The London School of Economics and Political Science

Essays in Applied Microeconomics

Vera Amanda Malin Dahlstrand Rudin

A thesis submitted to the Department of Economics
for the degree of Doctor of Philosophy

March 2022

Declaration

I certify that the thesis I have presented for examination for the MPhil/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 59,100 words.

Statement of conjoint work

I confirm that Chapter 3 was jointly co-authored with Guy Michaels, Dzhamilya Nigmatulina, Ferdinand Rauch, Tanner Regan and Neeraj Baruah and I contributed 33% of this work. I confirm that Chapter 4 was jointly co-authored with Oriana Bandiera and Greg Fischer and I contributed 33% of this work.

Abstract

This thesis studies service provision and organizations with a spatial perspective. Chapter 2 proposes that digitalisation makes the physical distance between service provider and user less relevant. I quantify the potential gains this flexibility offers in digital primary care in Sweden, harnessing nationwide conditional random assignment. I evaluate causal effects of matching patients of varying risks to doctors with different skills. Matching patients at high risk of avoidable hospitalizations to doctors skilled at triaging reduces avoidable hospitalizations by 20% on aggregate – without affecting other outcomes. Conversely, matching the best triaging doctors to the richest patients leads to more avoidable hospitalizations, since the most vulnerable patients are often the poorest. Hence, remote matching can sever the link between local area income and service quality.

Chapter 3 is a spatial study of infrastructure provision. Africa's demand for urban housing is soaring, even as it faces a proliferation of slums. In this setting, can modest infrastructure investments in greenfield areas where people subsequently build their own houses facilitate long-run neighborhood development? We study projects implemented in seven Tanzanian cities during the 1970s and 1980s, using a spatial regression discontinuity design to compare greenfield areas that were treated (de-novo) with nearby greenfield areas that were not. We find that by the 2010s, de-novo areas developed into neighborhoods with larger, more regularly laid-out buildings and better-quality housing.

Chapter 4 evaluates the effectiveness of performance incentives across locations. Performance rewards are a cornerstone of management practices in Western countries but rarely used elsewhere. We test the hypothesis that the effect of rewards depends on whether a society values individual achievements. To do so, we set up identical data-entry firms in three countries and randomize the incentives offered to workers. We find that the effect of incentives on productivity aligns with the country's rank on the individualism-collectivism scale, ranging from 0 in the least individualistic country to 20% in the most individualistic. We conclude that cultural norms must be embedded in the design of personnel policies.

Acknowledgments

This thesis would never have been written without the support of my advisors Oriana Bandiera, Robin Burgess and Guy Michaels, and the additional support of Nava Ashraf. Their dedication to detail while also seeing the whole picture, and allowing curiosity into a range of research ideas, will be forever an inspiration. I remain immensely grateful to them.

I am thankful for the input from several other members of the Economics Department at the LSE, and from helpful economists at various other universities. I thank UC Berkeley for hosting a research visit in California.

Chapter 2 has only been possible thanks to the collaboration with several employees and former employees at the healthcare provider, most importantly Nasim Bergman Farrokhnia, whose insistence on making the impossible possible has always impressed me. Chapter 3 and 4 have obviously had the benefit of collaborating with fantastic coauthors, all named in the declaration statement above.

Support and grants from STICERD and CEP at the London School of Economics have been instrumental to producing Chapter 2, while Chapter 3 has benefited from funding from the World Bank's Multi-Donor Trust Fund on Sustainable Urbanization, and the International Growth Centre has funded Chapter 4. Finally, I also gratefully acknowledge financial support through the PhD scholarship from the Economic and Social Research Council.

Innumerable discussions and coffees with colleagues, first and foremost Sacha Dray, Virginia Minni, Martina Zanella and Céline Zipfel, have contributed to both the work here and to enjoying the process. I am thankful to my family for always seeing the value in academic work, but also for promoting balance by emphasising the value in everything else in life. This thesis is dedicated to my husband, Anders Fridén, whose support is relentless.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis consists of three core chapters, which are distinct contributions in the fields of health, labor and development economics. They are unified by a common theme of a spatial perspective on the economics of service provision and firms. In particular, the different chapters consider how services like healthcare and infrastructure are provided across space, and how firms need to adapt to operating in different geographic locations.

The first core chapter (Chapter 2) considers changes in the spatial pattern of service provision through digitalization. Digital platforms are transforming services by making the physical distance between provider and user less relevant. I quantify the potential gains this flexibility offers in the context of digital primary care in Sweden, harnessing nationwide conditional random assignment between 200,000 patients and 150 doctors. I evaluate causal effects of matching patients of varying risks to doctors with different skills and assess counterfactual policies compared to random assignment. Matching patients at high risk of avoidable hospitalizations to doctors skilled at triaging reduces avoidable hospitalizations by 20% on aggregate – without affecting other adverse outcomes, such as counter-guideline antibiotics prescriptions. Conversely, matching the best triaging doctors to the richest patients leads to more avoidable hospitalizations, since the most vulnerable patients are often the poorest. Hence, remote matching can sever the link between local area income and service quality in favor of a needs-based assignment, improving the effectiveness and equity of service provision.

The second core chapter (Chapter 3, co-authored) is a spatial study of infrastructure provision. Africa's demand for urban housing is soaring, even as it faces a proliferation of slums. In this setting, can modest infrastructure investments in greenfield areas where people subsequently build their own houses facilitate long-run neighborhood development?

We study Sites and Services projects implemented in seven Tanzanian cities during the 1970s and 1980s, and we use a spatial regression discontinuity design to compare greenfield areas that were treated (de novo) with nearby greenfield areas that were not. We find that by the 2010s, de novo areas developed into neighborhoods with larger, more regularly laid-out buildings and better-quality housing.

The third core chapter (Chapter 4, co-authored) evaluates the effectiveness of performance incentives in different locations. Performance rewards are a cornerstone of management practices in Western countries but rarely used elsewhere. We test the hypothesis that the effect of rewards depends on whether a society values individual achievements. To do so, we set up identical data-entry firms in three countries and randomize the incentives offered to workers. We find that the effect of incentives on productivity aligns with the country's rank on the individualism-collectivism scale, ranging from 0 in the least individualistic country to 20% in the most individualistic. We conclude that cultural norms must be embedded in the design of personnel policies in organizations.

Chapter 2

Defying Distance? The Provision of Services in the Digital Age

2.1 Introduction

A range of services is moving online – including healthcare, banking, and education[1]. In many countries, digitalization started before the pandemic, and has been accelerated by it. A direct implication is that geographical distance no longer by necessity factors into which service provider meets which user – these services are *defying distance*. This creates new opportunities to transform how services are provided by improving the matching between service providers and users to make better use of variation in provider skills.

This paper asks: to what extent can matching patients to online primary care physicians improve healthcare outcomes? In particular, the matching policy I consider is on doctor task-specific skill and patient outcome-specific estimated need or risk. I consider a setting in which the first doctor you see when contacting primary care can be based anywhere in the country, instead of a local physical primary care provider. This setting is ideal to study the potential effects of new technology to improve matching, as primary care is the front line of healthcare with the largest patient pool and the most heterogeneous patients and tasks. Hence, physician specialization and division of labor have the potential to increase output (Smith, 1776). I consider policies matching doctors with high ability in specific tasks, for instance risk prediction and triaging, to the patients most in need[2]. Such improvements to patient-doctor matching, independent of either party's location, can represent a cost-neutral policy in the digital age.

In order to overcome the endogenous selection between physical primary care providers and patients, which normally confounds causal effects of doctors on patients[3], I assemble a novel dataset of consultations, patients and doctors in digital primary care, available across an entire country – Sweden – in 2016-2018.[4] The proprietary data on over 200,000 patients and 150 doctors[5] comes from Europe's largest digital primary care provider. The key feature of the digital care data is that the allocation of doctors to patients is random, conditional on time and date.[6] This is a by-product of the first-come-first-served

[1] Within education, this includes but is not limited to after-school tutoring, worker training programs and some university courses. Other services moving online are, e.g., therapy and counselling, exercise classes, real estate, financial advice and home improvement.

[2] Measured by their risk of each negative outcome, estimated from prior healthcare and demographic data.

[3] For instance, if in physical care some doctors meet patients with higher unobserved risk, selection bias would mean that those doctors appear worse.

[4] The fact that this is before the pandemic allows me to avoid any pandemic-related shocks to behavior, which particularly affected healthcare.

[5] The original dataset included approximately 380,000 patients. The sample restrictions are detailed in the Data section and in the Appendix.

[6] Conditional random assignment holds for 82% of the digital consultations, where patients have chosen to meet the first available doctor. Other options for patients include to book a specific time with a specified doctor. I only use the conditionally randomly assigned consultations for analysis. Moreover, I restrict

assignment procedure of patients to doctors, and neither party has the ability to intervene into this digital process.

To enable the analysis of healthcare outcomes in the physical care system, and to include patients' prior healthcare histories in physical care, this dataset is merged[7] on the individual patient level with physical healthcare data from the universal healthcare system. Here, patients are followed over six years, which allows me to measure patient risk in terms of past diagnoses and healthcare utilization history. Finally, the data is matched on the individual patient level with detailed socioeconomic and demographic variables from *Statistics Sweden*, to enable the addition of demographic variables to the patient risk estimation, and the study of redistributional effects of doctor reallocation across the income distribution.

In this paper, I compare counterfactual doctor skill-patient need matching policies to the most relevant other policies. These are, first, the status quo of digital time-conditional random matching between doctors and patients. Second, I simulate a second benchmark of positive assortative matching on patient income and doctor skill to approximate real-life existing healthcare inequalities in physical care. I provide evidence of such inequalities in physical primary care. Large location-based differences in healthcare outcomes persist within countries (see, e.g., Finkelstein, Gentzkow and Williams 2021) – even in countries with universal public health insurance such as Sweden (Chen, Persson and Polyakova 2021). I also study the *redistributional* effects of doctor skill-patient need matching policies along the patient income distribution. I provide evidence that changing doctor-patient matching can also allow us to address healthcare inequality by severing the link between the quality of local area service provision and patient income.

Estimating doctor ability in primary care has been a challenge, as important patient outcomes are often ambiguous, rare, or delayed.[8] Moreover, primary care physicians have multiple tasks, which opens the question of whether a single ability measure governs performance in all tasks, or whether doctors specialize. I address this by creating observable output measures of doctors in three key dimensions of a primary care physician's work: (1) identifying risky patients and triaging them to higher levels of care (2) providing

the data to one visit per patient, a patient's first visit in the service, to eliminate any concerns about endogeneity in any following consultations.

[7] Data merging and de-identification is done by *Statistics Sweden*.

[8] Mortality is the least ambiguous outcome, but the most rare and delayed as the conditions that people seek care for in primary care are often less serious. The main outcome I use (avoidable hospitalizations) can be seen as a proxy of mortality that is more commonly observed. Moreover, it is a preferable outcome to mortality as it is also more closely linked to the work of the primary care doctor, since this type of hospitalization is defined in the medical literature as preventable by primary care.

guideline-consistent treatment for common conditions and (3) leaving the patient informed and satisfied so that they do not seek additional, costly, care more than necessary. I measure the outcomes in each task by *negative* patient outcomes: in the case of providing guideline-consistent treatment, I measure whether the patient has received a counter-guideline antibiotic. In the case of risk prediction, the negative outcome is an *avoidable hospitalization*, i.e. a hospital admission that could have been avoided with sufficient primary care. For the third outcome, I measure whether the patient has sought additional in-person primary care in the week following the digital care visit, for a subsample. For each of these outcomes, I estimate patient risk. To measure risk for avoidable hospitalizations, I generate a propensity score using pre-determined demographic and healthcare variables, such as age, a disease index of chronic diagnoses, and previous hospitalizations. These are variables available to the doctors in the patients' medical records, meaning that the risk prediction does not use additional data.

I implement an empirical method that allows for both measurement of doctor task-specific skill and estimation of doctor-patient match effects, where the latter uses the measures of doctor skill interacted with patient risk. This method avoids overfitting and the (statistical) winner's curse (Andrews et al. 2021) in two ways: first, it is based on a split-sample strategy, where I split the conditionally randomly assigned data into two: Sample 1 (a "hold-out sample") and Sample 2 (the "main sample").[9] Sample 1 is used to estimate physician effectiveness in each task with a value-added framework. Sample 2 is used to estimate the complementarities between different patient risk types and doctors of varying estimated ability in each outcome[10]. The second step that I take to reduce the noise in the doctor skill estimates is to shrink them using an empirical Bayes method.

In all outcomes, I find large and statistically significant differences across physicians in their task-specific effectiveness. However, the evidence is not consistent with a single latent ability variable governing all of the skills, meaning that doctors even within general practice have individual "specializations"[11]. These specializations are usually not taken into account in the organization of primary care, as a primary care doctor is expected to deal with all types of tasks.

[9] I verify the conditionally random assignment of patients to doctors in both Sample 1 and Sample 2.

[10] The doctor ability was estimated on different patients than those present in Sample 2

[11] This could be due to different innate ability, for instance some are better at speaking with patients and reassuring them (so that they do not seek additional care when not necessary), while others are better at being strict with antibiotics guidelines even if a patient argues that they want antibiotics. Or the absence of a positive correlation between skills could suggest that doctors reach different balances in some trade-off between the prioritization of different tasks of a general practitioner, perhaps due to different risk aversion, preferences or training.

The next step is to understand whether physician-patient matching matters for patient outcomes in primary care. Indeed, the gains from matching are driven by another fact that I establish: that patients have predictable needs for different dimensions of doctor skills. In Sample 2 (the "main sample"), I estimate the effect of matching doctors with high skill in a task with patients who have a high estimated need for that task. One main result of this paper is that if we match a doctor who is among the top 10% at reducing avoidable hospitalizations, with a patient who is predicted to be among the top 1% risky for such adverse outcomes, we could reduce their number of such adverse outcomes by 90%. This is important, as avoidable hospitalizations are a sign of low quality primary care and are most common among low-income individuals. At the same time, patients who are not estimated as "risky" for this outcome have effects that are indistinguishable from zero from seeing a doctor among the top 10% at reducing avoidable hospitalizations. I will call this a complementarity between doctor and patient types.[12]

To increase the relevance of the causal treatment effects of some doctors on some patients, I assess the aggregate impacts of counterfactual policies of reallocations between doctors and patients, adapting a conceptual framework developed by Graham, Imbens and Ridder (2014). This framework enables us to answer a different question than the usual (what would the effect be of increasing a certain input?). In particular, as is especially relevant in healthcare, where the long education of doctors means inputs are difficult to increase, how can we reallocate existing inputs to get an output improvement? This conceptually simple framework relies on conditionally random matching to estimate an average match function (the average outcome when each doctor type meets each patient type), and then uses this function to evaluate effects of counterfactual reallocations. The reallocations chosen are based on an optimization problem, taking existing resources – doctor skills and work hours. The outcomes depend on the distribution and correlation of risks for each outcome in the patient population; the distribution and correlation of doctor skills; and the within-patient and within-doctor correlation of risk and skills across the different outcomes. The framework takes into account the externality on the patient *from* whom a high-skilled doctor is moved in a reallocation. Moreover, I add the consideration of possible correlations between doctor skills in different tasks, and study the effects on other outcomes than the main, in each reallocation.

A counterfactual simulation shows that we could reduce avoidable hospitalizations in the aggregate by 20% by matching doctors and patients, compared to random allocation. This

[12]This type of complementarity also exists for the other outcomes, even if the differences between patient groups are smaller.

reallocation does not negatively affect other main outcomes. The outcome is achieved by only reallocating of 2% of patients, since I show that I can accurately predict who the patients at risk for avoidable hospitalizations are using past healthcare data, and they are a small fraction of all patients. Moreover, the reallocation shifts this aspect of doctor skill (risk prediction and triaging) towards lower-income patients, who are the ones most in need.[13]

Matching is a resource-neutral policy that affects outcomes. However, its efficiency compared to resource-intence policy alternatives such as hiring and training, remains a priori ambiguous. To this end, I compare counterfactual doctor skill-patient risk matching policies to counterfactual physician hiring and selection policies, where doctors who have above median skill in three important tasks expand their hours of work at the expense of doctors with below median skill in these tasks. Even if these doctors expand their hours by as much as 70%, the gains are smaller[14] than from doctor-patient matching policies, and would moreover be more difficult to implement. Matching has larger effects as (1) patients in primary care have heterogeneous needs (and these needs can be identified with prior healthcare data) and (2) doctors have different skill sets that are important to some patients but not others.[15]

Matching of service providers to users is an under-utilized policy tool, which could be welfare improving at close to zero cost when distance is defied by digital services.[16] Algorithmic allocation means that machine prediction is used as a complement to human skill, as opposed to substitute[17]. The algorithm allocates patients to doctors, but the doctor makes the triage, diagnosis or treatment decision. This could make the policy less subject to "algorithm aversion" – that individuals trust recommendations from an algorithm less than from a human (Dietvorst et al. 2015, Yeomans et al. 2019). In fact, versions of matching are already being developed and used by digital platforms, including in digital primary care, without facing as much criticism as for instance artificial intelligence triaging. This paper establishes the potential impacts of such matching, and suggests new measures relevant for matching, such as doctor task-specific skill and patient risk.

[13]The estimated risk of having an avoidable hospitalization, as well as the number of prior avoidable hospitalizations, are concentrated in the lower end of the income distribution.

[14]No significant reduction in avoidable hospitalizations; 4% reduction in counter-guideline prescriptions.

[15]In the case of avoidable hospitalizations, it is also the case that the patients at risk are a very small subset of the total amount of patients. These patients are at risk for dangerous and costly complications, which is why focusing on them is important. The patients at risk for counter-guideline antibiotics are a much larger share of the total patient pool.

[16]The costs would be a small increase in waiting time for some patients, and the costs of importing data and developing the matching algorithm.

[17]If a substitute, the algorithm would make the medical decision. For a setting testing judges' predictions against algorithms, see Kleinberg et al. (2018).

The results on doctors' varying effects on heterogeneous patients could be generalizable also to physical care. The main reasons I focus on digital care are, first, that the policy of doctor-patient matching is feasible in digital care, due to the easing of shared location constraints, making some "nontradables" tradable (Muñoz 2021).[18] Second, digital services can be viewed as a "lab", which helps overcome endogeneity challenges endemic in physical primary care which have prevented the evaluation of causal effects of doctors. This is because, at least initially and for some parts of digital care, doctor-patient assignment has been conditionally random. In regular physical primary care, patient-doctor sorting confounds causal effects and all doctors do not meet all types of patients, meaning there is no common support for match effect estimators. The methods and conclusions of this study could speak also to other sectors, where the allocation of service providers, such as teachers, bank advisors, etc., to external clients could be key for effective production.

Digital provision services has become widespread in many sectors. This is the first paper to study nationwide digital service provision, and the first to show that digital services can defy inefficient and unequal matching due to distance and locational sorting. In addition, I bring a new source of conditionally random matching of service providers and external receivers to the literature. This complements the nascent empirical literature on reallocation and matching as mechanisms to improve outcomes instead of input augmentation (Aucejo et al. 2021, Bergeron et al. 2021, Fenizia 2020, Graham et al. 2021). These papers study teaching, tax collection and bureaucracies. I contribute by developing the ideas to a setting where there are lower obstacles and costs to matching on a large scale: digital service provision. Moreover, I add to this literature by studying matching in a medical setting, where the stakes are high and there is policy-relevant inequality in current resource allocation in many countries.

This paper also contributes to the literature on physician performance[19], by studying not only doctors' overall ability, but also specializations. Alsan et al. (2019) and Cabral and Dillender (2021) studied the effects of patient-doctor homophily on specific characteristics – gender and race, while the present paper is to the best of my knowledge the first to estimate causal effects of doctor skill on heterogeneous patients. Finally, I develop average reallocation effects to a new setting (Graham, Imbens and Ridder 2014, 2020), and implement them in a setting without pre-existing estimates of patient need and doctor skill.[20]

[18] Moreover, in digital care matching can be done by algorithms that quickly access patient and doctor data.

[19] See, e.g., Fadlon and van Parys 2020; Cutler et al. 2019; Abaluck et al. 2016; Doyle, Ewer and Wagner 2010; Grytten and Sørensen 2003.

[20] A more detailed literature overview is given in the appendix.

2.2 Institutional background

2.2.1 Digital and Physical Primary Care in Sweden

Sweden has a tax-financed universal public health insurance. Health expenditures accounted for 10.9% of GDP in 2016-2018.[21] Healthcare is provided by a mix of public (organized by 21 regions) and private providers. Only a small share of citizens - 6% in 2017 (Glenngard 2020) - have an additional private health insurance, mainly provided by employers. Private health insurance accounts for less than 1% of health expenditures (Glenngard 2020). Compared to other OECD countries, few people in Sweden (3.9%) skip a consultation due to cost (OECD 2017). Yet, patients complain of long waiting times for appointments in surveys, and the national goals of limiting waiting times are often unmet. In the few primary care outcomes that are measured and compared across countries, such as hospital admissions for asthma or chronic obstructive pulmonary disease, and congestive heart failure (related to avoidable hospitalizations), Sweden is above OECD average on one of the indicators and below on the other (OECD, 2017).

Primary care is the front line of healthcare, where the initial evaluation of a patient's condition, as well as cost-effective prevention takes place. In primary care in particular, patients are heterogeneous, as are the tasks facing primary care physicians/general practitioners (GPs), but the variation in doctor effectiveness with different patients is largely unknown. This is partly due to the endemic sorting between providers and patients in standard, physical primary care - sorting and selection has dominated this branch of healthcare more than others.[22] Yet, it is in primary care where patients and their health problems are most diverse, and the patient pool is largest, thus offering the largest possibilities for a better allocation.

Primary care physicians are institutionally positioned as a central gatekeeper in the access to healthcare. They are perhaps even more important in countries with universal health insurance, where access to specialists is more restricted, but they are central also in the US system (Fadlon and Van Parys 2020)[23].

Digital primary care, provided through smartphone video consultations, became widely

[21] This a is slightly higher share than the OECD average, but lower than in the US.

[22] Previous research has exploited plausible randomization in, e.g., emergency hospital care to evaluate doctor effectiveness, but randomization has been harder to come by in primary care.

[23] Differences in how primary care works varies both within and across countries. For instance, referrals from the primary care provider to a specialist take place in 3% of consultations in our data. This is comparable to the lower end of GP referrals in the UK physical primary care setting, where in a meta-analysis, they range from 1.5% to 24.5% (O'Donnell 2000).

available in Sweden in 2016. Digital primary care is not suitable for all conditions normally handled in primary care, since some conditions require physical examination or testing. However, many common conditions treated in primary care can be diagnosed and treated digitally. In Sweden, this is provided by private companies that are reimbursed by the regions, which are in turn responsible for the provision of healthcare from the universal public health insurance. Just as in physical primary care, which is provided by a mix of private (40%) and public providers (60%), doctors working in digital primary care are not paid fee for service but an hourly wage. The reimbursement level from the universal public health insurance for digital consultations has changed several times, while the fee faced by patients has remained at the level of fees for in-office primary care consultations during the study period 2016-2018. For children (under 18) and elderly (over 84 years old), the service is free from co-pay, just as in regular physical primary care.

How patients choose regular, physical primary care

Regular (physical) primary care is provided at GP clinics/primary care centers. Most patients are registered with one such clinic, but not registered with an individual doctor. Patients have the possibility to choose their GP clinic.[24] 92% of Swedish inhabitants live within 10 minutes' drive of the nearest physical primary care center, and 80% live within 5 minutes' drive of the second nearest (National Board of Health and Welfare, 2018). However, research indicates that a lower proportion (16% in 2011) of individuals with low education chose another center than their assigned default (compared to 29% among those with higher education) (Bendz 2011).[25] These results are in line with research showing that e.g. lower income students are less responsive to quality when choosing schools and need a larger quality increase to choose a school further away from them, than richer students (Bau 2021).

Physical care sorting: Lower-income areas have worse physical primary care scores

Aggregated public data indicate that patients across the country are less satisfied with their primary care in areas with lower income and higher share first-generation immigrants (Appendix Table 2a.1).[26] As Appendix Table 2a.1 does not directly connect a patient to a primary care center, I complement this with Table 2.1 below which includes the patients registered at the primary care clinics with non-missing data in one region, Skåne (there are around 150 primary care clinics in total in this region). In concordance with the country-

[24] In some regions, e.g. Stockholm, patients can remain unregistered with any GP clinic if they do not make an active choice, while in others, there is a default choice.

[25] This was early after the choice was introduced, and figures are likely to be different today.

[26] Table 2a.1 covers most of Sweden, using a matching between municipality and 4-digit postcode-level observations, and the outcome variable is a patient-reported primary care clinic score from the national patient survey (NPE, 2019).

wide evidence, Table 2.1 also shows that patients have a less positive experience with primary care[27] in areas with a higher deprivation index[28]. Moreover, in more deprived areas, patients are also less satisfied with the information they receive in physical primary care. There is also a marginally significant negative relationship between deprivation and the share of patients who get to see a doctor instead of another profession (e.g., a nurse) when they visit primary care (Column 3). Column 4 measures one aspect of objective quality of care: whether patients diagnosed with diabetes also receive a lipid-lowering treatment. Here, there is no significant correlation with the deprivation index.

Table 2.1: Quality measures of physical primary care centers, patient-reported (1,2) and objective (3,4)

	(1) Positive experience	(2) Satisfied with information	(3) Met physician rather than other profession	(4) Recommended treatment for diabetics
Deprivation index	-10.60***	-6.26***	-0.02*	-0.14
	(2.15)	(2.022)	(0.01)	(3.17)
Constant	89.61***	80.26***	0.42***	63.39***
	(2.21)	(2.02)	(0.011)	(3.11)
N	120	120	149	115
R^2	0.17	0.07	0.02	0.00

Robust SEs in parantheses. Sample is primary care centers in Skåne.

Deprivation index is winsorized. Source: Nationell Patientenkat and Region Skåne.

Sorting patterns into digital care

I assemble and analyze proprietary data from one digital primary care provider, which is the largest of providers in visit volume and which contributed with a majority of all such digital visits in the country during the study period. Patients sort freely into using the digital primary care service, and this is not the only option for primary care or digital primary care. When the service was started, advertisements were made on e.g. public transport, informing about the service and potential reasons to use it. To compare the sorting patterns into digital primary care to the sorting patterns into physical primary care, I study one Swedish region where I have the universe of physical primary care data.[29] This is Region Skåne, which is the southernmost region of 21 regions in Sweden, containing

[27]The outcome variable in Columns 1 and 2 are from the National Patient Survey, *Nationell Patientenkät (NPE)*, 2019, and the variables in Columns 3 and 4 are from Region Skåne's publicly reported data.

[28]The deprivation index is used by the Region and is a weighted average of the variables (1) Born outside EU (2) Unemployed 16-64 year old (3) Single parent with child under 18 years old (4) low education 25-64 years old (5) over 65 years old and in a single household (6) Person over 1 years old who has moved into the area (7) Age below 5 years old.

[29]Primary care data is not collected by the national body (the National Board of Health and Welfare) which contributes with the rest of the physical healthcare data to this study. To get access to physical primary care data in the entire country, separate applications and reviews have to be made to the 21 regions. I do not have data on individual socioeconomic variables of the patients in the region who do not use digital care, only their age.

around 10% of the digital care users and the third largest city in the country.

Using the same index of low socioeconomic status among the patients registered at the clinic as above, I find that the deprivation index (also called the Care Need Index) is similar among digital users and non-users (Figure 2a.3) (extensive margin). However, on the intensive margin (not comparing digital and physical anymore), individuals with higher deprivation index who use the digital service have more appointments in the digital service (Appendix Figure 2a.1). This is corroborated when looking at individual income: lower-income users use the digital service more intensively (Appendix Figure 2a.2). Figure 2a.5 shows that digital care users are younger than non-users. There is a similar level of prior disease among digital users and non-users who are under 60 years old, measured by the sum of comorbidities from the Elixhauser index, a commonly used measure for summarizing disease burden (Elixhauser et al. 1998).[30] For users over the age of 60, non-users seem to have less prior disease.

In results available on request, I have compared the digital care users to the average Swedish citizen (the above is a comparison with the primary care users in one region). This shows that digital care users are more likely to live in cities than the average Swedish citizen. They are less likely to be a first generation immigrant, but more likely to be a second generation immigrant than the average Swedish citizen. In terms of income, adult patients have a somewhat higher median income than the average citizen.

Patients' use of other healthcare while using digital care

Patients take up the service freely, and are not obliged to change their relationship with their regular physical primary care clinic. Using data on physical primary care from Region Skåne, I find that around 4% of digital care users have a nurse contact in physical primary care the week after their digital care visit.[31]

The digital care provider

The healthcare provider contributing with proprietary, de-identified data for this study (in collaboration with Statistics Sweden) provides on-demand primary care via video consultations with certified medical doctors. The physicians may have different specialties, but all are acting as general practitioners, and GP is the most common specialty. During the

[30] In this sorting analysis, the comorbidities are based only on data from primary care for both digital users and non-users, since I do not have data on other care for the digital non-users.

[31] This is consistent with evidence in Gabrielsson-Järhult et al. (2019), who find that 3.6% of digital care users in a different region (Jönköping) have a physical visit at a primary care centre within a week of using a digital care service.

study period, the healthcare provider employed or contracted with around 500 doctors, while around 700 doctors had ever done a consultation in the service (some only had a test period and did not work further).

Patients access healthcare appointments by downloading the company's smartphone application and log in via Sweden's electronic identification system (Bank ID) which is used for all digital bank and governmental agency interaction. Adult patients access the system via their own Bank ID, while child patients need one of their parents or guardians to log in via the parent or guardian's Bank ID.

Randomization

A key feature for this study is that doctors and patients are as good as randomly assigned to each other, conditional on calendar date and time of day. This has not been the intended purpose of the service, but is a by-product of the aim to provide care as quickly as possible nationally. Doctors can choose their time shifts, and often choose them around 2-3 weeks ahead. When they are not busy with a patient or with follow-up work (such as writing prescriptions), or have a booked patient, they are in the roster of available doctors.[32] Patients who enter the system can choose between two tracks: meet the first available doctor ("drop in"), or meet a specific doctor at a specified time. Patients who choose the first track (82%) are effectively randomized to a doctor within this time period. One exception to this is that if there is a doctor in the roster of available doctors who has a pediatric specialty, then this doctor will be more likely to be matched to a child patient if such a patient is in the line. Therefore, I remove all pediatric specialists and the patients they are matched with (see further below in the definition of the analysis sample).

Doctors' incentives and work pattern

Doctors who work for the service almost invariably work part time from home and also work for other healthcare services, such as public or privately run hospitals or clinics. Doctors are recruited across the spectrum of experience, with the conditions that they (1) have a certification as MD (legitimerad läkare) in Sweden from the National Board of Health and Welfare (Socialstyrelsen) which requires that they have finished the 18-21

[32]Data from a later period may not be randomized to as large an extent since the healthcare provider after the study period started trying both matching on language and matching on geographical region (so that e.g. a patient based in the capital would meet a doctor in the capital, even if the meeting was digital, in order to enable a physical meeting as well if needed).

months of intern period/residency (Allmäntjänstgöring, "AT") after medical school (2) that they have at least done 6 months of their intern period/residency (AT) in a Swedish GP clinic/primary care center *or* have at least 6 months of experience at a Swedish GP clinic after the intern period/residency (AT).

Table 2.2: Descriptive statistics of doctors included in the final sample.

	(1)				
	mean	sd	min	max	count
Specialist	0.31	0.47	0	1	143
In specialty training	0.36	0.48	0	1	143
MD + residency only	0.33	0.47	0	1	143
Speaks non EU15 language	0.36	0.48	0	1	143
GP specialist	0.40	0.49	0	1	143
Age	36.9	7.25	28	57	61
Female	0.43	0.50	0	1	61
Employed rather than contractor	0.38	0.49	0	1	52
Observations	143				

Doctors are paid per hour and there is no fee-for-service for the doctors, or bonus payments. Table 2a.2 in the Appendix describes the characteristics of all doctors, which includes doctors who have worked very few consultations for the service during the study period. Table 2.2 shows the same characteristics for doctors who have worked at least 600 randomized consultations for the service and are included in the final sample.[33]

Around 50% of doctors are employed and the rest are hired as contractors, billing from their private company. Doctors can choose either of these methods when starting working for the digital care company. There are benefits to each option, with different tax liabilities, paperwork and pension contributions. The costs for the company are similar: around USD 70-95 per hour. Most doctors work part-time, and most also work in another type of healthcare provision, for instance in a public hospital or in a public physical primary care centre. Doctors are evaluated yearly on key performance indicators, and good performance can lead to a pay increase. The main performance indicators are patients per hour, fraction of patients who are helped, and patient satisfaction. Fernemark et al. (2020) studied the motivations and impressions of doctors working in digital care with e.g. the company studied here. They found that doctors perceive this type of work as highly autonomous, and choose this partly because of the flexibility. They consider the stress level to be reasonably low, but want to complement this work with other types of work in order to continue developing their skills and abilities.

[33]These are 143 doctors. Data on the age and gender and employment status of these doctors are currently missing for a majority of doctors.

2.3 Data

Data	Timing	N	D
Digital care first visits	June 2016-Dec 2018	378,000	511
Hospital, acute & specialist	Jan 2013-Dec 2018	378,000	0
Prescriptions	Jan 2013-Dec 2018	378,000	0
Socioeconomics on adults	2013-2017	180,000	0
Demographics on patients	2013-2017	378,000	0
Primary care in 1 region	Jan 2013- Dec 2019	1.6mn	0

Table 2.3: Overview of data, timing and sample size.

2.3.1 Definition of analysis sample

The sample definition proceeds in three main steps. First, I start from the universe of patients who has had at least one digital consultation with the largest[34] provider of digital healthcare in Sweden, from the start of the service in mid-2016 to the end of 2018. I keep only the first visit for each patient, as these consultations are conditionally randomized, and I want to avoid any concern of endogeneity in following visits in terms of particular patients selecting in to a second visit. Hence, each patient has only one observation in digital care. I restrict the sample to "drop in" visits, that is visits where the patient had no way of specifying which doctor they want to meet, but rather meet the first available doctor. This is 82% of the first visit sample, and this is the sample where conditional randomization (conditional on time) holds. Moreover, I remove pediatricians and those children who are more likely to see a pediatrician (where randomization does not apply).

Second, I match this data to official registry data from Statistics Sweden on socioeconomic and demographic variables and data from the National Board of Health and Welfare (NBHW/ *Socialstyrelsen*) on diagnoses of chronic conditions from specialist, acute and inpatient care across the Swedish healthcare system in the three years preceding digital primary care, 2013-2015. In this physical healthcare dataset, there are many observations per patient. In addition, we match with data on physical primary care (2013-2019) from one Swedish region (Skåne), which matches for around 10% of the digital care sample.[35]

Third, for consistent definition of patient types according to their pre-digital physical healthcare utilization, I drop patients for whom I do not observe the full pre-period 2013-2016, i.e. patients who were born in or after 2013. Finally, I keep only doctors who have

[34] In terms of patient volumes in 2016-2020.

[35] Swedish physical primary care is devolved to 21 regions, which means that data from primary care is not included in the National Board of Health and Welfare data. Assembling primary care data in Sweden across all regions has eluded researchers, as policies, codings and applications vary across the country.

done >600 consultations and their patients, which leaves 210,171 patients (56% of original N) and 143 doctors (20% of original D). The reason is that many doctors were hired late in the sample period, since the service was expanding. These doctors have only done a few randomized consultations, many of them under 100. For more details on the sample definition, see the Data Appendix.

2.3.2 Measurement of outcomes

Estimating doctor performance in primary care has been a challenge, as important patient outcomes are often ambiguous, rare, or delayed. We might care most about mortality and quality of life, but also about costs to the rest of the healthcare system, where primary care physicians serve as gatekeepers. Mortality is the least ambiguous outcome, but the most rare and delayed as the conditions that people seek care for in primary care are often less serious. The main outcome I use (*avoidable hospitalizations*) can be seen as a proxy of mortality that is more commonly observed. Avoidable hospitalizations can even be seen as a preferable outcome to mortality as it is also more closely linked to the work of the primary care doctor, since this type of hospitalization is defined in the medical literature as preventable by primary care.

Moreover, primary care physicians have multiple tasks, which opens the question of whether a single ability measure governs performance in all tasks, or whether doctors specialize. I address this by creating observable output measures of doctors in three key dimensions of a primary care physician's work: (1) identifying risky patients and triaging them to higher levels of care (2) providing guideline-consistent treatment for common conditions and (3) leaving the patient informed and satisfied so that they do not seek additional, costly, care more than necessary. I measure the outcomes in each task by *negative* patient outcomes: in the case of providing guideline-consistent treatment, I measure whether the patient has received a counter-guideline antibiotic. In the case of risk prediction, the negative outcome is an avoidable hospitalization, i.e. a hospital admission that could have been avoided with sufficient primary care. For the third outcome, I measure whether the patient has sought additional in-person primary care in the week following the digital care visit, for a subsample.

Rare vs. frequent

The main outcome in this paper is Avoidable hospitalizations, which is a rare event (0.2% of patients have an avoidable hospitalization in the 3 months following the digital con-

Table 2.4: Summary of outcomes

Negative outcome	**Frequency**	**Non-missing data**
Data on full sample		
Avoidable hospitalization 3 months	0.2%	100%
Counter-guideline prescription	2%	100%
Data on part of sample		
Contacted physical nurse week after	4%	11%

sultation).[36] Yet, this is the most high stakes outcome of those which are measurable in the data and relatable to doctor inputs. The need to measure and understand rare and high-stakes events has been emphasized not least by the literature in financial economics (Dow and Bond 2021)[37] and the economics of disasters (e.g., Barro 2009)[38]. Another reason to focus on this outcome is that one of the main tasks of a primary care doctor is "looking for a needle in a haystack", i.e., sort the rare and seriously ill patients from the vast majority with minor complaints. Studies on healthcare in the United States often include mostly utilization and cost outcomes. A notable exception is Fadlon and Van Parys (2020) who include both outcomes such as avoidable hospitalizations and the doctor following guidelines, but not patient behavior outcomes or patient satisfaction.[39]

Avoidable hospitalizations (AH)

The number of avoidable hospitalizations[40] is a widely accepted measure of healthcare performance, as the hospital admission could be avoided if primary care was timely and adequate. Bacterial pneumonia, urinary tract infection and congestive heart failure accounted for 77% of the AH costs in US (Rocha et al 2020). The Center for Medicare and Medicaid Services (CMS) states that high rates of avoidable hospitalizations could indicate that "beneficiaries are not receiving high-quality ambulatory care," and that low rates of ACSCs at the provider level "may signify that the provider is providing better primary care and coordinating effectively with providers in the continuum of care" (CMS,

[36] That this adverse outcome is rare is, of course, a good outcome of primary care.

