%, 0.5%, and the baseline 0.1%) above which we flag synthetic innovation shocks. Reassuringly, larger levels τ, which require a lower marginal increase in patenting to flag a synthetic shock, lead to smaller ATEs. This is consistent with the idea that larger innovation shocks in the US should lead to larger innovation shocks in the UK. Second, we vary the threshold of emigration that we impose for a district to be considered exposed to the innovation shock. In our main analysis, we consider a district exposed to the innovation shock in a given county if it is in the top quartile of the distribution of emigration to that county. We consider two additional thresholds (top 50% and top 90%). We find that the baseline result is qualitatively robust to all such thresholds. Moreover, we confirm that larger exposure thresholds lead to larger estimated ATEs. This suggests that the more intense

the previous migration tie between a county and a district, the stronger the diffusion effect of county-level shocks on district-level innovation.

Robustness of the Influenza Shock Analysis

In Table 3.35, we perform several exercises to gauge the robustness of the effect of UK exposure to excess mortality in the US during the Great Influenza pandemic. In columns (1) and (2), we report the double-difference estimates that compare innovation in pharmaceuticals in districts with high and low exposure to counties with high excess mortality. Compared to the baseline triple-difference model, these estimates do not include non-pharmaceutical innovation in the control group. The results of this exercise are quantitatively comparable with the baseline model. In column (3), we report the result of a triple-difference model that does not discretize district-level exposure to US counties. Columns (4–7), instead, restrict the sample by excluding the top-patenting areas. The results remain consistent throughout these specifications.

Shock Falsification Checks

The rationale for the analysis discussed in the main text (table 3.3 and Figure 3.5) and thus far is that the influenza only impacted patenting in pharmaceutical patents in the US. If that is the case, then this would ignite an innovation shock that was localized in areas that were more exposed to the influenza, and that could reverberate in the UK to districts whose emigrants had settled in such areas.

We test this assumption in Figure 3.33(a). Each dot reports an estimated δ coefficient of equation (3.39), except that the treated technology is reported in each row. Thus, the exclusion restriction would require that each coefficient was not statistically different from zero, except for pharmaceuticals. This assumption is confirmed in the data. The ATE for pharmaceuticals is the only one that is positive, significant, and quantitatively large. Figure 3.33(a) thus implies that we expect to observe an increase in pharmaceutical patents only, and only in districts whose emigrants had settled in areas that were more

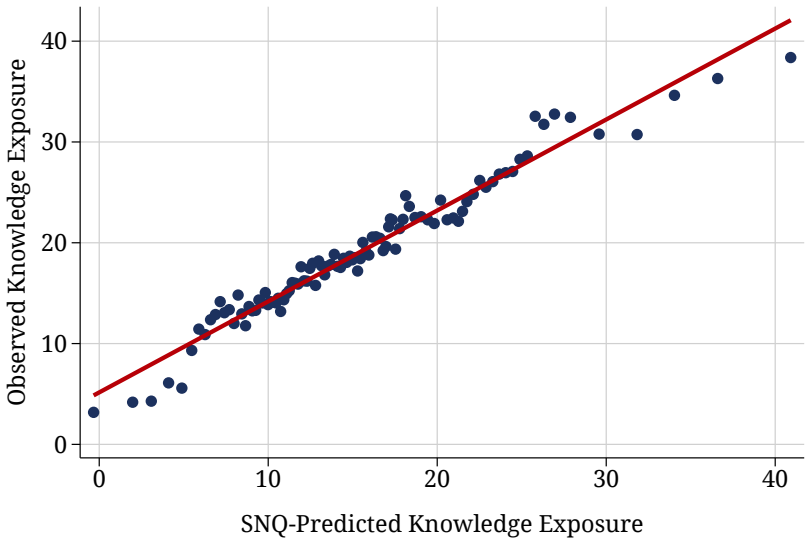
severely exposed to the pandemic.

We test this in Figure 3.33(b), in which we estimate the baseline triple-difference specification of the main text, except that the treated technology is reported in each row, as before. While estimates are noisier here, we confirm that the estimated ATE for pharmaceuticals is the largest and statistically significant across classes, as expected. Overall, Figure 3.33 thus provides convincing evidence that (i) the influenza fostered innovation in pharmaceuticals only in the US, and (ii) that districts whose emigrants had settled in areas that were more severely exposed to the influenza display higher innovation activity in pharmaceuticals.

3.14.4 Figures

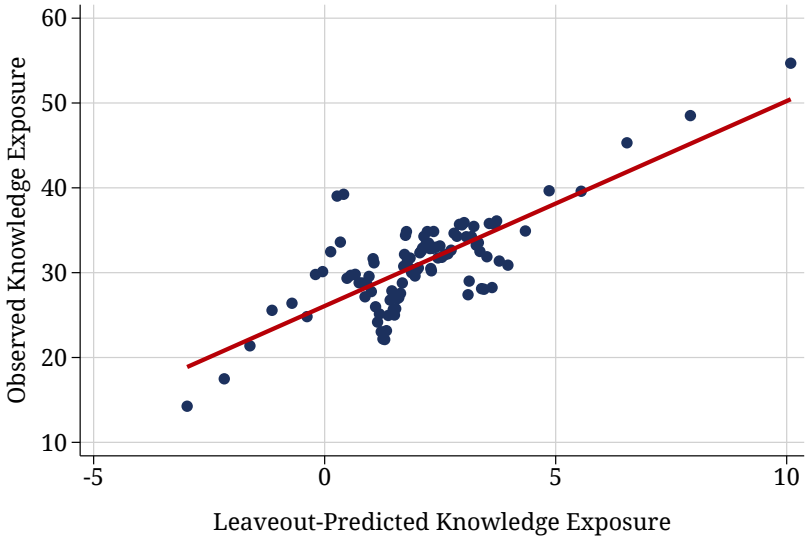
Figure 3.29: First Stage Binned Scatter Plot

((a)) Railway-Based Instrument



Notes. Coefficient = 0.453 (Clust. Std. Err. = 0.034). R^2 = 0.806.

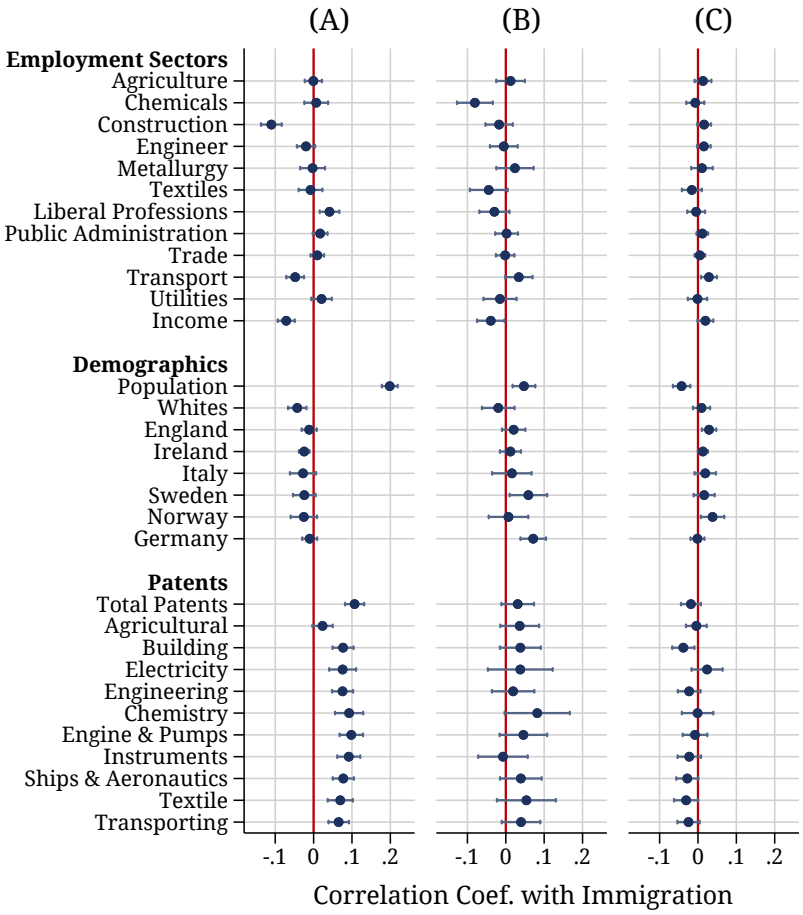
((b)) UK Innovation



Notes. Coefficient = 1.529 (Clust. Std. Err. = 0.256). R^2 = 0.773.

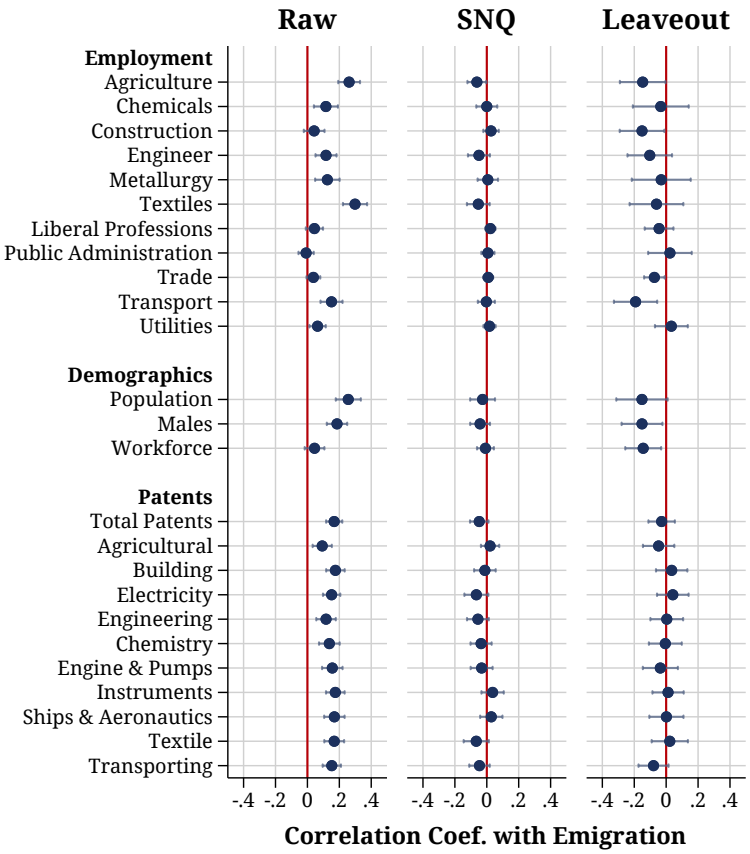
Notes. These figures are binned scatter plots of the association between actual and predicted knowledge exposure obtained using the railway-based instrument (Panel 3.29(a)) and the leave-out instrument (Panel 3.29(b)). The unit of observation is a district-technology class pair, at a decade frequency between 1880 and 1920. Graphs partial out district-by-decade and technology class fixed effects. We report the associated regression coefficients and standard errors, clustered at the district level, below each graph.

Figure 3.30: Shock-Level Balance Tests for Instrumental Variable Validity



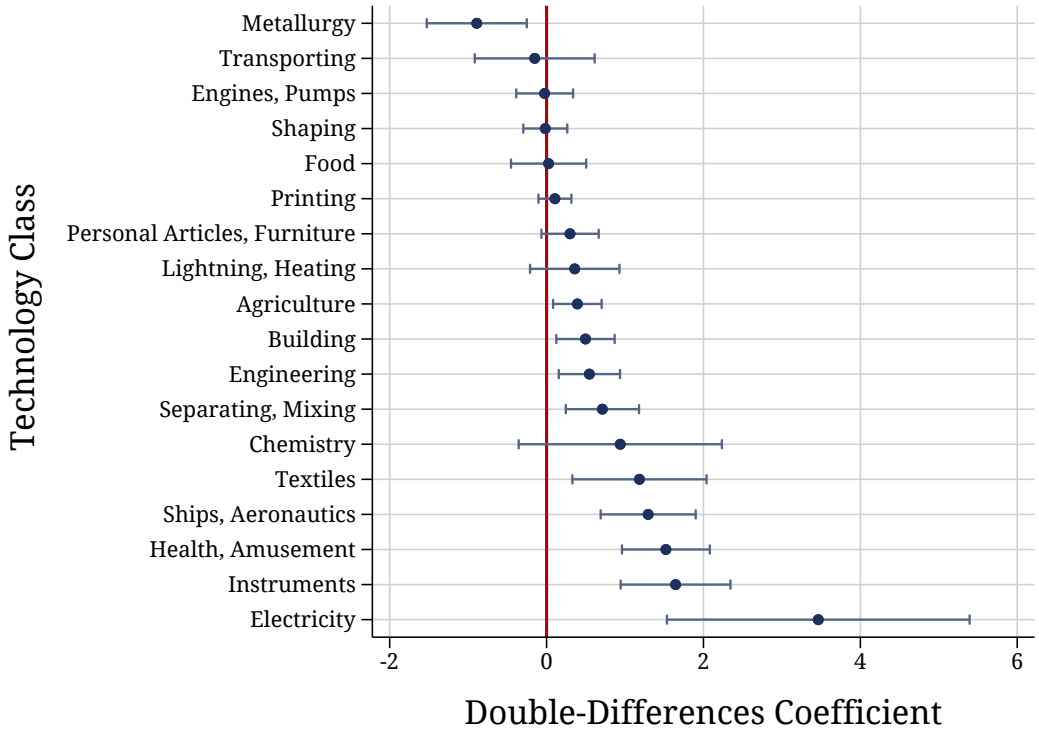
Notes. This figure reports the correlation between county-level observable characteristics and the (predicted) immigrant share. The unit of observation is a county observed at a decade frequency between 1870 and 1920. Panel (A) refers to the observed immigrant share; Panel (B) refers to the immigrant share predicted from the railway-based shock constructed from the zero-stage estimates *à la* Sequeira, Nunn and Qian (2020); Panel (C) refers to the leave-out shocks used to construct the alternative leave-out instrument. Each dot reports the correlation between the row variable and the immigrant share, lagged by one decade. Variables are standardized for the sake of readability. Each model includes county and state-by-decade fixed effects. Standard errors are clustered at the county level. Bands report 95% confidence intervals.

Figure 3.31: District-Level Balance Tests for Instrumental Variable Validity



Notes. This figure reports the correlation between district-level observable characteristics and the (predicted) number of emigrants. The unit of observation is a district observed at a decade frequency between 1870 and 1920. Panel (A) refers to the observed number of emigrants; Panel (B) refers to the predicted emigrant outflow obtained from the railway-based instrument; Panel (C) refers to the leave-out instrument. Each dot reports the correlation between the row variable and out-migration, lagged by one decade. Variables are standardized for the sake of readability. Each model includes district and decade fixed effects. Standard errors are clustered at the county level. Bands report 95% confidence intervals.

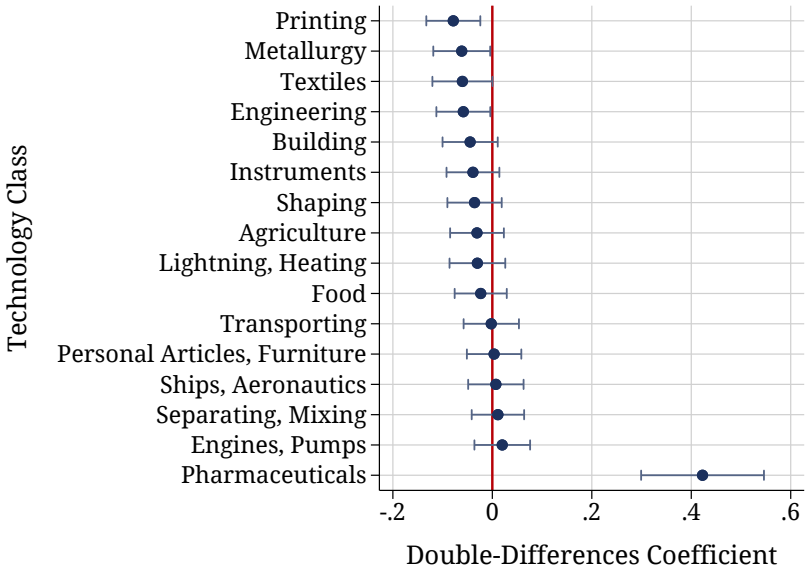
Figure 3.32: Effect of Synthetic Innovation Shocks Across Technology Classes



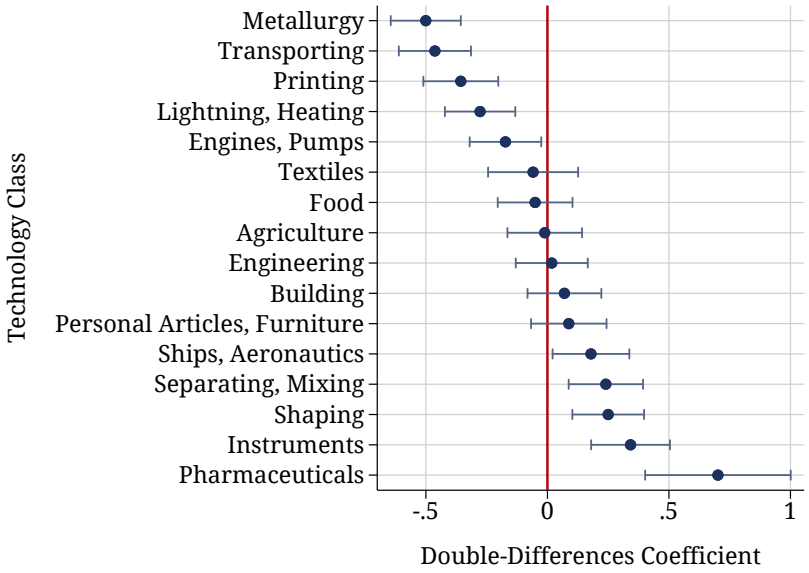
Notes. This figure reports the effect of synthetic innovation shocks on innovation in the UK by technology class. Each dot reports one double-differences estimated effect of the baseline exposure treatment with innovation; in each row, the treatment is activated whenever a district has above-median. The unit of observation is thus a district, observed at a yearly frequency between 1900 and 1993. Regressions include district and year fixed effects, and standard errors are clustered at the district level. Bands report 95% confidence intervals.