[37] This has also been at the forefront of public debate after the financial crisis and the pandemic. Dow and Bond cite Taleb (2007): "[w]hy do we keep focusing on the minutiae, not the possible significant large events, in spite of the obvious evidence of their huge influence?"

[38] Barro (2009) estimates the risk for disasters as 2% per year and shows that they have large welfare costs: society would be willing to reduce GDP by 20% each year to eliminate these rare adverse events. The rare events involved a contraction of GDP of 15-60%. An avoidable hospitalization involves not only the event per se, but can have large negative consequences as it is a negative health event that may lead to prolonged loss of productivity, and some risk of death.

[39] Only 0.24% of consultations in our sample end up in an avoidable hospitalization, as our sample is relatively young and seek for more minor conditions. In Fadlon and Van Parys (2020), 5% of the sample has an avoidable hospitalization each year. So in a sample like Fadlon and Van Parys', where the sample is over 65 years old and relatively sick.

[40] They are also called hospitalizations for ambulatory care sensitive conditions (ACSCs). This outcome is also used in Fadlon and Van Parys (2020) (using the 2017 CMS ACSCs).

2017). Avoidable hospitalizations are dangerous, both because of the inherent risks to a condition that has worsened unnecessarily, and because hospitalization in itself has risks such as hospital-acquired infections and risks from procedures carries out. It is estimated that 1.1 potential life year is lost from every AH (Rocha et al 2020). In both the United States and Sweden, AH decrease with income (McDermott and Jiang 2020), so reducing them could have an impact on health inequality.

Avoidable hospitalizations are also costly. In the US in 2017, 3.5 million adult AH (13% of hospitalizations) cost hospitals $33.7 billion (9% of costs for all adult non-childbirth hospital stays).[41] In Sweden, avoidable hospitalizations cost an estimated SEK 7.1 billion ($820 million) each year, and this represents 7% of all costs for inpatient curative and rehabilitative care.[42] The share of AH costs out of all national health expenditures is 1.3% in Sweden[43] (and also around 1% in the United States), and the share of these (purely hospital) costs out of GDP is 0.15% in Sweden.[44]

I create the variable measuring an avoidable hospitalization using the data from the National Board of Health and Welfare on all hospitalizations 2013-2018, where I code the hospitalization as an avoidable hospitalization if it has a diagnosis code (ICD 10) which is listed in Table A1 of Page et al. (2007). As a pre-digtial health risk factor, I use avoidable hospitalizations that took place within 3 years before the digital consultation. As an outcome of a digital consultation, I use avoidable hospitalizations that take place within 90 days of a digital consultation. Most of the avoidable hospitalizations within 90 days happen quite early after the digital consultation, and the mean is 33 days (see Figure 2a.10 in the Appendix for the distribution of number of days). I conduct several checks to determine whether the avoidable hospitalization can actually be considered as preventable in the digital consultation.

First, the most common diagnosis groups[45] which are registered at the hospital as the primary diagnosis for the avoidable hospitalization within 3 months after the digital consultation are respiratory and genitourinary (connected to kidneys and e.g. complications of urinary tract infections), see Figure 2a.7 in the Appendix. These are conditions which are commonly treated in digital care, for instance by prescribing antibiotics for urinary

[41] Jiang et al. (2009); McDermott and Jiang (2020).

[42] The number of hospital days for AH was around 1 million in Sweden in 2010 (Socialstyrelsen, 2011). The average cost per day in inpatient care is 7100 SEK (Socialstyrelsen, 2017). The exchange rate used as of 13 Sep 2021 is 8.64 SEK/USD. The costs for total inpatient and rehabilitative care are from Statistics Sweden Statistikdatabasen, 2021.

[43] Total expenditures were 528 billion SEK in 2018, from Statistics Sweden Statistikdatabasen, 2021

[44] GDP was 4 828 billion SEK in 2018, from Statistics Sweden Statistikdatabasen, 2021

[45] This is a medical grouping of the ICD diagnosis codes into 23 categories related to the type of disease

tract infections.[46] Second, patients who I have determined as risky for avoidable hospitalizations (see Sections 4.3 and 4.5) also are more likely to come to the digital service with symptoms that can later be related to avoidable hospitalizations: respiratory symptoms and urinary tract infection (see Figure 2a.6 in the Appendix). Moreover, I compare the diagnosis group[47] set by the digital care doctor to the diagnosis group set as primary diagnosis by the hospital, and find that 33% concord in respiratory system, 20% concord in genitourinary system, and 27% concord in symptomatic diagnosis (these are the 3 most common groups for these avoidable hospitalizations).

Counter-guideline prescriptions (CGP)

Widespread non-adherence to medical guidelines contributes to hospitalizations, deaths, and spending (Neiman 2017). Such non-adherence has recently been studied with growing interest in economics, see e.g. Finkelstein et al. (2021) and Frakes et al. (2021). Non-adherence to guidelines on antibiotics prescriptions is particularly interesting since excessive antibiotics prescriptions lead to the negative externality of bacterial resistance.

Bacterial resistance means that the antibiotics that are usually effective in treating a bacterial infection will no longer work, which can lead to prolonged infection and mortality. The guidelines serve to limit the use of antibiotics to where the benefit outweighs the social cost of using them. Bacteria adapt under pressure[48], and if there is less prescription of antibiotics, it is possible to decrease the number of resistant bacterial infections (Bergman et al 2004). Sizing the costs of antibiotics resistance has proved challenging, but a very conservative estimate of the additional healthcare and prescription costs only (not counting lost productivity etc.) is SEK 160 million yearly in Sweden (Folkhälsomyndigheten 2013). This is likely under-estimating the total costs, due to the externality: resistance developed in Sweden also affects people in other countries.

I code non-adherence to 16 guidelines from Swedish strategic programme against antibiotic resistance on digital care (Strama 2017, 2019). All the guidelines are intended to limit the use of antibiotics or use a more narrow-spectrum antibiotic as a first line of response (which contributes less to resistance than a broad-spectrum antibiotic). Thus, to follow the guidelines, doctors sometimes need to say no to patients who think that they need antibiotics. To define the variable, I combine the incidence of prescription in the digital care data, conditional on the diagnosis (ICD) code, with data on the drug code from the

[46] Figures 2a.8 and 2a.9 in the Appendix show that hospitalizations in general have a very different distribution of diagnosis groups.

[47] This is a medical grouping of the ICD diagnosis code that the doctor actually set into 23 categories related to the type of disease

[48] Penicillin was released in 1941, and a resistant germ to this antibiotic was identified in 1942. Another antibiotic, methicillin, was released in 1960 and the resistant germ was found in the same year.

NBHW's prescription register, which occurs once the patient has filled the presctiption.[49] The non-adherence measured in my sample is quite low by international standards. The Centers for Disease Control and Prevention (2019) estimate that 28% (47mn courses) of all antibiotics prescribed in doctors' offices and Emergency Departments in the United States are for infections that do not need antibiotics.

Contacted physical care within a week after the digital consultation

This is an outcome which is important for primary care costs and for patient satisfaction. If a patient contacts a physical primary care clinic in the week following the digital care consultation, this may indicate that they were not satisfied with the digital care consultation or the information given. This incurs additional costs to the universal health insurance in cases where the digital care consultation incurred a payment (which is not the case if the visit was deemed inappropriate for digital care by the doctor). I can measure this outcome in the region where I have primary care data, Region Scania, consisting of around 10% of the digital care sample.

2.4 Framework

This section lays out the econometric framework for estimating match functions between patients and doctors, and average reallocation effects from rearranging the doctor-patient matches. I follow Graham et al. (2014, 2020) in setting up the empirical framework, with some modifications and extensions. This framework takes into account the externality on the patient from whom the high skilled doctor is moved, and I add the consideration of the opportunity costs in terms of other outcomes when doctors are multitasking and skills are potentially correlated.

This study is complementary to the literature on mechanism design in matching markets (for an overview, see Roth 2012), where strategic incentives of agents are taken into account when studying matching problems. In this paper, I do not study strategic incentives of patients and doctors over whom they match with. There are two main reasons for this. First, in some settings (such as the new digital assignments in several markets), agents have little control over who they match with. Second, as Graham (2011) points out, the study of the effects of alternative assignments is the first step in a more complete policy

[49] Thus, there will be a slight under-estimation of the counter-guideline prescriptions, since if patients do not fill the prescription the prescription, we will not know if it was an antibiotic. I do robustness analysis using only the incidence of prescription for diagnosis codes where no antibiotics should be prescribed in digital care, available upon request.

formulation - before deciding if mechanism design of a decentralized system to implement a desired outcome is relevant, we need to know if there are large benefits to alternative allocations.

Consider D doctors and N patients. Doctors have observable characteristics W_j, which measure doctor effectiveness in different tasks, and patients have observable characteristics X_i which measure patient risk/need for different doctor inputs, and is predicted from patients' healthcare history. Patients also have unobserved attributes V_i and doctors have unobserved characteristics U_j. The potential healthcare output (healthcare outcome Y_{ij}) when patient i matches with doctor j:

$$Y_{ij} = g(W_j, U_j, X_i, V_i) \tag{2.1}$$

The research design is based on on random assignment (conditional on time[50]) of patients to doctors. Randomization of doctors to patients ensures that the joint density of patient observed characteristics X_i, unobserved characteristics V_i and doctor observed characteristics W_i and unobserved characteristics U_j can be factorized:

$$f_{X_i,V_i,W_i,U_i}(x, v, w, u) = f_{X_i,V_i}(x, v) f_{W_i,U_i}(w, u) \tag{2.2}$$

Under restriction (2) on the joint distribution of the characteristics of patients and doctors, the conditional mean of the outcome Y_i is:

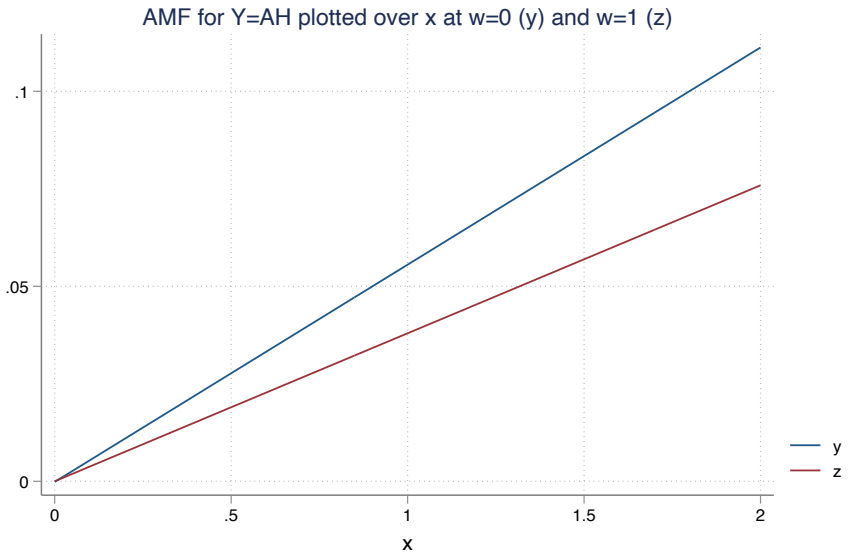
$$Y_{ij}|X_i = x, W_j = w = \iint [g(x, w, v, u) f_{V_i|X_i}(v|x) f_{U_j|W_j}(u|w)] dv du \equiv \beta(x, w) \tag{2.3}$$

The Average Match Function (AMF), $\beta(x, w)$, provides information on how match output varies across different types of agent pairings, when both doctor and patient are random draws from their respective subpopulations x and w. Figure 2.1 shows an example of the AMF in our context.

The Average Match Function is the main building block for conducting counterfactual

[50]The framework will omit the conditioning for simplicity, see Graham (2011, p. 989) for identification conditions under conditional random matching. The conditioning is on time of day (shift) and date of joining the queue for a consultation.

Figure 2.1: Illustration of the Average Match Function (AMF).



The y-axis measures Avoidable Hopsitalizations (AH) and x-axis measures patient risk. W is doctor quality, where w=1 is 1 sd better than w=0, and w=0 measures the worst doctor at this outcome. The positive slopes of both graphs show that a risky patient has higher risk of an avoidable hospitalization, and the flatter slope of the z-graph (where w=1, i.e. a 1sd better doctor) shows the risk is reduced more for risky patients when they meet a better doctor at this task.

analyses. Consider a counterfactual assignment of doctors to patients, i.e. a conditional distribution of doctor types $\tilde{W}_j$[51]:

$$\tilde{f}_{\tilde{W}_j|X_i}(w|x) \tag{2.4}$$

which satisfies the feasibility condition:

$$\int \tilde{f}_{\tilde{W}_j|X_i}(w|x) f_{X_i}(x) dx = f(w) \tag{2.5}$$

for all $w \in W$.

The distribution of patients is kept fixed, i.e. $f_{X_i}(x)$ is left unmodified. Average healthcare outcomes under a counterfactual patient-doctor assignment equal:

$$\tilde{Y} = \int \left[\int \beta(x,w) \tilde{f}_{\tilde{W}_j|X_i}(w|x) dw \right] f_{X_i}(x) dx \tag{2.6}$$

where the average match function is used as a building block. The Average Reallocation

[51] $\tilde{W}_j$ has an equal marginal distribution to W_j (due to the feasibility condition) but the distribution conditional on patient attributes will differ.

Effect (ARE) from the reallocation $\tilde{f}$ is $\tilde{Y}$ relative to the average outcome under the status quo allocation, $\bar{Y}^{sq}$:

$$ARE(\tilde{f}) = \tilde{Y} - \bar{Y}^{sq} \tag{2.7}$$

Since everything to the right of the equality in equations (6) and (7) is identified, so is the Average Reallocation Effect (Graham et al. 2014, 2020). To calculate this, I first compute the expected outcome for each type of patient (e.g., $X_i = x$) given their new doctor assignment (e.g., to type $\tilde{W}_i = w$ - the inner integral in equation (6). I then average over the status quo distribution of X_i, which is left unchanged (the outer integral in equation (6)). This yields average patient outcomes under the new assignment of doctors to patients.

2.4.1 Problem: Reallocation of Fixed Healthcare Resources

The objective of this problem[52] is to improve healthcare outcomes, under the constraint that resources are fixed. Here, the fixed resources are the doctors, including their abilities and number of consultations. Hence, I assume that in the relatively short run I am considering, it is not possible to hire more doctors or increase their abilities. In an extension, I consider selective hiring policies where I extend the working hours of the doctors who have above median skill in several tasks. However, such policies yield smaller gains than matching policies, and are not directly related to the increased flexibility in matching that digital care enables.

I will make one main simplification: to focus on one outcome k at a time, e.g. reducing avoidable hospitalizations. This is reasonable since it is unclear how a planner should weight the different outcomes against each other. I will study what happens to other outcomes when I reallocate to improve one outcome, e.g. avoidable hospitalizations. In fact, it turns out that doctor skills are not positively correlated across outcomes, so there are no important trade-off between the different outcomes measured here. This further strengthens the case for simplifying the problem by focusing on one outcome at a time.

To be realistic, I assume that the planner does not observe U_j or V_i, hence I are restricted

[52]It can be interpreted as a problem of a social planner, or of a planner of healthcare provision, either in a healthcare system such as Medicare or a national healthcare system, or a planner in e.g. a Health Maintenance Organization.

to consider only reallocations where unobserved traits are randomized. From now on, I let W_j and X_i be discretely-valued. This is motivated by the fact that I will reduce the dimensionality of doctor and patient types below in order to calculate a binary version of the average match function.

Suppose we know the AMF $\beta(w,x)\ \forall(w,x) \in W \times X$ (up to sampling uncertainty), and the marginal distributions of doctor and patient characteristics: $\rho = (\rho_1, ..., \rho_D)'$ for $\rho_d = Pr(W_j = w_d)$ and $\lambda = (\lambda_1, ..., \lambda_P)'$ for $\lambda_p = Pr(X^i = x_p)$. The planner chooses the assignment function $\pi_{ij} = Pr(W = w_j, X = x_i)$ to minimize a negative healthcare outcome such as avoidable hospitalizations:

$$min_\pi Y^k(\pi) = \sum_{i=1}^{I}\sum_{j=1}^{J} \beta^k(x_i, w_j)\pi_{ij} \tag{2.8}$$

s.t. feasibility/status quo constraints:

$$\sum_{j\in J} N_p \pi_{ij} = N_x \qquad \forall x \in X \tag{2.9}$$

(each patient gets 1 doctor)

$$\sum_{x\in X} N^\pi(x,w) = N^\pi_{SQ}(w) \qquad \forall w \in W \tag{2.10}$$

(same workload as in Status Quo (SQ))

where N_p = total number of patients, N_x = number of patients of type x, $N^\pi(x,w)$ = number of patients of type x that doctor w meets in any assignment π, $N^\pi_{SQ}(w)$ = total number of patients that doctor w are assigned to in the status quo (SQ). This problem is similar to those found in Graham, Imbens and Ridder (2020) and Bergeron et al. (2021).

The difference between a candidate assignment and the completely random matching (i.e., the status quo situation where both observed and unobserved characteristics are randomized) is given by:

$$ARE = Y(\pi') - Y(\pi^{rdm}) = \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} (\pi'_{ij} - \rho_j \lambda_i)(\beta(w_J, x_I) - \beta(w_J, x_i) - [\beta(w_j, x_I) - \beta(w_j, x_i)]) \tag{2.11}$$

where the last term is a measure of the average local complementarity between W and X.

Outcome-maximising assignments will tend to be assortative in regions of complementarity $[\beta(w_J, x_I) + \beta(w_j, x_i)] - [\beta(w_J, x_i) + \beta(w_j, x_I)] > 0$ (Becker 1973, Graham 2011). I will show evidence of complementarities and evaluate ARE of assortative matchings. The average reallocation effect (ARE) takes into account the externality on the patient from whom the high skilled doctor is moved. For each counterfactual reallocation, I will not only compute the average reallocation effect for the main outcome which was intended to be improved with this reallocation, but also compute AREs for other outcomes. The latter will shed light on the opportunity costs of reallocation in terms of other outcomes when doctors are multitasking and skills are potentially correlated.

2.5 Empirical strategy

The empirical strategy is founded on two building blocks. The first is nationwide conditional random assignment between patients and doctors in digital primary care service. This generates both variation in patient types that each doctor meets - variation in geographic location, age, socioeconomic status, previous healthcare utilization, etc. Moreover, the random allocation allows for causal identification of doctor effects, in contrast to the usual patient-doctor sorting in primary care.

The second building block of the empirical strategy is a split-sample approach, which is used to avoid overfitting and the (statistical) winner's curse (see Andrews et al. 2021). In particular, I evaluate doctor effectiveness using a fixed effects method in a hold-out sample of randomized digital care (Sample 1). In Sample 2, I use the estimates of doctor skill to estimate causal match effects with patients. This creates the average match function: the expected adverse outcomes conditional on the doctor and patient types. It is also in Sample 2 that I estimate the effects of counterfactual assignments. The samples are completely disjoint and no patients exist in both samples. Both samples have conditional random assignment between doctors and patients.

The patient risk factors are estimated from a third separate sample, which consists of

pre-digital (physical) healthcare data in 2013-2016 - the period preceding digital care. In this sample, I identify patient risk (X_i): using past healthcare records. I find logical ex ante patient characteristics which could indicate need for each outcome Y^k, and for the rare outcome avoidable hospitalizations, I predict the risk with a propensity score.

I choose to split the digital care sample in 40-60% shares for power reasons. Also, I want equal number of observations for doctors in Sample 1, thus using 600 visits for each doctor in the estimation of their effectiveness. I choose their *first* 600[53] randomized visits because that is how the procedure could be operationalized. It gives the employer ∼3 months of work by the doctor before they can evaluate the doctor and decide whom to match the doctor to.[54]

2.5.1 Validating the random assignment using pre-digital care administrative data

The identifying assumption is that within a time period (defined as a 3-hour shift, unique for each date), the allocation of doctors is orthogonal to any patient characteristics which affect the outcomes. To test this for observables, I regress doctor characteristics on patient characteristics when controlling for shift-by-date (randomization strata) fixed effects. Table 4 shows that characteristics are balanced.

2.5.2 Estimating doctor skill - in Sample 1

Primary care physician skill is challenging to evaluate for several reasons: (1) pervasive sorting between primary care physicians and patients, (2) a lack of linked patient-provider datasets followed over time (3) multitasking and the ambiguity of many measurable outcomes, (4) the delayed nature of the outcomes, and (5) the co-production of healthcare with the patient, where patient adherence, motivation and understanding is key. To overcome (1) and (2), I use the unique random patient-doctor allocation in a nationwide digital primary care service in Sweden in 2016-2018. I also match this with rich pre-digital care administrative data on both healthcare use and socioeconomics to validate the random assignment mechanism to doctors in digital care. For (3), I recognize that multitasking is at the core of possible specialization, and I will deal with this by defining several doctor tasks which stand in direct relation to measurable patient outcomes. I assemble data

[53] Robustness to 500 and 600 available on request.

[54] The median number of randomized appointments/doctor/calendar day is 10, and I assume 60 working days in 3 months.

Table 2.5: OLS of doctor on patient characteristics for dropin first visit

	(1) Doc foreign	(2) Only MD	(3) Specialist	(4) GP specialist
Female patient	-0.001	0.003	-0.005**	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)
Patient age	-0.000*	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
1st gen immigrant	0.001	-0.001	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)
2nd gen immigrant	0.000	-0.002	0.009*	0.000
	(0.006)	(0.006)	(0.005)	(0.006)
Municip. density	-0.000	0.000	0.000**	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Sthlm county	-0.002	-0.003	-0.002	0.002
	(0.004)	(0.003)	(0.003)	(0.003)
Self-employed	0.004	-0.004	0.010**	-0.004
	(0.005)	(0.005)	(0.005)	(0.005)
Unemployed	0.009	-0.001	0.003	0.003
	(0.007)	(0.006)	(0.006)	(0.006)
University	0.003	0.002	0.001	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)
Yearly income SEK	0.000*	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Welfare	-0.003	-0.001	0.005	0.000
	(0.004)	(0.004)	(0.004)	(0.004)
Hypertension pre-2016	0.010	0.000	-0.021	0.003
	(0.016)	(0.015)	(0.014)	(0.015)
Asthma pre-2016	-0.006	0.004	0.011	0.002
	(0.010)	(0.010)	(0.009)	(0.010)
Diabetes pre-2016	0.018	-0.036	0.026	0.034
	(0.027)	(0.026)	(0.024)	(0.026)
Anxiety pre-2016	-0.010	-0.001	-0.012*	-0.003
	(0.007)	(0.007)	(0.006)	(0.007)
Depres. pre-2016	-0.006	-0.004	0.006	-0.011
	(0.009)	(0.009)	(0.008)	(0.009)
Constant	0.443***	0.369***	0.278***	0.354***
	(0.005)	(0.005)	(0.005)	(0.005)
N	130941	130941	130941	130941
R^2	0.000	0.000	0.000	0.000

Year-month-date-time shift fixed effects included

on all patients' prescriptions, acute and secondary healthcare utilization in the national healthcare system, and define several important dimensions of the doctor's outputs, some of which are standard (e.g., preventing avoidable hospitalizations) and some novel (e.g., motivating the patient to fill prescriptions).

To deal with the delayed nature of the outcomes, (4), I use a variety of outcomes, ranging from frequent and lower-stakes, to rare and high-stakes, but all of which are measurable within 3 months. Some of these are closely related by the medical literature to longer-term outcomes such as mortality. I address (5) by specifically studying the varying effectiveness of different doctors with heterogeneous patient types. The co-production with the patient is at the core of possible complementarities, and I use a set of outcome measures that are at varying proximity to the locus of control of the doctor.

In a sample consisting of doctors' first 600 randomized consultations (40% of the sample), I estimate the doctor effect for each task as the average of the effect across all the patients.

$$Y_{ij} = Z_i\Pi + \lambda_t + u_j + \epsilon_{ij} \tag{2.12}$$

where $\hat{u}_j = \hat{W}_j^{EB}$ is estimated as the Empirical Bayes shrunk random effect of doctor j.[55] This regression is estimated separately for all the outcomes k. λ_t capture date-shift fixed effects (randomization strata) and Z_i is a vector of patient characteristics.[56]

Given a large enough sample size (creating common support in patient types for all doctors) and random allocation, all doctors have a similar patient pool. $\hat{W}_j$ is unbiased due to random assignment and a common support assumption. The common support assumption is that all doctors meet all type of patients in the sample where I estimate the doctor effectiveness. This sample consists of each doctors' first 600 randomized consultations, and this ensures that >95% of doctors have met a patient with an avoidable hospitalization in the past 3 years. However, $Var(\hat{W}_j)$ is positively biased due to (1) regular sampling noise (2) additional noise from sampling noise in the patient sensitivity. I perform an Empirical Bayes shrinkage procedure for the doctor estimates, which results in a best linear predictor of the random doctor effect (Morris 1983). The noisy estimate of doctor quality from a value added regression is multiplied by a measure of its reliability, which in turn is the ration of signal variance to signal plus noise variance. Similar shrinkage

[55] A Durbin Wu Hausman test between fixed and random effects does not reject random effects: $Prob > \chi^2 = 0.16$. Results with fixed effects instead of random are similar and are available upon request.

[56] In the case of Y_{ij} being the outcome variable avoidable hospitalizations, Z_{ik} includes the patients' past number of avoidable hospitalizations in the 3 years before digital care, as this is a rare outcome which is correlated over time within patient.

is common in studies of teacher value-added (see e.g. Kane and Staiger 2008; Chetty et al. 2014; Hjort et al. 2021.). Table 2a.7 in the Appendix shows this regression for the outcome negative number of avoidable hospitalization (so that the random effect is higher for a better doctor).

2.5.3 Defining patient types

We define patient types based on the corresponding traits which could logically complement the doctor observable tasks. These are listed in Table 2.6. For the rare outcome avoidable hospitalizations, I create a propensity score based on the lagged outcome variables from data before digital healthcare (2013-15):

$$P_i = X_i\Gamma + \upsilon_i \tag{2.13}$$

where P_i is the past number of avoidable hospitalizations and X_i are both demographic and healthcare-related variables. I generate a prediction $\hat{P}_i$ for each i, which is what I use as the patient risk variable in Sample 2 (a disjoint sample). Table 2.7 reports the regression used to create the propensity score for avoidable hospitalization risk.[57]

Dimensionality reduction for types regarding avoidable hospitalizations

To make fewer assumptions about the nature of complementarities in the match function, I reduce dimensionality of the patient types to a binary variable measuring high and low risk, and use this variable in a regression to estimate the average match function. Since around 1% of patients have an AH each year nationally, I characterize 1% of patients as risky ($X = 1$) based on the rank of the propensity score $\hat{P}_i$. To allow for a less time-restricted matching, I characterize 10% of doctors as high skilled in preventing avoidable hospitalizations ($W = 1$) based on the rank of $\hat{W}_{jk}^{EB}$.[58]

Figure 2.2 illustrates that the groups created based on the propensity score are closely related to the number of past avoidable hospitalizations of the patient. A patient in the risky group has had on average 0.35 AH in the past 3 years, while a patient classified as not risky has had on average 0.01 AH in the same period. Figure 2.3 shows how the risk groups (defined only based on past healthcare records and demographics) are highly

[57] This is done with a linear probability model, but robustness checks with ordered logit do not change the results.

[58] This will give a lower effect of the interaction effect than if I had also picked the top 1% of doctors in this skill, but since I do not want to make patients wait too long for the best doctor for them, I pick 10% so that there is a wider choice of good doctors in this skill in each time period.

predictive of future avoidable hospitalizations.

Table 2.6: Outcome variables and risk definition

Quality measure for doctor	**Prior outcome for patient (2013-15)**
1. Preventing avoidable hospitalizations	Risk score of nr. avoidable hospitalizations
2. No counter-guideline antibiotics prescription	Antibiotics share of total prescriptions
3. Visit to physical nurse following week	Patient visiting physical nurse week before

Notes: Overview of which patient prior variables are used to define target groups for different doctor quality measures.

2.5.4 Match effects: In Sample 2

By interacting doctor effectiveness with the relevant patient characteristic (X_i) in a second step, I estimate individual sensitivity to doctor input. Again, this is estimated in a different sample (Sample 2) from that where I estimated $\hat{W}_{jk}^{EB}$ (Sample 1). Sample 2 is each doctor's first visit randomized consultations *after* the 600th.

The estimation of the healthcare production function will be semi-parametric, to avoid making too restrictive assumptions on the form of complementarities. To do this, I reduce the dimensionality of doctor and patient characteristics, by making X and W binary. I characterize 1% of patients as risky ($X = 1$) based on the ranking of their propensity score $\hat{P}_i$ (around 1% of patients have an AH each year nationally). Moreover, I characterize 10% of doctors as high skilled in each outcome ($W = 1$) based on $\hat{W}_{jk}^{EB}$. Then, I estimate the effect of a top 10% doctor on top 1% risky patient:

$$Y_{ij} = \alpha + \beta_1 W_j + \beta_2 X_i + \beta_3 W_j X_i + \lambda_t + e_{ij} \tag{2.14}$$

where λ_t is date-time-shift (randomization strata) fixed effects. Standard errors are clustered on doctors. The main coefficient of interest is β_3. In addition, β_{2k} measures how different the patient group as I defined it is in the outcome variable on average.

2.5.5 Reallocation procedures and costs

Reallocation based on the binary AMF to reduce avoidable hospitalizations

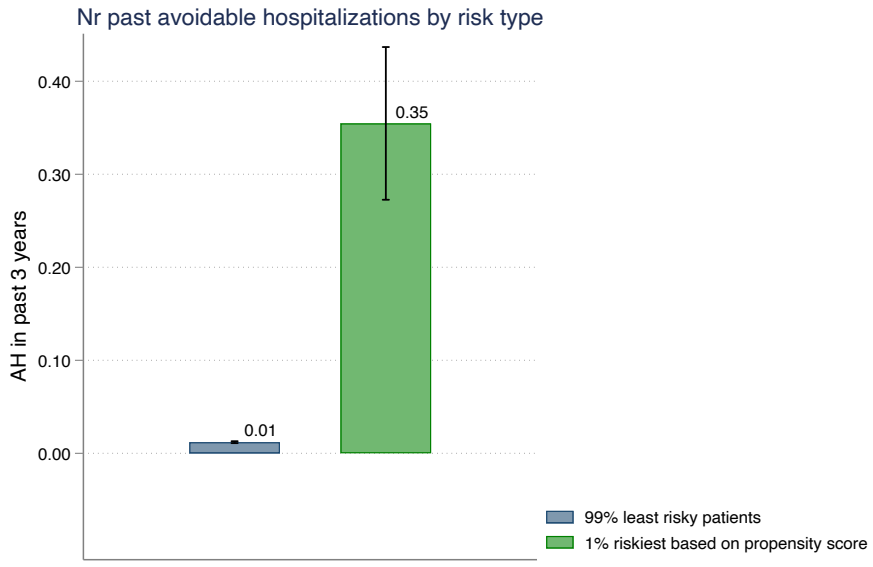
The simplest reallocation procedure I carry out uses the match function with a binary definition of doctor and patient types. Here, I reallocate the top 10% doctors randomly to top 1% high-risk patients and let them swap doctors with some non-risky patients. Costs of reallocations are small in the digital setting compared to the physical setting

Table 2.7: Nr. avoidable hosp. 3 years before consultation

	(1)
Disease index	0.0684***
	(0.0069)
Female	-0.0027*
	(0.0014)
Age	0.0001
	(0.0001)
2nd gen immigrant	0.0013
	(0.0021)
1st gen immigrant	0.0047*
	(0.0028)
Nr hosp 3 years before excl. AH	0.0049**
	(0.0020)
_cons	-0.0005
	(0.0018)
N	95816

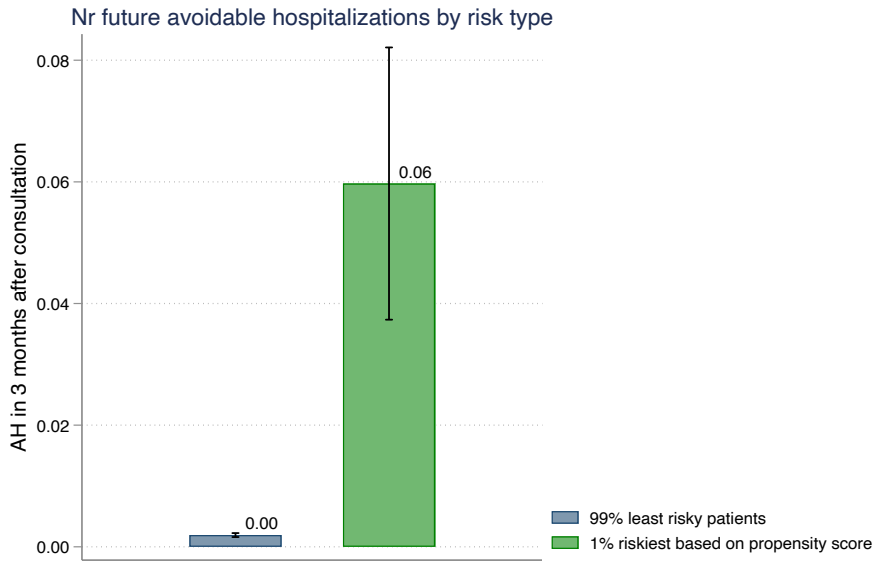
This is used to create patient propensity scores for AH risk.
Robust SEs in parantheses.
Main sample: patients who had visits after doctors' 600th visit.
Patients born before 2013, to allow 3 years pre-data.
Disease index is sum of Elixhauser comorbidities.

Figure 2.2: High and low risk patients: past 3 years' Avoidable Hospitalizations.



95% confidence intervals in black.

Figure 2.3: High and low risk patients: future 3 months' Avoidable Hospitalizations.



95% confidence intervals in black.

where geographic distances play a big role. One cost in the digital setting could be longer waiting time for patients to get the reallocated doctor. These costs are small as we are only reallocating 2% of consultations (= the top 1% risky patients and the patients they swap doctor with) in the reallocation mentioned above. Moreover, among these 2%, 55% of patients can be reallocated to a doctor within the same time shift, meaning there is a negligible additional time cost for them. Hence, any additional waiting from the reallocation procedure would occur only for 0.9% of patients, and only half of them are high-risk patients.[59]

Reallocation based on the continuous AMF to reduce counter-guideline prescriptions

The reallocation procedure for the continuous version of the match function, where I use continuous measures of patient risk and doctor skill, is positive assortative matching (PAM): allocate the highest effectiveness doctors to the highest need/risk patients. This builds on the standard Becker (1973) results that the outcome-maximizing allocation when there are complementarities is assortative.

Informational requirements

The informational requirements to carry out these reallocations consists in having access to patients' past healthcare records. This can be compared to earlier research showing that electronic medical records reduce deaths by making information accessible (Miller and

[59]Since I have defined high-skilled doctors as the top 10% in avoidable hospitalization reduction effectiveness, and depending on the year, day and time there are 10-30 doctors working at each shift, it is highly likely that if a high-risk patient cannot be allocated to a top 10% doctor in the same shift where she originally met a doctor, they would have to wait maximum 3 hours to the next shift.

Tucker 2011). Specifically, for the avoidable hospitalizations reallocation, data is needed on the past three years' avoidable hospitalizations as well as the age and gender of the patient. Demographic data about patients is readily available to the healthcare provider, while data on past avoidable hospitalizations can be accessed in theory, if the electronic medical records are built to flag these events.

The data needed on patients for the reallocation reducing counter-guideline prescription is data on their past three years' antibiotics prescriptions as a share out of total prescriptions. This data also exists in patients prescription histories which is part of their electronic medical records. The data needed on doctors is data on their first 600 patients' outcomes and histories. In the case of counter-guideline prescriptions, the outcomes data already exists within the medical provider as the diagnosis and prescription drug are recorded and can be used to determine guideline adherence. For avoidable hospitalizations, three months' follow up hospitalization data is needed for the doctor's first 600 randomized patients, and this could be achieved by an integration of medical records where only patients who have avoidable hospitalizations are flagged and reported back to the digital healthcare provider. Such follow up data would be useful even in the absence of a reallocation objective. Currently, the ownership of the means of prediction remains with the governmental agencies that host patient data, as well as with the providers that produce the data.

2.6 Results

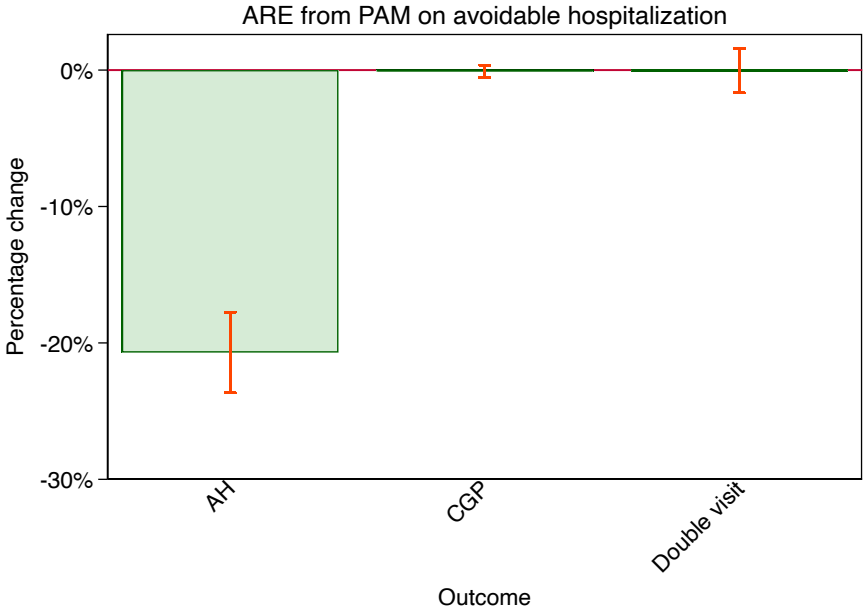
2.6.1 Reallocation results

The first part of the results covers counterfactual simulations: the Average Reallocation Effects (ARE). The following section presents results on what drives these effects in terms of causal match effects and stylized facts about skills. Finally, we study the healthcare production more in detail to make clear the mechanisms in terms of doctor actions.