Figure 3.33: Effect of the Influenza Shock Across Technology Classes

((a)) US Innovation

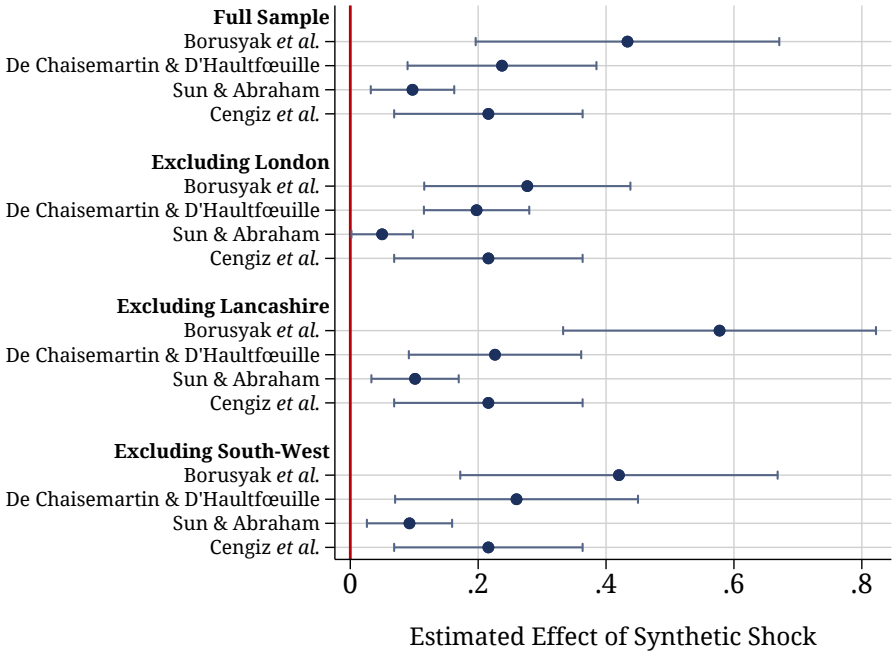


((b)) UK Innovation



Notes. This figure reports the effect of the Influenza shock on innovation, by technology classes, in the US (Panel 3.33(a)) and in the UK (Panel 3.33(b)). Each dot reports one triple-differences estimated effect of the baseline exposure treatment with innovation; in each row, exposure is interacted with a sector-specific dummy variable. If the shock only impacted innovation in pharmaceuticals, we would expect each coefficient but the pharmaceutical one to be statistically equal to zero. Regressions are saturated with fixed effects; standard errors are two-way clustered at the technology class and county (Panel 3.33(a)) or district (Panel 3.33(b)) level. Bands report 95% confidence intervals.

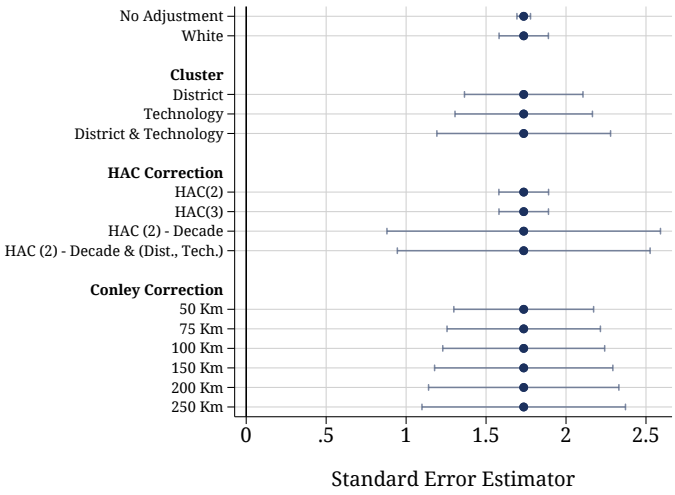
Figure 3.34: Alternative Staggered Triple Differences Estimators for the Effect of US Synthetic Shocks on Innovation



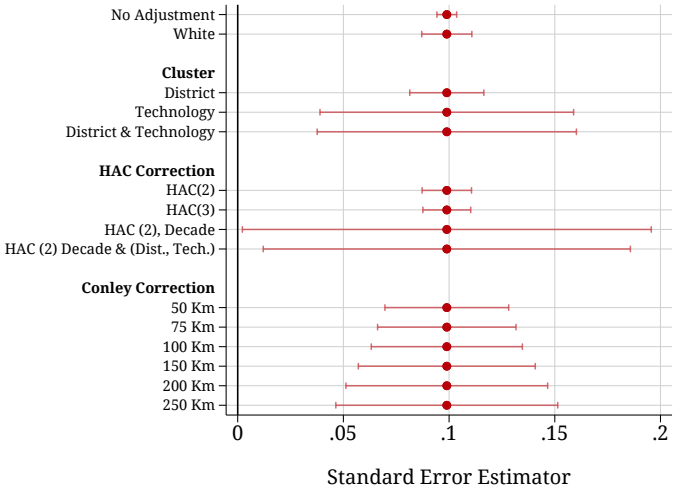
Notes. This figure reports the estimated effect of synthetic innovation shocks in US counties on innovation activity in the UK, using alternative estimators that explicitly allow for the staggered treatment roll-out design. The unit of observation is a district-technology class pair observed at a yearly frequency between 1900 and 1939. The dependent variable is the number of patents. The treatment variable is an indicator that, for a given district-technology, returns value one after a synthetic innovation shock in that technology class is observed in at least one county where the district has above-average out-migration. A synthetic innovation shock is observed whenever the residualized number of patents observed in the country is in the top 0.5% of the overall distribution. We estimate the models on the full sample of districts, as well as excluding the top three areas in terms of patents granted: London, Lancashire, and the South-West. We report the estimates obtained using four estimators that allow for the inclusion of all the triple differences interactions of the fixed effects: Borusyak, Jaravel and Spiess (2021), De Chaisemartin and D'Haultfœuille (2022), Cengiz, Dube, Lindner and Zentler-Munro (2022), and Sun and Abraham (2021). Standard errors are clustered at the district and technology class levels. Bands report 95% confidence intervals.

Figure 3.28: Alternative S.E. Estimators of the Return Innovation Result

((a)) Measured Knowledge Exposure

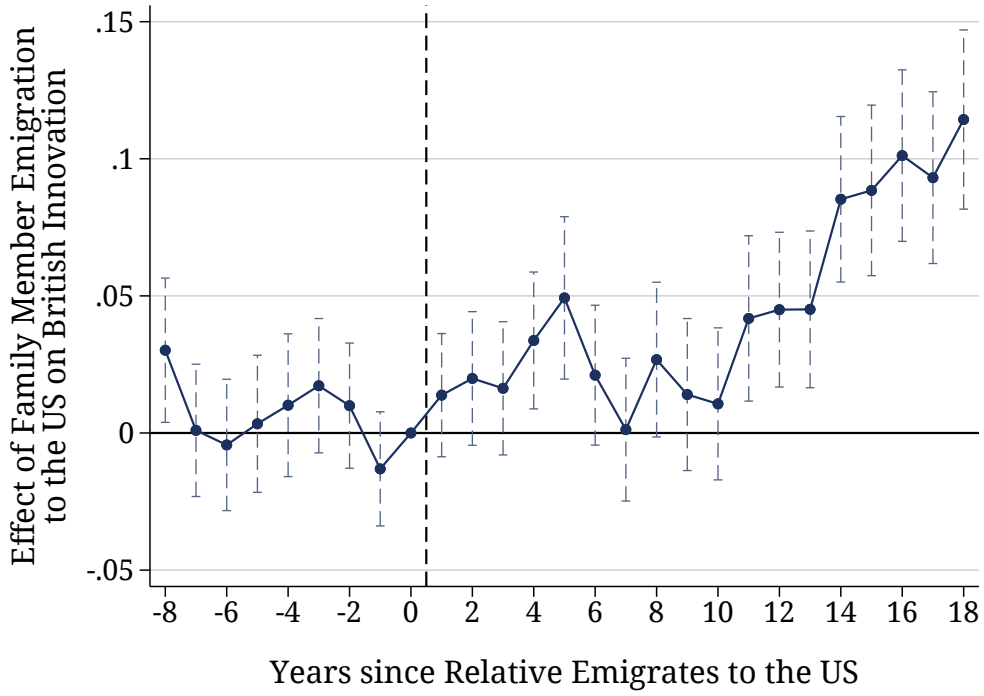


((b)) Railway-Based Knowledge Exposure



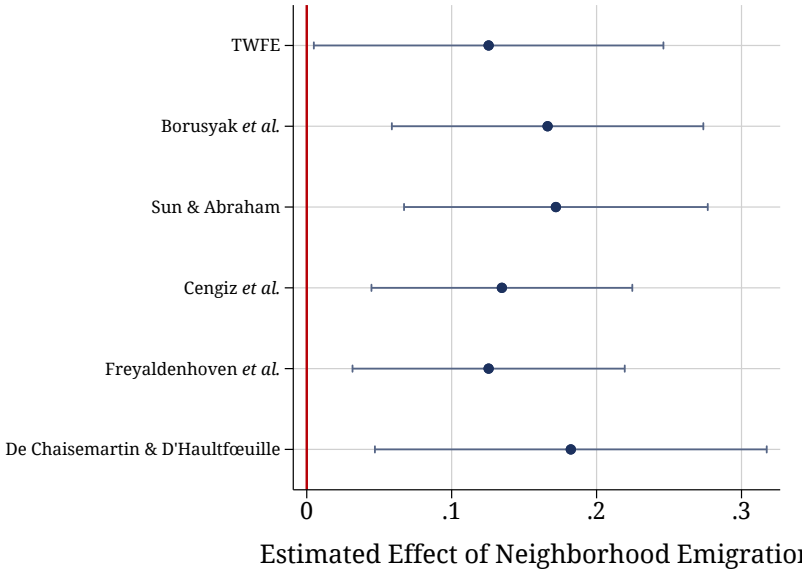
Notes. These figures report alternative estimates for the standard errors (SEs) of the regression between the number of patents and knowledge exposure. The unit of observation is a district-technology pair, observed at a decade frequency between 1880 and 1930. Models include district-by-technology and decade fixed effects. In Panel 3.28(a), the independent variable is measured knowledge exposure; Panel 3.28(b) reports the estimated reduced-form coefficient between patents and the railway-based instrument. We report unadjusted SEs, robust to heteroskedasticity (White); clustered at the district, technology class, and two-way by district and technology class; robust to heteroskedasticity and autocorrelation of order 2 (HAC (2)), order 3 (HAC (3)); robust to heteroskedasticity and autocorrelation, and clustered by decade (HAC (2) - Decade) and two-way by decade and district-by-technology class (HAC (2) - Decade & (Dist., Tech.). Finally, we also report SEs that account for spatial autocorrelation at various orders (between 50 and 250 kilometers) following Conley (1999). Bands report 95% confidence intervals.

Figure 3.35: Flexible Triple Differences Effect of Family Member Out-Migration on Innovation



Notes. This figure reports the effect of transatlantic emigration on innovation by inventors with the same surname as the emigrant. The unit of observation is a surname-county couple, observed at a year frequency between 1870 and 1929. The dependent variable is the number of patents granted to inventors with a given surname in a given county and year. The treatment is an interaction between year dummies and a variable that takes a value of one the first time at least one individual from a given county and with a given surname emigrates to the US, and zero otherwise. Each regression includes county-by-surname, surname-by-year, and county-by-year fixed effects. Standard errors are clustered at the surname level. Bands report 90% confidence intervals.

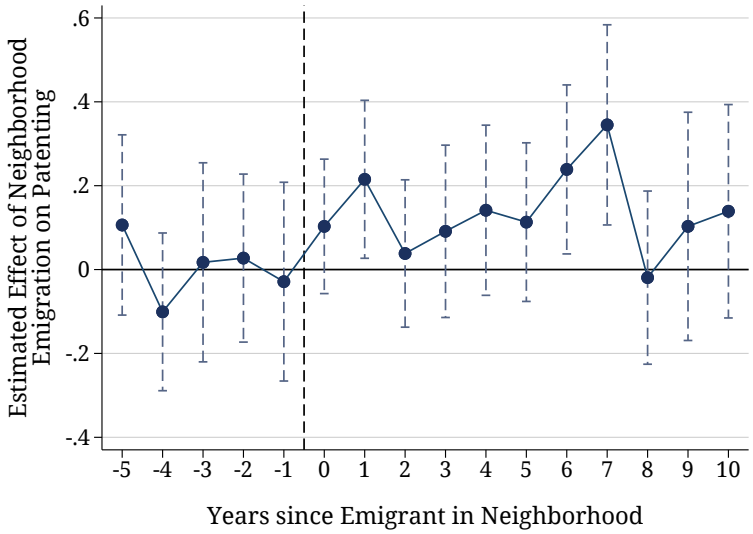
Figure 3.36: Alternative Staggered Double Differences Estimators for the Effect of Neighborhood Out-Migration on Innovation



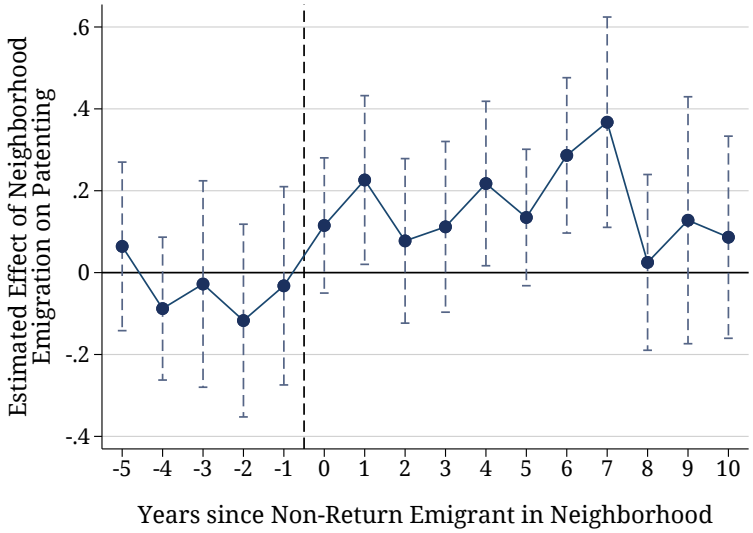
Notes. These figures report the effect of neighborhood out-migration on innovation. The units of observation are individuals observed at a yearly frequency between 1900 and 1920. The sample consists of all males who did not emigrate over the period and aged at least 18 in 1900. The dependent variable is the number of patents obtained every year. The treatment variable is an indicator that returns value one after at least one person living in the same neighborhood as the individual migrates to the United States. We report the estimates obtained using six estimators that allow staggered roll-out of treatment assignment: the baseline two-way fixed effects (TWFE) estimator, Borusyak, Jaravel and Spiess (2021), Sun and Abraham (2021), Cengiz, Dube, Lindner and Zentler-Munro (2022), Freyaldenhoven, Hansen and Shapiro (2019), and De Chaisemartin and D'Haultfœuille (2022). Standard errors are clustered at the district level. Bands report 90% confidence intervals.

Figure 3.37: Flexible Difference-in-Differences Effect of Neighborhood-Level Out-Migration on Innovation

((a)) All Emigrants

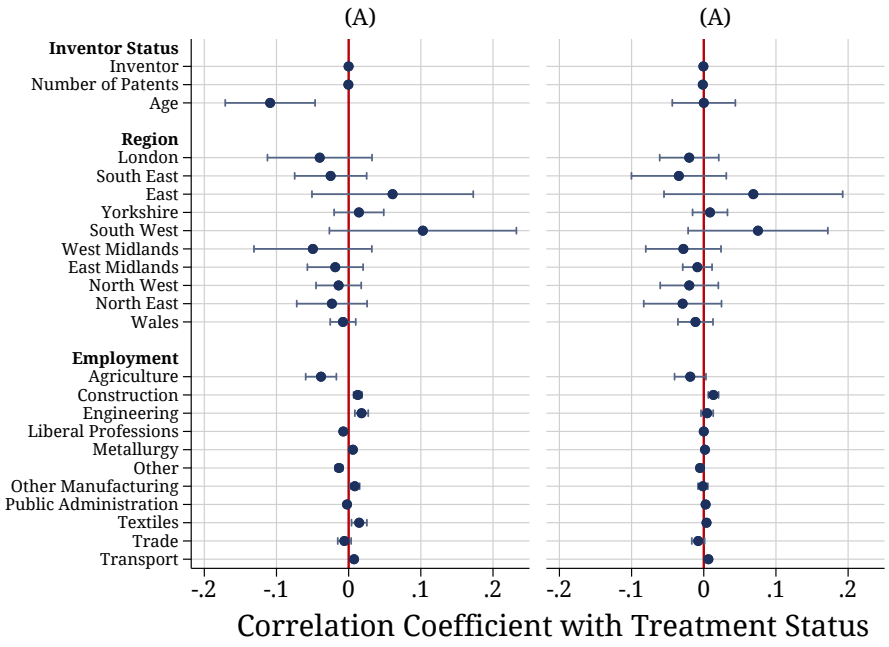


((b)) Non-Return Migrants



Notes. These figures report the effect of neighborhood out-migration on innovation. The units of observation are individuals observed at a yearly frequency between 1900 and 1920. The sample consists of all males who did not emigrate over the period and aged at least 18 in 1900. The dependent variable is the number of patents obtained every year. In Panel 3.37(a), the treatment variable is an indicator that returns value one after at least one person that was living in the same neighborhood as the individual migrates to the United States; in Panel 3.37(b), we restrict to emigrants that never return in the period of observation. Each model includes individual and parish-by-year fixed effects. Standard errors are clustered at the district level. The estimates are obtained using the estimator discussed in Borusyak, Jaravel and Spiess (2021). Bands report 95% confidence intervals.