The first set of Average Reallocation Effects are derived from the optimization problem in Section 4.1. This problem takes existing resources in terms of doctor skills and time worked as given, as it might be difficult if not impossible to increase all doctors' skills at several different tasks[60], and there are constraints to hiring new doctors. Moreover, retraining (and placing emphasis) on some skills may lead other skills to suffer in a multitasking setting. I evaluate the aggregate effects of reallocating doctors between patients with

[60] Although I will also consider a selective hiring policy, which creates smaller benefits than matching.

Figure 2.4: Average Reallocation Effects from minimising Avoidable Hospitalizations



These results come from the binary match function where a good doctor is defined as a top 10% in the Avoidable Hospitalization outcome, and a risky patient as the top 1% risky for the same outcome.

different needs, subject to the above-mentioned constraint. I consider reallocating doctors according to patients' risk for each outcome variable, as described above. I will focus here on reallocations to reduce the adverse outcome avoidable hospitalizations – other reallocations can be found in the Appendix.[61]

The first result (Figure 2.4) is that avoidable hospitalizations (AH) decrease by 20% when I have matched doctors and patients on doctor AH-prevention skill (skill in risk prediction/triaging) and patient AH risk as described in Section 5.5. At the same time, the aggregate number of counter-guideline prescriptions and double visits[62] do not change. Hence, the positive outcome (reducing AH) has been achieved without increasing other negative outcomes. There are also effects on healthcare inequality. Before reallocation, the probability of meeting a top 10% doctor in risk prediction/triaging was similar across patients' income distribution (Figure 2.5). After the reallocation, the chance of meeting a top 10% doctor in risk prediction/triaging increases for the bottom decile, without any large change for the rest of the income deciles (Figure 2.5). This is because the risk for avoidable hospitalizations is highest in the lowest income decile.

[61] Table 2a.13 in the Appendix shows that reallocating the doctors who are best at following antibiotics guidelines to patients who are intensive users of antibiotics reduces counter-guideline prescriptions by 10%, potentially contributing to the global battle against bacteria becoming resistant to antibiotics through externalities from over-prescription.

[62] Double visit means that the patient contacted a nurse in physical care the week after the digital visit

Figure 2.6 presents another way of understanding the income-health gradient aspect of doctor-patient matching. This figure shows Average Reallocation Effects from a reallocation where the highest-skilled doctors in reducing avoidable hospitalizations are matched with the highest-income patients (positive assortative matching (PAM) on doctor skill and patient income using the continuous match function[63]). This reallocation is compared to the random real-life digital assignment, and shows that avoidable hospitalizations would be around 5% higher if the highest income patients were matched with the highest skilled doctors in preventing avoidable hospitalizations[64].

These results can be interpreted in light of the results from the descriptive analysis earlier in this paper about a positive relationship between patient area-level income and perceived quality of local primary care, as well as results from other studies which indicate that higher-income patients get access to better doctors in physical care (e.g., Agency for Healthcare Research and Quality 2020). If this also applies to risk detection and triage skill for physical care doctors, Figure 2.6 suggests that avoidable hospitalizations after physical primary care could be lowered by up to 5% if patient-doctor matching changed to a random matching. Moreover, if we add together the results from Figures 2.4 and 2.6, they suggest that moving to an needs-based allocation on avoidable hospitalizations from a positive assortative matching on patient income and doctor skill could reduce the number of avoidable hospitalizations by 25%.

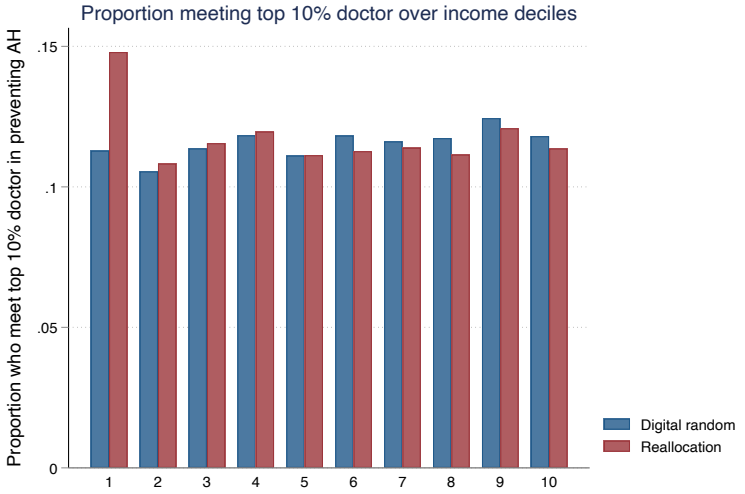
The gains from matching are much larger than the gains from a more selective doctor hiring policy, which I simulate by increasing the work hours of the doctors who are have above median skill ("good") in all three outcome measures (preventing avoidable hospitalizations, following antibiotics guidelines, and reducing double visits in digital and physical primary care in the same week), and commensurately reducing the hours of the remaining doctors. However, Figure 2.7 illustrates that even if we improved doctor selection by increasing these doctors' work hours by as much as 70%, we still would see no significant improvement in aggregate avoidable hospitalizations, and only a 4% reduction in counter-guideline prescriptions (less than half of the reduction from matching doctors and patients to reduce counter-guideline prescriptions). Moreover, a 70% increase in these doctors' work hours would hard to achieve, even if digitalization can be expected to give room for some increase in hours for the best doctors.[65].

[63]The continuous match function means that I use each doctor's estimated effect from the hold-out sample, and each patient's estimated risk, instead of the binary dimensionality reduction of high/low types that I use in the binary match function

[64]The figure also shows that counter-guideline prescriptions would remain unchanged compared to the random allocation, which is expected given the zero correlation in those skills within doctors.

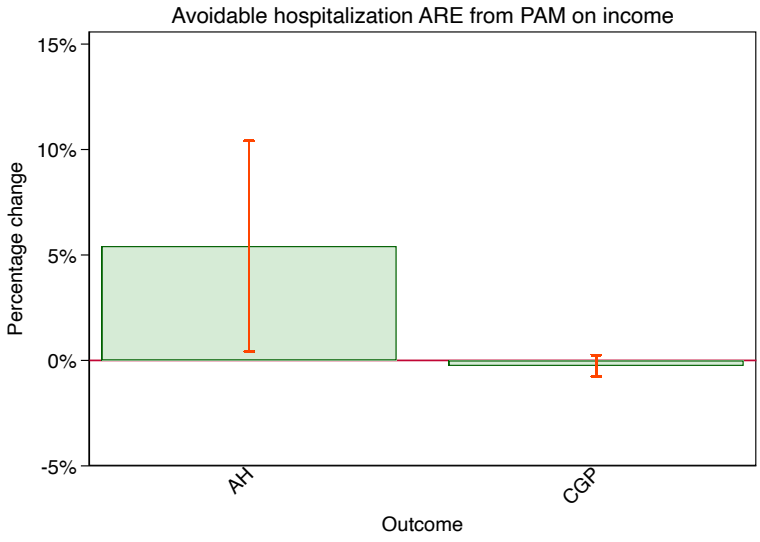
[65]For instance, if digital care saves commuting time for the doctors, we could imagine increasing the "good" doctors' working hours by 10-20%, but not by 70%. An average round-trip commute in Sweden is

Figure 2.5: Redistributional effects of the AH-minimising reallocation



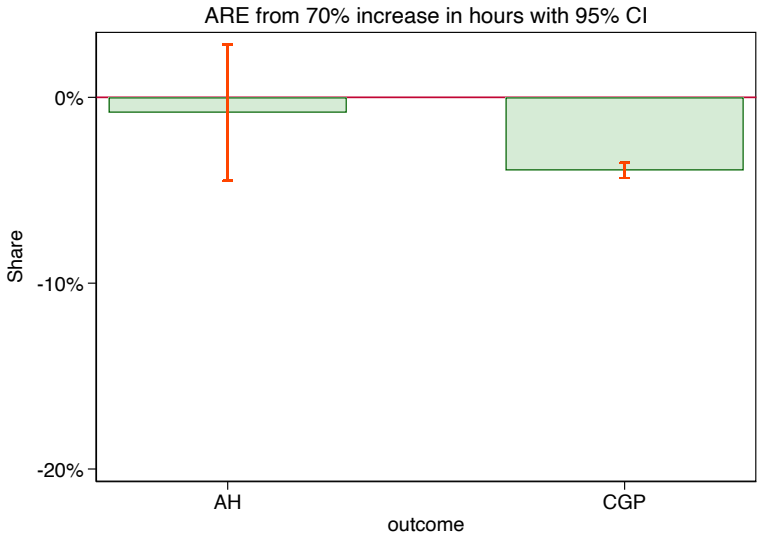
This figure shows what proportion of patients across income deciles who meet a doctor who is classified as top 10% in reducing avoidable hospitalizations, in the reallocation to reduce avoidable hospitalizations as well as in the random allocation. All income deciles have a slightly higher than 10% proportion of top doctors, which is because the top doctors work more consultations than other doctors. The patient income is the income of adult patients in 2017.

Figure 2.6: Reallocation effects from matching higher income patients with more AH-skilled doctors



Average Reallocation Effects from Positive Assortative Matching (PAM) on doctor skill in reducing avoidable hospitalizations, and patient income, using the continuous match function.

Figure 2.7: Alternative policy of selective doctor hiring



Average Reallocation Effects from an increase in work hours by 70% for the doctors who are above median in all three measures (preventing avoidable hospitalizations, double visits and counter-guideline prescriptions) without any matching to patients, using the continuous match function.

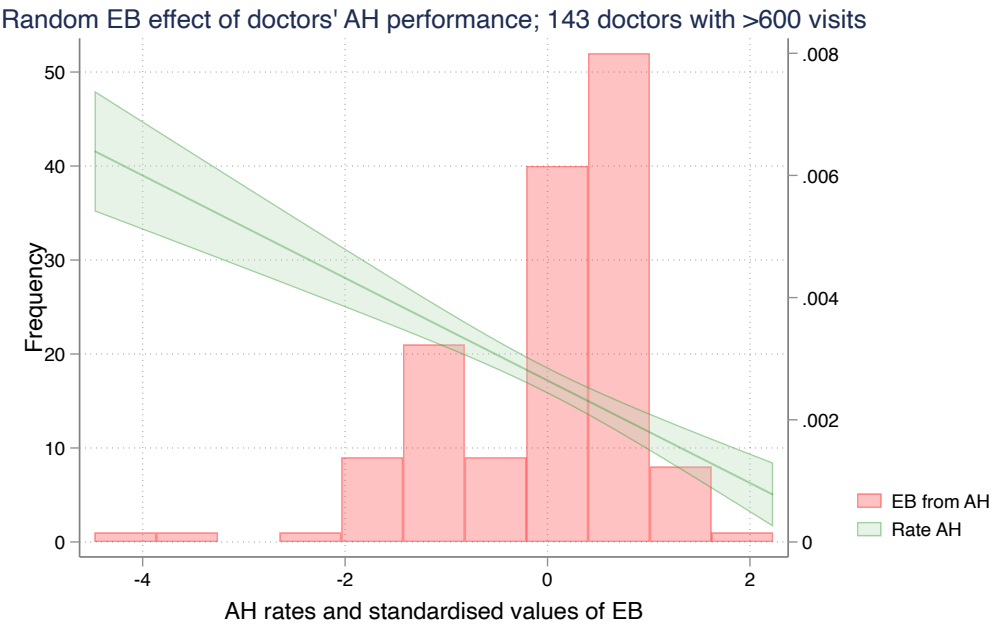
around 40 minutes and doctors work shorter shifts than 8 hours.

2.6.2 What drives the gains from matching?

Variation in doctor effectiveness within task

The first driver behind the gains from reallocation is that there is variation in doctor effectiveness in each task. Figure 2.8 shows that the share of a doctor's patients who end up having an avoidable hospitalization within 3 months after the consultation ranges from virtually 0% to 0.6% .

Figure 2.8: Variation in doctor performance in preventing avoidable hospitalizations



Histogram of the standardized (over doctors) Empirical Bayes (EB) estimates of doctor quality with the outcome *negative* avoidable hospitalizations (AH) in the 3 months after the visit (in red). Included are the 143 doctors with more than 600 randomized first visit consultations. Overlaid in green is the predicted rate of AH out of the doctor's total consultations, from a regression of the rate on the EB random effect, with a 95% confidence interval.

No positive correlation in effectiveness across tasks: specialization

If some doctors are better at all tasks, then reallocation would be more difficult as the planner would need to prioritize more between different patients who have needs for different doctor skills. However, Figure 2.9 and Table 2.8 show that there is no positive correlation between doctors' effectiveness in different tasks. In fact, that there is a negative within-doctor relationship between certain skills. For instance, a doctor who is better at following antibiotics guidelines is slightly worse at preventing double visits (when the patient seeks physical nurse care the week after the digital doctor appointment) (see also Appendix

Figure 2.9: Scatterplots of different doctor skills.

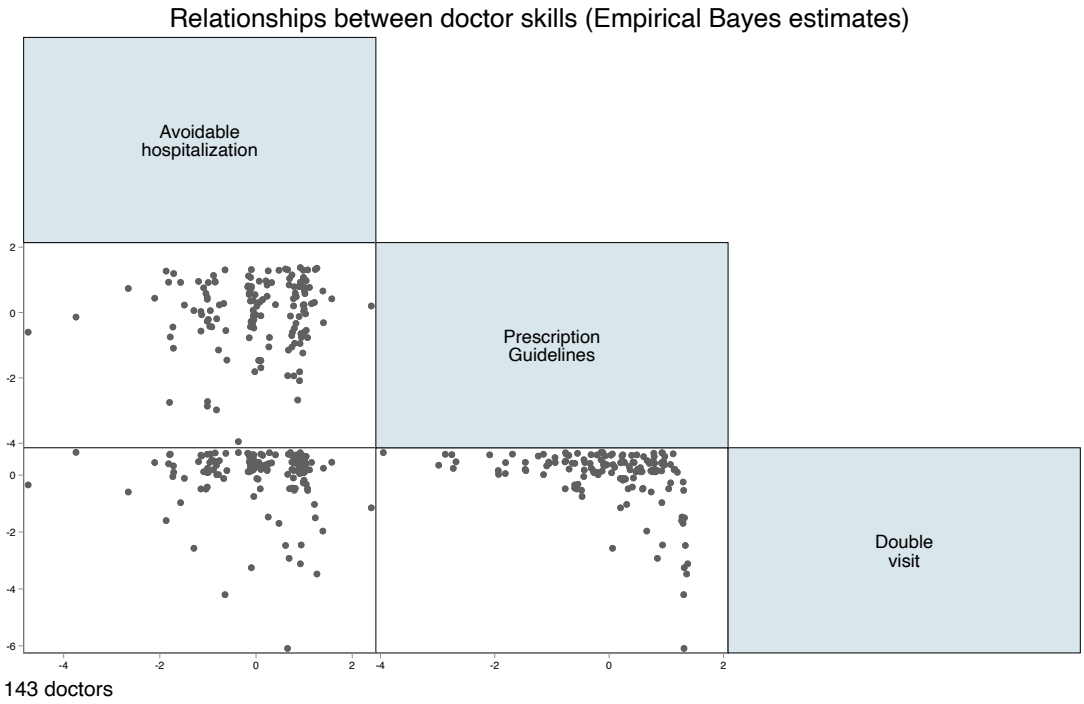


Table 2.8: Spearman's rank-order correlation coefficient between doctor skills

Doctor skill	AH	CGP
CGP	0.0655	
	(0.4368)	
Double visit	-0.0861	-0.3528
	(0.3068)	(0.0000)

In parantheses: **p-value** from test of H^0: the two effectiveness measures are independent. N= 143.

Figure 2a.12). This can be conceptualized as specialization. It may also be related to patient behavior. Some patients may particularly want an antibiotic. If they do not get it from the digital doctor, because the doctor adheres to guidelines, then they might be more likely to call the nurse at the physical healthcare clinic the following week, to try to get antibiotics from there. But even in this case, it reflects a different balance struck by the doctor in the trade-off between following guidelines and satisfying the patient.

2.6.3 Match Effects

The final driver of the reallocation effects is evidence of strong complementarities or "match effects": causal treatment effects of matching doctors of higher effectiveness in outcome k

to patients with higher estimated need/risk in outcome k.[66] A doctor who is among the top 10% at reducing avoidable hospitalizations (AH) in the hold-out sample reduces AH by as much as 90% for the top 1% risky patients in the main sample, but has no effect on the rest of patients (Table 2a.12). These complementarities in patient-doctor matching are illustrated graphically in Figure 2.10. Table 2a.8 in the Appendix shows results from the continuous version of the match regression, and includes robustness checks.

Table 2.9: Number Avoidable Hospitalizations within 3 mo. after visit

	(1) Clustered SEs	(2) Bootstrapped SEs
Top 10% doctor X top 1% risky patient	-0.060***	-0.060***
	(0.014)	(0.016)
Top 10% doctor on AH	0.000	0.000
	(0.001)	(0.001)
Riskiest 1% patient in AH	0.067***	0.067***
	(0.012)	(0.014)
N	95816	95816
Mean	0.003	0.003
Mean_risky	0.062	0.062

All columns have date-time shift fixed effects. Sample is all doctors' randomized visits after the 600th' consultation

Table 2.10: Definitive counter guideline prescription

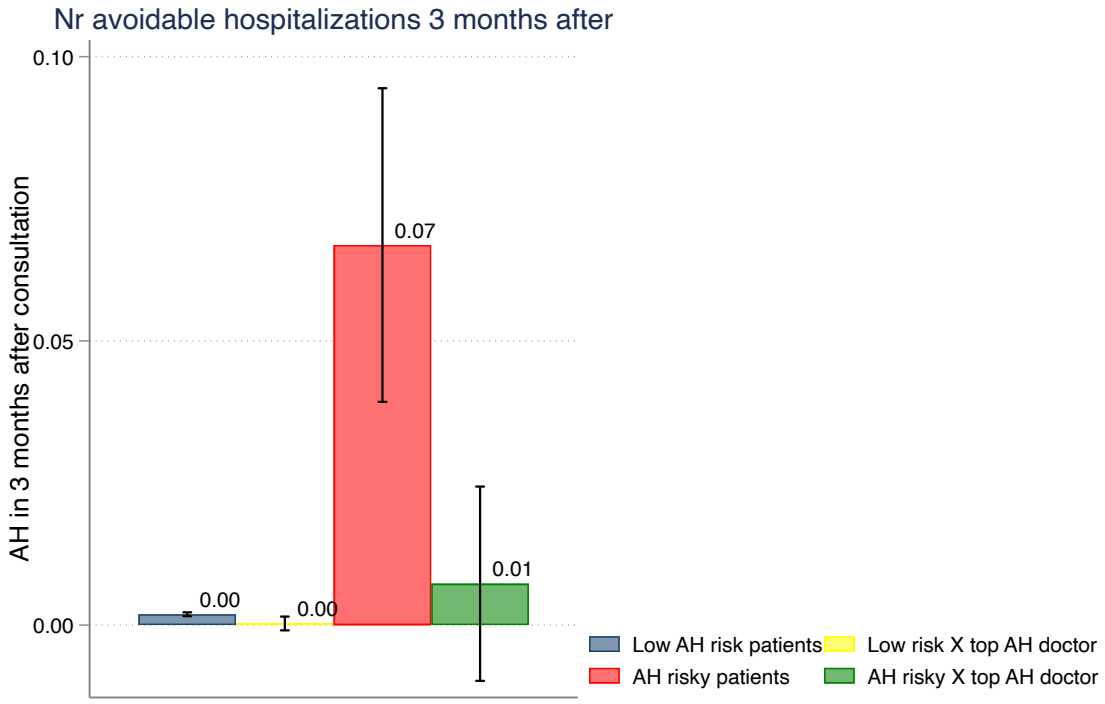
	(1) OLS Simple	(2) Controls	(3) Bootstrap	(4) Time shift FE
Std doctor FE based on no CGP	-0.0028***	-0.0028***	-0.0028***	-0.0028***
	(0.0010)	(0.0010)	(0.0011)	(0.0008)
Share antib 3yrs b4	0.0489***	0.0529***	0.0529***	0.0526***
	(0.0031)	(0.0034)	(0.0033)	(0.0033)
Std doc FE X share antib 3yrs b4	-0.0131***	-0.0131***	-0.0131***	-0.0126***
	(0.0020)	(0.0020)	(0.0022)	(0.0020)
Controls	No	Yes	Yes	Yes
N	116396	116391	116391	116391
R^2	0.008	0.009	0.009	0.009
Mean	0.0172	0.0172	0.0172	0.0172

All columns: SEs in parantheses clustered on doctors. Col 3 has booststrapped SEs.
Sample is dropin first visits after doctor's 600th such visit. Only patients born before 2013.
Controls are Elixhauser sum of comorbidities, female, age, first- and second-generation immigrant.

While the targeting of patients who are at risk for avoidable hospitalizations may be most important, there are also effects of matching patients on who have had a higher

[66] This is not ex ante evident - it could have been that high risk patients are simply not possible to help from the bad outcome, and that it would be best to allocate the most effective doctors to patients who had less risk and were more amenable to change.

Figure 2.10: Match effects on avoidable hospitalizations.



Bootstrap 95% confidence intervals in black.

share of antibiotics prescriptions in the past to doctors better at following antibiotics guidelines.[67] A patient who had a 50% higher share of antibiotics out of their total pre-digital care prescriptions has 2.4%-2.6% higher risk of receiving a counter-guideline prescription, suggesting that the patient may want or need more antibiotics (Table 2.10).[68] Ex ante, it is not obvious that a doctor who has had a good track record in the holdout sample of not prescribing a counter-guideline prescription (CGP), would also be more restrictive with antibiotics when meeting a patient in the main sample who has a higher share of antibiotics in the past[69]. It could be the case that such a patient needs more antibiotics and any doctor would be willing to surpass the guidelines with such a patient.

However, it turns out that if a patient who has a 50% higher share of antibiotics out of their total pre-digital care prescriptions is matched with a doctor who is one standard deviation better in the hold-out sample at following guidelines, their risk of getting a CGP is reduced by 24-27%.[70]

[67] In this regression, I have not reduced the dimensionality of doctor and patients types to binary. Instead, the regression specification has the continuous doctor skill and patient risk and their interaction

[68] Either that the patient is particularly fragile so that any doctor would prescribe a little more over cautiously for them - but I am controlling for age, gender and Elixhauser sum of comorbidities in Columns 2-4 which controls for their disease level. Otherwise it suggests that the patient is particularly keen on antibiotics, and potentially tries to pressure the doctor to get them.)

[69] We do not know if the antibiotics in physical healthcare in the patient's history were according to guidelines or not.

[70] Table 2a.9 in the appendix uses the number of antibiotics instead of the share for the patient risk variable, and the results are similar.

2.6.4 Mechanisms for preventing avoidable hospitalizations

To shed light on the mechanisms though which some doctors are particularly effective at preventing avoidable hospitalizations, I study the actions that the doctors take during the digital care consultation. Table 2.11 shows the most common outcomes (in terms of doctor actions) of a consultation, together capturing 98% of the consultations' outcomes. These outcomes are prescription, advice only, redirection, referral, and sick note. To redirect a patient means to tell them that their condition is not suitable for digital primary care, and that they should go to, e.g., a physical primary care center, possibly one with extended opening hours, or in some cases the Emergency Department.[71] The main takeaway from Table 2.11 is that doctors who are among the top 10% at preventing avoidable hospitalizations (AH) are more likely to identify that the AH-risky patients need other care than digital and redirect them. At the same time, they are less likely than other doctors to only give advice to these patients. There are no significant differences in how the top 10% doctors at AH treat other patients than the risky – meaning that it is not the case that these doctors are simply more cautious and avoid false negatives at the expense of increasing false positives. False negatives in this case would be that patients who need additional checkups in physical care are not redirected, while false positives would be that patients who do not need additional checkups in physical care are redirected for these checkups, which would be costly.

These results indicate that the AH-skilled doctors are better at identifying the patients at high risk and determining that they (and not other patients) need more care (triaging), which can possibly prevent an avoidable hospitalization. Triaging is one of the key components of a primary care physician's job and can make the difference between appropriate, cost-effective care and poor outcomes at high cost (Vasilik 2021). Triaging is difficult and requires separating the few urgent patients from the many non-urgent patients. The medical literature indicates that while triage handbooks exist, they may be difficult to use in practice and there are no explicit guidelines at many primary care centers (Vasilik 2021). Hence, the triaging process requires experience and knowledge within several fields of medicine (Göransson et al. 2021).

We have seen that the doctors who prevent more avoidable hospitalizations for risky patients do not do this at the expense of redirecting a higher share of non-risky patients

[71] A referral, on the other hand, means that the doctor writes a letter to a specialist clinic and the patient will in due course be called by the clinic. This can take weeks or months depending on the condition and wait list. In our data, the share of consultations ending in referrals from the primary care provider to a specialist clinic (3% of consultations) are comparable to the lower end of GP referrals in the UK physical setting (where in a meta-analysis, they range between 1.5% and 24.5% (O'Donnell 2000)).

for additional care. But are there other downsides to these doctors' work, potentially that they spend longer time with the patients, thus decreasing the time available for other patients? Column 1 of Table 2.12 shows that the consultation duration is no different when an AH-risky patient meets a top 10% doctor in preventing AH. Column 2 of Table 2.12 also shows that there are no significant differences in the administration time - the time that the doctor spends after the consultation on writing notes and prescriptions, etc.

A final question which bears on future possible strategic incentives and mechanism design, is whether patients recognize which doctors are most appropriate for their needs. Column 3 of Table 2.12 shows that patients in general are more satisfied with the top 10% doctors in AH prevention. However, patients who are at risk for avoidable hospitalizations are not differentially more satisfied with these doctors.

Table 2.11: Doctor actions during digital visit

	(1) Redirected	(2) Advice only	(3) Prescription	(4) Referral	(5) Sick note
Top 10% doctor on AH	0.00	-0.00	-0.00	0.01	0.00
	(0.02)	(0.02)	(0.02)	(0.01)	(0.00)
Riskiest 1% patient in AH	0.07***	0.00	-0.09***	-0.00	-0.01
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Top 10% doc.X 1% riskiest	0.11**	-0.09**	-0.03	0.00	-0.01
	(0.06)	(0.04)	(0.07)	(0.01)	(0.01)
N	91519	91519	91519	91519	91519
Mean	0.12	0.26	0.53	0.03	0.04

Date time shift FE included. SEs in parantheses clustered on doctors.

Table 2.12: Process outcomes during digital visit

	(1) Duration, mins	(2) Admin time, mins	(3) Score, 1-5
Top 10% doctor on AH	-0.0693	-0.2043	0.0812**
	(0.3180)	(0.9744)	(0.0385)
Riskiest 1% patient in AH	0.1956	0.1098	-0.1567***
	(0.1263)	(0.3186)	(0.0389)
Top 10% doctorX 1% riskiest patient	-0.0942	0.8547	-0.0452
	(0.1613)	(1.0113)	(0.1381)
N	93869	93868	70607
Mean	4.5226	11.7034	4.6331

Date time shift FE included. SEs in parantheses clustered on doctors.
In column 1 the outcome variable is patient-doctor consultation duration; in column 2 it is the doctor's administration time after the meeting, spent on e.g. issuing prescriptions and writing notes; and in column 3 it is the patient's satisfaction rating of the doctor, ranging between 1 and 5.

2.7 Conclusion

The digitalization of services, such as parts of healthcare and education, has many implications, three of which are especially important for the topic of this paper. First, new possibilities for matching service providers and users emerge as the number of potential providers that any given user could meet has increased. Video consultations mean that the constraint of doctors and patients sharing the same geographic location is less binding. Second, the digitalization of services results in detailed data about each agent's work and outcomes, which can be used to measure providers' task-specific skills and users' needs and used in algorithmic matching. Finally, this technology disrupts geographically related sorting patterns between service providers and receivers, which often resulted in inequality in service quality. Together, this creates new opportunities to transform the organization of service production by changing the matching between service providers and users to make better use of provider skills. This is a process that has already started with digital healthcare firms using algorithms to match users and patients, so we need to better understand this development.

Within healthcare, many countries face pressing cost challenges and human capital shortages. At the same time, digital technology and video consultations have grown fast in healthcare, just as in other sectors. I quantify the potential gains from a better allocation of scarce resources: doctor skills, by first using detailed data about both patients and doctors to classify types. Such a reallocation procedure is potentially cost-neutral, as opposed to training and hiring additional doctors, which is costly. Moreover, I show that the gains from a selective hiring policy are not nearly as large as those from a matching policy.

The matching gains are driven by another fact that I establish: that there is heterogeneity in skill among doctors in dimensions that vary in importance for heterogeneous patients, even within general practice which is studied in this paper. I have shown that physician effectiveness varies considerably in different tasks. The evidence is not consistent with a single latent ability variable governing doctor effectiveness on all the outcome measures, but rather with specialization. Moreover, doctors' effectiveness varies with different patients who have varying pre-existing risk relevant for the different doctor tasks. If we match a doctor who is among the top 10% at reducing the main outcome *avoidable hospitalizations*, with a patient who is predicted to be among the top 1% risky for such adverse outcomes, we could reduce their number of such adverse outcomes by 90%. However, we need to take these doctors from other patients who may themselves also have a small risk for the adverse outcome.

To understand the trade-off between the positive and the negative effects from this reallocation, I calculate the aggregate effects of reallocating doctors. Reallocating the doctors who are best at preventing avoidable hospitalizations (AH) to the patients at risk reduces AH by 20% without making other main outcomes worse, and by reallocating only 2% of patients. A back-of-the-envelope calculation shows that an AH reduction of 20% scaled up nationally could hypothetically save up to 2% of total hospital costs in Sweden (USD 160 million in Sweden)[72] and the US (USD 6.7 billion in the US for only adults in purely hospital costs), apart from lives saved.[73] Moreover, reallocating the doctors who are best at following antibiotics guidelines to patients who are intensive users of antibiotics reduces counter-guideline prescriptions by 10%, potentially contributing to the global battle against bacteria becoming resistant to antibiotics through externalities from over-prescription.

A main take-away is that in primary care, doctor heterogeneity in skill and patients' varying needs matter: there are gains to be made from a doctor-patient reallocation where provider specialized skills are put to better use. It is highly likely that this could also be the case in other service sectors, and when services move to digital, this is becoming a feasible and resource-neutral, low-cost way of increasing effectiveness of service provision. Moreover, this could have effects on inequality in access to high-quality services across the income distribution.

[72] Calculated from an estimate of the total costs of avoidable hospitalizations: 820 million USD per year in Sweden. The number of hospital days for AH was around 1 million in Sweden in 2010 (Socialstyrelsen, 2011, p.51). The average cost per day in inpatient care is 7100 SEK (Socialstyrelsen, 2017). The US figure on the total costs of avoidable hospitalizations is USD 33.7 for adults only (Jiang et al 2009; McDermott and Jiang 2020)

[73] In both countries, this saving represents around 0.03% of GDP.

2.8 Appendix I: Additional Tables and Figures

Table 2a.1: Physical Primary Care Clinic Scores, standardized

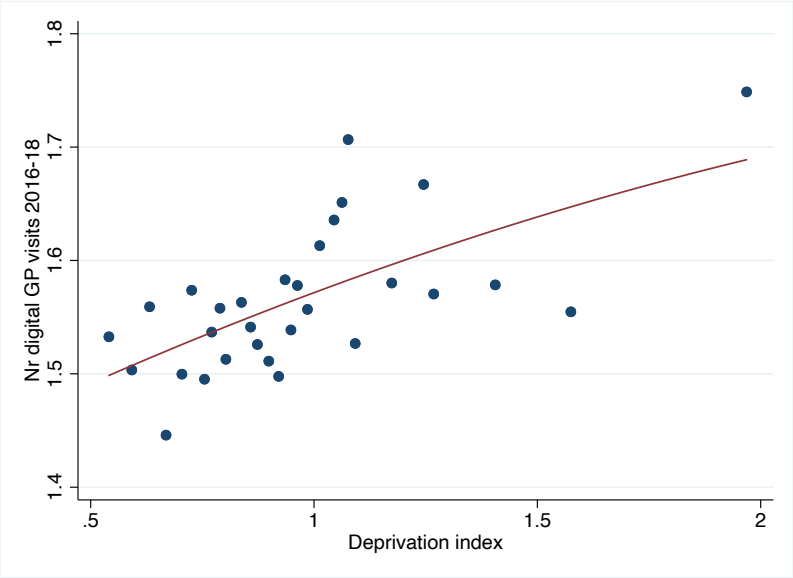
	(1)	(2)	(3)	(4)	(5)
Std(Foreign)	-0.24***		-0.25***	-0.24***	-0.24***
	(0.033)		(0.037)	(0.049)	(0.049)
Std(Income)		0.18***	0.14***	0.14***	0.15***
		(0.031)	(0.030)	(0.030)	(0.041)
Avg. age				-0.00	-0.00
				(0.032)	(0.032)
Gender				-1.55	-1.46
				(2.465)	(2.501)
Std(foreign)XStd(Income)					-0.02
					(0.028)
Region FE					
Robust SE					
N	1298	943	943	943	943

Robust standard errors in parantheses. The unit of observation is a 4-digit postcode matched with municipality. Region fixed effects are included. Std(Foreign) measures the standardised share of foreign-born inhabitants in the area. Std(Income) measures the standardised mean income in the area. The outcome variable comes from the National Patient Survey (NPE).

Table 2a.2: Descriptive statistics of all doctors

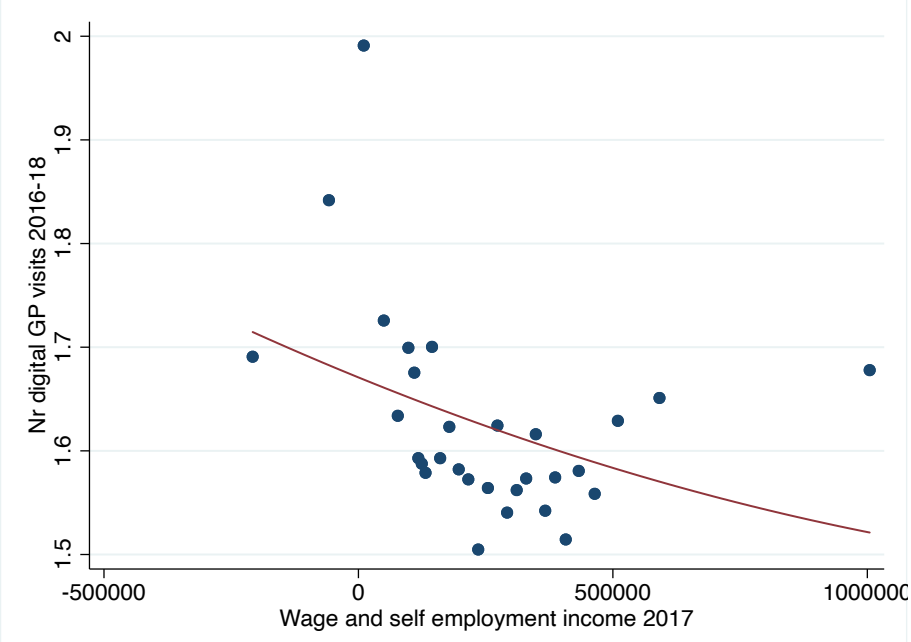
	(1)				
	mean	sd	min	max	count
Specialist	0.36	0.48	0	1	619
In specialty training	0.30	0.46	0	1	619
MD + residency only	0.33	0.47	0	1	619
Speaks non EU15 language	0.21	0.41	0	1	693
GP specialist	0.35	0.48	0	1	619
Age	41.2	10.9	25	80	262
Female	0.41	0.49	0	1	262
Employed rather than contractor	0.43	0.50	0	1	131
Observations	693				

Figure 2a.1: Number of digital visits by negative socioeconomic status



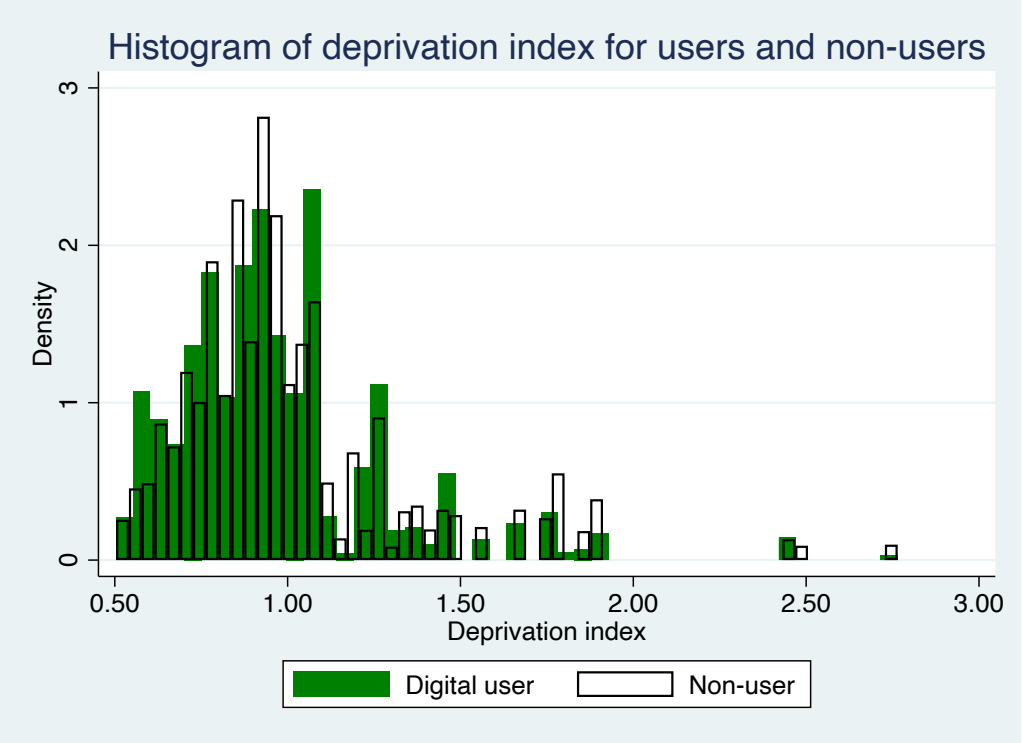
Binned scatterplot of number of digital GP visits in 2016-18 (individual level), controlling for age, against index of social deprivation (Care Need Index) of the individual's GP clinic in Scania.

Figure 2a.2: Number of digital visits by income



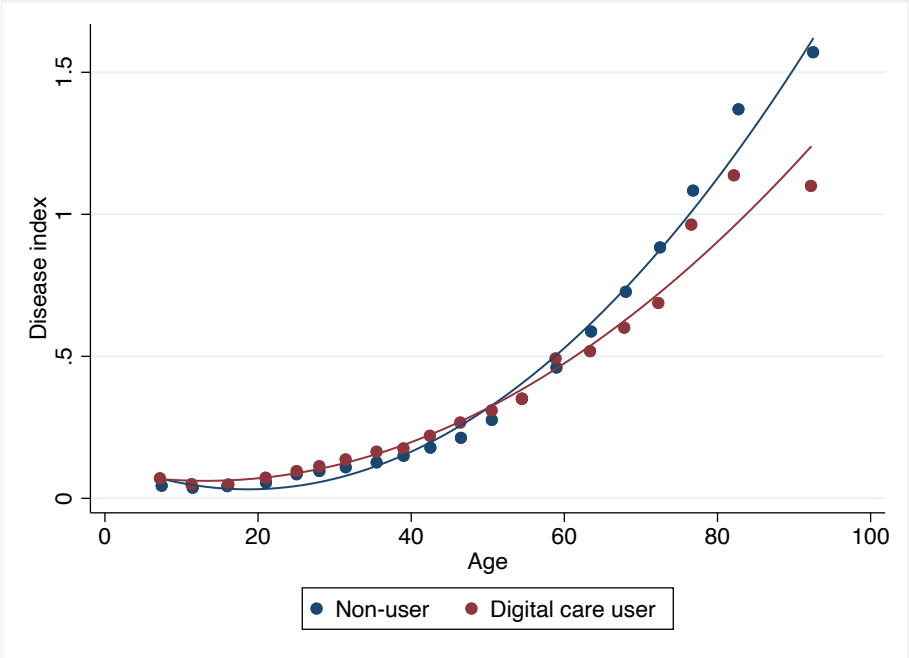
Binned scatterplot of number of digital GP visits (individual level) in 2016-18 against individual total income 2017, controlling for age.

Figure 2a.3: Selection into digital care by socioeconomic status



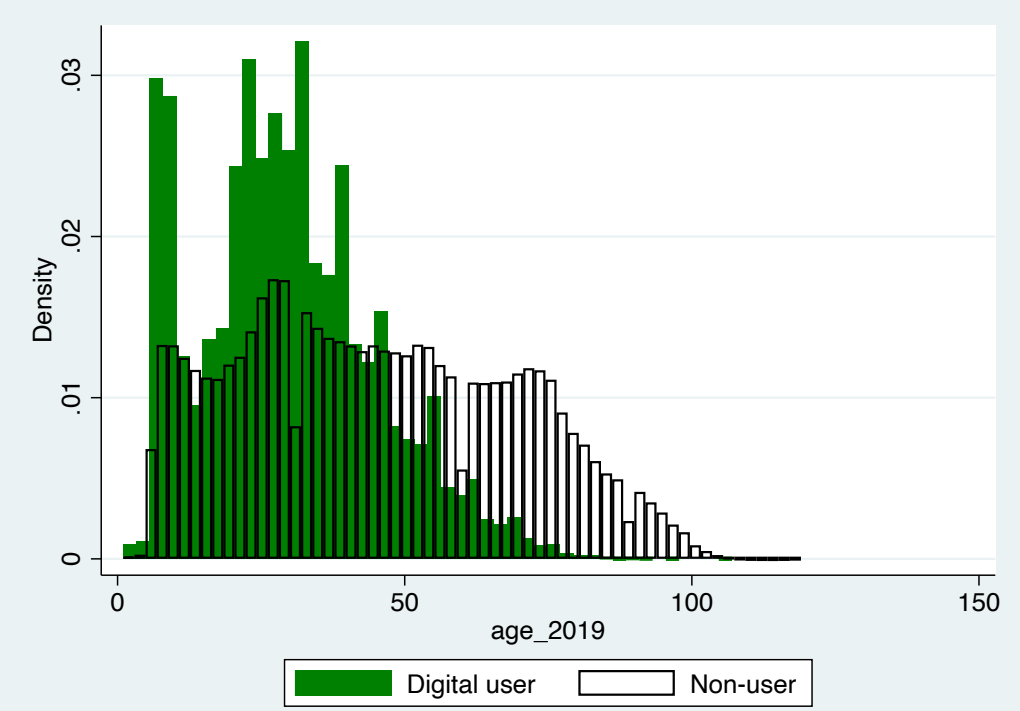
Notes: Data from Region Skåne. Deprivation index is the Region's *Care Need Index*. It is a weighted average of the variables (1) over 65 years old and in a single household (2) Born outside EU (3) Unemployed 16-64 year old (4) Single parent with child under 18 years old (5) Person over 1 years old who has moved into the area (6) low education 25-64 years old (7) Age below 5 years old.

Figure 2a.4: Relationship between age and number of chronic diseases for digital care users vs. non-users



Notes: Elixhauser's comorbidity index using data from 2013-15 in Scania.

Figure 2a.5: Selection of different age groups into digital care



Notes:

Data from Region Skåne.

Table 2a.3: Comparison of doctors included in the final analysis and those who are not.

Variable	(1) Not included	(2) Included	(3) Difference
Total number of first and revisit consultations	323.9	3346.0	3022.2***
	(303.8)	(2584.7)	(138.9)
Seniority	1.1	1.0	-0.1
	(0.8)	(0.8)	(0.1)
Specialty	2.9	2.1	-0.8*
	(4.6)	(3.8)	(0.4)
Speaks non EU15 language	0.2	0.4	0.1***
	(0.4)	(0.5)	(0.0)
Average admin duration	13.4	11.3	-2.1***
	(4.3)	(2.3)	(0.4)
Average consultation duration	6.1	5.0	-1.2***
	(1.9)	(1.2)	(0.2)
Observations	357	143	500

Table 2a.4: Summary statistics of different groups of doctors

Variable	(1) Not included	(2) Included	(3) Difference
Total number of first and revisit consultations	112.0	1188.2	1076.1***
	(489.3)	(1958.5)	(142.0)
Seniority	1.0	1.0	0.1
	(1.0)	(0.8)	(0.1)
Specialty	2.9	2.7	-0.2
	(4.4)	(4.4)	(0.5)
Speaks non EU15 language	0.0	0.3	0.2***
	(0.2)	(0.4)	(0.0)
Average admin duration	20.2	12.8	-7.4***
	(22.1)	(4.0)	(1.1)
Average consultation duration	4.9	5.8	0.9***
	(2.6)	(1.8)	(0.2)
Observations	195	500	780

Notes: This table shows summary statistics of (Column 2:) the doctors who are (a) not pediatricians (who have a different assignment protocol to patients) (b) who have worked a sufficient number of consultations to merit inclusion in the sample of 500 doctors, compared to (Column 1:) the doctors who are either pediatricians or have worked very few consultations and are thus not included in any sample.

Table 2a.5: Socioeconomic status correlates of having had a previous avoidable hospitalization

	(1) No AH	(2) Past AH	(3) Difference
below median income	0.452	0.619	0.167***
	(0.498)	(0.486)	(0.009)
adult without income	0.064	0.201	0.138***
	(0.244)	(0.401)	(0.005)
age	36.491	40.461	3.970***
	(12.456)	(16.043)	(0.227)
patient female	0.630	0.674	0.044***
	(0.483)	(0.469)	(0.009)
any benefit	0.134	0.305	0.172***
	(0.340)	(0.461)	(0.006)
disability insur	0.013	0.066	0.054***
	(0.111)	(0.249)	(0.002)
housing subsidy	0.039	0.065	0.026***
	(0.193)	(0.247)	(0.004)
employed	0.870	0.771	-0.100***
	(0.336)	(0.420)	(0.006)
self-employed	0.073	0.065	-0.008*
	(0.260)	(0.246)	(0.005)
unemployed (20-67)	0.047	0.127	0.080***
	(0.211)	(0.333)	(0.004)
minority	0.168	0.197	0.029***
	(0.374)	(0.398)	(0.007)
foreign born	0.108	0.135	0.026***
	(0.311)	(0.341)	(0.006)
born outside EU15,Scandi.	0.087	0.111	0.024***
	(0.282)	(0.315)	(0.005)
married	0.343	0.328	-0.015*
	(0.475)	(0.469)	(0.009)
inhabitants per km2 munic.	1,649.564	1,373.369	-276.195***
	(2,062.333)	(1,951.354)	(37.415)
Observations	157,475	3,115	160,590

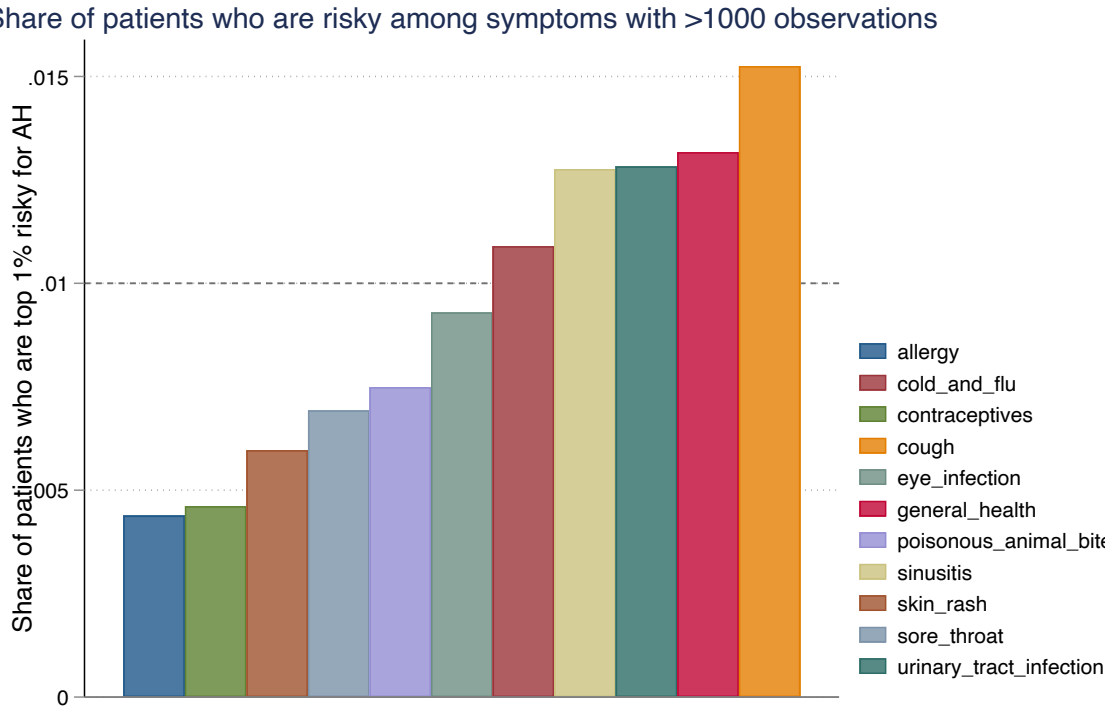
This table shows the difference in socioeconomic covariates (measured in 2017) for patients who have had no previous avoidable hospitalization (AH) in 2013-2016, compared with patients who have had at least one such hospitalization in the period before digital care. The socioeconomic variables do not exist for child patients.

Table 2a.6: Medical correlates of having had a previous avoidable hospitalization

	(1) No AH	(2) Past AH	(3) Difference
hypertension pre 2016	0.008	0.078	0.070***
	(0.089)	(0.269)	(0.002)
asthma pre 2016	0.018	0.056	0.038***
	(0.133)	(0.230)	(0.002)
diabetes pre 2016	0.002	0.083	0.081***
	(0.039)	(0.276)	(0.001)
depression pre 2016	0.025	0.057	0.032***
	(0.157)	(0.232)	(0.003)
anxiety pre 2016	0.040	0.095	0.056***
	(0.195)	(0.294)	(0.004)
hyperactivity pre 2016	0.019	0.041	0.022***
	(0.138)	(0.199)	(0.003)
had any visit pre 2016	0.733	0.937	0.205***
	(0.443)	(0.243)	(0.008)
nr acute visits pre 2016	0.069	0.263	0.193***
	(0.349)	(0.901)	(0.007)
never filled presc. 2013-16	0.096	0.016	-0.080***
	(0.295)	(0.127)	(0.005)
nr presc. filled201316	21.314	88.221	66.907***
	(54.234)	(189.340)	(1.083)
above median presc. 2013-16	0.499	0.834	0.336***
	(0.500)	(0.372)	(0.009)
Observations	157,475	3,115	160,590

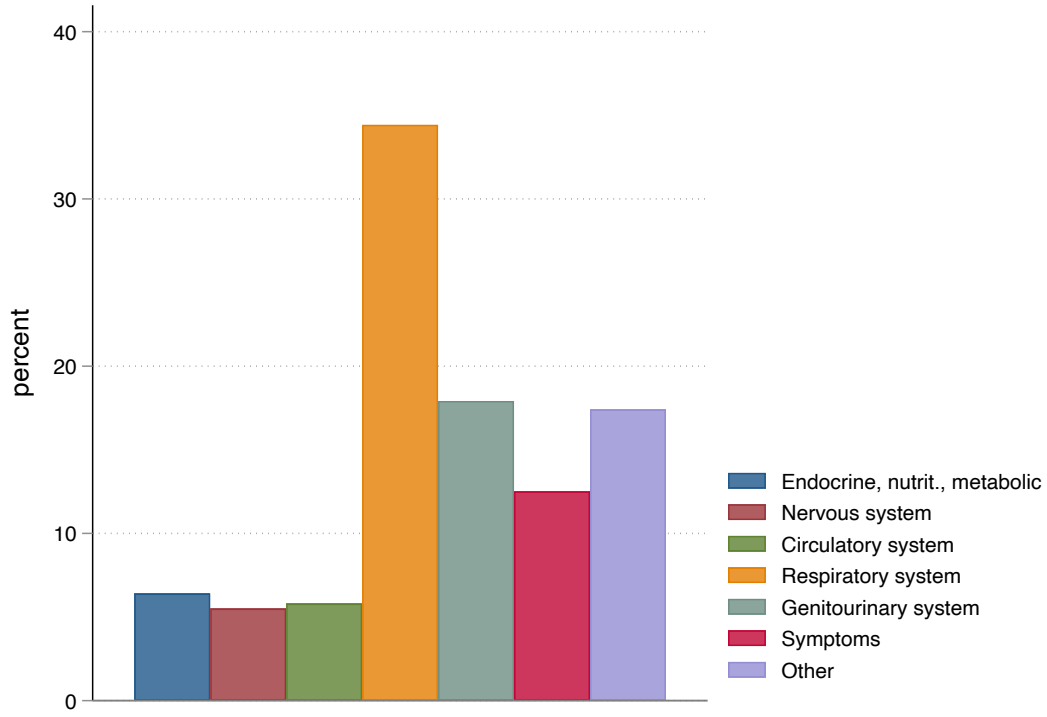
Notes: This table shows the difference in pre-digital care diagnosis and healthcare utilization covariates for patients who have had no previous avoidable hospitalization (AH) in 2013-2016, compared with patients who have had at least one such hospitalization in the period before digital care. The socioeconomic variables do not exist for child patients.

Figure 2a.6: Patients classified as risky have more of the symptoms connected with later avoidable hospitalizations



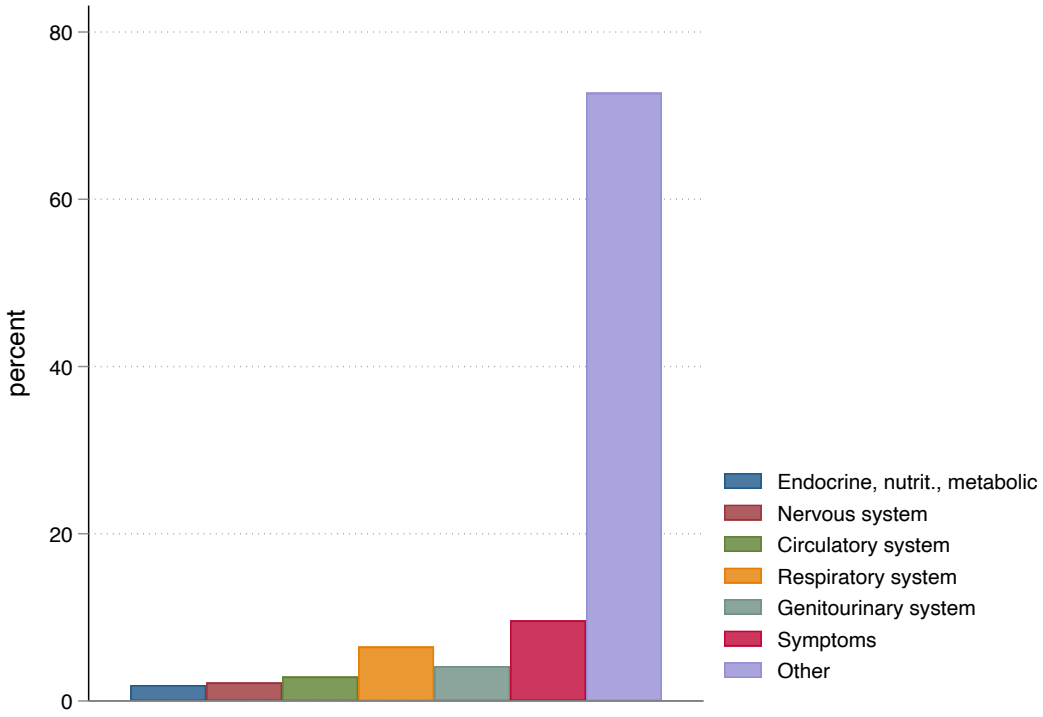
Notes: Patients are asked to fill in their main symptom/reason for seeking care just before the consultation. Exactly 1% of patients are risky for AH in the overall sample, so symptoms with over 1% risky patients are over- represented for risky patients and vice versa.

Figure 2a.7: Diagnosis groups of avoidable hospitalizations



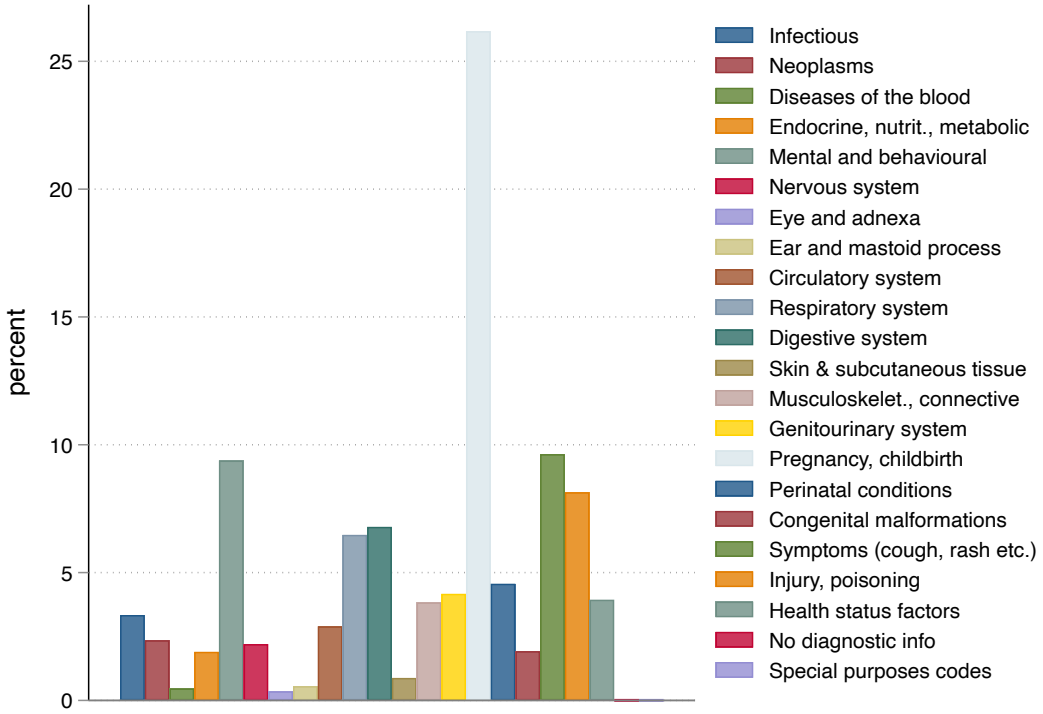
Notes: Grouping of the primary diagnosis code among the avoidable hospitalizations within 3 months after the digital visit.

Figure 2a.8: Diagnosis groups of all hospitalizations, AH relevant groups



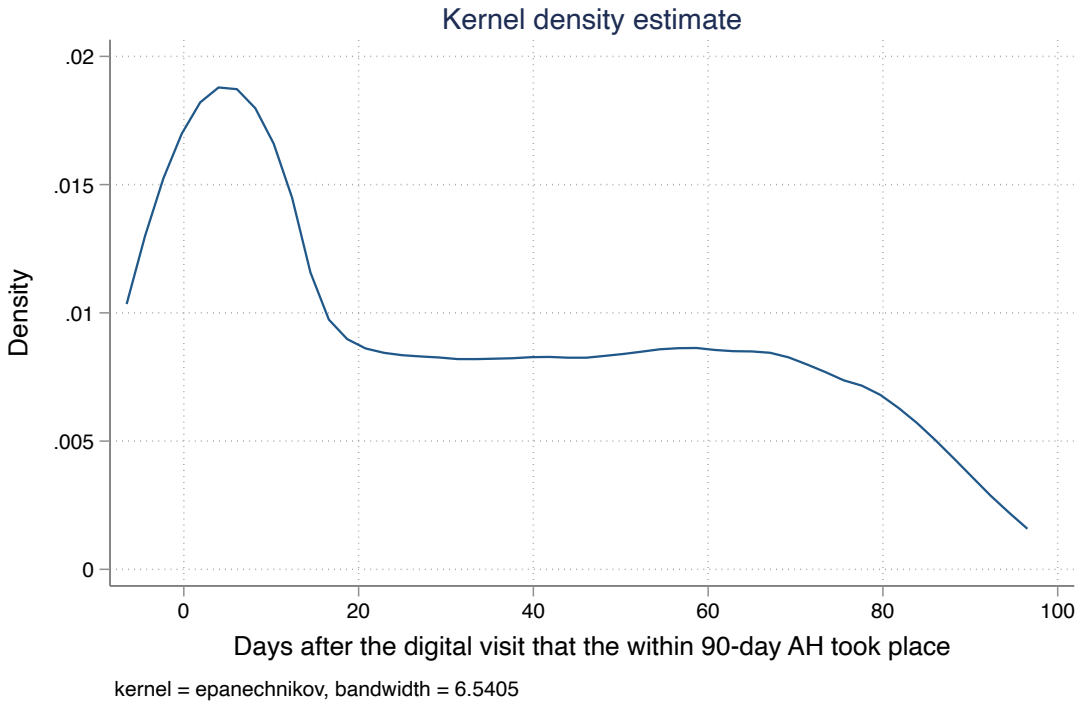
Notes: Grouping of the primary diagnosis code among ALL hospitalizations (not just avoidable hospitalizations) with the same groups in 2a.7

Figure 2a.9: Diagnosis groups of all hospitalizations, all groups



Grouping of the primary diagnosis code among all hospitalizations with all diagnosis groups.

Figure 2a.10: Length of time between consultation and avoidable hospitalization



Notes: Distribution of days after digital visit that the AH within 90 days happened

Table 2a.7: Neg. nr. avoidable hosp. in 3 months after consultation, re

	(1)
negative_ah	
Nr AH 3 years before	-0.0277
	(0.0196)
Disease index	-0.0076***
	(0.0020)
Female	-0.0002
	(0.0004)
Age	-0.0000*
	(0.0000)
2nd gen immigrant	-0.0005
	(0.0008)
1st gen immigrant	-0.0020**
	(0.0010)
_cons	0.0007
	(0.0005)

With date time shift FE, doctor RE.
SEs in parantheses clustered on doctors.
Sample born before 2013, doctors' visits before 600th.
Sample before Oct 2018 to allow 3 month follow up.
Disease index is sum of Elixhauser comorbidities.

Table 2a.8: Nr AH 3 months after first digital visit

	(1) OLS Simple	(2) Controls	(3) Bootstrap	(4) Time shift FE	(5) ZI Poisson
Std doctor FE	0.0002	0.0001	0.0001	0.0002	-0.0275
	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0644)
Nr AH 3 years before	0.0586***	0.0551***	0.0551***	0.0552***	-0.0157
	(0.0103)	(0.0101)	(0.0099)	(0.0102)	(0.0440)
Doc FEX AH 3yrs b4	-0.0188**	-0.0190**	-0.0190**	-0.0193**	-0.0602**
	(0.0088)	(0.0087)	(0.0091)	(0.0088)	(0.0247)
Inflation for the ZIP:					
Nr AH 3 years before					-1.6412***
					(0.4146)
Controls	No	Yes	Yes	Yes	Yes
N	122662	122564	122564	122564	122564
R^2	0.051	0.056	0.056	0.056	
Mean	0.0023	0.0023	0.0023	0.0023	0.0023
Mean_risky	0.0589	0.0589	0.0589	0.0589	0.0589

All columns: SEs in parantheses clustered on doctors. Col 3 has booststrapped SEs.
Sample is dropin first visits after doctor's 600th such visit. Only patients born before 2013.
Controls are Elixhauser sum of comorbidities, female, age, first- and second-generation immigrant.
The 6th column shows results from a Zero-Inflated Poisson model.

Table 2a.9: Definitive counter guideline prescription

	(1) OLS Simple	(2) Controls	(3) Bootstrap	(4) Time shift FE
Std doctor FE based on no CGP	-0.0035***	-0.0035***	-0.0035***	-0.0035***
	(0.0008)	(0.0008)	(0.0008)	(0.0006)
Nr antib filled 3yrs before	0.0021***	0.0022***	0.0022***	0.0021***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Std doc FE X nr antib 3yrs b4.	-0.0005**	-0.0005**	-0.0005**	-0.0005**
	(0.0002)	(0.0002)	(0.0003)	(0.0002)
Controls	No	Yes	Yes	Yes
N	116396	116391	116391	116391
R^2	0.006	0.006	0.006	0.006
Mean	0.0172	0.0172	0.0172	0.0172

All columns: SEs in parantheses clustered on doctors. Col 3 has booststrapped SEs.
Sample is dropin first visits after doctor's 600th such visit. Only patients born before 2013.
Controls are Elixhauser sum of comorbidities, female, age, first- and second-generation immigrant.

2.8.1 Additional Results

Additional results: Good doctors at all three measures are less senior and have worked less in the service.

Table 2a.10: Explaining quality with doctor characteristics

	(1) Std CGP skill	(2) Std AH skill	(3) Std double skill visit skill	(4) Over median at all 3
100s randomized consultations	0.00	-0.00	-0.02***	-0.00*
	(0.00)	(0.00)	(0.01)	(0.00)
In specialty training	-0.02	-0.39**	-0.14	-0.14**
	(0.22)	(0.19)	(0.20)	(0.07)
Specialist	-0.03	-0.38	-0.11	-0.15**
	(0.22)	(0.25)	(0.19)	(0.07)
Non EU15 language	-0.35*	-0.20	0.25*	0.01
	(0.20)	(0.19)	(0.15)	(0.05)
_cons	0.02	0.36*	0.57**	0.24***
	(0.19)	(0.19)	(0.22)	(0.07)
N	143	143	143	143
R^2	0.03	0.04	0.20	0.06

This corroborates studies e.g. Newhouse et al. (2017) showing that younger hospital doctors have lower mortality and costs than older doctors. Older doctors have more experience, but are less up to date with recent medical knowledge.

Additional result: Female doctors are 0.7sd better at following guidelines.

Table 2a.11: Gender and doctor characteristics

	(1) Std CGP skill	(2) Std AH skill	(3) Std double visit skill	(4) Over median at all 3
Female doctor	0.71***	0.11	0.23	0.09
	(0.24)	(0.27)	(0.31)	(0.07)
_cons	-0.30	0.03	-0.33	0.03
	(0.19)	(0.19)	(0.25)	(0.03)
N	61	61	61	61
R^2	0.12	0.00	0.01	0.03

This corroborates studies e.g. Kim et al. 2005; Berthold et al. 2008; Baumhäkel et al. 2009, which show that female doctors adhere more to other guidelines. Note that I have data on gender on only 43% of doctors.

Figure 2a.11: Histogram of the share of antibiotics out of all prescriptions of patients before digital care.

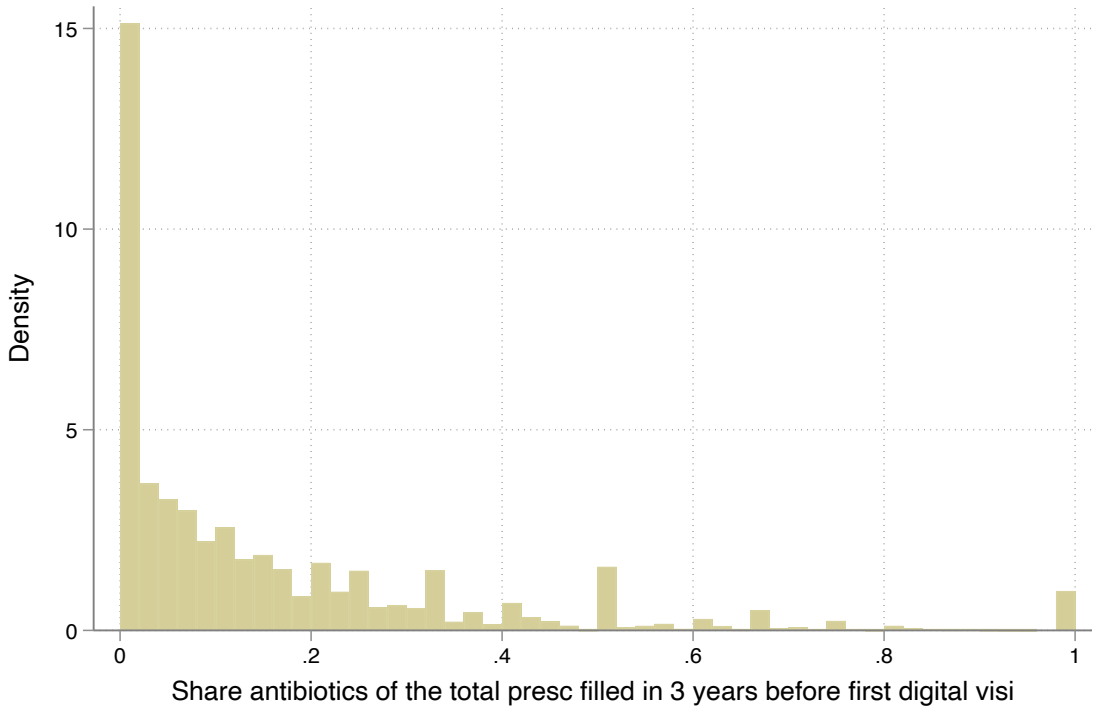
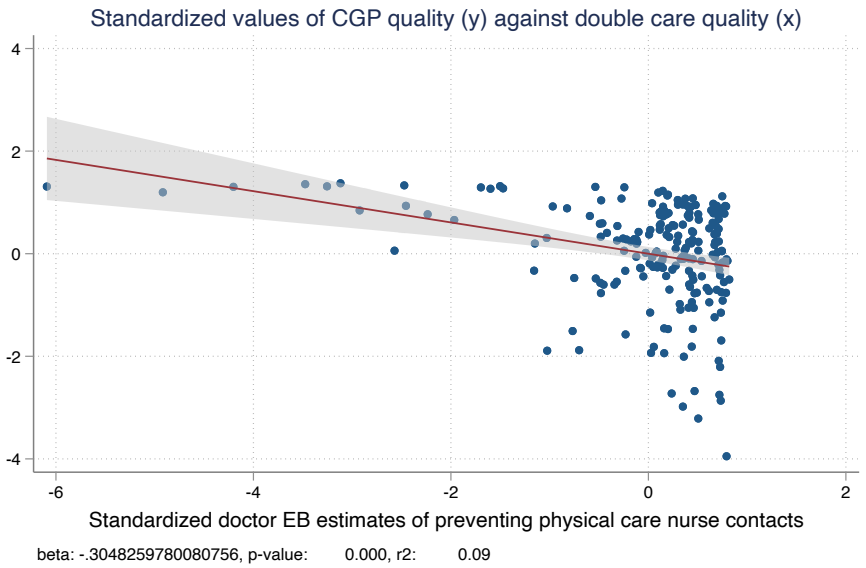


Figure 2a.12: Correlation between two different quality measures within doctors.



Match effects

Table 2a.12: Number Avoidable Hospitalizations within 3 mo. after visit

	(1) Clustered SEs	(2) Bootstrapped SEs
Top 10% doctor X top 1% risky patient	-0.060***	-0.060***
	(0.014)	(0.016)
Top 10% doctor on AH	0.000	0.000
	(0.001)	(0.001)
Riskiest 1% patient in AH	0.067***	0.067***
	(0.012)	(0.014)
N	95816	95816
Mean	0.003	0.003
Mean_risky	0.062	0.062

All columns have date-time shift fixed effects. Sample is all doctors' randomized visits after the 600th' consultation

A person who has had 1 more AH in the past 3 years has 6 percentage points higher risk of getting one again within 3 months, but only 4 percentage points higher risk if they meet a 1 standard deviation better doctor (a reduction of 29%).

2.8.2 ARE for Counter-Guideline Antibiotics Prescriptions

Reallocation on patients' previous share of antibiotics and doctor CGP effect

Table 2a.13: Average Reallocation Effects from minimizing Counter-Guideline Antibiotics Prescriptions

	Status quo	PAM	PAM-SQ
CGP	3417	3059	-358
SE		(2.722)	
N	199,867	199,867	199,867

Notes: Average Reallocation Effects for Counter-Guideline Antibiotics Prescriptions (CGP)), with Positive Assortative Matching (PAM) on patient previous antibiotic share out of total prescriptions and doctor CGP quality. Standard errors are bootstrapped in the regression computing the average match function, which is calculated over the appointments after the 600th, then using 'predict' for a counterfactual allocation and then aggregating all individual effects using 'total'. The measure of doctor CGP quality here is the standardized shrunk Empirical Bayes estimate calculated over doctors' first 600 appointments.

Reallocation on patient income and doctor CGP effect

When patients with lower income are matched with doctors with higher quality on CGP (i.e. who have been more likely to follow guidelines in the auxiliary sample), the result is more counter-guideline prescriptions, but only by a very small amount, 46, corresponding to a 1.6% increase. However, we also see an increase in CGP when patients with higher income are matched with doctors of higher quality, but by an even smaller amount.

Table 2a.14: Average Reallocation Effects for Counter-Guideline Antibiotics when matching high income patients with the best doctors

	Status quo	PAM	PAM-SQ	NAM	NAM-SQ
CGP	2792	2807	15	2838	46
SE		(4.776)		(5.329)	
N	163,443	163,443	163,443	163,443	163,443

Notes: Negative Assortative Matching (PAM) on patient income and continuous doctor CGP quality.

2.9 Appendix II: Data Appendix

2.9.1 Datasets

All datasets are proprietary and confidential, and were accessed after applications to the Stockholm Regional Ethics Council (2018, number 2108/2318-31 and Swedish Ethics Authority (2019, number 2019-06062) had been approved. Additionally, Statistics Sweden and the other entities carried out their own confidentiality assessments before approving the sharing of data. Statistics Sweden anonymized the personal identifiers and matched with other datasets, and then shared only an anonymized version of the data with the researcher.

Definition of analysis sample

I start from the universe of patients who has had at least one digital consultation with one of the largest[74] providers of digital healthcare in Sweden, from the start of the service in mid-2016 to the end of 2018. There are 631,681 consultations in the dataset. I keep only the first visit for each patient, as these consultations are conditionally randomized, but there is a bigger concern of endogeneity in any following visits. Hence, each patient has only one observation in digital care.

There are 378,627 unique patients, who have on average had 1.67 consultations. We match this data to official registry data from Statistics Sweden on socioeconomic and demographic variables[75] and data from the National Board of Health and Welfare (NBHW / *Socialstyrelsen*) on diagnoses of chronic conditions from specialist, acute and inpatient care across the Swedish healthcare system in the three years preceding digital primary care, 2013-2015. In this physical healthcare dataset, there are many observations per patient. In addition, we match with data on physical primary care (2013-2019) from one Swedish region (Skåne), which matches for around 10% of the digital care sample.[76]

[74] In terms of patient volumes in 2016-2020.

[75] In total 847 people (0.22% of the initial sample) could not be matched to the Statistics Sweden or NBHW records. Of these, there are 262 individuals with an incorrect personal identification number (PIN) according to Statistics Sweden. In addition, there are 112 people with a re-used PIN, which are dropped. An additional 473 people could not be matched for other reasons.

[76] Swedish physical primary care is devolved to 21 regions, which means that data from primary care is not included in the National Board of Health and Welfare data. Receiving primary care data in Sweden across all regions has eluded researchers, as policies, codings and applications vary across the country.

The sample now consists of all individuals (377,780) who have had a digital video consultation with a medical doctor from the start of the service in 2016 until the end of 2018, and can be matched with registry data. For most of the analysis, we restrict the sample to "drop in" visits, that is visits where the patient had no way of specifying which doctor they want to meet, but rather meet the first available doctor. This is 82% of the first visit sample (310,000 patients), and this is the sample where conditional (on time) randomization holds. Moreover, we remove pediatricians and those children who are more likely to see a pediatrician (where randomization did not hold), which leaves 302,883 patients and 511 doctors.

For consistent definition of patient types according to their pre-digital physical healthcare utilization, we drop patients for whom we do not observe the full pre-period 2013-2016, i.e. patients who were born in or after 2013. This leaves 233,489 patients and 499 doctors. Finally, we keep only doctors who have done >600 consultations and their patients, which leaves 210,171 patients (56% of original N) and 143 doctors (20% of original D). The reason is that many doctors were hired late in the sample period, since the service was expanding. These doctors have only done a few randomized consultations, many of them under 100. This is not a sufficient sample to base the analysis on. For the outcome avoidable hospitalization, we need a post-digital consultation period of 3 months, which means we drop all consultations which took place in October-December 2018, as our follow up data in physical healthcare ends on 31 December 2018, just as the digital care dataset does.

Statistics Sweden dataset

We can measure most socioeconomic characteristics for adults only, since the variables on e.g. income and education do not exist for minors. The socioeconomic variables from Statistics Sweden reported here are all measured at the same time for all individuals, irrespective of the year when they started using the digital service (income and employment variables in 2017, education in 2018).

Digital care dataset

This dataset was created from the internal logs of the digital healthcare company.