Figure 3.38: Co-variate Balance for Individual-Level Design



Notes. These figures report the correlations between individual-level observable characteristics and treatment status in the individual-level analysis. The units of observation are individuals observed at a yearly frequency between 1900 and 1920. The sample consists of all males who did not emigrate over the period and aged at least 18 in 1900. Variables are observed in the 1911 census. Hence some of them are not pre-determined when the treatment initiates. Each dot reports the correlation between the row variable and a dummy variable equal to one if the individual is treated in the observation period and zero otherwise. Variables are standardized for readability. Panel (A) reports the unweighted correlation; in Panel (B), individuals are weighted by their CEM weights. Standard errors are clustered by division. Bands report 95% confidence intervals.

3.14.5 Tables

Table 3.24: Knowledge Exposure and Innovation: Alternative Dependent Variables

	Level of Patents					Share of Patents				
	(1) Baseline	(2) $\ln(\cdot)$	(3) $\ln(1+\cdot)$	(4) $\ln(\varepsilon+\cdot)$	(5) $\text{arcsinh}(\cdot)$	(6) Share	(7) $\ln(\cdot)$	(8) $\ln(1+\cdot)$	(9) $\ln(\varepsilon+\cdot)$	(10) $\text{arcsinh}(\cdot)$
Panel A. OLS Estimates										
Knowledge Exposure	1.342***	0.015***	0.067***	0.142***	0.082***	0.005***	0.015***	0.004***	0.169***	0.005***
	(0.143)	(0.002)	(0.007)	(0.016)	(0.009)	(0.001)	(0.002)	(0.000)	(0.020)	(0.001)
R^2	0.772	0.802	0.824	0.766	0.813	0.330	0.625	0.344	0.523	0.334
Std. Beta Coef.	0.299	0.101	0.396	0.495	0.407	0.411	0.139	0.439	0.629	0.418
Panel B. Reduced-Form Estimates										
Knowledge Exposure	0.037***	0.001***	0.001***	0.001***	0.001***	0.000***	0.001***	0.000***	0.001***	0.000***
	(0.007)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
R^2	0.800	0.811	0.816	0.752	0.805	0.347	0.651	0.358	0.510	0.350
Std. Beta Coef.	0.075	0.032	0.043	0.027	0.039	0.046	0.043	0.046	0.022	0.046
Panel C. Two-Stage Least Square Estimates										
Knowledge Exposure	1.224***	0.018***	0.031***	0.034***	0.034***	0.002***	0.018***	0.002***	0.027***	0.002***
	(0.195)	(0.005)	(0.005)	(0.007)	(0.006)	(0.000)	(0.005)	(0.000)	(0.008)	(0.000)
K-P F-stat	109.826	83.266	109.826	109.826	109.826	109.826	83.266	109.826	109.826	109.826
Std. Beta Coef.	0.296	0.116	0.171	0.108	0.153	0.181	0.158	0.181	0.087	0.181
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	8475	11268	11268	11268	11268	8475	11268	11268	11268
N. of Observations	67549	36290	67549	67549	67549	67549	36290	67549	67549	67549
Mean Dep. Var.	10.392	1.795	1.137	-0.005	1.400	0.051	-2.946	0.046	-4.636	0.050

Notes. This table displays the association between innovation and exposure to US knowledge using alternative transformations of the dependent variable. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. In columns (1–5), the dependent variable is the number of patents; in columns (6–10), the dependent variable is the share of patents in a given technology, normalized by the total number of patents. In columns (1) and (6), we do not transform the dependent variable; in columns (2) and (7), we take the log; columns (3) and (8) report the estimates using log(1+), which avoids dropping zeroes; in columns (4) and (9) we take log(0.01+) of the dependent variable; columns (5) and (10) report the estimates using the inverse hyperbolic sine. The main explanatory variable is knowledge exposure. In Panel A, we estimate the correlation through OLS; in Panel B, we report the reduced-form association between the instrument for knowledge exposure and innovation; in Panel C, we display the two-stage least-squares estimates. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.25: Knowledge Exposure and Innovation: Alternative Measures of Knowledge Exposure

	Dependent Variable: N. of Patents				
	(1)	(2)	(3)	(4)	(5)
Knowledge Exposure	1.342***				
	(0.143)				
ln(1 + Knowledge Exposure)		4.175***			
		(0.228)			
Fixed-Emigrants Knowledge Exposure			2.610***		
			(0.300)		
Fixed-Patents Knowledge Exposure				0.063***	
				(0.015)	
Cumulative Knowledge Exposure					0.136**
					(0.067)
District-Decade FE	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	11268	11268	11268	11268
N. of Observations	67549	67549	67547	67555	67549
R^2	0.772	0.766	0.770	0.765	0.764
Mean Dep. Var.	10.392	10.392	10.393	10.392	10.392
Std. Beta Coef.	0.299	0.119	0.249	0.104	0.017

Notes. This table displays the association between innovation and exposure to US knowledge, using alternative transformations of knowledge exposure. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. In column (1), we report the baseline estimate. In column (2), we take knowledge exposure in log terms, adding one to avoid dropping zeros since the baseline measure is defined as non-negative. In column (3), we fix bilateral district-county bilateral exposure shares as the number of emigrants from the given district to the given county in the decade 1870-1880. In column (3), instead, we fix county-level specialization as the share of patents in a given field granted in the decade 1870-1880 only. In column (5), for a given decade, we measure specialization as the sum of patents obtained until the end of that decade relative to the total number of patents obtained until the end of that decade. The measure used in column (5) thus considers the cumulative patent stock instead of its decade-on-decade flow. The main explanatory variable is knowledge exposure. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.26: Knowledge Exposure and Innovation: Alternative Sets of Fixed Effects

	Dependent Variable: N. of Patents									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	10)
Panel A. Correlational Estimates										
	OLS					Poisson				
Knowledge Exposure	2.393***	1.947***	1.936***	1.942***	1.241***	0.038***	0.011***	0.014***	0.014***	0.010***
	(0.212)	(0.194)	(0.184)	(0.191)	(0.149)	(0.004)	(0.002)	(0.003)	(0.003)	(0.002)
R^2	0.284	0.431	0.539	0.547	0.781	0.226	0.718	0.749	0.760	0.864
Std. Beta Coef.	0.533	0.433	0.431	0.432	0.276	0.303	0.084	0.112	0.109	0.080
Panel C. Instrumental Variable Estimates										
	Reduced Form					Two-Stage Least Squares				
$\widehat{\text{Knowledge Exposure}}$	0.158***	0.080***	0.041***	0.041***	0.033***					
	(0.012)	(0.009)	(0.010)	(0.012)	(0.009)					
Knowledge Exposure						2.053***	1.490***	0.867***	0.779***	1.097***
						(0.157)	(0.167)	(0.192)	(0.208)	(0.271)
R^2	0.104	0.385	0.501	0.509	0.808	0.293	0.111	0.065	0.059	0.028
K-P F-stat						184.700	96.871	169.304	132.033	46.312
Std. Beta Coef.	0.322	0.164	0.083	0.083	0.068	4.196	3.045	1.772	1.591	2.243
District FE	No	Yes	–	–	–	No	Yes	–	–	–
Decade FE	No	Yes	–	–	–	No	Yes	–	–	–
Class FE	No	Yes	Yes	–	–	No	Yes	Yes	–	–
District-Decade FE	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Class-Decade FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
District-Class FE	No	No	No	No	Yes	No	No	No	No	Yes
N. of District-Class	11268	11268	11268	11268	11268	11268	11250	11250	11250	10081
N. of Observations	67549	67549	67549	67549	67549	67549	67474	65946	65946	59703
Mean Dep. Var.	10.392	10.392	10.392	10.392	10.392	10.392	10.404	10.645	10.645	11.758

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. The dependent variable is the number of patents. The main explanatory variable is knowledge exposure. In Panel A, in columns (1–5), we estimate the correlation through OLS; in columns (6–10), we estimate the model as a Poisson regression to account for the many zeros in the data; columns (1–5) in Panel B report the reduced-form association between the instrument for knowledge exposure and innovation; columns (6–10) report the two-stages least square estimates. Columns (1) and (6) reports the unconditional regressions; in columns (2) and (7), we include district, decade, and technology class fixed effects; columns (3) and (8) add district-by-decade fixed effects; in columns (4) and (9) we include district-by-decade and class-by-decade fixed effects; models in columns (5) and (10) are saturated with all couples of fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.27: Return Innovation Accounting for Patent Quality

	Quality Weight		Breakthrough 20%		Breakthrough 10%		Breakthrough 5%	
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
Panel A. Unadjusted Quality Indicator								
Knowledge Exposure$_t$	1.433***	3.812***	0.706***	1.330***	0.387***	0.697***	0.228***	0.328***
	(0.254)	(1.206)	(0.134)	(0.342)	(0.078)	(0.194)	(0.048)	(0.100)
R^2	0.822	-0.038	0.660	-0.049	0.585	-0.033	0.519	-0.037
Mean Dep. Var.	16.327	12.647	3.842	2.103	1.994	1.027	1.003	0.409
Std. Beta Coef.	0.265	0.908	0.408	1.491	0.366	1.443	0.324	1.270
Panel B. Adjusted Quality Indicator (Net of Class-Year FE)								
Knowledge Exposure$_t$	0.051***	0.119***	0.389***	0.808***	0.229***	0.449***	0.110***	0.436***
	(0.017)	(0.043)	(0.083)	(0.186)	(0.059)	(0.126)	(0.033)	(0.100)
R^2	0.464	-0.055	0.649	-0.106	0.580	-0.108	0.519	-0.228
Mean Dep. Var.	0.284	0.168	2.698	1.688	1.621	0.903	0.850	0.460
Std. Beta Coef.	0.180	0.991	0.321	1.261	0.261	1.209	0.188	1.966
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11263	11198	11263	11198	11263	11198	11263	11198
N. of Observations	33764	22396	33764	22396	33764	22396	33764	22396
K-P F-stat		44.652		44.652		44.652		44.652