Sample definition:

- Keeping only the first visits: from ∼630,000 visits (1.67 visits/ person) to ∼378,000; 715 doctors.
- Randomly assigned ("drop in"): 310,000 patients (82%), 526 doctors.
- Removing pediatricians and those children who are more likely to see a pediatrician: 302,883 patients, 511 doctors.
- Dropping patients born before 2013 (when pre-data starts): 233,489 patients (62% of original N), 499 doctors (70% of original D).
- Keeping only doctors who have done >600 consultations **210,171 patients** (56% of original N), **143 doctors** (20% of original D).

Chapter 3

Planning Ahead for Better Neighborhoods: Long Run Evidence from Tanzania

3.1 Introduction

Africa's cities are growing rapidly. With its expanding population (United Nations 2015) and rising urbanization rate (Freire et al. 2014), we expect that almost a billion people will join the continent's cities by 2050. But many of these cities, especially in Sub-Saharan Africa, already face problems of poor infrastructure and low quality housing (Henderson et al. 2016 and Castells-Quintana 2017). According to UN Habitat (2012), as many as 62% of this region's urban dwellers live in slums, whose population was expected to double within 15 years. The poor living conditions in those slums have important consequences for residents' lives (Marx et al. 2013).

There are various policy options for addressing the immense challenges posed by African urbanization. One option, which is often the default, is to allow neighborhoods to develop organically without much planning or infrastructure. At the other end of the spectrum, a second option is for the state to not only plan but actually build public housing. This is expensive, but has been done for example in South Africa (Franklin 2020). Between these two alternatives lies a third option of laying out basic infrastructure on the fringes of cities, and allowing people to build their own homes, an option advocated by Romer (2012) and Angel (2012). A fourth option is to improve infrastructure in areas where low quality housing develops.

Understanding the implications of these options is important for current policy discussions. For example, there are debates about the respective merits of upgrading and starting anew (e.g. Duranton and Venables 2020). But we know relatively little about these options' implications for private investments and the survival of infrastructure, or about their distributional consequences. One of the main contributions of this paper is to shed light on these issues. To do so, we study the long run development of neighborhoods, which were part of the "Sites and Services" projects (described below).[1] These took place not only in Tanzania's biggest city, Dar es Salaam, but also in six of its secondary cities.[2]

Our paper focuses on the long run consequences of the third option discussed above (*de novo*) compared to the first (default) option of unregulated development. We study de

[1] Throughout the paper we refer interchangeably to: "areas" and to the "neighborhoods" that develop in them; houses and housing units; and squatter settlements and slums. Finally, we refer to "owners" as those with de-facto rights to reside in a house or rent it out. Legally, even formal ownership consists of a long and renewable lease from the state.

[2] This is important because Africa's secondary cities are relatively understudied, despite being home to the majority of its urban population. See for example Brinkhoff (2017), Agence Française de Développement (2011), National Oceanic and Atmospheric Administration (2012), and Tanzania National Bureau of Statistics (2011).

novo neighborhoods, which were developed in greenfield areas on what were then the fringes of Tanzanian cities. The developments included the delineation of formal residential plots and the provision of basic infrastructure, consisting primarily of roads and water mains. People were then offered an opportunity to build homes on these plots in exchange for a fee. To provide a counterfactual, we use nearby *control* areas that were also greenfields before the Sites and Services projects began. We also provide descriptive evidence on the fourth approach discussed above by studying *upgrading* areas, which received infrastructure investments similar to the de novo areas, but only after people had built low quality housing.[3] We compare these upgrading areas to nearby areas and also, for Dar es Salaam only, to slums that were not upgraded.

We investigate how different neighborhoods developed over more than three decades, and we ask a number of questions. First, do de novo investments solve coordination failures and facilitate neighborhood development in the long run? Second, how do they shape private housing investments and the survival of public infrastructure? And finally, what characterizes the sorting of owners and residents across neighborhoods, and to what extent can owners' sorting account for the differences in outcomes across neighborhoods?

Concretely, we study Sites and Services projects, which were co-funded by the World Bank and the Tanzanian government, and were similar to projects carried out in other countries. In Tanzania they were implemented in two rounds: one began in the 1970s and the other in the early 1980s. Altogether, 12 de novo neighborhoods and 12 upgrading neighborhoods were developed in Dar es Salaam, Iringa, Morogoro, Mbeya, Mwanza, Tabora, and Tanga. (World Bank 1974a,b, 1977a,b, 1984, and 1987).[4]

To study the consequences of de novo investments, we combine high resolution spatial imagery on all seven cities and building-level survey data on three of the cities with historical imagery and maps. We analyze these data using a spatial regression discontinuity (RD) design. We find that in de novo areas, houses are larger and more densely and regularly laid out, are better connected to electricity, and (in some specifications) also have better sanitation. A "family of outcomes" index and a hedonic measure of house values show that de novo areas have higher quality housing. These results, which are robust to the inclusion of various controls and robustness checks, demonstrate the crowding-in of private investment in response to the public de novo infrastructure investments. We also find that de novo areas have better access to roads and water mains, reflecting the

[3] Unlike de novo areas, however, upgrading areas did not receive formal plots.

[4] Until recent years the government maintained sole authority for creating new formal plots in Tanzania, so we cannot study the long run consequences of privately provided plots.

persistence of the Sites and Services infrastructure investments over several decades.

To shed light on the mechanisms that underlie our findings, we develop a simple model of owners' investment decisions, which features complementarity between public and private investments. In de novo neighborhoods, where a sufficient fraction of owners can invest in housing quality, infrastructure investment crowds in private investment in housing quality, which in turn preserves infrastructure quality. This virtuous feedback, however, does not occur in upgrading areas, both because the existing stock of (low quality) housing disincentivizes wholesale reconstruction of housing, and because owners' credit constraints prevent them from investing sufficiently, and as a result infrastructure deteriorates. At the same time, in control areas the infrastructure investments are lower, so no high quality housing is built.

The model helps us interpret our empirical findings in two important ways. First, it allows us to separate the roles of owners' different credit constraints from the effect of infrastructure investments, when comparing de novo and control areas. In practice, we find that adding owner fixed effects reduces the quality differences between de novo and control areas by up to one-third, but these differences remain large and precisely estimated. Second, the model allows us to infer land value differences across neighborhoods from differences in housing quality. Our calculations suggest that the local gains in land value from de novo were, at least in Dar es Salaam, no less than $75-100 per square meter of plot (in 2017 prices). These gains far exceed the costs of the project, which amounted to no more than $8-13.

In our empirical analysis we also use census micro data to characterize the sorting of residents across neighborhoods. We find that as of 2012, de novo neighborhoods attracted better educated residents, who likely had higher incomes to pay for better amenities. The sorting on education across neighborhoods is, however, only partial: about 45 percent of the adults in de novo areas had no more than a primary school education. Furthermore, even less educated people who initially owned de novo plots and eventually sold them likely gained from some of the land value appreciation.[5]

In contrast to our findings on de novo areas and in line with our model, our descriptive analysis of upgrading areas suggests that their housing quality is either similar to that of nearby areas or non-upgraded slums, or in some cases even worse. Our findings also suggest that upgrading areas do not enjoy better access to water mains or roads than

[5] As we discuss below, a few years after Sites and Services were implemented, most of the residents in de novo neighborhoods in Dar es Salaam were still those targeted by the policy, many of whom were poor.

the control areas, so the Sites and Services investments in these areas likely deteriorated. These results should be interpreted cautiously, however, since it is harder to find a clean counterfactual for upgrading areas (which were populated to begin with) than for de novo areas.

The economic evaluation of de novo Sites and Services areas is thus the focal point of our paper. Previous studies of Sites and Services around the world include surveys (e.g. Laquian 1983) and critical discussions (e.g. Mayo and Gross 1987 and Buckley and Kalarickal 2006). In the Tanzanian context, there are descriptive studies of Sites and Services in Dar es Salaam (Kironde 1991 and 1992 and Owens 2012). Other work on Dar es Salaam studies different interventions, including the short-term impact of more recent slum upgrading projects on health, schooling, and income (Coville and Su 2014); descriptive analyses of a more recent episode of serviced plot provision, known as the "20,000 plots" project, which suggests sizeable short-run gains in land values (Tiba et al. 2005 and Kironde 2015); and willingness to pay for land titling in poor neighborhoods (Ali et al. 2016 and Manara and Regan 2019). But as far as we are aware, ours is the first long run econometric evaluation of de novo Sites and Services areas.

Our study is related to research on the role of coordinating land institutions (Libecap and Lueck 2011) - in our case formal plots - in underpinning economic development. It is also related to studies of housing externalities in cities (Hornbeck and Keniston 2017 and Rossi-Hansberg et al. 2010). Another recent and related paper - on Indonesia rather than Tanzania - is Harari and Wong (2017). They, like us, find that upgraded slums do not perform well economically in the long run. Our paper, however, differs from theirs since we focus on de novo neighborhoods, which are not part of the context they study.

Our paper is also related to the literature on the economics of African cities (Freire et al. 2014). Like Gollin et al. (2016) we study not only the largest African cities (such as Dar es Salaam in Tanzania), but also secondary cities. Our contribution to this literature comes from studying these cities at a fine spatial scale, examining individual neighborhoods and buildings, using a combination of very high resolution daylight satellite images, building-level survey data, and precisely georeferenced census data.

A few recent papers study outcomes not only across African cities but within them (see for example Henderson et al. 2016). Our study differs not only in our focus on secondary African cities, but also in the longer time horizon we cover. We use historical satellite images and highly detailed maps going back over 50 years, which allow us to evaluate long run changes on historically undeveloped land in response to specific infrastructure

investments. By combining these with data on individuals, we also provide more evidence about the sorting across neighborhoods.

Also related to our paper is a broader literature on the economics of slums (e.g. Castells-Quintana 2017 and Marx et al. 2019). Our contribution to this literature is to illustrate conditions under which housing of better quality forms and persists, and the limitations of upgrading existing slums. Poor neighborhoods have also been studied in other settings, especially in Latin America and South Asia. For example, Field (2005) and Galiani and Schargrodsky (2010) find that providing more secure property rights to slum dwellers in Latin America increases their investments in residential quality.[6] Our paper differs in its setting (Tanzania is considerably poorer than Latin America) and its focus on early infrastructure provision.

While our paper's focus is on new neighborhoods rather than new cities, it is also related to Romer (2010), who investigates the potential for new Charter Cities as pathways for urban development in poor countries. Our work is also related to the position advocated by Shlomo Angel, that Sites and Services may be a relevant model for residential development in some circumstances.[7]

Methodologically, we contribute to the nascent literature using very high resolution daylight images (e.g. Jean et al. 2016). Like Marx et al. (2019) we study roof quality as a measure of residential quality. Our measure of quality differs, however; instead of measuring luminosity, we assess whether roofs are painted, since paint protects the roofs from rust. We also use the imagery data to develop a set of measures of residential quality, including building size, access to roads, and a measure of regularity of neighborhood layout, which we combine with survey data on building quality.

The remainder of our paper is organized as follows. Section 2 discusses the institutional background and data we use; Section 3 presents the research design and our empirical findings; Section 4 contains a model of investments in infrastructure and housing in different neighborhoods; and Section 5 concludes.

[6] In another paper, Galiani et al. (2013) study an intervention that provides pre-fabricated homes costing around US$1,000 each in Latin America, but come without any infrastructure.

[7] See for example this interview with Angel, which discusses this idea: http://www.smartcitiesdive.com/ex/sustainablecitiescollective/conversation-dr-shlomo-angel/216636/

3.2 Institutional background and data

3.2.1 Institutional background

What were Sites and Services projects?

This paper studies the long term consequences of ambitious projects that were designed to improve the quality of residential neighborhoods in Tanzania. These projects, called "Sites and Services", formed an important part of the World Bank's urban development strategy during the 1970s and 1980s. Sites and Services projects were implemented not only in Tanzania, but also in other countries such as Senegal, Jamaica, Zambia, El Salvador, Peru, Thailand, and Brazil (Cohen et al. 1983). Of the World Bank's total Shelter Lending of $4.4 billion (2001 US$) from 1972-1986, Sites and Services accounted for almost 50 percent, and separate slum upgrading accounted for over 20 percent.

In Tanzania, Sites and Services were implemented in two rounds – the first began in the 1970s (World Bank 1974b and 1984) and the second in the 1980s (World Bank 1977b and 1987). These projects were co-financed by the World Bank and the Tanzanian government (World Bank 1974a and 1977a).

Sites and Services projects in Tanzania fell into two broad classes. The first involved de novo development of previously unpopulated areas. The second involved upgrading of pre-existing squatter settlements (sometimes referred to as "slum upgrading"). In total across both rounds, the program laid the groundwork for 12 de novo neighborhoods and 12 upgrading neighborhoods spread across seven cities (World Bank 1974b, 1977b, 1984, and 1987).

The overall cost of the Sites and Services projects in Tanzania was approximately $130 million (in US$2017), of which $83 million were direct costs, covering for infrastructure, land compensation, equipment and consultancy (World Bank 1974b, 1977b, 1984 and 1987).[8] The direct costs per square meter in the first round in de novo ($2.20) and upgrading ($2.37) were similar (World Bank 1974b, 1977b, 1984 and 1987). To compare these costs to present-day land values (see below), we focus on costs per square meter of plot, excluding public areas. As we explain in the Data Appendix, we estimate that the direct costs per square meter of plot were no more than $8, and the total costs were no more than $13 per square meter.

[8]The remainder of the costs covered a loan scheme and community buildings.

What were the treatment and counterfactual?

Our main empirical analysis compares de novo (our main treatment) to nearby control areas (our counterfactual). As we explain in Section 3, we implement a spatial regression discontinuity design, focusing on the difference in outcomes close to the boundary of de novo areas and adjacent control areas, which were (like de novo) unbuilt before the Sites and Services projects began. In Section 3 we also discuss and address potential threats to our identification strategy. Here we explain why we focus on the comparison between de novo and control areas and what we learn from it.

De novo areas received roads, which were mostly unpaved, and water mains, as well as formal plots.[9] The combination of these three infrastructure elements (formal plots, roads, and water mains) constitutes the main treatment for de novo areas.[10] Roads reduce travel costs for both work and leisure for residents, customers and visitors. Water mains may improve the quality and reliability of water consumed, and reduce the transaction costs of purchasing water (e.g. from water trucks). They may also improve the residents' health and help them grow food. Formal plots reduce the risk of full expropriation, and of infringements onto parts of owners' plots and public spaces (such as roads and areas required to maintain water mains). They may also reduce conflicts over ownership, and the need to engage in costly defensive actions (such as building fences or walls). Moreover, the formal and regular plots may mitigate coordination problems, lead to easier access and better use of space, and make plots more easily tradeable, increasing the incentives to invest in them. Long-term gains in land values from a regular grid of plots (compared to a decentralized and irregular system) have been documented in the US context (Libecap and Lueck 2011).[11]

In addition to the main treatment components, both de novo and upgrading areas received a small number of public buildings, which were designated as schools, health clinics, and markets.[12] While these could have had an impact, we think that they matter less than the plots, the roads and the water mains. First, the total cost of the public buildings

[9]Formal plots are delineations of land, which meet local surveying and town planning standards. They increase tenure security, and are a prerequisite in any application for a Certificate of Right of Occupancy (the highest land tenure document in Tanzania).

[10]Upgrading areas also received roads and water mains, but no formal plots. The Data Appendix contains more information about the precise timing and more details the investments cost breakdown. The second round investments were generally lower - in some cases they may have excluded water mains, and for one of the de novo areas (the one in Tanga), we have some uncertainty as to the extent of infrastructure that was actually provided (World Bank 1987). Most of the de novo plots were, however, laid out in the first round.

[11]Hornbeck and Keniston (2017) similarly emphasize that starting afresh can lead to higher local land values in an urban setting.

[12]The first round buildings public buildings were also surrounded by street lighting

was lower than either the roads or the water mains; and second, even if Sites and Services areas received more buildings than other areas (which we don't know), there is no evidence that access to them ends discontinuously at the project boundaries. In addition to the infrastructure investments, some Sites and Services residents were offered loans, which were not fully repaid. We think of these loans as relaxing some owners' budget constraints, and below we explain our strategy for studying the implications of differences across neighborhoods in owners' credit constraints.

As we discuss in more detail in the Data Appendix, control areas appear to have received significantly less infrastructure investments, although our data do suggest that they have some roads and connections to water mains.

For upgrading areas we do not have a clean counterfactual, because those areas were built on by squatters before Sites and Services began. Thus any present-day differences between them and other areas may reflect a combination of preexisting differences and the treatment effect of upgrading. In Section 3 we explain what we can nevertheless learn about those areas, at least descriptively, by comparing them to nearby areas or to other preexisting squatted areas that were not treated by Sites and Services.

How were treatment areas selected?

While our regression discontinuity design helps to mitigate concerns about selection of areas, it is nonetheless important to explain how the locations of the treatment and control areas were selected. For de novo neighborhoods, the planners intended to purchase mostly empty (greenfield) land parcels measuring at least 50 hectares each, although in practice this criterion appears to have been met only for seven of the twelve de novo areas. The planners also sought land suitable for construction (e.g. with natural drainage) with access to off-site water mains, trunk roads, and employment opportunities. For upgrading the planners looked for squatter settlements that were large, well-defined, hazard-safe, and suitable for infrastructure investments (World Bank 1974a, 1977b). In most but not all cities, de novo and upgrading areas were adjacent to each other. We discuss our selection of the control areas in more detail below. All the areas are depicted in Figure 3a.1.

Who took part in the program?

Another important aspect of the Sites and Services projects was the characteristics of the population they targeted. The planners had intended for the plots to be allocated following a point system, which prioritized applicants who met certain criteria. Different sources do not agree precisely on the criteria used, although it seems that a preference was given to the poor – but not the poorest – urban residents (World Bank 1974 and World Bank 1977). Laquian (1983) explains that the de novo projects in Tanzania were intended for income groups between the 20th and 60th income percentile of a country. In similar vein, Kironde (1991) argues that eligibility for de novo sites in Dar es Salaam excluded the poorest and richest households, but targeted an intermediate range of earners which covered over 60% of all urban households. It seems that the opportunity to purchase de novo plots was initially given to low income households, including those displaced from upgrading areas, presumably as a result of building new infrastructure (World Bank 1984 and Kironde 1991).[13]

There is some disagreement as to how this process was implemented in practice. One report (World Bank 1984) argues that there were irregularities in this process, which allowed some richer households to sort into de novo neighborhoods. But in discussing the de novo sites in Dar es Salaam in the late 1980s, Kironde (1991) argues that most plots were awarded to the targeted income groups, and as of the late 1980s: "The majority of the occupants (57.9 percent) are still the original inhabitants but there are many 'new' ones who were either given plots after the original awardees had failed to develop them, or who were given 'created' plots. A few, however, obtained plots through purchase or bequeathment". Taken together, the evidence suggests that de novo locations attracted some households with modest means, but gradually also richer ones. As our model below illustrates, this type of sorting would likely have occurred even if the project had been administered flawlessly.

How relevant are Sites and Services today?

The difficulty of recouping Sites and Services costs, and criticism that they excluded the poorest urban population, appear to have motivated a shift away from them during the 1980s (World Bank 1987, Mayo and Gross 1987, and Buckley and Kalarickal 2006). As

[13]The planners had intended for the plots to be allocated following a point system, which prioritized applicants who met certain criteria. But different sources (e.g. World Bank 1974, World Bank 1977, and Kironde 1991) differ in their accounts of what these precise criteria were.

a result, the share of Sites and Services (including slum upgrading) in the World Bank's Shelter Lending fell from around 70% from 1972-1986 to around 15% from 1987-2005 (Buckley and Kalarickal 2006).

Nevertheless, Sites and Services projects deserve renewed attention for several reasons. First, as mentioned above, Africa's urban population is growing rapidly, and adding pressure to its congested cities. Second, Africa's GDP per capita has grown in recent decades, so more Africans can now afford better housing, and an important question is how to deliver this. Alternative solutions, such as government provision of public housing, are considerably more expensive than a de novo approach of the type we study.[14] Third, cost recoupment and administration have since improved through increased use of digital record keeping, as evidenced by the Tanzanian Strategic Cities Project (TSCP - World Bank 2013).[15] For example, the "20,000 Plots" project, a de novo program implemented in Tanzania in the early 2000s appears to have reduced the cost per plot by about half compared to the historical Sites and Services projects, even though the new plots were bigger (Tiba et al. 2005). Finally, land on the fringes of Tanzanian cities remains inexpensive (Tanzania Ministry of Lands 2012), so there are still opportunities for more de novo developments.[16]

To shed light on the motivations of urban planners in considering de novo projects, we turn to the above-mentioned "20,000 Plots" project. Among the concerns that lay in the background to this program were the ongoing expansion of unplanned squatter areas, which suffer from poor waste management, an inadequate supply of urban services and infrastructure, and transportation problems. These unplanned areas also hamper the government's ability to collect tax revenues (Tiba et al. 2005). It is in this context that the "20,000 Plots" project aimed to alleviate the shortage of surveyed and serviced plots and to reduce the rapid increase of informal settlements, as well as to restrict land speculation and corruption (Tiba et al. 2005). At the same time, distributional concerns regarding de novo projects remain relevant (Kironde 2015), and we revisit those in Section 5.

[14] According to correspondence with Simon Franklin, from the experience of housing programs in cities such as Addis-Ababa, four room apartments (with a bathroom) in five-story buildings entail construction cost of around $10,000, plus a further $3,000-4,000 for infrastructure and administration. This figure excludes land costs.

[15] The TSCP was approved by the World Bank in May 2010 (see http://projects.worldbank.org/P111153/tanzania-strategic-cities-project?lang=en).

[16] Even cheap land on the city fringes is likely to have some residents, however, and ensuring that de novo programs treat them inclusively is an important issue, which we revisit in the conclusions.

3.2.2 Data description

This section outlines how we construct the datasets that we use in our empirical analysis, leaving further details to the Data Appendix. First, we explain how we measure the treatment and control areas. Second, we explain our choice of units of analysis. Third, we explain how we construct the variables that we use in our analysis. Lastly, we discuss summary statistics for key outcomes.

How do we measure treatment and control areas?

For five of the seven Sites and Services cities (Dar es Salaam, Iringa, Tabora, Tanga, and Morogoro) we have maps showing the program area boundaries (World Bank 1974a,b, 1977a,b, 1984, 1987). For the two remaining cities we use information from local experts (for Mbeya) and other historical maps (for Mwanza), as we explain in the Data Appendix. Tables 3a.1 and 3a.2 list all 24 areas (12 de novo and 12 upgrading) with some information on the data we have on each.

Having defined the treated areas, we now explain how we construct our control areas. In much of our analysis, we use all initially unbuilt (greenfield) land within 500 meters of the boundary of de novo, as control areas.[17] We exclude areas that were uninhabitable (e.g. off the coast), built up, or designated for non-residential use prior to the start of the Sites and Services projects. In order to infer what had been previously built up, we use historical maps and imagery collected as close as possible to the start of the Sites and Services project, and where possible before its start date, as discussed in the Data Appendix.[18]

To construct control areas for the upgrading areas we similarly use greenfield areas within 500 meters of upgrading; or alternatively 21 slums that were delineated in the 1979 Dar es Salaam Masterplan (Marshall, Macklin, Monaghan Ltd. 1979) and were not upgraded as part of Sites and Services. Comparisons across slums should be taken with caution, since in accordance with the planners' intention to target larger slums (see Section 2), the upgraded slums covered an average area about four times larger than the control slums. Both upgraded and non-upgraded slums, however, had similar initial population densities (195 people per hectare in the upgraded slums and 234 in non-upgraded slums in 1979).

[17] Note that throughout our paper the control areas always exclude de novo and upgrading areas.

[18] For some of the analysis we also study untreated areas further than 500 meters from the treatment areas, in which case we again excluded areas that were built up before Sites and Services began.

Figure 3a.1 shows the de novo, upgrading, and control areas in all seven cities.[19]

Our empirical approach described below assumes that both the de novo and the control areas were unbuilt (greenfields) before the onset of Sites and Services. To provide evidence that this was indeed the case, we use a subsample of the TSCP survey data, which provides construction years for buildings in Mbeya and Mwanza (see Data Appendix). We report results from using these data cautiously, since they involve a fairly small sample and a variable (construction year), which is measured with noise, and only observed for surviving houses. With these caveats in mind, we note that only about 0.5 percent of the housing units in de novo areas and about 1.3 percent of the housing units in the nearby control areas were built before the start of Sites and Services, suggesting that the control and de novo areas were probably very sparsely populated.

How do we construct the units of analysis?

Our research design (discussed below) uses as its main units of analysis a grid of 50 x 50 meter "blocks", each of which is assigned to novo, upgrading, or control area depending on where its centroid falls. This creates a fine partition of our study area, which allows us to account for empty areas at the block level and within blocks. As we explain below, however, data constraints compel us to conduct some of the analysis at the level of individual housing units, or at the level of 2012 census enumeration areas (EAs) or subunits of EAs (Tanzania National Bureau of Statistics, 2014, 2017).

What are the key variables we measure?

To study the quality of housing across all 24 Sites and Services locations we use high resolution Worldview satellite images (DigitalGlobe 2016).[20] We employed a company (Ramani Geosystems) to trace out the building footprints from these data for six of the seven cities. For the final city, Dar es Salaam, we used separate building outlines from a freely available source – Dar Ramani Huria (2016). For all seven cities we then assembled more information on outcomes and control variables, as we explain in the Data Appendix. Here we describe some of the key variables.

For the purpose of measuring private housing quality using imagery data, we think of

[19] To keep the maps on a fixed and legible scale, we do not show the locations of the non-upgraded slums in Dar es Salaam.

[20] The images' resolution is 50 x 50 centimeters for greyscale, and a little coarser for color.

slums as typically containing small and irregularly laid out buildings, made of low quality materials and with poor access to roads. We therefore define as positive outcomes the opposite of this image of slums: buildings with large footprints, which are regularly laid out, and have good roofs and access to roads. We use three outcomes which we think of as largely reflecting private complementary investments. First is the logarithm of building footprint size, derived directly from the processed imagery. Second, we use the color satellite imagery to assess whether each roof is likely painted, and therefore less prone to rust. Third, we calculate the orientation of each building using the main axis of the minimum bounding rectangle that contains it. We then calculate the difference in orientation between each building and its nearest neighboring building, modulo 90 degrees, with more similar orientations representing a more regular layout.[21] Finally, we construct an indicator for buildings that are within no more than 10 meters from the nearest road. Unlike the three previous measures, we think of this measure of road access as largely representing persistence of Sites and Services infrastructure investments.

While the imagery and the outcomes we derive from it have the advantage of broad coverage, we complement them with detailed survey data on all the buildings in three of the Sites and Services cities, Mbeya (in southwest Tanzania), Tanga (in northeast Tanzania), and Mwanza (in northwest Tanzania). These data are derived from the TSCP survey, which was conducted from 2010-2013 (World Bank 2010). We use these data to build a more detailed picture of building quality in the areas we study. The TSCP data allow us to identify outbuildings (e.g. sheds, garages, and animal pens), which are generally smaller, and which we exclude from the analysis.[22] This leaves us with a sample of buildings that are used mostly for residential purposes, although a small fraction may also serve commercial or public uses.

We use the TSCP survey data to measure the logarithm of building footprint, and create indicators for buildings which have more than one story, good (durable) roof materials, connection to electricity, and at least basic sanitation.[23] These measures likely reflect private investments, since they were not part of the Sites and Services investments. In

[21] When we regress the log hedonic price index (discussed below) on the three imagery measures using a block-level regression, the coefficients on each of the three measures is positive and significant. This provides further support for our use of these measures of housing quality. Where applicable we standardize and pool the three quality measures together to construct a "family of outcomes" z-index (Kling et al. 2007; Banerjee et al. 2015).

[22] Outbuildings account for around 10-30% of buildings in the areas we consider, where the fraction varies by city. Their mean size is typically around one third that of the average regular building size.

[23] In the de novo, upgrading, and control areas we classify as "basic sanitation" having either a septic tank (30% of buildings) or sewerage connection (0.5% of buildings). Not having basic sanitation usually means a pit latrine (67% of buildings) or "other" or none. As before, we construct a "family of outcomes" measure based on non-missing observations for each variable.

addition, we measure connection to water mains and having road access as largely reflecting persistence of Sites and Services investments. The TSCP data also provide the full names of owners of housing units, which we use as explained below.

We also use separate TSCP (World Bank 2013) valuation data for Arusha, a city where Sites and Services were not implemented, to construct a hedonic measure of building quality, as we explain in the Data Appendix.[24] Another separate data source (Tanzanian Ministry of Lands 2012) provides us information about land values in Dar es Salaam, although at a coarser level.

In addition to these variables we construct geographic variables (distance to the nearest shore; an indicator for rivers or streams; and a measure of ruggedness), and other variables, which we use in our analysis below. All these are again explained in the Data Appendix.

We complement all these measures of the physical environment, with some data on people, including indicators for owners (identified by their full name and the city), taken from the TSCP survey, and population density and measures of schooling and literacy, which we calculate from the 2002 and 2012 censuses at the level of enumeration areas. We sometimes split enumeration areas to allocate them across treatment and control areas, as we explain in the Data Appendix.

How do the different areas compare using raw data?

Table 3a.3 summarizes information on the number of plots and the population density, as of 2002, in de novo and upgrading areas, and their respective control areas. As the table shows, de novo areas were more densely populated than nearby control areas. Upgrading areas were very densely populated, and again denser than control areas near them. As we shall see below, the higher density in upgrading areas did not correspond to more multistory buildings, but in fact the opposite.

Figure 3a.2 shows visual examples of parts of a de novo area, a control area near de novo, and an upgrading area, all in the same district of Dar es Salaam. The differences between the most orderly location (de novo) area and the least orderly one (upgrading) are visibly clear, and the control area lies somewhere in between.

The impression that de novo areas have higher quality housing is corroborated in the

[24] Our approach of using characteristics linearly in a hedonic regression follows Giglio et al. (2014). There is also some evidence that in the case of housing, using the imputed hedonic values as dependent variables does not lead to much bias in the inference (McMillen et al. 2010 and Diewert et al. 2015).

summary statistics table (Table 3a.4). The imagery data shows that compared to the control areas, de novo areas have buildings with larger footprints, a higher fraction of painted roofs, more regularly laid out buildings, and better access to roads. The survey data shows that de novo areas are also more likely to have multiple stories, good roof materials, connection to electricity, basic sanitation, and connection to water mains, as well as a much higher hedonic value. On almost all these measures, including the fraction of buildings with multiple stories, upgrading areas look worse, and control areas are somewhere in between de novo and upgrading. The log hedonic price differences suggest that on average, de novo housing units are about 63 percent more valuable than those in control areas and about 92 percent more valuable than those in upgrading areas.

3.3 Research design and empirical findings

3.3.1 Research design

The differences in outcomes described in Table 3a.4 suggest that housing quality in de novo areas is considerably better than in control areas. The higher quality of housing in de novo areas reflects both elements that Sites and Services invested in directly, such as roads and water, and elements that it did not, such as electricity. But in order to study whether the de novo investments did in fact crowd in private investments (and if so - how much), we need to move beyond the descriptive statistics, as this section explains.

Our identification strategy compares de novo areas to nearby control areas, which (like de novo areas) were largely empty before the onset of Sites and Services. In our main analysis, we follow Gelman and Imbens (2017), by implementing a semi-parametric regression discontinuity design:

$$\begin{aligned} y_i = \beta_0 + \beta_1 Denovo_i + \beta_2 Dist_i + \beta_3 Dist_i \times Denovo_i + \beta_4 \mathbf{Nearest\_Denovo}_i + \\ + \beta_5 Dist\_CBD_i + \beta_6 \mathbf{Controls}_i + \epsilon_i, \end{aligned} \tag{3.1}$$

where y_i measure various outcomes, as described in Section 2 and the Data Appendix; $Denovo_i$ is the main regressor of interest, which indicates whether the centroid of i is in de novo areas, where control areas are the omitted category; $Dist_i$ is the distance in kilometers to the boundary between de novo and control areas; $\mathbf{Nearest\_Denovo}_i$ is a vector of fixed effects for the nearest de novo areas; $Dist\_CBD_i$ measures the distance in kilometers of unit i from the Central Business District (CBD) of the city in which it

is located; $\mathbf{Controls}_i$ is a vector of additional controls, which we discuss below; and ϵ_i denotes the error term. The role of distance to the central business district is emphasized in many urban economics models (see Duranton and Puga 2015 for an overview), and adding $\mathbf{Nearest\_Denovo}_i$ ensures that we only compare control areas to their nearest de novo area. In our baseline specification, each observation is a 50 x 50 meter block, but later on, as we explain, we also use housing units within buildings and enumeration areas as units of analysis.

Our baseline analysis uses data from within 500 meters of the boundary between de novo and control areas. Using this fixed distance allows us to analyze all our outcomes across imagery and TSCP survey data (World Bank 2013) consistently. As we discuss further below, 500 meters also turns out to be fairly close to the optimal bandwidth we find for our key outcomes using the survey data. Finally, we also present below alternative specifications using more - and less - data.

In our baseline estimates we cluster the standard errors on 850 x 850 meter blocks, following the approach of Bester et al. (2011) and Bleakly and Lin (2012). The size of the blocks on which we cluster reflects the size of the Sites and Services neighborhoods. The median size of the 12 de novo neighborhoods was approximately 0.538 square kilometers, and the median size of all 24 neighborhoods was around 0.718 square kilometers. This last figure is just a little smaller than the area of a square whose sides are 850 meters, which we chose as a conservative benchmark for clustering.[25]

Addressing threats to identification

Our identification strategy assumes that conditional on the controls in specifications (3.1), the potential expected outcome functions are continuous at the discontinuity threshold. Our spatial regression discontinuity approach is similar to Dell (2010), and much of our analysis likewise applies a semi-parametric RD, which combines both controls, as in equation (3.1), and a focus on areas that are close to (within 500 meters of) the boundary of de novo and control areas.[26]

[25] In earlier versions of this paper we also reported specifications using Conley (1999) standard errors with a decay area equal to the size of the above-mentioned blocks, and the results were similar. To mitigate concerns about the variation in neighborhood size, we also experimented with modifying our baseline clustering blocks to treat each Sites and Services neighborhood as a separate clustering unit, with the remainder of the cluster units based on the grid (cut where necessary by the Sites and Services neighborhoods). Once again the estimated standard errors were quite similar.

[26] Since we have variation within several cities, we use functions of distance to the de novo boundary in our main specification, and functions of longitude and latitude only in our robustness checks, as we discuss below.

One potential concern is that the areas selected for de novo differed in their "first nature" location fundamentals. But in our setting the geographic distances are much smaller than in most other settings, so we are less concerned with larger scale changes in geography, such as climate or soil fertility. Further, in our empirical analysis, we report balancing tests, which use specification (3.1) to compare the geographic variables as outcomes as we cross the de novo - control boundary. Some of the geographic variables - the land's ruggedness and the presence of rivers or streams - may be endogenous to housing development. Therefore below we report estimates both with and without the geographic controls.

Our identification strategy also assumes that both de novo and control areas were essentially empty (greenfields) before the start of Sites and Services. As we discuss in the Appendix, our classification of areas relies on historical aerial images and topographic maps, which allow us to detect pre-existing buildings. And in Section 2 we provide support for this assumption using a subsample of buildings for which we have construction dates.

Another relevant question is whether administrative boundaries correspond to some of the de novo - control boundaries, leading to different municipal policies on either side of the boundary. To address this question, we verified that in none of the cases do the boundaries between any treatment areas and the control areas coincide with the ward or district boundaries.[27]

A different type of concern is that there may be spillovers across neighborhoods.[28] So, for example, it is possible that proximity to de novo areas improves nearby control areas, or that proximity to control areas worsens de novo areas; both would attenuate our estimates. To mitigate this concern we report "doughnut RD" specifications, which exclude bands of 100 meters around the boundary between de novo and control areas. To mitigate a related concern that upgrading areas may be affecting our estimates, we also report specifications, which exclude all blocks within 100 meters of upgrading areas. In a similar vein, since the TSCP data, but not the imagery data, cover entire cities, we also report specifications that use wider control areas, rather than only those near de novo areas. In those cases we use the same specification as in the baseline, but also report some results using second- and third-order polynomials in distance to the boundary.[29]

[27] The closest case is Mwanza in 2012, where one district (Nyamangana) cuts into less than a quarter of the control area, while another (Ilemela) contains all of the treatment and most of the control area. However, this boundary was only observed in the 2012 census and not in the 2002 census, so it is almost certainly either unrelated to the Sites and Services project, or an indirect outcome of it. In the 2002 census, Ilemela district fully contained the Mwanza treatment and control areas.

[28] See related discussions in (Turner et al. 2014), Hornbeck and Keniston (2017) and Redding and Sturm (2016).

[29] The full city data also allow us to estimate regressions using an optimal bandwidth (Imbens and

A related concern is that Sites and Services may have reshaped cities, and even affected the location of their CBD, and the distance to it. To address this concern we report robustness checks, which use distances to historically central locations - mostly railway stations, as discussed in Section 2 and the Data Appendix.

Did Sites and Services create or displace value?

Another question that we consider is whether Sites and Services created value or merely displaced it. Like many studies of place-based policies, it is difficult for us to answer this question definitively, since we do not have counterfactual cities of similar size, which were untreated by Sites and Services. And even if such cities had existed, one might still have worried about displacement of activity across cities. Nevertheless, our findings below suggest that de novo areas are relatively regularly laid out, and preserve good access to roads. It therefore seems likely that by solving coordination failures they created value and not merely displaced it.[30]

Exploring mechanisms: sorting across neighborhoods and infrastructure persistence

Our setting allows us to explore another important issue - the role of sorting of owners across neighborhoods. As we discuss above, initial ownership criteria in de novo areas excluded the poorest, and program loans may have further alleviated credit constraints for some of these owners (as well as for some of the owners in upgrading areas). The model characterizes sufficient conditions under which including owner fixed effects overcomes the potential differences in credit constraints of owners who rent out multiple housing units.[31] We note that renting is fairly common in our setting: as of 2007, renters accounted for a small majority of Dar es Salaam's residents, and over a third of the residents in other urban areas; back in 1992, the share of renters was even higher (Komu 2013).

To shed light on the sorting across neighborhoods of residents, we also use census data to characterize residents by measures of education, which are the best proxies we have for

Kalyanaraman 2012), which we also report.

[30] As we also discuss below, our findings suggest that Sites and Services not only had positive effects on local land values, they may also have generated positive spillovers on nearby areas, an issue that we revisit in Section 3.

[31] To be precise, we consider a full name as different if it appears in more than one city. In practice this does not seem to make much difference. Since this strategy uses variation within owners, it only employs part of the data, so in this case we need to use control areas from the rest of the city to ensure sufficient variation. We also acknowledge that some units may be owner-occupied, while others may be rented out, but we cannot separate the two with our data.

lifetime earnings.

Our model also highlights the role of persistently better infrastructure in de novo neighborhoods as a mechanism for crowding-in investments in housing quality. Empirically, we estimate regressions of the same form of as equation (3.1), using measures of water connection and access to roads as outcomes, since these closely relate to the investments made in the Sites and Services projects.

Studying upgrading areas

Finally, we repeat our analysis for upgrading areas, comparing them to proximate control areas, following the procedure outlined above.[32] Finding appropriate counterfactuals for upgrading areas (which were populated before the program began) is harder than for de novo areas (which were essentially empty). To mitigate concerns about different starting conditions, we also report regressions that compare upgrading areas to 21 other slums that existed in Dar es Salaam in 1979, and which were not upgraded as part of Sites and Services. The slums that were not upgraded were on average smaller in area (see Section 2), but had similar, or even slightly higher, population density in 1979. The comparisons of upgrading areas to non-upgraded slums come with two caveats: first, this analysis is not a spatial RD, since the non-upgraded slums were not adjacent to the upgraded ones, although for consistency we still use specification (3.1); and these comparisons are only possible for the imagery data, since Dar es Salaam is not covered by the TSCP survey data.

3.3.2 Empirical findings

Balancing tests

We begin the discussion of our findings by reporting balancing tests on geographic characteristics. As Table 3a.5 shows, when we compare geographic characteristics in de novo areas to nearby control areas, both distance to the shore and ruggedness differ in de novo areas (Panel A), but after including our baseline controls as in equation (3.1) (Panel B) de novo and control areas look balanced. We also report balancing tests using TSCP data, which also look balanced (with the exception of rivers and streams in the sample adjacent to the de novo areas). We note, however, that rivers and ruggedness may be endogenous

[32] In upgrading area regressions we measure distance to the upgrading (instead of de novo) - control boundary, and fixed effects for the nearest upgrading (instead of de novo) area.

to the de novo development, which may have flattened the soil and buried or diverted some streams. For completeness we report below estimates both with and without the geographic controls.

Crowding in of private investments

We now turn to our main results. In Table 3.1 we report estimates using specification (3.1) and our imagery sample. Panel A shows that de novo areas have footprints that are roughly 12 percent larger and have more regular layout, but their roof quality is not better. The z-index aggregating all three measures indicates that de novo areas have higher quality housing than nearby areas, and other estimates show that they have fewer empty blocks and a higher fraction of their area is built up.[33] Panel B reports robustness checks for the z-index using geographic controls, longitude and latitude polynomials, an alternative measure of CBDs that predates Sites and Services, and excluding blocks near upgrading areas - all are similar to our baseline estimate. When we use doughnut RD specifications to exclude areas near the boundary of de novo and control the estimates increase somewhat, suggesting that our baseline estimates may be a little attenuated due to spillovers (positive ones from de novo to controls, or negative ones from controls to de novo, or both). This finding also suggests that the higher quality housing in de novo areas may generate positive spillovers on neighboring areas (see Hornbeck and Keniston (2017) and Turner et al. (2014) for related discussions of local spillovers).

In sum, results for all seven cities using the satellite image data suggest that de novo areas have larger and more regularly oriented buildings. To get a more detailed picture of the differences in residential quality we turn to the TSCP survey data for Mbeya, Mwanza, and Tanga. In Panel A of Table 3.2 we report results again using specification (3.1). One advantage of the survey data is that unlike the imagery data they allow us to focus on residential buildings by excluding outbuildings, which we do. As Panel A shows, buildings in de novo areas have footprints that are about 50 percent (or 0.41 log points) larger than the control areas. They are also about 23 percentage points (or 48 percent) more likely to be connected to electricity. The regressions also show economically large but statistically imprecise differences in favor of de novo areas in the share of buildings with multiple stories and with at least basic sanitation, but again almost no difference in roof quality.

We aggregate the measures of quality in the survey data in two ways: first using a z-index,

[33] To visualize our results, Panel A of Figure 3a.3 shows a regression discontinuity plot of binned values of the z-index.

and second using the predicted log hedonic value. Regressions using either as an outcome indicate significantly higher residential quality in de novo areas than in control areas.[34] Specifically, the regressions suggest that the hedonic price is around 56 percent. This may understate the actual differences in house values, since the hedonics do not directly account for all housing characteristics, nor for the full impact of local neighborhoods' infrastructure. Noting this caveat, a result from the model in section 4.4 below suggests that land value differences in de novo (compared to control areas) are about 50 percent larger than house price differences. This result, combined with our hedonic estimates, suggests that land values in de novo areas are at least 86 percent higher than in control areas. To interpret this difference, we note that in Dar es Salaam, land values in de novo neighborhoods is in the range of $160-220 per square meter (in 2017 prices).[35] Combined with our estimates above, this suggest that de novo may have increased local land values by at least $75-100 per square meter.

These values are high compared to the cost of investments per unit of treated plot area which we estimate above to be no more than $8 per square meter of plot area, or no more than $13 per square meter if we include indirect costs (in US$2017). While these estimates should be interpreted with caution, they suggest that the gains from de novo investments were large, at least in Dar es Salaam. That said, we acknowledge that the gains in other cities, where prices are lower, may not be quite as high.[36]

In Panel B of Table 3.2 we report results from a series of robustness checks, focusing for brevity on the z-index and the log hedonic price. The estimates with geographic controls in column (1) are a little lower than the baseline; this could be either because the baseline regressions overstate the difference due to better geographic fundamentals in de novo location, or that the geographic controls are themselves outcomes and adding them understates the impact of de novo. Columns (2) and (3) show that controlling for the polynomial of longitude and latitude or using distance to historical (instead of contemporary) CBDs makes little difference compared to Panel A. The doughnut specification in column (4) is larger than the baseline, suggesting (as in Table 3.1) that the baseline estimates may be too small due to positive spillovers from de novo to controls (or negative ones going the other way). Column (5) excludes blocks near upgrading areas, and the results are similar to the baseline. Columns (6) uses control areas from the rest of the city, and the estimates are again larger, possibly because we are comparing de novo areas to a control group that

[34] Panels B and C of Figure 3a.3 show regression discontinuity plots for the Z-index and log hedonic prices.

[35] The coarse data we have on land values do not separately identify the control areas near de novo.

[36] Unfortunately, our land value data for other cities are either missing or not detailed enough to give a credible picture.

is on average further away, and less affected by local spillovers.[37] Finally, column (7) uses an optimal bandwidth, following Imbens and Kalyanaraman (2012), and the estimates is again quite similar to the baseline.

The results using hedonic values as outcomes in Panel C follow a similar pattern, where adding geographic controls reduces the estimate a little, and excluding areas near the boundary increases them a little. The main message, however, is that our baseline estimates are robust to using different specifications.

The role of sorting

The results discussed so far are silent on the respective role of the de novo treatment and the endogenous sorting across neighborhoods of owners with different levels of credit constraints. As our model below (in Section 4) shows, we can account for differences across areas in owners' credit constraints by adding owner fixed effects, which allow us to isolate the impact of de novo areas compared to control areas for owners with multiple housing units. The units of analysis used in these regressions are individual housing units, since this is the level at which ownership is defined. The housing units we focus on are those owned by owners of multiple units, which account for about 13 percent of all housing units. To ensure a sufficiently large sample, we reestimate specifications as in (3.1) for the full city TSCP sample, but now focusing on housing units whose owners have more than one unit. Table 3.3 reports estimates of these regressions with owner fixed effects (Panel A) and without them (Panel B). The estimates show that in this sample, housing units in de novo areas are considerably larger, and much more likely to have electricity and basic sanitation. Without owner fixed effects they also are more likely to be in multistory buildings, although this difference vanishes once we control for owner fixed effects. As reported previously, de novo housing units do not have better roof materials. The difference in quality between de novo and control areas, as reflected in the z-index and the hedonic value, suggests that de novo areas may be about 60 log points (or about 83 percent) more valuable; as discussed above, this may understate the actual differences since it is unlikely to reflect all the amenity differences. Panels C and D of the table report robustness checks for the specifications with and without fixed effects, using the z-index as an outcome. Across a range of specifications reported in Table 3.3, roughly a third of the quality advantage of de novo areas is accounted for by the different ownership, and the rest likely reflects the impact of de novo on quality for owners who are relatively

[37] The estimates are robust to using second- and third order polynomials, although in the latter case they are smaller.

unconstrained in terms of investment.[38]

The characteristics of residents in de novo areas, compared to control areas, likely reflect their willingness to pay for higher quality housing. In Table 3a.6 we report regressions using 2012 census data with "cut" enumeration areas as units of analysis (see Section 2 and Data Appendix for details).[39] Consistent with the results discussed above, residents in de novo areas are better educated and more likely to be literate in English. The higher schooling of de novo residents is consistent with sorting across neighborhoods and a higher willingness of the more educated to pay for better housing quality, although it is also possible that some of it is the result of better access to schooling of existing residents. Still, as Table 3a.6 shows, only about 55 percent of adults in de novo areas had more than primary school education, so the other 45 percent had no more than primary school education. This means that many less educated Tanzanians are still benefitting from de novo amenities.

The persistence infrastructure

To conclude our empirical analysis of the de novo areas, we explore whether their better housing quality corresponds to persistently better infrastructure. Here we focus on two of the main investments in Sites and Services, roads and water mains, and we again use specification (3.1). As Panel A of Table 3.4 shows, across both our imagery and TSCP data, de novo areas enjoy better access to roads, and the TSCP data also show that they are more likely to be connected to water mains.[40] Panels B-D report robustness checks using the same specifications as in Table 3.2. Again the estimates are a little smaller when we control for geographic covariates, and a little larger when we focus on control areas that are further from de novo, with our main estimates in between. And all the estimates are positive and statistically significant, showing that de novo investments translated into better infrastructure in the long run.

[38] When we use the hedonic measure as an outcome, the regressions estimates with and without owner fixed effects are more similar to each other (results available on request).

[39] In this case number of units of analysis is small and they are uneven (some are whole EAs and some are cut), which makes it difficult to get a good measure of distance to the boundary. Therefore in these specifications we use non-parametric regression discontinuity, without controls for distance to the boundary.

[40] This last result is robust to excluding Tanga, where we have some uncertainty about the nature of de novo investments.

Upgrading areas

Having discussed the de novo areas, we now briefly discuss what we can learn from similar regressions for upgrading areas. As Table 3a.7 suggests, upgrading areas look fairly similar to nearby control areas in terms of the geographic controls, except that in most specifications they are less likely to have rivers or streams. When compared to the non-upgraded slums, and conditional on our baseline controls, the upgrading areas are closer to the shore but not significantly different in the other two geographic controls (results available on request).

Table 3a.8 reports estimates using imagery data for all seven cities. Panel A suggests that housing quality in upgrading areas is similar to that of nearby control areas. The only significant differences are that upgrading areas have fewer empty areas and are more densely built up. Panel B shows that this conclusion is robust to a range of different specifications.

In Panels C and D we compare upgrading areas in Dar es Salaam only to the preexisting ("old") slums that were not upgraded as part of Sites and Services. Once again the results suggest that upgrading areas are no different, except perhaps in a slightly more regular orientation of buildings than their control areas. Upgrading areas also seem to have fewer empty blocks and a larger fraction of built up area.

Next, in Table 3a.9, we use TSCP survey data outcomes. Here the upgrading areas look somewhat worse than nearby control areas: they have fewer multistory buildings, worse roofs, and possibly worse sanitation, and their overall quality seems lower. This conclusion is reinforced in most of the robustness checks in Panels B and C, although not all the estimates are precise.

In Table 3a.10 we examine the role of ownership in accounting for the worse quality in upgrading areas. The results suggest that ownership differences may partially explain the worse housing quality in upgrading areas, since controlling for owner fixed effects results in estimates that are small and in most cases imprecise.

A comparison of infrastructure persistence measures in upgrading areas may also help to explain why their housing is no better than that of nearby control areas. As Table 3a.11 shows, upgrading areas look similar to nearby areas in their access to roads and water; the coefficients on upgrading areas are small, imprecise, and mostly negative. Adding the coefficients and the control means and comparing them to the estimates in de novo

areas (Table 3.4) suggest that upgrading areas have worse infrastructure than de novo areas. As we discussed in Section 2, upgrading areas did receive roads and water mains, and investments measured in dollars per square meter were similar to those of de novo areas. A likely explanation for the poor state of upgrading areas' infrastructure today is that those areas' infrastructure deteriorated more than that of de novo areas. Kironde (1994, page 464) and Theodory and Malipula (2012) discuss evidence that infrastructure did in fact deteriorate in upgrading slums in Dar es Salaam. Kironde (1994) mentions, for example, the deterioration of roadside drainage due to lack of maintenance; private construction on land that was intended for public use; and the degradation of water provision infrastructure.

Finally, Table 3a.12 shows that residents of upgrading areas are less educated than those of nearby areas, consistent with the lower housing quality in these neighborhoods.

3.4 Model

3.4.1 Assumptions and their relationship to the institutional setting

To frame our empirical analysis we present a model, which characterizes conditions under which investment in infrastructure (as defined below) incentivizes owners to build higher quality housing. The model captures key aspects to our description of the Sites and Services projects in Section 2. It also connects to our econometric analysis in Section 3, by relating gains in house values to gains in land values, and motivating our use of owner fixed effects to account for owner sorting across neighborhoods.[41]

We consider a population of infinitely lived, profit maximizing owners, with formal or informal rights to build on their plot(s), which are organized into neighborhoods (areas). In each plot, the owner can build a house and rent it out.[42] The model is in discrete time, and in each period $t \geq 1$, owners maximize their expected present discounted stream of

[41] Our model builds on Hornbeck and Keniston (2017), but differs from theirs in several ways. We add to the model infrastructure and variation across owners in credit constraints, and we derive new analytical results. We also model spillovers across houses differently, and for simplicity we exclude the exogenous time trends.

[42] A "house" in the model denotes is a shorthand for a housing unit that we consider in the empirical analysis. Unlike Bayer et al. (2007), our model does not account for renter heterogeneity, because we have no data on the rents paid and have little information on the residents. Knowing more about renters would have allowed to build a better picture of the welfare gains from de novo areas.

rents, net of house construction costs, on each plot they own:

$$E\left[\sum_{s=t}^{\infty}\delta^{t}\left[r\left(q_{t},I_{t}\right)-B_{t}c\left(q\left(I_{t}\right)\right)\right]\right],s.t.\Pr\left(q_{t+1}=q_{t}\right)=1-d,\Pr\left(q_{t+1}=0\right)=d. \quad (3.2)$$

The expectations are defined over the exogenous destruction probability of houses in each period, as discussed below. Owners are assumed to have a time preference $\delta \in (0,1)$. The rent that each owner receives on each house in each period is $r(q,I) = q^{\alpha}I^{1-\alpha}$, where q and I denote the quality of the house and the neighborhood infrastructure, and $\alpha \in (0,1)$. B_t is an indicator equal to one if a house is built in period t and zero otherwise. The construction costs of a house of quality q are: $c(q) = cq^{\gamma}$, where $c > 0, \gamma > 1$. This convex cost function generalizes Hornbeck and Keniston (2017), who assume $\gamma = 2$. In a different context, Combes, Duranton and Gobillon (2016) finds that the production function for housing can be approximated by a constant returns to scale Cobb-Douglas function using land and other inputs, where the coefficient on non-land inputs is approximately 0.65. Holding land constant, this production function is consistent with a cost function $c(q) = cq^{\gamma} = cq^{1/0.65}$, or $\gamma \simeq 1.54$.

Infrastructure captures a broad set of neighborhood characteristics, including formal and regularly laid out plots, which reduce coordination failures and protect owners' property rights; roads, which reduce the cost of travel and trade; and water mains, which contribute to living standards and health.[43] Infrastructure also reflects other neighborhood level effects.[44] For tractability, we consider three types of infrastructure: high quality (I_H), medium quality (I_M), and low quality (I_L), where $I_H > I_M > I_L > 0$. High quality describes the bundle that Sites and Services offered - mostly formal plots, roads, and water mains. We assume that high quality infrastructure deteriorates to medium quality unless the fraction of high quality housing is larger than a constant $\phi > 0$.[45] Medium quality infrastructure is basic and unmaintained (e.g. bumpy dirt roads). It may be either high quality infrastructure that has deteriorated or it may start out as medium

[43] Property rights protection may reduce the risk of outright expropriation, as we discuss below, as well as the risk of partial expropriation, when part of an owner's plot is built without authorization, which we do not model explicitly.

[44] In practice, other types of neighborhood effects may also matter. For example, the absence of proper sewerage may increase the risk of contagious diseases. Consistent with this, Jaupart et al. (2016) show that cholera outbreaks in Dar es Salaam were much more severe in slum areas with poor infrastructure. Another possibility is that neighborhoods with poor electrification and lighting (Painter and Farrington 1997) and high population density (Gollin et al. 2017) may attract crime. While we think that both of these channels could amplify the land value differentials between neighborhoods, we do not have the data to study them in our context.

[45] High quality housing is $q_H = q(I_H)$, as defined below. The potential for infrastructure deterioration means that owners' housing quality can be indirectly affected by those of their neighbors, through the effect on infrastructure. This mechanism is different from the direct impact of neighbors' housing quality in Hornbeck and Keniston (2017).

quality. We assume that medium quality infrastructure does not deteriorate.[46] Low quality infrastructure corresponds to the level that prevails without any infrastructure investments in the neighborhood.

There are two types of owners in the model. Unconstrained owners may each own any finite number of plots and afford any level of investment in each plot, while Constrained owners may own no more than a single plot, and may afford to build at most low quality housing $q_L = q(I_L)$, as defined below.[47] Consistent with our setting, we assume that no single owner has a sufficiently large number of plots to exert market power or to solve coordination problems that arise from neighborhood-level externalities.[48]

We consider three types of areas (neighborhoods), each with a continuum of plots.[49] De novo areas start with empty plots ($q = 0$) and high quality infrastructure (I_H); control areas start with empty plots and medium quality infrastructure (I_M); and upgrading areas start out with low quality housing (q_L) and high quality infrastructure.[50] This reflects the situation at the time when Sites and Services was implemented.[51]

The initial fractions of unconstrained owners are: θ_D in de novo areas, θ_C in control areas, and θ_U in upgrading areas. We assume that (as in the real world) upgrading areas are targeted for their relatively poor population, so they have few unconstrained owners, and therefore $\theta_U < \phi$.

In every period, the following sequence of events takes place. First, each owner decides whether to build (or rebuild) a house on each plot they own.[52] Second, if the neigh-

[46] Our assumption that medium infrastructure and deteriorated infrastructure are equal in quality is a simplifying assumption, motivated by our empirical finding that upgrading areas are no better than nearby control areas in terms of access to roads and water. Adding further parameters for deteriorated high quality and deteriorated medium quality infrastructure would not add much insight to the model.

[47] The distinction between two types of owners allows us to analyze owner sorting in a simple way. The results would have been similar if we had assumed that constrained owners could build up to any quality that is strictly lower than $q(I_M)$, as defined below.

[48] Our TSCP data indicate that only a small share of housing units are owned by those with more than a handful of plots. It is true that in principle a rich individual or a firm could buy up an entire neighborhood and internalize the externalities involved. But until recent years the Tanzanian government exerted strict control that prevented the concentration of neighborhood ownership.

[49] Our model does not account for other types of neighborhoods, such as former colonial areas (which typically constitute a small and wealthy part of cities), nor do we consider movements between different neighborhoods within the city.

[50] In Section 2 we discuss the investments that were made as part of the Sites and Services projects. These suggest that though the investment per total land area in de novo and upgrading were similar.

[51] We also note that while the control areas we use were by definition empty to begin with, other areas looked like control areas but had a stock of low quality housing by the time they received infrastructure I_M.

[52] Following Hornbeck and Keniston (2017) and Henderson et al. (2017), we assume that owners cannot renovate incrementally, and that houses do not depreciate. The assumption that rebuilding a higher quality house requires a fresh start is particularly relevant for low quality housing that characterizes poorer neighborhoods in East African cities. It may be possible to make minor improvements to a house built of tin or mud walls. However, demolition and construction from scratch is required to make meaningful improvements such as adding brick walls, multiple stories, or plumbing. For simplicity, we maintain the

borhood's housing quality is insufficiently high, infrastructure quality deteriorates, as we discuss below. Third, each owner collects the rent on each house they own. Finally, there is an exogenous probability $d > 0$ that each house is destroyed, resetting housing quality to zero.[53]

We assume that the risk that houses are destroyed and the fraction of owners of each type in each neighborhood are common knowledge, as is the understanding that all unconstrained owners will build high quality housing if the share of unconstrained owners is at least ϕ. In Nash Equilibrium, each owner solves her maximization problem in each period, assuming that all other owners do the same.

3.4.2 Solving the model

This section characterizes the optimal level of investment by owners, beginning with unconstrained owners and then by constrained owners.

Unconstrained owners maximize profits on each plot they own by solving the following Bellman equation:

$$V(q,I) = Max \begin{cases} r(q,I) + \delta E[V(q,I)] \\ r(q(I),I) + \delta E[V(q(I),I)] - c(q(I)), \end{cases} \tag{3.3}$$

where r is return on house (e.g. rent), $q \geq 0$ is the house quality; $I \geq 0$ is the infrastructure quality which is expected when rents are collected and from that point onward; $q(I)$ is the optimal house quality; and $c(q(I))$ is the cost of building a house of quality $q(I)$.[54]

The infrastructure quality which is anticipated when rents are collected and from that point onward is equal to the existing level, except where infrastructure of quality I_H deteriorates to I_M. This deterioration happens when the fraction of high quality housing ($q_H = q(I_H)$, as described below) is strictly lower than ϕ.

assumption that no incremental improvement is possible. Relaxing this would reduce the benefit of early (de novo) investments.

[53] If a house is destroyed, the owner retains their plot. Given the paucity of construction dates in our data, it is difficult to assess d. But Henderson et al. (2017) estimate it at 3.2 percent per year using data from Tanzania's neighbor, Kenya.

[54] We could have included a probability $(1-\psi)$ that a plot is fully expropriated at the end of each period. If that were the case we would need to substitute $\psi\delta$ instead of δ throughout the analysis, but for simplicity we focus on the case without expropriation, namely $\psi = 1$. Higher patience may reflect, at least in part, a lower risk of expropriation. Collin et al. (2015) elicit owners' perceived expropriation risk in Temeke, an informal area close to the CBD of Dar es Salaam, which implies a risk of around 8% per year. Given the setting, this is likely an upper bound to the perceived expropriation risk in the locations we study.

The model reflects a tradeoff between keeping the current house quality q and improving it to $q(I)$. But if an unconstrained owner's house is exogenously destroyed it is always rebuilt at the optimal quality $q(I)$. Starting from an empty plot, the optimal house quality for an unconstrained owner anticipating infrastructure I at the time of rent collection is:

$$q(I) = \left[\frac{\alpha I^{1-\alpha}}{\gamma c(1-\delta+d\delta)}\right]^{\frac{1}{\gamma-\alpha}}. \tag{3.4}$$

The quality of housing is characterized by the following comparative statics. First, $\frac{\partial q(I)}{\partial \delta} > 0$, so more patient people invest more. Second, $\frac{\partial q(I)}{\partial d} < 0$, so a higher probability of house destruction leads to lower quality housing. And finally, $\frac{\partial q(I)}{\partial c} < 0$, so a higher construction cost reduces housing quality.

If an unconstrained owner starts with housing $q_1 \equiv q(I_1)$ but with infrastructure I_2 (where $I_2 > I_1$), they choose between two options.[55] They can replace their house with a higher quality house, in which case their expected payoff is equal to the expected value of an unbuilt a plot of land:

$$\pi(0, I_2) = \frac{q_2^{\alpha} I_2^{1-\alpha} - cdq_2^{\gamma}}{1-\delta} - (1-d)cq_2^{\gamma}, \tag{3.5}$$

where $\pi(q, I)$ is the maximized expected payoff from an existing house of quality q and infrastructure quality I. Alternatively, they can keep the current quality q_1 and only build a better house when their house needs rebuilding. In this case their expected payoff is:

$$\pi(q_1, I_2) = q_1^{\alpha} I_2^{1-\alpha} + \delta\left[(1-d)\pi(q_1, I_2) + d\pi(0, I_2)\right]. \tag{3.6}$$

Solving this expression we get:

$$\pi(q_1, I_2) = \frac{q_1^{\alpha} I_2^{1-\alpha} + d\delta\pi(0, I_2)}{1-\delta+d\delta}. \tag{3.7}$$

Proposition 1. *For each level of infrastructure $I_1 > 0$, there exists a unique value $I_1^{crit} = \left(\frac{\gamma}{\gamma-\alpha}\right)^{\frac{\gamma-\alpha}{\alpha(\alpha-1)}} I_1$, such that unconstrained owners starting with $q_1 = q(I_1)$ and infrastructure $I_2 = I_1^{crit}$ are indifferent between rebuilding and not rebuilding, and owners rebuild if and only if $I_2 > I_1^{crit}$.*

Proof. To obtain $I_2 = I_1^{crit}$, combine the condition $\pi\left(q_1, I_1^{crit}\right) = \pi\left(0, I_1^{crit}\right)$ with (3.5) and (3.6), where housing quality $q_2 = q(I_2)$ comes from (3.4). To show that owners rebuild if

[55] We assume that if owners are indifferent they do not improve their houses.

and only if $I_2 > I_1^{crit}$, note that $\frac{\partial}{\partial I_2}(\pi_{I_2} - \pi_{q_1,I_2}) > 0$. □

This result implies that unconstrained owners face what we refer to as an "inaction zone", $(I_1, I_1^{crit}]$. If infrastructure is upgraded from I_1 to a level in the inaction zone, owners will not improve their house right away, but only when it is exogenously destroyed. But if the infrastructure upgrade is to $I_2 > I_1^{crit}$, unconstrained owners will rebuild at a higher quality q_2 right away.

The investment problem for constrained owners is similar to that of unconstrained owners, except that the maximum quality they can build is q_L. As a result, in equilibrium they build q_L if their plot is empty, and otherwise they do not rebuild.

3.4.3 Neighborhood development

De novo areas

De novo areas begin empty with infrastructure (I_H). Constrained owners build q_L, so that the share of low quality houses is $1 - \theta_D$. If $\theta_D \geq \phi$ then there is no deterioration in equilibrium, anticipated infrastructure is I_H, and the unconstrained owners build q_H. If $\theta_D < \phi$ then there is deterioration in equilibrium, anticipated infrastructure is I_M, and unconstrained owners build $q_M = q(I_M)$. In practice it seems that de novo areas' infrastructure is better than other areas' (Table 3.4), suggesting that at least some higher quality infrastructure survived.

Control areas

Control areas begin empty and with medium quality infrastructure (I_M). Unconstrained owners build housing quality q_M, while constrained owners build q_L.

As discussed above, Tanzanian cities also contained areas (which are not part of our main analysis), which are similar to control areas but had a stock of low quality housing by the time they received infrastructure I_M. In those areas the constrained owners keep q_L, while the unconstrained owners either build q_M right away (if $I_M > I_L^{crit}$) or otherwise build q_M only when their house is destroyed.

Upgrading areas

Upgrading areas begin with housing quality q_L and infrastructure I_H, and we consider four different cases. In the first case $I_L^{crit} > I_H$, so the upgrading is minimal and all owners initially keep q_L, and infrastructure deteriorates to I_M; in later periods, as houses are exogenously destroyed, unconstrained owners build to q_M. In the second case $I_H \geq I_L^{crit} > I_M$ and $\theta_U < \phi$, in which case everyone initially keeps q_L, infrastructure deteriorates to I_M; and unconstrained owners improve their houses to q_M when they are destroyed. In the third case $I_M \geq I_L^{crit}$ and $\theta_U < \phi$, in which case unconstrained owners build q_M right away while constrained owners keep q_L, and infrastructure deteriorates to I_M. In the final case $I_H \geq I_L^{crit}$ and $\theta_U \geq \phi$, so unconstrained owners build q_H and infrastructure remains I_H, while constrained owners keep q_L. But in practice this final case is unlikely to be relevant, because upgrading areas were targeted as poor.

3.4.4 Relating the model to the empirical analysis

The model demonstrates the role of differing infrastructure investment and owner sorting in accounting for neighborhood quality. For example, consider the following scenario. De novo areas had enough unconstrained owners to ensure that their higher quality infrastructure (I_H) survived. In this case, the difference in the logarithm of mean housing quality between de novo and control areas is:

$$\ln(\theta_D q_H + (1 - \theta_D)\, q_L) - \ln(\theta_C q_M + (1 - \theta_C)\, q_L). \tag{3.8}$$

This quality difference reflects both the effect of the higher infrastructure quality in de novo areas and the different composition of owners in those areas. But controlling for owner fixed effects allows us to focus on houses owned by unconstrained owners, for whom the difference in log mean housing quality between de novo and control areas is:

$$\ln(q\,(I_H)) - \ln(q\,(I_M)). \tag{3.9}$$

In other words, under the model's assumptions, adding owner fixed effects allows us to identify the effect of de novo investments on housing quality for unconstrained owners. We acknowledge that in practice adding owner fixed effects may not solve all the potential problems, if for example some owners are constrained in investing in a second house (but

not in the first), or have some different preferences for investing across areas. Nevertheless, the model shows that adding owner fixed effects is useful in the context of Sites and Services, where owners in different areas may have had different levels of wealth, due both to sorting and to the program's loans scheme.

The model also allows us to relate differences in infrastructure and housing quality, which we cannot measure directly, to the estimated differences in the value of housing, which are approximated by the hedonic regressions, subject to the limitations discussed in Section 2. Specifically, our model predicts the following:

Proposition 2. *For unconstrained owners who face no risk of exogenous house destruction* $(d = 0)$

$$\ln(I_H) - \ln(I_M) = \frac{\gamma - \alpha}{\gamma - \alpha\gamma} \left(\ln(\pi(q(I_H), I_H) - \ln(\pi(q(I_M), I_M)\right), \tag{3.10}$$

and

$$\ln(q(I_H)) - \ln(q(I_M)) = \frac{1}{\gamma} \left(\ln(\pi(q(I_H), I_H) - \ln(\pi(q(I_M), I_M)\right) \tag{3.11}$$

Proof. To derive the expression for $\ln(I_H) - \ln(I_M)$, use (3.5) and the fact that $\pi(q_2, I_2) = \pi(0, I_2) + c(q_2)$, and plug in $d = 0$ to obtain $\ln(\pi(q_2, I_2)) = \ln(q_2^{\alpha} I^{\frac{1-\alpha}{2}}) - \ln(1-\delta)$. Next apply a similar calculation for $\ln(\pi(q_1, I_1))$ and plug in (3.4) to calculate $\ln \pi(q_2, I_2) - \ln(\pi(q_1, I_1)$. Now combine the expression for $\ln(I_H) - \ln(I_M)$ with (3.4) to derive the expression for $\ln(q(I_H)) - \ln(q(I_M))$. □

This result indicates that the difference across areas in log housing quality are smaller than the differences in log values. Taking the above-mentioned estimate of γ suggests that the quality differences across neighborhoods are about $\frac{1}{\gamma} = 0.65$ times the value differences for unconstrained owners, for low values of d, which seem empirically relevant. Our baseline estimate of the hedonic log value differences between de novo and control areas, with owner fixed effects, are around 0.5, suggesting log quality differences of around one third.[56]

The model also allows us to consider differences between upgrading and control areas. As discussed in Section 3, upgrading areas look similar, or in some cases worse than control areas. Table 3a.10 suggests that the worse housing in upgrading areas may in part be explained by owner fixed effects. In the context of the model, this may reflect persistence in upgrading areas of some of the initial owners (or their descendants), who were targeted by the program, and may have been poorer than their counterparts in control areas ($\theta_U < \theta_C$).

[56] As discussed in Section 2, the log value differences in the hedonic regressions may understate the actual value differences.

Finally, the similarity of housing quality in upgraded and non-upgraded slums is also consistent with the model, if we think of the non-upgraded slums as control areas with constrained owners.

3.4.5 Implications of the model

The model offers several implications for thinking about infrastructure investments for housing. First, an important theme of the paper is that infrastructure investment may crowd in private investments. The model helps us to think about the conditions under which this takes place. In the model, infrastructure investments crowd in more private investments when its quality is sufficiently high and owners can afford to invest in housing quality. In these cases, private investment in housing quality takes place when there is a sufficient fraction of unconstrained owners, either due to their own wealth or through loans that allow them to invest. This also suggests a note of caution: if de novo investments were expanded widely, poor and credit constrained residents may be unable to make full use of them, since infrastructure may deteriorate without sufficient complementary private investment.

Second, the model helps us think about the benefits of early infrastructure investments compared to ex-post infrastructure upgrading. Upgrading areas do not always fully benefit from high levels of infrastructure investments, since in those settings infrastructure either deteriorates or leads to the scrapping of existing houses.[57]

Finally, turning back to our empirical findings, the model can help explain why infrastructure survived better in de novo areas, but not in upgrading areas. The model highlights the importance of feedback from owner investments to infrastructure, which can be seen as a neighborhood externality, and is sometimes overlooked when infrastructure investments are made.

3.5 Concluding remarks

This paper examines the consequences of different strategies for developing basic infrastructure for residential neighborhoods. Specifically, we study the Sites and Services

[57] In reality there are other costs of delivering infrastructure in a dense settlement that has developed organically, because it is difficult to resolve coordination failures and negative externalities once they have been put in place. In Sites and Services, for example, the cost per square meter was similar in de novo and upgrading areas, even though de novo areas received formal plots, which upgrading areas did not.

projects implemented in seven Tanzanian cities during the 1970s and 1980s. These projects provided basic infrastructure, leaving it to the residents to build their own houses. We examine the long run development of these neighborhoods, emphasizing the comparison between de novo neighborhoods and other nearby areas that were greenfields when the Sites and Services program started. We also provide descriptive evidence on the development of neighborhoods whose infrastructure was upgraded.

We use high-resolution imagery and building level survey data to study housing quality and infrastructure in the de novo neighborhoods and other areas in their vicinity that were also greenfields to begin with. We find that the de novo neighborhoods developed significantly higher quality housing than other initially unbuilt areas. Our findings reflect complementary private investments that were made in response to the Sites and Services programs. We also present evidence that the initial infrastructure investments in roads and water mains were more likely to persist in de novo areas. For three cities where we have survey data, we find sizeable gains in quality from de novo even when we control for owner fixed effects, although these fixed effects account for up to a third of the average housing quality. Our findings suggest that de novo areas increased local land values by at least 75-100 USD per square meter, compared to total costs of no more than 8-13 USD per square meter (all in 2017 prices).

We also report evidence that de novo neighborhoods attract more educated residents, who can afford to pay for the higher quality on offer. But as of 2012 almost half of the adults in de novo areas still had no more than primary school education, suggesting that some people with lower lifetime incomes also benefitted from the de novo investments. But we also note that de novo areas were unaffordable to the poorest of the urban poor, a consideration that future projects may want to take into account, perhaps by creating some smaller and more affordable plots. Such plots may also benefit the few people who may be displaced by such projects, even when they target largely empty areas.

Our paper also reports descriptive evidence on upgrading areas, comparing them to nearby control areas, or where the data permit to slums that were not upgraded. The results suggest that upgrading areas now have either similar, or worse, housing quality, and the program's investments in roads and water mains did not survive well in upgrading areas. While we should be cautious in interpreting these results, they suggest that upgrading, at least as implemented in Sites and Services, was not a panacea for pre-existing squatter areas. We cannot rule out that other upgrading efforts may be prove more successful, but in order to provide long lasting benefits, upgrading programs should aim to address the

risk of infrastructure deterioration.

Taken together, our findings suggest that de novo investments are a policy tool worthy of consideration for growing African cities. They are considerably cheaper than building public housing, and therefore more affordable for poor countries. They also offer important advantages to residents, who can invest in higher quality housing. Our findings also suggest that it is important to ensure that the infrastructure investments do not deteriorate as a result of poor private investments. While the implementation of Sites and Services projects in Tanzania in the 1970s and 1980s was not flawless, it has taught us important lessons. We hope that these lessons can inform future planning and investment decisions in a continent that is growing in both population and income per capita, but where many poor people still live in poor quality buildings and neighborhoods.

3.6 Main Tables

Table 3.1: De novo regressions using imagery data for all seven cities

	(1)	(2)	(3)	(4)	(5)	(6)
	Mean log building footprint area	Share of buildings with painted roof	Mean similarity of building orientation	Mean z-index	Share of empty blocks	Share of area built up
Panel A: 500m bandwidth						
De novo	0.114	-0.013	2.821	0.168	-0.152	0.094
	(0.051)	(0.012)	(0.722)	(0.057)	(0.037)	(0.013)
Observations	6,562	6,500	6,562	6,562	8,440	8,440
Mean (control)	4.457	0.184	-8.669	0.042	0.306	0.155

	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade
Panel B: robustness (mean z-index only as outcome)					
De novo	0.143	0.156	0.168	0.241	0.175
	(0.053)	(0.057)	(0.057)	(0.100)	(0.059)
Observations	6,562	6,562	6,562	4,568	6,158
Mean (control)	0.042	0.042	0.042	0.015	0.047

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities. The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(3) in Panel A). In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panel B, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, and column (5) excludes areas within 100 meters of the boundary between upgrade and control areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 90 clusters, except in columns (5) and (6) of Panel A, which have 92 clusters, and column (4) of Panel B, which has 89 clusters, and column (5) of Panel B, which has 88 clusters.