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1900 and 1939. The main explanatory variable is knowledge exposure. The dependent variables capture the "quality" of patents. To measure quality, we adapt the text-based indicator developed by Kelly, Papanikolaou, Seru and Taddy (2021). The sample excludes the years 1880–1899 because, for the subsequent years, we only have abstracts. The quality measure is thus evaluated on abstracts. In columns (1–2), the dependent variable is the number of patents, weighted by their quality. In columns (3–8), following Kelly, Papanikolaou, Seru and Taddy (2021), we only count patents in the top 20%, 10%, and 5% of the overall quality distribution. Odd columns display the OLS correlation between knowledge exposure and the dependent variables; Even columns report the associated two-stage least-square estimates. In panel (A), the quality measure is not adjusted; in panel (B), we remove class-by-year fixed effects in the quality measure following Kelly, Papanikolaou, Seru and Taddy (2021) to control for fashion effects in language. All regressions include district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.28: Return Innovation Accounting for Patent Quality of US Patents

	Quality Weight		Breakthrough 20%		Breakthrough 10%		Breakthrough 5%	
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
Panel A. Unadjusted Quality–Weighted Knowledge Exposure								
Knowledge Exposure (Weighted)	0.003**	0.233						
	(0.001)	(0.179)						
Knowledge Exposure (top 20%)			0.532***	0.278***				
			(0.048)	(0.081)				
Knowledge Exposure (top 10%)					0.500***	0.189**		
					(0.047)	(0.079)		
Knowledge Exposure (top 5%)							0.446***	-0.037
							(0.047)	(0.072)
Panel B. Adjusted Quality-Weighted Knowledge Exposure (Net of Class-Year FE)								
Knowledge Exposure (Weighted)	0.002	0.205***						
	(0.002)	(0.070)						
Knowledge Exposure (top 20%)			0.531***	0.332***				
			(0.052)	(0.084)				
Knowledge Exposure (top 10%)					0.495***	0.265***		
					(0.049)	(0.077)		
Knowledge Exposure (top 5%)							0.486***	0.093
							(0.050)	(0.075)
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11160	11106	11160	11106	11160	11106	11160	11106
N. of Observations	66901	55507	66901	55507	66901	55507	66901	55507
K-P F-stat		1.579		117.459		136.236		156.128

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1870 and 1939. The main explanatory variable is knowledge exposure, weighted by the quality of US patents. The dependent variable is the number of patents, in logs. To measure quality, we adapt the text-based indicator developed by Kelly, Papanikolaou, Seru and Taddy (2021). The first row in each panel weights patents by their quality; the second, third, and fourth rows only count patents in the top 20%, 10%, and 5% of the overall quality distribution, respectively. In panel (A), the quality measure is not adjusted; in panel (B), we remove class-by-year fixed effects in the quality measure following Kelly, Papanikolaou, Seru and Taddy (2021) to control for fashion effects in language. All regressions include district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.29: Effect of Exposure to US Technology on Innovation in Great Britain: Patents with Firm Assignee

	OLS		Reduced Form		Two-Stage Least-Squares	
	(1)	(2)	(3)	(4)	(5)	(6)
Knowledge Exposure$_t$	0.170***				0.068***	
	(0.012)				(0.021)	
Knowledge Exposure$_{t-1}$		0.176***				0.327***
		(0.021)				(0.046)
$\widehat{\text{Knowledge Exposure}}_t$			0.004***			
			(0.001)			
$\widehat{\text{Knowledge Exposure}}_{t-1}$				0.023***		
				(0.003)		
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	11266	11214	11214	11214	11214
N. of Observations	33798	22532	33642	22428	33640	22428
K-P F-stat					134.004	154.011
R^2	0.841	0.914	0.835	0.912	0.021	0.014
Mean Dep. Var.	1.250	1.759	1.256	1.767	1.256	1.767
Std. Beta Coef.	0.179	0.165	0.026	0.133	0.071	0.306

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1900. The main explanatory variable is knowledge exposure. The dependent variable is the number of patents with at least one firm listed as an assignee. In columns (1–2), we estimate the OLS correlation with the observed measure of knowledge exposure; in columns (3–4), we estimate the reduced-form association with the railway-based instrument of knowledge exposure through OLS; columns (5–6) report the two-stage least-squares estimate. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.30: First Stage of the Instrumental Variable Estimation

	Railway-Based (SNQ) Instrument				Leaveout Instrument			
	Baseline	Dropping Districts in...			Baseline	Dropping Districts in...		
	(1)	(2) London	(3) Lancs	(4) S-W	(5)	(6) London	(7) Lancs	(8) S-W
Panel A. Dependent Variable: Bilateral Flows								
SNQ Migrants	0.007***	0.008***	0.007***	0.007***				
	(0.001)	(0.001)	(0.001)	(0.001)				
Leaveout Migrants					0.006***	0.005**	0.005***	0.005***
					(0.002)	(0.002)	(0.002)	(0.002)
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Counties	1736040	1653240	1518000	1625640	1786360	1701160	1562000	1672760
N. of Observations	8399666	7999046	7344700	7865506	10403031	9906861	9096450	9741471
Panel B. Dependent Variable: Knowledge Exposure								
SNQ Knowledge Exposure	0.005***	0.005***	0.005***	0.005***				
	(0.000)	(0.000)	(0.000)	(0.000)				
Leaveout Knowledge Exposure					0.169***	0.157***	0.159***	0.145***
					(0.034)	(0.036)	(0.034)	(0.032)
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Classes	11322	10782	9900	10602	11304	10764	9882	10584
N. of Observations	56587	53887	49488	52987	67801	64561	59280	63481

Notes. This table reports the first-stage estimates of the two shift-share instruments we propose. In Panel A, the observation units are district-county pairs, observed at a decade frequency between 1870 and 1920 (columns 1–4) and 1930 (columns 5–8). In Panel B, the observation units are district-technology classes, at decade frequency between 1870 and 1920 (columns 1–4) and 1930 (columns 5–8). In columns (1–4), the predicted number of emigrants constructed using the railway-based instrument that leverages shocks *à la* Sequeira, Nunn and Qian (2020); in columns (5–8), predicted emigrants are constructed using the leave-out instrument. Columns (1) and (5) report the full-sample estimates; in columns (2) and (6), we exclude districts in the London area; columns (3) and (7) exclude districts in the Lancashire area; in columns (4) and (8) we drop districts in the South-West. In Panel A, all models include district-by-decade and county-by-decade fixed effects; in Panel B, regressions include district-by-decade and technology class-by-decade fixed effects. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.31: Return Innovation Result Using the Leaveout Instrument

	Reduced Form			TSLS			Overidentified TSLS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\widehat{\text{Knowledge Exposure}}$	0.007*								
	(0.004)								
$\widehat{\text{Knowledge Exposure}}_{t-1}$		0.018***							
		(0.006)							
$\widehat{\text{Knowledge Exposure}}_{t-2}$			0.029**						
			(0.012)						
Knowledge Exposure				0.093*			0.322***		
				(0.051)			(0.052)		
Knowledge Exposure$_{t-1}$					0.180***			0.082*	
					(0.065)			(0.044)	
Knowledge Exposure$_{t-2}$						0.103**			0.032
						(0.041)			(0.038)
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11196	11196	11196	11196	11196	11196	11196	11196	11196
N. of Observations	55980	44784	33588	55957	44761	33586	55957	44761	33586
R^2	0.816	0.831	0.850	0.018	-0.006	-0.002	0.054	-0.001	-0.000
Mean Dep. Var.	1.079	1.202	1.312	1.079	1.203	1.312	1.079	1.203	1.312
Std. Beta Coef.	0.007	0.017	0.015	0.051	0.103	0.055	0.175	0.047	0.017
K-P F-stat				27.303	23.391	248.916	62.737	59.305	198.950
Sargan-Hansen J							24.274	6.131	31.636

Notes. This table reports the estimated return innovation effect estimated using the leave-out instrument. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. The dependent variable is the number of patents. The main explanatory variable is knowledge exposure. In columns (1–3), we report the reduced-form association between knowledge exposure constructed using predicted emigration flows using the leave-out instrument and the dependent variable; in columns (4–6), we report the associated two-stage least-squares estimates. In columns (7–9), instead, we exploit the railway and the leave-out instruments to estimate an over-identified instrumental variable regression. This allows us to report the associated Sargan-Hansen J-statistic to test the validity of the over-identifying restrictions. The Sargan-Hansen test does not refute the null that the instruments are valid. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.32: Return Innovation Result Using the Modified Leaveout Instruments