Table 3.2: De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean log building footprint area	Share of buildings with multiple storeys	Share of buildings with a good roof	Share of buildings connected to electricity	Share of buildings with sewerage or septic tank	Mean z-index	Mean log hedonic value
Panel A: 500m bandwidth							
De novo	0.405	0.081	-0.010	0.226	0.142	0.342	0.446
	(0.070)	(0.066)	(0.008)	(0.039)	(0.091)	(0.091)	(0.081)
Observations	2,009	1,975	2,009	2,009	2,008	2,009	2,009
Mean (control)	4.739	0.096	0.984	0.466	0.381	0.033	17.234
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Full City	Optimal bandwidth
Panel B: robustness (mean z-index only as outcome)							
De novo	0.263	0.323	0.342	0.408	0.375	0.588	0.312
	(0.091)	(0.076)	(0.090)	(0.190)	(0.093)	(0.079)	(0.082)
Observations	2,009	2,009	2,009	1,410	1,887	34,602	34,602
Mean (control)	0.033	0.033	0.033	0.001	0.022	-0.149	0.038
Panel C: robustness (mean log hedonic value only as outcome)							
De novo	0.329	0.431	0.446	0.541	0.427	0.505	0.411
	(0.081)	(0.059)	(0.077)	(0.190)	(0.077)	(0.089)	(0.063)
Observations	2,009	2,009	2,009	1,410	1,887	34,602	34,602
Mean (control)	17.234	17.234	17.234	17.231	17.229	17.113	17.239

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A), and in Panel C the dependent variable is the predicted log value from hedonic regressions. In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B and C, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, column (5) excludes areas within 100 meters of the boundary between upgrade and control areas, column (6) changes the control area to the sample of blocks covering the whole city excluding de novo areas, and column (7) uses 2033 observations inside the optimal bandwith for panel B and 1882 observations inside the optimal bandwith for panel C based on Imbens and Kalyanaraman (2012). The 'Full City' in column (6) is robust to higher order polynomials in distance to boundary: In panel B: second order polynomial gives an estimate of 0.571 and standard error of 0.087, and third order polynomial gives an estimate of 0.381 and standard error of 0.098. In panel C: second order polynomial gives an estimate of 0.498 and standard error of 0.104, and third order polynomial gives an estimate of 0.296 and standard error of 0.117. The control mean in column (7) reports the mean for the control areas inside the optimal bandwidth (Imbens and Kalyanaraman 2012). Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 29 clusters, except in column (5) of Panels B and C, which has 28 clusters, and column (6) of Panels B and C, which has 439 clusters.

Table 3.3: De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with owner name fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log building footprint area	Multistorey building	Good roof	Connected to electricity	Sewerage or septic tank	Z-index	Log hedonic value
Panel A: Full City, Owner FE							
De novo	0.553	0.119	-0.002	0.417	0.123	0.447	0.604
	(0.145)	(0.062)	(0.038)	(0.078)	(0.091)	(0.088)	(0.133)
Observations	20,177	16,605	20,054	20,139	19,595	20,177	20,177
Mean (control)	4.573	0.164	0.968	0.404	0.249	-0.016	17.016
Panel B: Full City, no Owner FE, same sample as A							
De novo	0.594	0.514	-0.010	0.405	0.122	0.642	0.612
	(0.177)	(0.138)	(0.015)	(0.066)	(0.069)	(0.086)	(0.185)
Observations	20,177	16,605	20,054	20,139	19,595	20,177	20,177
Mean (control)	4.573	0.164	0.968	0.404	0.249	-0.016	17.016
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Second Order Polynomial	Third Order Polynomial
Panel C: robustness owner FE (z-index only as outcome)							
De novo	0.422	0.411	0.434	0.336	0.471	0.431	0.372
	(0.085)	(0.092)	(0.087)	(0.166)	(0.093)	(0.112)	(0.140)
Observations	20,177	20,177	20,177	19,694	19,729	20,177	20,177
Mean (control)	-0.016	-0.016	-0.016	-0.019	-0.018	-0.016	-0.016
Panel D: robustness, no owner FE, same sample as C (z-index only as outcome)							
De novo	0.616	0.648	0.631	0.654	0.675	0.674	0.627
	(0.086)	(0.089)	(0.084)	(0.134)	(0.102)	(0.081)	(0.092)
Observations	20,177	20,177	20,177	19,694	19,729	20,177	20,177
Mean (control)	-0.016	-0.016	-0.016	-0.019	-0.018	-0.016	-0.016

Notes: This table reports estimates from regressions using specification (1) and unit level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and the entire city as control areas. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a property unit in a building, and only multi-unit owners are used. Units are assigned to de novo or control areas based on where their building's centroid falls. Outcomes are measured at the building level (see Data Appendix for further details). In Panels A and B the outcomes vary, while in Panels C and D the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A). Panels A and C display results with unit owner last name fixed effects, including units inside de novo and control areas but restricting the sample by keeping only last name owners that appear more than once in the sample. Panel B (D) displays results with the same sample as in A (C) but without owner last name fixed effects. In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, column (5) excludes areas within 100 meters of the boundary between upgrade and control areas, and columns (6) and (7) control for second and third order polynomials in distance to the boundary, respectively. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 342 clusters, except in column (2) of Panels A and B and in columns (4) and (5) of Panels C and D, which all have 341 clusters.

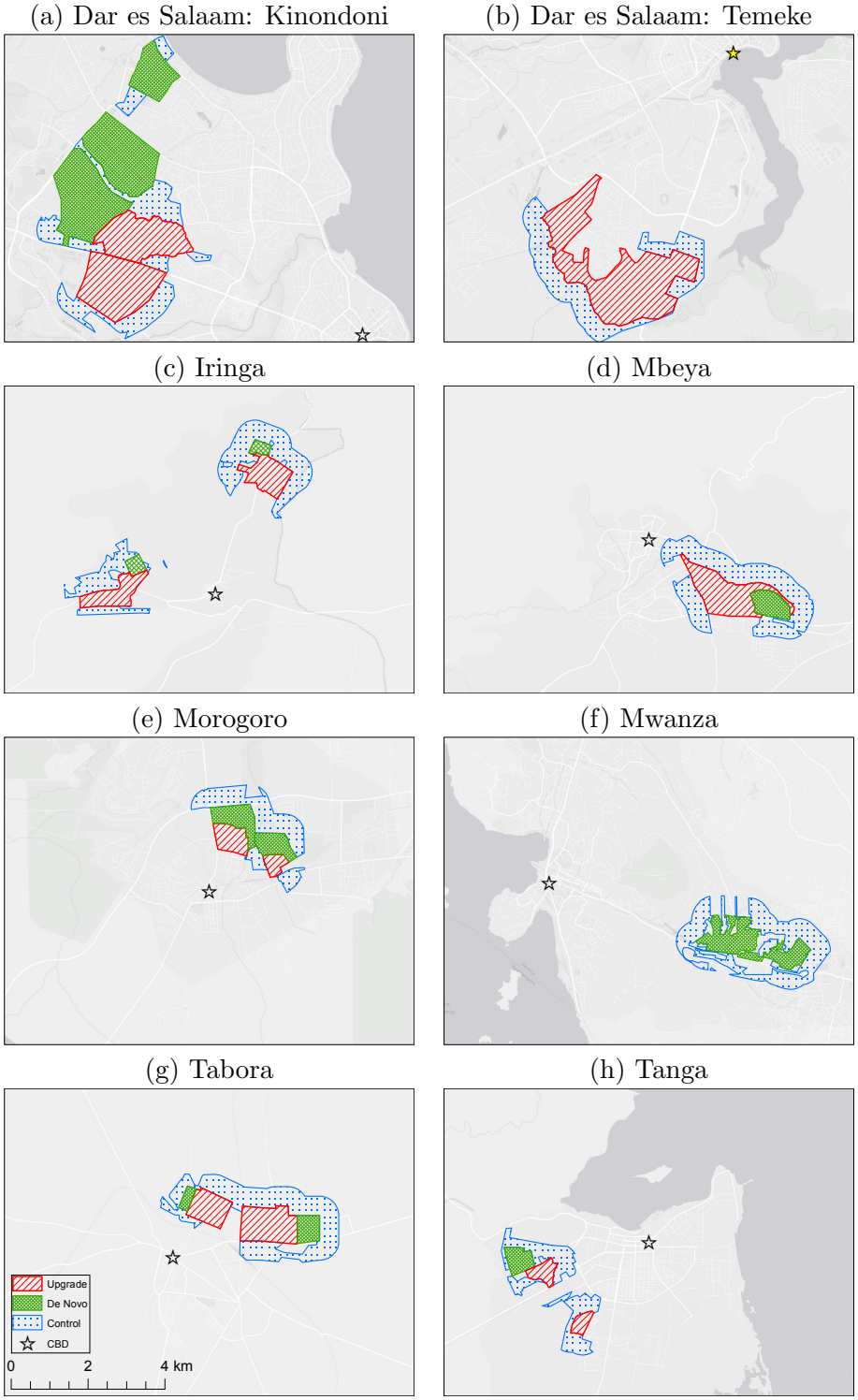
Table 3.4: De novo regressions on persistence measures using imagery and TSCP survey data

	(1)	(2)	(3)	(4)	(5)	(6)
	Imagery	TSCP Survey		TSCP Survey, Excl. Tanga		
(lr)2-2 (lr)3-4 (lr)5-5	Share of buildings with road within 10m	Share of buildings with road access	Share of buildings connected to water mains	Share of buildings connected to water mains		
Panel A: 500m bandwidth						
De novo	0.141	0.197	0.211	0.233		
	(0.028)	(0.050)	(0.057)	(0.060)		
Observations	6,562	2,008	2,009	1,952		
Mean (control)	0.202	0.477	0.547	0.547		
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	
Panel B: robustness for share of buildings with road within 10m (Imagery)						
De novo	0.129	0.142	0.142	0.185	0.150	
	(0.025)	(0.029)	(0.028)	(0.056)	(0.029)	
Observations	6,562	6,562	6,562	4,568	6,158	
Mean (control)	0.202	0.202	0.202	0.205	0.197	
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Full City
Panel C: robustness for share of buildings with road access (TSCP)						
De novo	0.134	0.190	0.199	0.191	0.206	0.170
	(0.039)	(0.049)	(0.050)	(0.159)	(0.051)	(0.056)
Observations	2,008	2,008	2,008	1,409	1,886	34,578
Mean (control)	0.477	0.477	0.477	0.485	0.449	0.573
Panel D: robustness for share of buildings connected to water mains (TSCP)						
De novo	0.164	0.188	0.209	0.319	0.204	0.403
	(0.060)	(0.051)	(0.057)	(0.128)	(0.062)	(0.042)
Observations	2,009	2,009	2,009	1,410	1,887	34,588
Mean (control)	0.547	0.547	0.547	0.535	0.534	0.433

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities (road within 10m) and TSCP survey data for Mbeya, Mwanza, and Tanga (road access and connection to water mains). The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of persistence of infrastructure treatment. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the share of buildings with a road within 10 meters (from imagery data), in Panel C the dependent variable in all columns is the share of buildings with road access (from TSCP data), and in Panel D the dependent variable is the share of buildings connected to water mains (from TSCP data). In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B, C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, and column (5) excludes areas within 100 meters of the boundary between upgrade and control areas. Moreover, in Panels C and D, column (6) changes the control area to the sample of blocks covering the whole city excluding upgrade areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services area. There are 29 clusters in TSCP data, except in column (5) of Panels C and D, which have 28 clusters, and in column (6) of Panels C and D, which have 439 clusters. There are 90 clusters in imagery data, except in column (4) of Panel B which has 88 clusters, and column (5) of Panel B which has 89 clusters.

3.7 Appendix tables and figures

Figure 3a.1: Locations of de novo, upgrading, and control areas by city



Notes: This figure maps de novo (green cross-hatch), upgrading (red hatch), control areas (blue dots), and the CBD (yellow star) for each city. Panel (a) shows the northern part of Dar es Salaam (Kinondoni), while the southern part (Temeke) is shown in panel (b). Control areas are all 500m buffers of study areas, excluding land that was determined uninhabitable, built-up, or designated for specific use prior to the program. Each map is set to the same scale. Background imagery from ArcGIS is for context only and was not used for analysis, it depicts modern day roads (white lines), heavily vegetated areas (green-grey) and water bodies (dark grey).

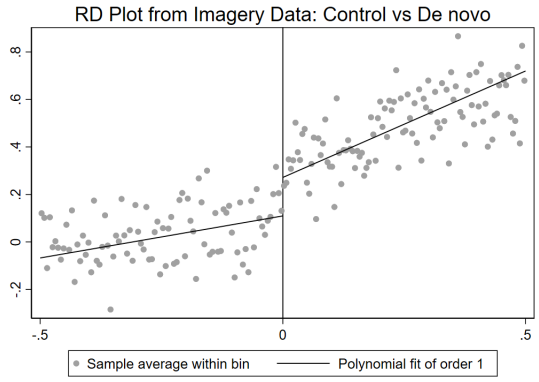
Figure 3a.2: Example images of de novo, upgrade and control areas

De-novo | Control (for de-novo) | Upgrading

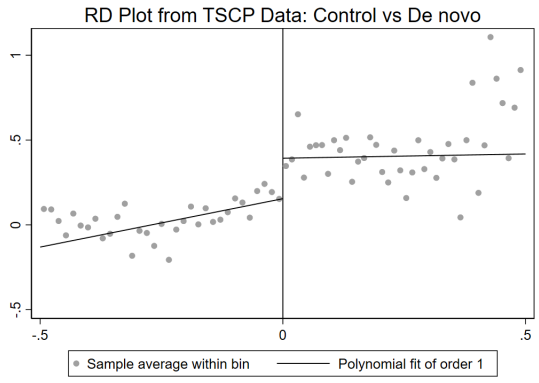
Notes: Each of the three images covers an area of approximately 440 x 360 meters. Source: Google earth V 7.1.2. (2018). Kinondoni District, Dar-es-Salaam, Tanzania.

Figure 3a.3: Regression discontinuity plots of summary outcomes from Tables 1 and 2

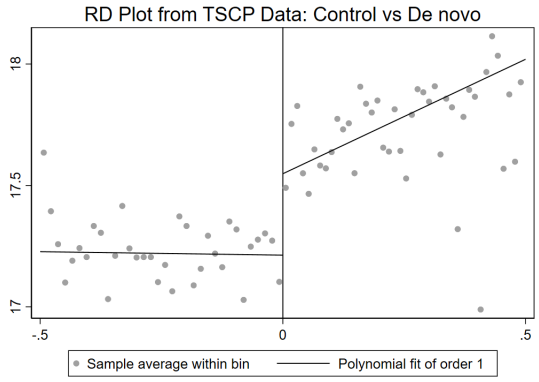
(a) Z-index from Table 1



(b) Z-index from Table 2



(c) Log hedonic value from Table 2



Notes: This figure plots the raw summary outcomes (z-index and log hedonic value) on the y-axis against the running variable (x-axis) being distance to boundary between de novo and control areas in kilometers. Observations from control areas are to the left of the cutoff (marked 0) and de novo areas to the right. The graphs are created with the command rdplot (Calonico et al 2017), using a triangular kernel. The sample is defined by a 500m bandwidth on either side of the boundary. For subfigure (a), the variable definition and data is the same as in Table 1, panel A, column 4 and the number of observations in (a) is 6562, with 3147 to the left (control) and 3415 to the right (de novo) of the cutoff. The average bin length is 6m to the left of the cutoff and 5m to the right for subfigure (a). For subfigure (b), the variable definition and data is the same as in Table 2, panel A, column 6 and the number of observations in (b) is 2009, with 1177 to the left (control) and 832 to the right (de novo) of the cutoff. The average bin length is 16m to the left of the cutoff and 12m to the right for subfigure (b). For subfigure (c), the variable definition and data is the same as in Table 2, panel A, column 7 and the number of observations in (c) is 2009, with 1177 to the left (control) and 832 to the right (de novo) of the cutoff. The average bin length is 14m to the left of the cutoff and 12m to the right for subfigure (c).

Table 3a.1: De novo neighborhoods

City	Area within city	Round	Pre-treatment satellite photos	Pre-treatment topographic map
Dar es Salaam	Sinza	1	1966	N
Dar es Salaam	Kijitonyama	1	1966	N
Dar es Salaam	Mikocheni	1	1966	N
Mbeya	Mwanjelwa (*)	1	1966	N
Mwanza	Nyakato (**)	1	1966	N
Tanga	Nguvu Mali (***)	2	1966	N
Tabora	Isebya	2	1978	1967
Tabora	Kiloleni	2	1978	1967
Morogoro	Kichangani	2	N	1974
Morogoro	Msamvu	2	N	1974
Iringa	Kihesa & Mtuiwila	2	1966	1982
Iringa	Mwangata	2	1966	1982

Notes: This table reports information about the 12 de novo neighborhoods, the round in which the Sites and Services projects were implemented, and the data we have on the areas before the program was implemented. (*) Treatment area maps were unavailable, so areas were drawn by experts that were involved in the projects, as explained in the Data Appendix. (**) Treatment area maps were unavailable, so we inferred from the detailed Mwanza central plan. (***) We have some uncertainty as to the extent of infrastructure that was actually provided in Nguvu Mali.

Table 3a.2: Upgrading neighborhoods

City	Area within city	Round	Pre-treatment satellite photos	Pre-treatment topographic map
Dar es Salaam	Manzese A	1	1966 & 1969	N
Dar es Salaam	Manzese B	1	1966 & 1969	N
Mbeya	Mwanjelwa (*)	1	1966	N
Dar es Salaam	Mtoni & Tandika	2	1966	N
Iringa	Kihesa	2	1966	1982
Iringa	Mwangata	2	1966	1982
Morogoro	Kichangani	2	N	1974
Morogoro	Msamvu	2	N	1974
Tabora	Isebya	2	1978	1967
Tabora	Kiloleni	2	1978	1967
Tanga	Gofu Juu & Mtuiwila	2	1966	N
Tanga	Mwakizaro	2	1966	N

Notes: this table reports information about the 12 upgrading neighborhoods, the round in which the Sites and Services projects were implemented, and the data we have on the areas before the program was implemented. (*) Treatment area maps were unavailable, so areas were drawn by experts that were involved in the projects, as explained in the Data Appendix.

Table 3a.3: Plot counts and population by project type

		Plots completed by 1980s	Population in 2002	Ratio of population to plots completed	Area (sq-km)	Population density (people per sq-km)	Built area (building footprints, sq-km)	Crowding (people per sq-km of) built area)
Round 1	De novo	8,527	89,207	10.5	8.6	10,400	2.7	32,975
	Control (DN)		44,846		6.7	6,723	1.5	29,151
	Upgrading	14,634	200,630	13.7	6.5	31,064	2.9	68,084
	Control (U)		89,920		6.2	14,415	2.0	44,849
Round 2	Denovo	1,978	17,927	9.1	2.5	7,158	0.5	36,883
	Control (DN)		14,708		6.5	2,253	0.6	23,976
	Upgrading	20,128	204,074	10.1	10.5	19,483	3.2	64,721
	Control (U)		67,871		11.7	5,801	1.9	36,593
Total	Denovo	10,505	107,134	10.2	11.1	9,667	3.2	33,570
	Control (DN)		59,554		13.2	4,512	2.2	27,676
	Upgrading	34,762	404,704	11.6	16.9	23,900	6.1	66,346
	Control (U)		157,791		17.9	8,796	3.9	40,882

Notes: This table reports completed plot counts and population in 2002 by treatment type and round. Control (DN) is control for de novo, while Control (U) is control for upgrading area.

Table 3a.4: Summary statistics

	Imagery data (Blocks)			
	De novo	Upgrade	Control	Total
Mean log building footprint area	4.580 (0.569)	4.243 (0.503)	4.381 (0.699)	4.394 (0.625)
Share of buildings with painted roof	0.337 (0.314)	0.186 (0.222)	0.174 (0.266)	0.221 (0.277)
Mean similarity of building orien-tation	-4.735 (5.751)	-6.981 (5.208)	-8.202 (7.638)	-6.911 (6.657)
Share of buildings with road within 10m	0.288 (0.322)	0.213 (0.277)	0.202 (0.307)	0.228 (0.305)
Obs.	3,925	4,341	6,380	14,646

	TSCP data (Blocks)			
	De novo	Upgrade	Control (Full City)	Total
Mean log building footprint area	5.134 (0.464)	4.612 (0.456)	4.706 (0.688)	4.712 (0.684)
Share of buildings with multiple storeys	0.202 (0.384)	0.015 (0.100)	0.071 (0.240)	0.072 (0.243)
Share of buildings with a good roof	0.975 (0.109)	0.868 (0.268)	0.951 (0.174)	0.950 (0.175)
Share of buildings connected to electricity	0.713 (0.344)	0.423 (0.322)	0.425 (0.431)	0.430 (0.429)
Share of buildings with sewerage or septic tank	0.547 (0.412)	0.227 (0.328)	0.387 (0.431)	0.387 (0.430)
Share of buildings connected to water mains	0.767 (0.320)	0.493 (0.329)	0.483 (0.434)	0.488 (0.433)
Share of buildings with road access	0.676 (0.440)	0.748 (0.341)	0.611 (0.453)	0.615 (0.451)
Mean log hedonic value	17.689 (0.496)	17.039 (0.468)	17.200 (0.723)	17.207 (0.719)
Obs.	798	729	40,563	42,090

Notes: Summary statistics are estimates of the sample mean and its standard deviation in parentheses. The first panel displays summary statistics for outcomes derived from satellite imagery for all seven Sites and Services cities over the sample of observations with their centroid in either a de novo, upgrading, or control area. The second panel displays summary statistics for outcomes derived from TSCP survey data for Mbeya, Mwanza, and Tanga over the whole city sample. Observations are blocks based on an arbitrary grid of 50x50 meter blocks for both imagery and TSCP data. All columns report the maximum populated number of observations. Block outcomes are derived from all buildings with a centroid in the block. Blocks that fall between two treatment types are assigned according to where their centroid falls. The imagery variable painted roof has 14530 observations for the Total column, i.e. 116 less than the other variables. This is due to measurement error in assigning roof type to a building (outlines of some buildings in Dar es Salaam did not correspond to an actual building on the satellite image). Similarly, due to the survey nature of the TSCP data, in the Total column, the following TSCP variables have fewer than 42,090 observations: multiple storeys has 40,990 observations, good roof has 42,047 observations, sewerage or septic tank has 41,948 observations, water mains has 42,063 observations, and road access has 42,062 observations.

Table 3a.5: De novo regressions balancing first geography

	(1)	(2)	(3)
	Distance to Shore (km)	Block contains river or stream	Ruggedness within 50m
Panel A: no controls, 500m bandwidth (Imagery)			
De novo	-0.167	-0.007	-0.646
	(0.087)	(0.013)	(0.211)
Observations	8,440	8,440	8,440
Mean (control)	7.292	0.050	2.930
Panel B: baseline controls, 500m bandwidth (Imagery)			
De novo	-0.080	-0.017	-0.266
	(0.063)	(0.017)	(0.223)
Observations	8,440	8,440	8,440
Mean (control)	7.292	0.050	2.930
Panel C: baseline controls, 500m bandwidth (TSCP)			
De novo	-0.064	-0.074	-0.760
	(0.056)	(0.025)	(0.555)
Observations	2,693	2,693	2,693
Mean (control)	5.512	0.062	3.721
Panel D: baseline controls, Full City (TSCP)			
De Novo	-0.819	0.009	-0.364
	(0.222)	(0.011)	(0.337)
Observations	35,662	35,662	35,662
Mean (control)	4.850	0.016	3.236

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities in Panels A and B, while in Panels C and D the outcomes are derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and control areas within 500 meters of their boundary in Panels A, B and C. In Panel D, the sample includes de novo areas and the full city as control areas. In all panels, all blocks, including empty ones, are used. The outcomes are measures of geographical fundamentals and can be interpreted as quantifying any imbalance in selection of de novo and control areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A, the controls are only nearest de novo fixed effects. In Panels B, C and D, the controls are the regular ones: a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city.

Table 3a.6: De novo regressions of adult census outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean years of schooling	Share with exactly primary education	Share with more than primary education	Share attending school	Share literate in any language	Share literate in Swahili	Share literate in English
De novo	0.566	-0.041	0.051	0.018	0.010	0.004	0.053
	(0.121)	(0.016)	(0.015)	(0.009)	(0.006)	(0.010)	(0.023)
Observations	814	814	814	814	814	814	814
Mean (control)	9.343	0.412	0.497	0.128	0.960	0.936	0.449

Notes: This table reports estimates from regressions using cut Enumeration Area (EA) level observations with outcomes derived from Tanzania 2012 Census microdata for all seven Sites and Services cities. In each specification the regressor of interest is de novo, and the control variables include city fixed effects (separate for Temeke and Kinondoni in Dar es Salaam), and distance to the Central Business District (CBD) of each city. The sample includes de novo observations and control areas which are near de novo areas. The outcomes are measures of sorting into the treatment and control areas. Outcomes are the EA mean over the set of all adults at least 18 years old enumerated in the EA. Each observation is an EA of varying size, or a cut EA if the EA intersects both de novo and control areas. Cut EAs are assigned to de novo, and/or control areas if more than 5 percent of the cut EA lies inside the respective area. Analytic weights for the cut EA observations used in the regression are based on the proportion of the EA area that lies inside each treatment or control area. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares. There are 90 clusters.

Table 3a.7: Upgrading regressions balancing first geography

	(1)	(2)	(3)
	Distance to Shore (km)	Block contains river or stream	Ruggedness within 50m
Panel A: no controls, 500m bandwidth (Imagery)			
Upgrade	-0.057	-0.029	-0.389
	(0.075)	(0.012)	(0.161)
Observations	12,854	12,854	12,854
Mean (control)	6.778	0.060	2.663
Panel B: baseline controls, 500m bandwidth (Imagery)			
Upgrade	0.021	-0.050	0.099
	(0.049)	(0.019)	(0.233)
Observations	12,854	12,854	12,854
Mean (control)	6.778	0.060	2.663
Panel C: baseline controls, 500m bandwidth (TSCP)			
Upgrade	0.058	-0.075	-0.370
	(0.039)	(0.041)	(0.328)
Observations	2,576	2,576	2,576
Mean (control)	7.873	0.063	2.386
Panel D: baseline controls, Full City (TSCP)			
Upgrade	0.045	0.042	-1.002
	(0.126)	(0.024)	(0.315)
Observations	11,798	11,798	11,798
Mean (control)	7.079	0.019	2.422

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities in Panels A and B, while in Panels C and D the outcomes are derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and control areas within 500 meters of their boundary in Panels A, B and C. In Panel D, the sample includes upgrading areas and the full city as control areas. In all panels, all blocks, including empty ones, are used. The outcomes are measures of geographical fundamentals and can be interpreted as quantifying any imbalance in selection of upgrading and control areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A, the controls are only nearest upgrading fixed effects. In Panels B, C and D, the controls are the regular ones: a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city.

Table 3a.8: Upgrading regressions using imagery data for all seven cities

	(1)	(2)	(3)	(4)	(5)	(6)
	Mean log building footprint area	Share of buildings with painted roof	Mean similarity of building orien-tation	Mean z-index	Share of empty blocks	Share of area built up
Panel A: 500m bandwidth						
Upgrade	-0.053	-0.005	0.500	-0.010	-0.139	0.076
	(0.042)	(0.010)	(0.389)	(0.034)	(0.033)	(0.016)
Observations	10,909	10,837	10,909	10,909	12,854	12,854
Mean (control)	4.333	0.146	-7.352	-0.008	0.234	0.219

	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo
Panel B: robustness (mean z-index only as outcome)					
Upgrade	-0.014	-0.014	-0.011	-0.049	-0.007
	(0.035)	(0.035)	(0.034)	(0.061)	(0.035)
Observations	10,909	10,909	10,909	7,573	10,531
Mean (control)	-0.008	-0.008	-0.008	0.008	-0.017

	Mean log building footprint area	Mean similarity of building orien-tation	Share of empty blocks	Share of area built up
Panel C: upgrade vs old slums				
Upgrade	-0.152	0.801	-0.278	0.122
	(0.073)	(0.396)	(0.106)	(0.047)
Observations	8,000	8,000	9,319	9,319
Mean (control)	4.214	-6.195	0.231	0.303
Panel D: upgrade vs old slums, first geography controls				
Upgrade	-0.139	0.755	-0.233	0.123
	(0.065)	(0.264)	(0.091)	(0.045)
Observations	8,000	8,000	9,319	9,319
Mean (control)	4.214	-6.195	0.231	0.303

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities. The sample in Panels A and B includes the upgrading areas and control areas within 500 meters of their boundary. The sample in Panels C and D includes the upgrading areas in Dar es Salaam and the areas of that city which could be identified as slums before Sites and Services and that were not treated (see the Data Appendix for more details). The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panels A, C and D the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(3) in Panel A). In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panel B, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 117-125 clusters in Panel A, 117 clusters in Panel B, and 104-105 clusters in Panels C and D.

Table 3a.9: Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean log building footprint area	Share of buildings with multiple storeys	Share of buildings with a good roof	Share of buildings connected to electricity	Share of buildings with sewerage or septic tank	Mean z-index	Mean log hedonic value
Panel A: 500m bandwidth							
Upgrade	-0.111	-0.112	-0.178	-0.082	-0.130	-0.569	-0.181
	(0.101)	(0.050)	(0.084)	(0.078)	(0.079)	(0.253)	(0.133)
Observations	2,066	1,863	2,062	2,066	2,059	2,066	2,066
Mean (control)	4.801	0.094	0.972	0.524	0.350	0.041	17.281
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Full City	Optimal bandwidth
Panel B: robustness (mean z-index only as outcome)							
Upgrade	-0.572	-0.631	-0.597	-0.456	-0.578	-0.681	-0.633
	(0.246)	(0.243)	(0.242)	(0.323)	(0.264)	(0.209)	(0.265)
Observations	2,066	2,066	2,066	1,462	2,001	11,225	11,225
Mean (control)	0.041	0.041	0.041	0.046	0.030	-0.084	0.045
Panel C: robustness (mean log hedonic value only as outcome)							
Upgrade	-0.211	-0.286	-0.236	-0.200	-0.189	-0.370	-0.217
	(0.129)	(0.119)	(0.119)	(0.224)	(0.135)	(0.083)	(0.149)
Observations	2,066	2,066	2,066	1,462	2,001	11,225	11,225
Mean (control)	17.281	17.281	17.281	17.300	17.273	17.259	17.267

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A), and in Panel C the dependent variable is the predicted log value from hedonic regressions. In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B and C, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas, column (6) changes the control area to the sample of blocks covering the whole city excluding treatment areas, and column (7) uses 2889 observations inside the optimal bandwith for panel B and 1699 observations inside the optimal bandwith for panel C based on Imbens and Kalyanaraman (2012). The control mean in column (7) reports the mean for the control areas inside the optimal bandwidth. The 'Full City' in column (6) is robust to higher order polynomials in distance to boundary: In panel B: second order polynomial gives an estimate of -0.737 and standard error of 0.243, and third order polynomial gives an estimate of -0.662 and standard error of 0.243. In panel C: second order polynomial gives an estimate of -0.361 and standard error of 0.117, and third order polynomial gives an estimate of -0.237 and standard error of 0.143. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 30 clusters in Panel A, and 28-30 clusters in Panels B and C, except in column (6) of Panels B and C, which have 132 clusters.

Table 3a.10: Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with Owner Name Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log building footprint area	Multistorey building	Good roof	Connected to electricity	Sewerage or septic tank	Z-index	Log hedonic value
Panel A: Full City, Owner FE							
Upgrade	-0.225	-0.186	-0.021	-0.058	-0.039	-0.243	-0.218
	(0.150)	(0.067)	(0.047)	(0.094)	(0.076)	(0.140)	(0.144)
Observations	18,843	14,227	18,708	18,805	18,231	18,843	18,843
Mean (control)	4.601	0.205	0.966	0.416	0.221	0.002	17.026
Panel B: Full City, no Owner FE, same sample as A							
Upgrade	-0.243	-0.123	-0.011	-0.098	-0.172	-0.256	-0.310
	(0.146)	(0.084)	(0.012)	(0.054)	(0.082)	(0.105)	(0.152)
Observations	18,843	14,227	18,708	18,805	18,231	18,843	18,843
Mean (control)	4.601	0.205	0.966	0.416	0.221	0.002	17.026
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Second Order Polynomial	Third Order Polynomial
Panel C: robustness owner FE (z-index only as outcome)							
Upgrade	-0.244	-0.264	-0.250	-0.475	-0.228	-0.066	0.040
	(0.136)	(0.137)	(0.140)	(0.161)	(0.141)	(0.155)	(0.180)
Observations	18,843	18,843	18,843	17,914	18,780	18,843	18,843
Mean (control)	0.002	0.002	0.002	0.000	0.000	0.002	0.002
Panel D: robustness, no owner FE, same sample as C (z-index only as outcome)							
Upgrade	-0.329	-0.218	-0.272	-0.390	-0.255	-0.060	0.116
	(0.101)	(0.105)	(0.106)	(0.094)	(0.105)	(0.133)	(0.137)
Observations	18,843	18,843	18,843	17,914	18,780	18,843	18,843
Mean (control)	0.002	0.002	0.002	0.000	0.000	0.002	0.002

Notes: This table reports estimates from regressions using specification (1) and unit level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and the entire city as control areas. The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a property unit in a building, and only multi-unit owners are used. Units are assigned to upgrading or control areas based on where their building's centroid falls. Outcomes are measured at the building level (see Data Appendix for further details). In Panels A and B the outcomes vary, while in Panels C and D the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A). Panels A and C display results with unit owner last name fixed effects, including units inside upgrading and control areas but restricting the sample by keeping only last name owners that appear more than once in the sample. Panel B (D) displays results with the same sample as in A (C) but without owner last name fixed effects. In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas, and columns (6) and (7) control for second and third order polynomials in distance to the boundary, respectively Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 111-112 clusters.

Table 3a.11: Upgrading regressions on persistence measures using imagery and TSCP survey data

	(1)	(2)	(3)	(4)	(5)	(6)
	Imagery	Imagery Slums 1979	TSCP Survey		TSCP Survey, Excl. Tanga	
(lr)2-2 (lr)3-3 (lr)4-5 (lr)6-6	Share of buildings with road within 10m	Share of buildings with road within 10m	Share of buildings with road access	Share of buildings connected to water mains	Share of buildings connected to water mains	
Panel A: 500m bandwidth						
Upgrade	-0.019	0.018	0.004	-0.078	-0.059	
	(0.018)	(0.039)	(0.056)	(0.087)	(0.109)	
Observations	10,909	8,000	2,065	2,066	1,923	
Mean (control)	0.190	0.032	0.775	0.586	0.586	
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	
Panel B: robustness for share of buildings with road within 10m (Imagery)						
Upgrade	-0.021	-0.015	-0.019	-0.008	-0.022	
	(0.019)	(0.018)	(0.019)	(0.038)	(0.019)	
Observations	10,909	10,909	10,909	7,573	10,531	
Mean (control)	0.190	0.190	0.190	0.197	0.190	
	Geography	Lat-Long 2^{nd} Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Full City
Panel C: robustness for share of buildings with road access (TSCP)						
Upgrade	0.006	-0.014	-0.007	-0.053	0.012	-0.013
	(0.046)	(0.052)	(0.053)	(0.098)	(0.057)	(0.048)
Observations	2,065	2,065	2,065	1,461	2,000	11,207
Mean (control)	0.775	0.775	0.775	0.764	0.768	0.771
Panel D: robustness for share of buildings connected to water mains (TSCP)						
Upgrade	-0.079	-0.102	-0.089	-0.045	-0.076	-0.181
	(0.083)	(0.081)	(0.081)	(0.132)	(0.089)	(0.058)
Observations	2,066	2,066	2,066	1,462	2,001	11,214
Mean (control)	0.586	0.586	0.586	0.587	0.579	0.586

Notes: This table reports estimates from regressions using specification (1) and block level observations. The outcomes in both columns (1) and (2) of Panel A, and in Panel B (road within 10m) are derived from imagery for all seven Sites and Services cities. The outcomes in columns (3) and (4) of Panel A, and in Panels C and D (road access and connection to water mains) are derived from TSCP survey data for Mbeya, Mwanza, and Tanga. In column (5) of Panel A, Tanga is excluded from the TSCP survey data because of the uncertainty about water mains in that city (see Data Appendix). The sample in Panels B-D and in Panel A columns (1) and (3)-(5) includes the upgrading areas and control areas within 500 meters of their boundary. The sample in column (2) in Panel A includes upgrading areas and the areas of Dar es Salaam that could be identified as slums in 1979 but excluded from the Sites and Services projects (see Data Appendix). The outcomes are measures of persistence of infrastructure treatment. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the share of buildings with a road within 10 meters (from imagery data), in Panel C the dependent variable in all columns is the share of buildings with road access (from TSCP data), and in Panel D the dependent variable is the share of buildings connected to water mains (from TSCP data). In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B, C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas. Moreover, in Panels C and D, column (6) changes the control area to the sample of blocks covering the whole city excluding de novo areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services area. There are 28-30 clusters in TSCP data, except in column (6) of Panels C and D, which have 132 clusters. There are 117 clusters in imagery data, except in column (2) of Panel A which has 104 clusters.

Table 3a.12: Upgrading regressions of adult census outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean years of schooling	Share with exactly primary education	Share with more than primary education	Share attending school	Share literate in any language	Share literate in Swahili	Share literate in English
Upgrade	-0.469	0.049	-0.060	-0.018	-0.012	-0.011	-0.066
	(0.131)	(0.012)	(0.016)	(0.004)	(0.004)	(0.004)	(0.017)
Observations	2,842	2,842	2,842	2,842	2,842	2,842	2,842
Mean (control)	8.349	0.533	0.357	0.084	0.955	0.934	0.315

Notes: This table reports estimates from regressions using cut Enumeration Area (EA) level observations with outcomes derived from Tanzania 2012 Census microdata for all seven Sites and Services cities. In each specification the regressor of interest is upgrade, and the control variables include city fixed effects (separate for Temeke and Kinondoni in Dar es Salaam), and distance to the Central Business District (CBD) of each city. The sample includes upgrading observations and control areas which are near upgrading areas. The outcomes are measures of sorting into the treatment and control areas. Outcomes are the EA mean over the set of all adults at least 18 years old enumerated in the EA. Each observation is an EA of varying size, or a cut EA if the EA intersects both treatment and control areas. Cut EAs are assigned to upgrading and/or control areas if more than 5 percent of the cut EA lies inside the respective area. Analytic weights for the cut EA observations used in the regression are based on the proportion of the EA area that lies inside each treatment or control area. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares. There are 124 clusters.