	Baseline	Excluding Immigrants from...			
	(1)	(2) UK	(3) UK + North Eu.	(4) UK + South Eu.	(5) UK + East Eu.	(6) UK + Europe
Panel A. Second-Stage Estimates						
Knowledge Exposure	1.849***	0.454***	0.589***	0.428***	0.315**	0.487***
	(0.174)	(0.125)	(0.170)	(0.096)	(0.160)	(0.139)
N. of Observations	78876	78876	78876	78876	78876	78876
Mean Dep. Var.	11.768	11.768	11.768	11.768	11.768	11.768
K-P F-statistic		39.267	30.074	346.557	14.672	52.663
Panel B. First-Stage Estimates						
Knowledge Exposure (No Northern Europe + UK)			0.215***			
			(0.039)			
Knowledge Exposure (No Southern Europe + UK)				0.889***		
				(0.048)		
Knowledge Exposure (No Eastern Europe + UK)					0.481***	
					(0.126)	
Knowledge Exposure (No Europe + UK)						3.103***
						(0.428)
N. of Observations		78876	78876	78876	78876	78876
Mean Dep. Var.		3.063	3.063	3.063	3.063	3.063
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes
Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes. This table reports the instrumental variable estimates of the effect of knowledge exposure on innovation using modified versions of the leave-out instrument. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. The dependent variable is the number of patents. The explanatory variable is knowledge exposure. In column (1), we report the OLS correlation. In columns (2–6), we construct predicted bilateral emigrant flows using county-level immigration shocks that exclude immigrants from different parts of the world: in (2), we exclude only immigrants from UK nations; in (3), we exclude the UK immigrants along with those from other Northern Europe countries; in (4), we exclude immigrants from the UK and Southern Europe; in (5), UK and Eastern Europe immigrants are excluded; in (6), we exclude all European immigrants. Panel A reports the second-stage estimates; Panel B reports the associated first-stage estimates. All regressions include district-by-decade and technology class fixed effects. Standard errors, clustered at the district level, are displayed in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.33: Robustness Analysis on the Effect of Exposure to US Technology on the Similarity and Originality of Innovation in Great Britain

	10-Year Similarity			log(Similarity)			Net of Year-Technology FE		
	(1) OLS	(2) RF	(3) TSLS	(4) OLS	(5) RF	(6) TSLS	(7) OLS	(8) RF	(9) TSLS
Panel A. Dependent variable: "Copying" (Similarity with Previous US Patents)									
Knowledge Exposure$_t$	0.296***		0.122	0.953***		0.935***	0.340***		0.334***
	(0.045)		(0.096)	(0.106)		(0.151)	(0.042)		(0.057)
$\widehat{\text{Knowledge Exposure}}_t$		0.036			0.276***			0.099***	
		(0.029)			(0.051)			(0.019)	
R^2	0.733	0.767	0.003	0.752	0.795	0.046	0.681	0.692	0.023
Mean Dep. Var.	36.221	32.577	32.585	62.691	49.809	49.818	22.720	17.962	17.964
Std. Beta Coef.	0.165	0.016	0.065	0.340	0.101	0.399	0.303	0.089	0.352
Panel B. Dependent Variable: "Originality" (Similarity with Subsequent US Patents w.r.t. Previous US Patents)									
Knowledge Exposure$_t$	0.195***		0.270***	0.146***		0.158***	0.053***		0.076***
	(0.022)		(0.038)	(0.016)		(0.025)	(0.006)		(0.011)
$\widehat{\text{Knowledge Exposure}}_t$		0.080***			0.046***			0.022***	
		(0.013)			(0.008)			(0.004)	
R^2	0.684	0.748	0.034	0.752	0.791	0.049	0.571	0.522	0.011
Mean Dep. Var.	11.709	9.049	9.051	9.572	7.628	7.630	3.305	2.635	2.635
Std. Beta Coef.	0.338	0.145	0.583	0.344	0.109	0.439	0.285	0.105	0.422
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	11214	11214	11268	11214	11214	11268	11214	11214
N. of Observations	67553	56070	56050	67553	56070	56050	67553	56070	56050
K-P F-stat			103.344			103.344			103.344

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. In Panel A, the dependent variable is the text similarity between UK patents and US patents issued before ("copying"); in Panel B, the dependent variable is the similarity of UK patents with US patents granted in the subsequent years, over the similarity of UK patents with US patents granted in the preceding years ("originality"). The similarity measure is akin to Kelly, Papanikolaou, Seru and Taddy (2021). In columns (1), (4), and (7) we report the OLS regressions; columns (2), (5), and (8) report the reduced-form regressions; columns (3), (6), and (9) display the two-stage least-squares coefficients. In columns (1–3), the dependent variable is the baseline, except that we compute similarities over a ten-year window compared to the baseline five; in columns (4–6), we take the log of the patent-level similarity measure; in columns (7–9), remove year-by-technology class fixed effects from the patent-level similarity metrics. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.34: Triple Differences Estimated Effect of US Synthetic Shocks on UK Innovation: Alternative Thresholds

	Top 1% Synthetic Shocks			Top 0.5% Synthetic Shocks			Top 0.1% Synthetic Shocks		
	(1) Top 50%	(2) Top 75%	(3) Top 90%	(4) Top 50%	(5) Top 75%	(6) Top 90%	(7) Top 50%	(8) Top 75%	(9) Top 90%
Innovation Shock (Above 50%) × Post	0.187**			0.299***			0.617***		
	(0.083)			(0.098)			(0.126)		
Innovation Shock (Above 75%) × Post		0.224***			0.377***			0.617***	
		(0.080)			(0.081)			(0.126)	
Innovation Shock (Above 90%) × Post			0.326			0.825***			0.532***
			(0.269)			(0.229)			(0.175)
District-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
District-by-Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Class-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Number of Counties	189586	362024	426147	217975	381438	434687	431467	431467	445120
Number of Observations	5247	9187	10671	6762	9786	10900	10834	10834	11128
Mean Dep. Var.	1.022	1.106	1.586	1.410	1.3	1.686	2.064	2.064	2.020

Notes. This table displays the effect of US innovation shocks on innovation activity in the UK. The unit of observation is a district-technology class pair observed at a yearly frequency between 1900 and 1939. The dependent variable is the number of patents. The treatment variable is equal to one for district-technology class pairs after a synthetic innovation shock in a given technology class is observed in counties where the district has above k-percentile emigrants. We consider three different thresholds for k: above the median, above the top 25%, and above the top 10%. A synthetic shock is observed whenever the residualized patenting activity in a given county-technology class pair is in the top ℓ-percentile of the residualized patenting activity distribution. We consider three such *ell*: top 1%, in columns (1–3), top 0.5%, in columns (4–6), and top 0.1%, in columns (7–9). Since the treatment timing is staggered, we estimate the models using the imputation estimator developed by Borusyak, Jaravel and Spiess (2021). All models include district-by-year, district-by-technology class, and technology class-by-year fixed effects; standard errors, clustered two-way by district and technology class, are shown in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.35: Double and Triple Differences Estimated Effect of the Great Influenza Pandemic in the US on Innovation in the UK: Robustness Analysis

	Double Differences		Triple Differences				
	(1)	(2)	(3)	(4)	(5) No London	(6) No Lancs	(7) No S/W
Influenza Emigration × Post	0.008*						
	(0.004)						
1(Q. of Influenza Emigration ¿ 75) × Post		0.980**					
		(0.463)					
Influenza Emigration × Post × Pharma			0.004**				
			(0.002)				
1(Q. of Influenza Emigration ¿ 75) × Post × Pharma				0.584***	0.396**	0.671***	0.423**
				(0.163)	(0.139)	(0.173)	(0.156)
District FE	Yes	Yes	–	–	–	–	–
Year FE	Yes	Yes	–	–	–	–	–
District-Year FE	–	–	Yes	Yes	Yes	Yes	Yes
District-Class FE	–	–	Yes	Yes	Yes	Yes	Yes
Class-Year FE	–	–	Yes	Yes	Yes	Yes	Yes
N. of District-Class	631	631	10727	10727	10217	9384	10047
N. of Observations	18930	18930	321810	321810	306510	281520	301410
Classes in Sample	Pharma	Pharma	All	All	All	All	All
R^2	0.544	0.544	0.668	0.668	0.616	0.653	0.679
Mean Dep. Var.	0.927	0.927	0.763	0.763	0.559	0.721	0.706
Std. Beta Coef.	0.082	0.082	0.014	0.016	0.015	0.018	0.010

Notes. This table displays the effect of the Great Influenza Pandemic shock on innovation activity in the UK. In columns (1–2), the observation unit is a district; in columns (3–7), the observation unit is a pair district-technology class; units are observed at a yearly frequency between 1900 and 1939. The dependent variable is the number of patents. In column (1), the treatment variable is an interaction between an influenza exposure term equal to the share of emigrants to counties in the top 25% of the flu-related excess mortality distribution and a post-Influenza indicator; in column (2), we code exposure as a binary variable equal to one for districts in the top 25% of the continuous exposure distribution. In columns (3) and (5–7), the treatment term in column (1) is interacted with an indicator variable for pharmaceutical patents; in column (4), we interact the treatment term in column (2) with the same pharmaceutical indicator. Regressions in (1–4) report full-sample estimates; in columns (5), (6), and (7), instead, we drop districts in the London, Lancashire, and South-West areas, respectively. Regressions in columns (1–2) include district and year fixed effects; regressions in columns (3–7) include district-by-year, technology class-by-year, and district-by-technology class fixed effects. Standard errors, reported in parentheses, are clustered by district in columns (1–2) and two-way by district and technology class in (3–7). *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.36: Effect of Exposure to US Technology on the Similarity and Originality of Innovation in Great Britain

	Ordinary Least Squares			Reduced Form			Two-Stages Least-Squares		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Dependent variable: "Copying" (Similarity with Previous US Patents)									
Knowledge Exposure$_t$	0.117***			0.029*			0.098*		
	(0.024)			(0.017)			(0.053)		
Knowledge Exposure$_{t-1}$		0.161***			0.029***			0.098***	
		(0.026)			(0.006)			(0.023)	
Knowledge Exposure$_{t-2}$			0.196***			0.013			0.044
			(0.027)			(0.015)			(0.050)
Mean Dep. Var.	17.043	17.981	19.842	14.906	14.278	19.919	14.910	14.283	19.925
Std. Beta Coef.	0.126	0.140	0.139	0.025	0.030	0.008	0.101	0.121	0.031
R^2	0.694	0.569	0.601	0.712	0.712	0.599	-0.002	0.007	0.002
Panel B. Dependent Variable: "Originality" (Similarity with Subsequent US Patents w.r.t. Previous US Patents)									
Knowledge Exposure$_t$	0.175***			0.065***			0.219***		
	(0.019)			(0.011)			(0.032)		
Knowledge Exposure$_{t-1}$		0.084***			0.015*			0.050*	
		(0.014)			(0.008)			(0.027)	
Knowledge Exposure$_{t-2}$			0.097***			-0.029			-0.099
			(0.017)			(0.025)			(0.087)
Mean Dep. Var.	10.885	13.815	15.851	8.584	12.687	15.913	8.586	12.691	15.918
Std. Beta Coef.	0.341	0.127	0.119	0.132	0.020	-0.030	0.531	0.079	-0.121
R^2	0.699	0.672	0.710	0.775	0.718	0.708	0.047	0.006	-0.013
District-Decade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-Technology Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. of District-Class	11268	11268	11268	11214	11214	11214	11214	11214	11214
N. of Observations	67553	67553	56299	56070	56070	56070	56050	56050	56050
K-P F-stat							103.344	103.344	103.344

Notes. This table displays the association between innovation and exposure to US knowledge. The unit of observation is a district-technology class pair, observed at a decade frequency between 1880 and 1939. In Panel A, the dependent variable is the text similarity between UK patents and US patents issued five years before ("copying"); in Panel B, the dependent variable is the similarity of UK patents with US patents granted in the subsequent five years, over the similarity of UK patents with US patents granted in the preceding five years ("originality"). The similarity measure is akin to Kelly, Papanikolaou, Seru and Taddy (2021). In columns (1–3), we estimate the OLS correlation with the observed measure of knowledge exposure; in columns (4–6), we estimate the reduced-form association with the railway-based instrument of knowledge exposure through OLS; columns (7–9) report the two-stage least-squares estimate. Each model includes district-by-decade and district-by-technology class fixed effects. Standard errors are reported in parentheses and are clustered at the district level. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.37: Triple Differences Effect of Exposure to US Shocks on the Similarity Between UK and US Innovation

	Synthetic Shocks				Great Influenza Pandemic Shock			
	(1) Full Sample	(2) No London	(3) No Lancs	(4) No S/W	(5) Full Sample	(6) No London	(7) No Lancs	(8) No S/W
Panel A. Copying (Similarity with Previous US Patents)								
Synth. Shock × Post × Emigrants	0.266***	0.169***	0.328***	0.259***				
	(0.063)	(0.039)	(0.066)	(0.066)				
Pharma × Post × Emigrants					0.763***	0.450***	0.884***	0.664***
					(0.210)	(0.179)	(0.199)	(0.206)
Panel B. Originality (Similarity with Subsequent US Patents w.r.t. Similarity with Previous US Patents)								
Synth. Shock × Post × Emigrants	0.416***	0.238***	0.580***	0.375***				
	(0.140)	(0.094)	(0.148)	(0.144)				
Pharma × Post × Emigrants					5.249***	3.827***	5.538***	4.009***
					(0.990)	(0.868)	(1.068)	(0.925)
District-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District-by-Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Units	393046	382153	343450	375850	429080	408680	375360	401880
Number of Observations	10029	9697	8760	9547	10727	10217	9384	10047
Mean Dep. Var.	0.498	0.376	0.460	0.472	0.791	0.577	0.737	0.733

Notes. This table displays the effect of US innovation shocks on the similarity and originality of innovation activity in the UK compared to US patents. The unit of observation is a district-technology class pair observed at a yearly frequency between 1900 and 1939. In Panel A, the dependent variable is the text similarity between UK patents and US patents issued five years before ("copying"); in Panel B, the dependent variable is the similarity of UK patents with US patents granted in the subsequent five years, over the similarity of UK patents with US patents granted in the preceding five years ("originality"). The similarity measure is akin to Kelly, Papanikolaou, Seru and Taddy (2021). In columns (1–4), the independent variable is an indicator that, for a given district–technology, returns value one after a synthetic innovation shock in that technology class is observed in at least one county where the district has above-average out-migration. A synthetic innovation shock is observed whenever the residualized number of patents observed in the country is in the top 0.5% of the overall distribution. In columns (5–8), the independent variable is an indicator that returns value one for pharmaceutical patents only and only if emigration from the observed district to counties in the top quartile of the influenza mortality distribution is in the top quartile across districts. Both models are triple-difference designs. Models in columns (1–4) are staggered designs and are estimated using the imputation estimator by Borusyak, Jaravel and Spiess (2021). In columns (2) and (6), we drop districts in the London area; in columns (3) and (7), we exclude districts in the Lancashire area; in columns (4) and (8), we drop districts in the South-West area. All models include district-by-year, district-by-technology class, and technology class-by-year fixed effects; standard errors, clustered two-way by district and technology class, are shown in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 3.38: Difference-in-Differences Effect of Neighborhood-Level Out-Migration on Innovation: Alternative Proximity Threshold

	Baseline Sample				Dropping Individuals in...		
	(1)	(2)	(3)	(4)	(5) London	(6) Lancashire	(7) South-West
Panel A. All Emigrants							
Neighborhood Emigrant × Post	0.120**	0.146**	0.133**	11.846*	0.079	0.142**	0.167***
	(0.059)	(0.068)	(0.059)	(6.144)	(0.065)	(0.062)	(0.061)
Std. Beta Coef.	0.016	0.019	0.018	0.155	0.011	0.020	0.022
Panel B. Only Non-Return Emigrants							
Non-Return Neighborhood Emigrant × Post	0.148***	0.199***	0.160***	14.694**	0.061	0.172***	0.226***
	(0.056)	(0.062)	(0.058)	(6.293)	(0.061)	(0.059)	(0.058)
Std. Beta Coef.	0.019	0.025	0.020	0.186	0.008	0.023	0.028
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	–	Yes	Yes	Yes	Yes	Yes
Parish × Year FE	No	Yes	No	No	No	No	No
Matching	No	No	Yes	No	No	No	No
Sample	Full	Full	Full	Inventors	Full	Full	Full
N. of Individuals	473112	473112	469585	4224	410327	422230	352064
N. of Observations	9462240	9412502	9391700	84480	8206540	8444600	7041280
Mean Dep. Var.	0.890	0.892	0.893	99.716	0.794	0.836	0.893
S.D. Dep. Var.	40.291	40.337	40.351	414.695	37.439	39.126	41.333

Notes. This table reports the effect of neighborhood out-migration on innovation. The units of observation are individuals who are observed yearly between 1900 and 1920. In columns (1–3) and (5–7), the sample consists of the universe of males who did not emigrate over the period and that were at least 18 years old in 1900; in columns (4) and (8), we restrict the sample to inventors. The dependent variable is the number of patents obtained annually. In columns (1–4), the sample consists of individuals residing in all England and Wales divisions; in columns (5–7), we exclude the top tree-patents producing areas: London, Lancashire, and the South-West. In Panel A, the independent variable is an indicator that, for a given individual, returns value one after at least one person that was living in the same neighborhood as the individual migrates to the United States; in Panel B, we restrict to emigrants that never return in the period of observation. In this context, "neighborhood" refers to emigrants within a range of 100 meters from the individual in the sample. Each model includes individual and—at least—year fixed effects; in column (2), we include parish-by-year fixed effects; in column (3), individuals are weighted by their coarsened exact matching weight. The estimates are obtained using the method discussed in Borusyak, Jaravel and Spiess (2021) to account for the staggered roll-out of the treatment across individuals. Standard errors, clustered at the district level, are reported in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.