Table 3a.13: Details on the selection of control areas by city

Dar es Salaam	• Sources: the 1974 (World Bank 1974a) and 1977 (World Bank 1977b) project proposal maps. • De novo and upgrading: the 1974 map is used to trace areas in the north of Dar es Salaam (Kinondoni Municipality), and the 1977 map is used in the south of Dar es Salaam (Temeke municipality). • Exclusions: the 1974 map is used to exclude areas in Kinondoni where we identify previously established residential areas and land reserved for special institutions and industry. The 1977 map is used to exclude areas in Temeke where there are low density residential areas and special institutions.
Iringa	• Sources: the 1977 project proposal map (World Bank 1977b), and a 1978 topographic map (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions from control areas: the 1977 project proposal map is used to exclude industrial and established residential areas east of Mwangata. The 1978 topographic map is used to exclude already developed areas west and east of Mwangata, and also north, south and east of Kihesa. Additionally, north of Mwangata is excluded because of a power plant.
Mbeya	• Sources: a 1966 satellite image (United States Geological Survey, 2015), and drawings by experts on the Sites and Services projects in Mbeya. Those experts are Shaoban Sheuya, Anna Mtani, and Amulike Mahenge and were all interviewed by the authors in Dar es Salaam, June 30, 2016. • De novo and upgrading: the drawings from our experts were used to trace areas. • Exclusions: the 1966 satellite image is used to exclude already built-up areas at the center of the city and areas with shops along the highway southeast of Mwanjelwa, already developed areas northwest of Mwanjelwa, and the airport. • For consistency across TSCP and imagery data, we kept all TSCP buildings in Mbeya within the minimum bounding rectangle of the Worldview imagery for Mbeya, this excluded a very small fraction of buildings at the fringes.
Morogoro	• Sources: the 1977 project proposal map (World Bank 1977b), and a 1974 topographic map (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the 1977 project proposal map is used to exclude a large industrial area southwest of Msamvu and a large previously developed area to the south of Msamvu. The 1974 topographic map is used to exclude a previously developed area south of Kichangani, and to confirm the exclusions from the 1977 project proposal map. Finally 0.07km^2 of undeveloped farm land is excluded from the area to the adjacent to the railway station.
Mwanza	• Sources: a 1973 cadastral map (Mwanza City Municipality, 1973). • De novo: the cadastral map is used to trace areas, it delineates all surveyed plots and so contains a few that are outside of the actual Sites and Services treatment. We include plots that are small (288m^2 is the known treated plot area) and recorded with a plot number, and community buildings. We do not include plots that are large or that are small but do not have a recorded plot number. • Exclusions: the cadastral map is used to exclude areas with large plots or plots without a recorded number. Also excluded are previously developed areas along the road in the southeast of Mwanza, as well as areas to the north that are off of the map. The 1966 satellite imagery was used to exclude built-up center of the city.
Tabora	• Sources: the 1977 project proposal map (World Bank 1977b), a 1967 topographic map (Directorate of Overseas Surveys, 2015), and 1978 aerial imagery (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the project proposal map is used to excluded previously built areas to the west and southwest of the Kiloleni. The 1967 topographic map is used to exclude an industrial area to the south of Isebeya in between the two of upgrading area. The 1978 aerial image is used to confirm the exclusions.
Tanga	• Sources: the 1977 project proposal map (World Bank 1977b), and a 1966 satellite image (United States Geological Survey, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the 1966 satellite image is used to exclude already developed areas south, southwest, north and east of Gofu Juu and east of Mwakizaro, as well as the center of the city near the coast. The 1977 project proposal map is used to exclude industrial area between Gofu Juu and Mwakizaro.

Notes: This table explains what imagery and maps were used to (a) delineate the de novo and upgrading areas, and (b) create exclusion areas (i.e. areas to be excluded from the control areas) among areas that are within 500 meters of Sites and Services, as explained in the Data Appendix. Sources are all georeferenced maps of the city in question. Almost all areas in the studied cities were covered by these maps, with minor exceptions in the western areas of Tabora, and north of the northern treatment area (Kihesa neighborhood) in Iringa.

Table 3a.14: Description of variables derived from imagery data

Variable label	Definition
Log building footprint area	Calculated directly for the shape file (calculated as a direct measure for the building, or a sample average of that measure for each block.)
Painted roof	Indicator for painted as opposed to tin or rusted tin (an indicator for the building or a share of buildings with painted roofs for each block). Please see the Data Appendix.
Similarity of orientation	Calculated using the main axis of the minimum bounding box that contains each building. We then calculated the difference in orientation between each building and its neighboring building, modulo 90 degrees, with more similar orientations representing a more regular layout (an indicator for the building or a sample average for each block).
Z-index	We construct a family of outcomes measure following Kling et al. 2007 and Banerjee et al. 2014. We integrate all "good" variables into one index. We subtract the mean in the control group and divide the result by the standard deviation in the control group. Then we create the index by taking a simple average of the normalized variables (a measure for the building or a sample average for each block). Please refer to the Data Appendix for more details.
Road within 10m	An indicator that the distance form the boundary of the building to the nearest roads is no more than 10m).
Distance to the CBD	The CBD for each city is the centroid of the most lit pixel in 1992 from the NOAA "Average Visible and Stable Lights, Cloud Free Coverage" dataset. The distance to the CBD is calculated from the centroids of each building or block.
Empty block indicator	Indicator for a block that has no buildings.
Share of area built up	Share of the area of the block that is built.
Number of buildings	Count of buildings in a 50x50m block.

Note: this table describes the variables derived from imagery data.

Table 3a.15: Description of TSCP variables and how they are created

Variable label	Definition
Connected to electricity	Indicator for whether a building is connected to electricity.
Sewerage or septic tank	Indicator for good sanitation, i.e. having sewerage or a septic tank as opposed to an alternative of pit latrine, no sanitation at all, or other.
Good roof	Indicator for roof being made of concrete, metal sheets, clay tiles or cement tiles as opposed to an alternative of grass/palm, asbestos, timber or other. This is a different measure from the "Painted roof" variable in Table A14.
Multistorey building	Indicator for one or more storeys above the ground floor.
Z-index	We construct a family of outcomes measure following Kling et al. 2007 and Banerjee et al. 2014. We integrate all "good" variables into one index. We subtract the mean in the control group and divide the result by the standard deviation in the control group. Then we create the index by taking a simple average of the normalized variables.
Hedonic Value	We run a hedonic regression using property values of 3663 buildings in Arusha based on log area, electricity, and indicators for good sanitation, good roof, and multi-story. We predict this value in our three TSCP cities (Tanga, Mbeya, and Mwanza).
Connected to water mains	Indicator for good water supply (metered/mains as opposed to borehole; stand tap; river; rain; water trucks; or other/none).
Road access	Indicator for access to tarmac; gravel; or earth road.

Note: this table describes the variables the we derived from TSCP building data.

Table 3a.16: Hedonic housing value regressions using TSCP survey data

	(1) ln value
Log building footprint area	0.797
	(0.019)
Connected to electricity	0.235
	(0.040)
Sewerage or septic tank	0.524
	(0.041)
Good roof	0.0474
	(0.090)
Multiple storeys	-0.0359
	(0.178)
Intercept	13.11
	(0.221)
Observations	3,663
R^2	0.416

This table reports estimates from a hedonic regression with buildings as units of observation using property values of 3,663 buildings in Arusha. The dependent variable is property value. This sample is selected because these buildings had both valuation data and data from the TSCP survey. Regressors are the buildings' log area, electricity, and indicators for good sanitation, good roof, and multi-storey. We then use the coefficient estimates to construct measures of hedonic values, as we explain in the Data Appendix.

Table 3a.17: Description of variables from Tanzanian census 2012

Variable label	Definition
Years of schooling	How many years of schooling the adult respondent has obtained. Missing values in the microdata are coded as 0 since there was no category for "Never attended school", and since the missing values were found to match reasonably well with the proportion of people with no schooling in the IPUMS 2012 Tanzanian Census data (which does not, however, have low level geographical identifiers). Moreover, the proportion of missing values in the microdata increased with age and with gender and age, which corresponds to the pattern of people lacking any schooling in Tanzania. Respondents with Training after primary school/Pre-secondary school or Training after secondary school are coded as 8 or 12 years respectively, i.e. one more year than primary or secondary schooling. Respondents with university education, are coded as 15, i.e. one more year than the maximum number of secondary schooling.
Exactly primary school	Binary indicator that takes the value 1 if the adult respondent has completed exactly 7 years of schooling, 0 otherwise. Missing values coded as 0 as in the variable above.
More than primary school	Binary indicator that takes the value 1 if the respondent has completed more than 7 years of schooling, 0 otherwise. Missing values coded as 0 as in the variables above.
Literate in any language	Binary indicator that takes the value 1 if the adult respondent is literate in any language.
Literate Swahili	Binary indicator that takes the value 1 if the adult respondent is literate in Swahili.
Literate English	Binary indicator that takes the value 1 if the adult respondent is literate in English.

Note: this table describes the variables we derived from the Tanzanian Census 2012 microdata.

3.8 Data appendix

This data appendix is organized as follows. We begin by describing the Sites and Services projects, the nature of the treatment, selection of the treated areas, and how the de novo plots were allocated. We then explain how we measure the treatment and control areas in the seven cities. We then describe the three main datasets: the first comes from imagery data; the second from the Tanzania Strategic Cities Project Survey (TSCP, World Bank 2013); and the third comes from 2012 Tanzanian census micro data. Finally, we discuss other auxiliary datasets, including: geographic variables; additional census data; land values data; data on project costs; and population data for 2002. Finally, we explain how we make currency conversions.

3.8.1 Project background and treatment

Background

The Sites and Services projects were implemented in seven Tanzanian cities. The projects treated 12 de novo areas (greenfield investments) and 12 slum upgrading areas (involving the upgrading of squatter settlements). The projects were rolled out in two rounds. The first round was implemented from 1974-1977, with infrastructure construction taking place in 1975-1976; and the second round was implemented from 1977-1984, with infrastructure construction taking place from 1980-1984. In the First Round, the World Bank treated the northwest of Dar es Salaam (Kinondoni district) and Mbeya with both de novo and upgrading and Mwanza with de novo investment only. In the Second Round the two types of treatment took place in the southeast of Dar es Salaam (Temeke district), Tanga, Tabora, Morogoro and Iringa. The number of de novo and upgrading plot surveyed in each round is reported in Table 3a.3.[58] Details of the projects are discussed in World Bank (1974a,b, 1977a,b, 1984, and 1987).

Sites and Services projects in Tanzania fell into two broad classes. The first involved de novo development of previously unpopulated areas. The second involved upgrading of pre-existing squatter settlements (sometimes referred to as "slum upgrading").

We provide a more detailed breakdown of the project costs below, but we note that among

[58] An additional upgrade was planned for the area Hanna Nassif in Dar es Salaam, but it was not implemented as part of Sites and Services. This area was nevertheless upgraded later on in a separate intervention (Lupala et al. 1997), but it is excluded from our analysis. Two additional areas, Mbagala and Tabata, were considered for the Second Round of Sites and Services, but it appears that they were eventually excluded from the project (World Bank 1987 and Kironde 2017).

the infrastructure costs, the two main components were roads and water mains, and the cost of surveying the plots formal de novo plots was also important. Other investments, which covered public buildings (schools, clinics, and markets) were minor parts of the overall scheme.[59] It is also unlikely that access to these services ends discontinuously at the program boundaries, so our regression discontinuity design should mitigate any effect from such services. The indirect costs of the project mainly consisted of loans, which we discuss below. Taken together, it seems that roads, water mains, and plot surveys were the most relevant elements of the program. The roads and water mains were implemented in both de novo and upgrading, but the formal plots were only implemented in de novo areas.[60]

In addition to the three elements discussed above, both de novo and upgrading areas received a small number of public buildings, which were designated as schools, health clinics, and markets. While these could have had an impact, we think that they matter less than the plots, the roads and the water. First, the total cost of the public buildings was lower than both the roads and the water mains (separately); and second, even if Sites and Services received more buildings than other areas, there is no evidence that access to those facilities ended discontinuously at the project boundaries, which is relevant for the empirical strategy that we explain below. And some Sites and Services residents were offered loans, which were not fully repaid. We think of these loans as relaxing some owners' budget constraints, so we explain in the main body of the paper our strategy for addressing this channel.

The control areas (see more details below) mostly developed in an informal way. We have traced back the history of the control areas near de novo using various reports, at least for Dar es Salaam, whose urban evolution seems better documented. For example, according to the 1968 Dar Masterplan (Project Planning Associates Ltd. 1968) the De-novo control areas appear to be "Vacant land and land used for agriculture", and according to the 1979 Dar Masterplan (Marshall, Macklin, Monaghan Ltd. 1979), the de-novo areas are not indicated as being squatted; by the late 1980s, however, it seems that all the control areas have some unplanned sections (Kironde 1994). Finally, the Transport Policy and System Development Master Plan (Dar es Salaam City Council 2008) in Dar indicates all de-novo, control and upgrading areas as "built up" by 1992. But we note that our data gives a more disaggregated picture on the extent of built up area, and it appears that at

[59] The first round buildings public buildings were also surrounded by street lighting.

[60] The Second Round investments were generally lower, and for one of the de novo areas (the one in Tanga), we have some uncertainty as to the extent of infrastructure that was actually provided (World Bank 1987).

this more fine-grained resolution not all the control areas were built up.

For the six secondary cities in which Sites and Services projects were implemented, we have not found evidence that any parts of the control areas were made formal under any planning scheme. In Dar es Salaam, however, Kironde (1994) documents that one planning scheme (Mbezi Beach) took place after the Sites and Services project. While we do not have precise maps, looking at present-day neighborhood boundaries this planned area may overlap with around 10-15% of our control area in Dar es Salaam. For the Mbezi scheme it seems that there was very little, if any, government provision of infrastructure, at least in the initial stages. As we discuss in the paper, however, eventually it seems that some investments in water mains and roads were made, but these were modest at best.

Treatment and control areas

We use a variety of historical maps and imagery from satellites and aerial photographs to define the exact boundaries of treatment and control areas. For Dar es Salaam, Iringa, Tabora, Tanga, and Morogoro, the World Bank Project Appraisals (World Bank, 1974a and World Bank, 1977b) provide maps with resolutions from 1:10,000 to 1:30,000 of the planned boundaries of the upgrading and de novo sites. In Dar, two maps were available, from 1974 and 1977, differing slightly for Mikocheni area. For all the areas except Tandika and Mtoni, we used the 1974 map, which appeared more precise. However, for Tandika and Mtoni we had to use the 1977 map, since these areas were not covered by the 1974 map.

For the two remaining cities, Mbeya and Mwanza, the maps from the project appraisal were unavailable. Therefore, for Mbeya we asked three experts to draw the boundaries of treatment. These experts were Anna Mtani and Shaoban Sheuya from Ardhi University, who both worked on the first round of Sites and Services project, and Amulike Mahenge from the Ministry of Land, who was the Municipal Director in Mbeya.

To delineate the treatment areas in Mwanza we obtained from the city municipality cadastral maps, dating back to 1973, at a resolution of 1:2,500. Since in Mwanza the treatment included only de novo plots, the cadastral map was sufficient to get the information for the intended treatment areas. We define the treatment area as covering the numbered plots that were of a size that (approximately) fitted the project descriptions (288 square meters); we also include public buildings into the treatment areas, to be consistent with the procedure in other cities. This procedure gives us a comprehensive picture of the

twelve de novo and twelve upgrading neighborhoods across all seven cities.

To define our control areas, along with the historical World Bank maps from the Appraisal reports (World Bank, 1974a and World Bank, 1977b), we use historical topographic maps, and satellite and aerial images taken just before the dates of the treatment. We assign all undeveloped ("greenfield") land within 500 meters of any treatment border to our set of control areas. However, as we explain in more detail below, we exclude areas that were either designated for non-residential use, or that were developed prior to treatment, or that are uninhabitable. Our rationale for looking at greenfield areas as controls because we want a clear counterfactual for the de novo areas. We have no "natural" counterfactual for the upgraded squatter areas, because we do not observe untreated squatter areas in the vicinity. The 500 meter cut-off reduces the risk of substantial heterogeneity in locational fundamentals. As part of our analysis we also focus on areas that are even closer to the boundaries between areas.

In order to know what had been previously developed, we used historical maps or imagery as close in time to the treatment date as we could find. We used all planned treatment maps. These include the 1974 and 1977 maps for Dar es Salaam and the 1977 maps for Morogoro, Iringa, Tanga and Tabora (World Bank 1987); the 1973 cadastral map of Mwanza (Mwanza City Municipality, 1973); satellite images from 1966 (United States Geological Survey 2015); aerial imagery from 1978 for Tabora and topographic maps from 1967, 1974, and 1978 for Tabora, Iringa and Morogoro (Directorate of Overseas Surveys 2015).[61] All areas (with some minor exceptions described below) were covered by at least one source. Satellite images and maps also confirm that the areas designated as de novo were indeed unbuilt before the Sites and Services program was implemented.

We use all these data to determine which areas within 500 meter of Sites and Services areas to exclude from our baseline control group. Our rules for exclusion from the control areas are as follows. First, we exclude areas that were planned for non-residential use. These were indicated on the planned treatment map for industrial or governmental use. Second, we exclude areas that were developed before the Sites and Services projects began. These were either indicated as houses or industrial areas on topographic maps, or visibly built in the historical satellite images. Third, we exclude uninhabitable areas, for example, those off the coast. Finally, in the case of Mwanza (where we had to infer the treatment areas) we applied additional criteria for exclusion. In this case we exclude large numbered plots and all unnumbered plots, which do not seem to fit the description of de novo plots.

[61] The resolutions of these maps range from 1:2,500 to 1:50,000.

We also exclude areas where the treatment areas are truncated at the edge, since we do not know where the exact boundary of treatment is. In this case we drew rectangles perpendicular to the map edge where the treatment area is truncated, and exclude the area within them.[62] Further details on defining exclusion areas in each city are outlined in Table 3a.13.

It is possible that some of the areas that were unbuilt in 1966 were built up from 1966 until the start of Sites and Services. But from partial evidence on construction dates in the TSCP data for two cities - Mbeya and Mwanza - it seems that only a very small share (about 1.3 percent) of the buildings with construction dates in control areas near de novo were built before 1974.

Our treatment maps (Figure 3a.1) show upgrading, de novo and control areas, as well as excluded areas. Moreover, with these appropriately defined control areas net of excluded locations, we can analyze present day outcomes using boundaries between control areas and de novo areas, and between control areas and upgrading areas.

We also note that for some of the analysis using the TSCP survey data (more below) we also used data further than 500 meters from Sites and Services. For the three Sites and Services cities with TSCP data (Mbeya, Mwanza, and Tanga), we used imagery from 1966 to exclude areas that were built up at the time.

Allocation of de novo plots

Plots were allocated to beneficiaries according whose i) houses were demolished in the upgrading areas ii) income was in the range of 400-1000 Tanzanian shilling (Tsh) a month. The income range was meant to target the 20th-60th percentiles of countrywide incomes (Kironde, 1991). According to project completion reports (World Bank 1984 and World Bank, 1987), between 50% and 70% of all project beneficiaries belonged to the target population. There was some evidence (World Bank, 1987) that a number of more affluent individuals obtained some of the plots after they had not been developed by initial beneficiaries.

[62] We include in the baseline control areas (minor) areas where there is no pre-treatment data, because they are very sparse and are located near other empty areas.

3.8.2 Outcome variables derived from imagery data

A summary of the outcome variables we construct using the imagery data can be found in Table 3a.14. Here we provide more detail on some of the key variables.

Buildings

To study the quality of housing we use Worldview satellite images (DigitalGlobe 2016), which provide greyscale data at resolution of approximately 0.5 meters along with multispectral data at a resolution of approximately 2.5 meters.[63] We employed a company (Ramani Geosystems) to trace out the building footprints from these data for six of the seven cities. For the final city, Dar es Salaam, we used building outlines from a different, freely available, source - Dar Ramani Huria (2016).[64]

We derive the following indicators of building quality using the building outlines: the logarithm of building footprint, building orientation relative to its neighbors and, finally the distance to the nearest road using ArcGIS tools.

For block outcomes we average each measure and indicator to get averages and shares. To do that, we begin with an arbitrary grid of 50 x 50 meter blocks. If a block is divided between de novo, upgrading, and control areas, we attribute the block to the area where its centroid lies. Finally, we match into each block the buildings whose centroids fall within it. This allows us to additionally measure three variables: the share of built up area in the block, the count of buildings in a block and whether the block is empty.

Roofs

To study the quality of roofs, we use the same Worldview satellite images as we did for the building outcomes above. Our aim was to separate painted roofs (which are less prone to rust) from unpainted tin roofs (rusted or not), in order to get a measure for roof quality that captures more variation than the TSCP survey indicator for good quality roofs. The cut-off between painted and unpainted roofs was chosen also because we had evidence

[63] The images were taken at different dates: Iringa (2013), Mbeya (2014), Morogoro (2012), Mwanza (2014), Tabora (2011), Tanga (2012) and there are two separate images for two districts in Dar es Salaam: Kinondoni (2015) and Temeke(2014)

[64] We have checked a sample of buildings traced out from the imagery data to the buildings in the TSCP survey data. Incidence of splitting or merging of buildings are fairly rare, occurring around 10 percent of the time, and more so in slum areas. This may also be in part due to a gap of a few years between the datasets. Therefore splitting or merging of buildings does not seem like much of a problem, especially when we focus on de novo areas.

from our initial field investigation that the painted roofs are considerably more expensive.

To this end, we create an algorithm through which ArcGIS and Python can separate painted from unpainted roofs for each satellite image of the seven Sites and Services cities. Before running the algorithm, we created unique color bins which would identify each type of roof material. These bins are three-dimensional sections of the red-green-blue space that correspond to different colors, which we think of as either painted roofs (e.g. painted red, green, or blue[65]) or unpainted ones (e.g. tin, rusted, and bright tin[66]). We defined the bins through a process of sampling pixels from each roof material type, identifying the color bins to which the pixels belong, and iteratively narrowing the bins for each roof type until they were mutually exclusive. Since each satellite image was slightly different in terms of sharpness, brightness and saturation, we sampled pixels from each image and created city-specific bands.

The algorithm is then applied to each city with its unique color bins. The algorithm works by reading the values of the color spectrum for red, green and blue of each pixel of a roof, and comparing these values to the above-mentioned unique bands of the color spectrum identifying painted, rusted and tin roofs. We assign to each roof the color bin that contains the plurality of pixels, and this indicates whether we classify it as a painted roof or not.

Roads

For all seven cities we used road data from Openstreetmap (2017). We had to clean these data in some locations using ArcGIS and Python, so that we only use roads that seem wide enough for a single car to pass through (we eliminated "roads" between buildings that were less than one meter apart). Following this automated procedure, we cleaned the road data manually to identify roads that appear passable to a single car.

3.8.3 Tanzanian Strategic Cities Project survey data

For three cities, Mbeya (in southwest Tanzania), Tanga (in northeast Tanzania), and Mwanza (in northwest Tanzania) we have detailed building-level data from the Tanzanian Strategic Cities Project (TSCP) which is a World Bank project implemented by the Prime

[65] Apart from red, green and blue we also had a bin for brown painted roofs in Kinondoni, since only in that image we noticed a large number of painted roofs that had a brown color, either due to image particularities or geographically varying preferences for brown painted roofs.

[66] In Iringa and Mwanza we did not have the category bright tin since the particularities of the image or the conditions of the day when the image was taken resulted in other roofs than tin also being very bright in these cities.

Minister's Office of Regional Administration and Local Government (World Bank 2010). These surveys were carried out by the Tanzanian government from 2010-2013. We use these data to build a more detailed picture of building quality in the areas we study. Table 3a.15 summarizes the key outcome variables that we derive from the TSCP data. Here we explain in more detail some of the issues relating the to dataset and how we use it.

The data arrived in raw format, with multiple duplicated records of each building and unit and many of these duplicate observations with missing data. We used the following rules to identify the unique observations. Buildings are identified by 'Building Reference Numbers' (BRN) and building units by BRN-units.

Rules for excluding buildings

1. Drop exact duplicates. i.e. if multiple buildings have all the same variables (including IDs) only keep one of them (dropped 1,202,669 observations).

2. Of all remaining observations with a duplicate BRN, drop all where all 'variables of interest' are missing. Variables of interest are an extensive list and comprise much more than what is used in the analysis of this paper (dropped 166,131 observations).

3. Of all remaining observations with a duplicate BRN, keep the observations with strictly more non-missing variables of interest (dropped 12,842 observations).

4. Of all remaining observations with a duplicate BRN, rank by 'information provider' and keep the observations with a strictly higher rank (dropped 15,486 observations).

5. Of all remaining observations with a duplicate BRN, for a set of observations with the same BRN, replace with missing all variables where the records are inconsistent. For example, if there are two observations with the same BRN and both have '2' for number of stories there is no inconsistency. But if one has '1' number of rooms while the other has '2': replace the number of rooms with missing for both.

6. Of all remaining observations with a duplicate BRN all duplicate BRNs will have exactly the same records, keep only one record for each BRN (dropped 27,483 observations).

7. There are no longer any duplicate BRNs. We drop 35,912 unique buildings from the records that do not match a building in one of the city shapefiles of building footprints.

8. We drop 38,180 buildings from the records that are coded as outbuildings.
9. We drop 596 buildings that do not match to a unit.
10. Finally, we are left with 119,914 buildings all with at least one corresponding unit.

Rules for excluding building units

1. Drop exact duplicates, for example, if multiple units have all the same variables (including IDs) only keep one of them (dropped 1,288,430 observations).
2. Of all remaining observations with a duplicate BRN-unit, drop all where all variables of interest are missing. Variables of interest are an extensive list and comprise much more than what is used in the analysis of this paper (dropped 221,134 observations).
3. Of all remaining observations with a duplicate BRN-unit, keep the observations with strictly more non-missing variables of interest (dropped 6,383 observations)
4. Of all remaining observations with a duplicate BRN-unit, for a set of observations with the same BRN-unit, replace with missing all variables with mismatched records within the set. i.e. if there are two observations with the same BRN-unit and both have '2' for number of toilets: do nothing, if one has '1' number of rooms while the other has '2': replace the number of rooms with missing for both.
5. There are no longer any duplicate BRN-units. We drop 32,322 units from the records that do not match a building in one of the city shapefiles of building footprints.
6. We drop 3,216 units from the records that are coded as outbuildings.
7. We do not need to drop any more units, since all remaining units match to a building.
8. Finally, we are left with 154,734 units all with a corresponding building.

From the building data set we exclude all buildings categorized as "Outbuildings" (sheds, garages, and animal pens). This leaves us with a sample of buildings that are used mostly for residential purposes, although a small fraction also serve commercial or public uses.

For these buildings in analysis we use the logarithm of building footprint; connection to electricity; connection to water mains; having at least basic sanitation (usually a septic tank and in rare cases sewerage); having good (durable) roof materials; having more than one story; and having road access.

Hedonic values

To calculate hedonic building values we use an auxiliary TSCP dataset covering 57,136 buildings from Arusha, which is not one of the seven Sites and Services cities, but is the only one for which we have valuation data at the level of individual buildings. Specifically, we have valuations for 6,837 buildings. The buildings for which we have valuations are concentrated near the city center.

The intention of the valuations is to determine the rateable value (annual rental value of a property) of each property as a basis for collecting property tax. This is estimated by professional valuers under a set of formal guidelines. The valuer is given building-level characteristics, a photograph of the property, and where possible, property transaction records (see figure below). The valuer uses these inputs along with a standard set of guidelines that give bounds on how much each characteristic of the building is worth, but ultimately makes a subjective valuation of the property based on the information provided.

Of the valued buildings, 3,663 also have building-level characteristics (log area, electricity, and indicators for good sanitation, good roof, and multi-story) from the TSCP survey. We use these to perform hedonic regressions and make out-of-sample predictions of the valuations in the three TSCP cities (Tanga, Mbeya, and Mwanza) where Sites and Services was implemented. For buildings in our out-of-sample prediction that are missing some, but not all, characteristics we fill these missing values with the average of their respective characteristic. Consequently, 6 percent of the buildings with hedonic values in our TSCP dataset have had missing data filled for at least one of their characteristics.

The results of the hedonic regressions are shown in Table 3a.16.[67] Buildings with larger footprints, electricity connection, and some sanitation, have higher hedonic values; conditional on these factors, roof materials and multistory buildings are uncorrelated with value, perhaps due to the sample size.

Construction dates

For two cities (Mbeya and Mwanza) we have building dates for less than 10 percent of the housing units in the de-novo and control areas within 500 meters. In absolute terms, this means we have construction dates for 215 de novo units and 300 control units close to the boundary. In both cities the de-novo areas were part of Round 1, so the infrastructure was

[67] We follow Giglio et al. (2014) in including observable characteristics linearly in a hedonic regression.

built from 1975-76, and for both we have pre-treatment imagery from 1966. According to the TSCP data, the fraction of units that existed as of 2013 that were built before 1975 was 0.5 percent in de-novo and (1 of the 215 units with construction dates) 1.3 percent (4 of the 300 units with construction dates) in control areas close to the boundary. Admittedly these data are imperfect, and some buildings may have been replaced over time, but the data do not suggest that old buildings that pre-date the Sites and Services are a major concern.

3.8.4 Geographic control variables

Distance to shore and rivers and streams indicators

We use as geographic controls the distance in kilometers to the nearest shore (either the Indian Ocean or Lake Victoria) and an indicator for rivers or streams.[68] These variables are derived from Openstreetmap - we use current data since historical data are unavailable. We consider proximity to the coast an amenity, while rivers or streams may be an amenity if their water is usable, or a disamenity if they increase flood risk.

Ruggedness

Ruggedness is calculated using SRTM elevation at a horizontal resolution of 1 arc-second (United States Geological Survey 2000). We use those data to compute the standard deviation of elevation of each 50m X 50m block relative to its eight neighbors.[69] We again use current data since historical data are unavailable.

Distance to historical CBD

For some of the robustness analysis we use measures of distance to historical CBDs, to mitigate concerns that our main measure of the CBD may be endogenous to Sites and Services. To construct these measures we use data on the location of railway stations in six of the cities, since these stations' locations were generally determined before the onset of Sites and Services, as we discuss below. Iringa does not have a railway station, so the coordinates of the Iringa municipal office were used instead. We then calculate distance

[68] The distance to the shore is winsorized at 10 kilometers, hence the distances to other water bodies, such as Lake Tanganika, are irrelevant in our seven cities.

[69] For a small fraction of blocks that are at the border of our study area, we instead use the mean of the standard deviation for those blocks for which it is calculated.

in kilometers to these coordinates in the same way as we do with the light-based CBDs and then use this as an alternative measure in some regression specifications.

To justify our argument that railroad stations existed even before Sites and Services, and hence can be used as ex ante markers of the centers of the cities, we refer to a map of the railways from 1948, which shows that five the seven cities had railways in 1948, and the location of railway stations is unlikely to be moved.[70] Of the remaining two cities, Mbeya's railway was built from 1970 and completed and opened in 1975 (Edson 1978), while Iringa does not have railway, as mentioned above.

3.8.5 Land values

Matching land value data to enumeration areas

We obtained an Excel sheet titled "RATES LAND VALUE MIKOA 10 2012.xls", which we received from the Kinondoni Municipal council, but were told that it was created by the Ministry of Lands, with minimum, mean, and maximum land values for different neighborhoods in Tanzania. We can identify these neighborhoods by four string identifiers: region, district, location, and streets. To locate neighborhoods we match them based on the 2002 enumeration area (EA) shapefile, which contains string identifiers for region, district, location, and vill_stree (we consider 'vill_stree' comparable with 'streets' from the land values table).

Land use

The Excel table has different minimum, mean, and maximum land values by land use. There are typically four categories: Residential, commercial, commercial/residential, and institutional. Though the differentiation of land values across uses is mechanical (commercial is 1.4* res, com/res is 1.1*res, institutional is the same as res), the variation across areas is not mechanical. Throughout we use mean land values from the residential categories only.

[70] Britishempire.co.uk. (1948). [online] Available at: https://www.britishempire.co.uk/images2/tanganyikamap1948.jp [Accessed 3 Jul. 2019].

Spatially mapping land values

We merge EA boundaries to land value observations using the four identifiers: region, district, location, and streets. Each entry in the land value table we treat as an observation, often this contains a group of 'streets'. Typically there are many EAs per land value observation, so each observation in the land values table is matched to a large group of EA boundaries. Then we dissolve the EA boundaries to have a single spatial unit for each entry in the land value sheet. We then plot the mean residential land rate for each spatial unit.

Results

The merged areas are quite large. Some roughly match our treatment areas:

1. Sinza – one unit at 240,000TSh
2. Manzese A – three partial units all at 65,000TSh
3. Manzese B – split in half, one at 65,000TSh the other at 50,000TSh
4. Kijitonyama – one unit at 325,000TSh

The other two do not match as well:

1. Mikocheni – contained by a much larger unit at 125,000TSh
2. Tandika/Mtoni – overlaps many areas of values; 40,000TSh, 30,000TSh, 50,000TSh, and 18,000TSh

These values per square meter put us in the range of 125,000-325,000 TSh (2017 US$80-220) in de novo and 18,000-65,000 TSh (2017 US$10-40) in upgrading. For the areas where we have better matched data the ranges are 240,000-325,000 TSh (2017 US$160-220) in de novo and 50,000-65,000 TSh (2017 US$30-40) in upgrading.

3.8.6 Project costs

The total cost of First Round of Sites and Services was $60m in US$2017, of which just over half was due to direct costs (World Bank 1984): infrastructure (38% of total costs),

consultants (9%), land compensation (6%). Other costs (45%) included the community centers (14%), mentioned above, and a loan scheme (29%), which later failed because of poor repayment rates, and a few other costs. This investment covered a total of 23,161 plots: 8,527 de novo plots and 14,634 upgrading plots. The Second Round of Sites and Services cost $70m in US$2017 where 70% was spent on direct costs, paying for a total of 22,106 plots: 1,978 de novo plots and 20,128 upgrading plots (World Bank 1987).

The First Round project reports (World Bank 1974a and 1984) indicate that the total infrastructure investment costs per area in de novo and upgrading were very similar. The project report for Round 1 provided costs separately for de novo and upgrading areas (World Bank, 1984). However only infrastructure investment differed for the two types of treatment, while land compensation, equipment, and consultancy costs were reported as split 50-50 between de novo and upgrading. Direct costs by treatment were $19 million in de novo and $15 million in upgrading areas (in US$2017). To get costs per unit area we normalize by total area covered by each treatment type in Round 1 (8.5 square kilometers in de novo and 6.5 square kilometers in upgrading). This gave costs for de novo and upgrading areas of $2.20 and $2.37 per square meter respectively (in US$2017).

Further, in order to compare with present day land values (per plot area) we would like an estimate of costs per unit of treated plot area. Due to data limitations we can only do that for de novo neighborhoods where the reports give both plot counts and plot areas. We estimate that the direct costs per square of plot were no more than $8 per square meter, and total costs were no more than $13 per square meter (in US$2017).[71]

An alternative way to look at costs is to break them down by plot which we can do for both de novo and upgrading areas. According to the report there were 8,527 de novo plots and 14,634 upgrading plots in Round 1. We can divide the direct costs of de novo and upgrading areas by their plot counts to get $2,200 and $1,000 per plot respectively (in US$2017). The difference in costs reflects both the larger size of the de novo plots and the larger share of allocated to public amenities (such as roads).

Cost recovery

Costs were meant to be recovered through land rent (4% of land value a year) and service charge (the cost of infrastructure provider), but assessment of parcels was long and interim

[71] To calculate the costs per square meter of each plot, we use the planned areas of de novo plots from Appraisal report 1 (World Bank, 1974a); the planned area was 288 square meters, except for 8.56% of the plots (those in Mikocheni) where it was 370 square meters. Taking the weighted average at 295 square meters, we can divide the de novo direct costs by total plot area treated to get $7.5 per square meter.

charge well below the adequate amount to cover the costs (100 Tsh/year or 2017 US$51) was imposed. Collection rates were low and not timely.

3.8.7 Additional data

Outcomes in 2012 Tanzanian census micro data Extract

This extract was obtained through a contact from Tanzanian Census Bureau. Unlike the Tanzanian census data, which can be obtained online at IPUMS (2017), these data are at the level of individuals. We match these census observations from this extract to geographical areas using EA identifiers in the census extract. Using shapefiles of EAs (with the same identifiers) from the Tanzanian Census 2012, also obtained from the same contact, we match the census data observations to our treatment and control areas. The process of matching EAs to treatment areas (de novo, control and upgrading) was done through Python and ArcGIS.

In case an EA straddled two (or more) of the treatment and control areas, we cut that EA in ArcGIS into multiple parts, each part belongs to a treatment or a control area. We then use this information to remove the census data observations which belonged to EAs whose area inside a treatment and control area was less than 5% of the entire EA area. We also use the information on how large a part of the EA was inside a treatment or control area to create analytic weights (the weight is higher when the relevant overlap is higher) for some of the robustness checks.

Our variables are discussed in Table 3a.17, and include years of schooling and indicators for different schooling thresholds (exactly primary and more than primary school education; the omitted category is less than primary school). We also create indicators for literacy in any language; literacy in Swahili; and literacy in English. We then calculate means of each of these variables across adults in each "cut" EA.

Population data for 2002

To calculate the population density in each of the neighborhoods, we use data on population by enumeration areas from the 2002 Tanzanian Census (Tanzania National Bureau of Statistics 2011). In cases where an entire enumeration area falls into a Sites and Services neighborhood, we assign its entire population to that neighborhood. When only a fraction of an enumeration area falls into a Sites and Services neighborhood, we assign to

the neighborhood the fraction of the enumeration area population that corresponds to the fraction of the land area that lies within the neighborhood. The mean number of enumeration areas matched to each neighborhood is 33 for de novo areas and 35 for upgrading areas.[72] Population counts for 2002 are outlined in Table 3a.3.

IPUMS 2012 Tanzanian census by region

We use data downloaded from the IPUMS online repository of country censuses, in order to check the correctness of the above-mentioned microdata extract from the same census. This was done in particular for the education variable which had been cleaned by IPUMS staff to include many observations recorded as having "never attended" school. The microdata that we had received directly from the Tanzanian Census Bureau had many missing values for the education variable, and none coded as never having attended school. The missing values in the micro-data followed the same pattern as the "never attended" in the IPUMS data, which contributed to our decision to code them as zero years of schooling. We also checked age and gender patterns in the microdata which confirmed our interpretation of the data.

Conversion to 2017 US dollars

All monetary values in the paper are reported in their source units and also converted to 2017 US dollars (2017 US$). To calculate the dollar values we used the exchange rates to contemporaneous year US$ from Penn World Tables 9.0 (Feenstra et al., 2015). Then we used the US CPI factors to bring the value to 2017 US$.

[72] We are unable to report the population counts from 2012 census, because we only have a sample from the census, and in this sample, not every 2012 enumeration area is populated.

Chapter 4

Incentives and Culture – Evidence from a Multi-Country Field Experiment

4.1 Introduction

Performance rewards are the classic solution to agency problems in firms and a cornerstone of modern management practices. Yet their popularity is limited outside a group of, mostly Anglo-Saxon, high-income countries. The correlation between income and the use of performance rewards can be easily seen in Figure 4.1a.[1] Do firms in lower-income countries face institutional constraints, such as poor contract enforcement, that limit the use of performance pay? Or do they optimally choose not to offer incentives because they would not be effective, as different cultural norms govern agents' responses to incentives? We set up a multi-country field experiment to provide answers.

Our starting point is that the effectiveness of performance pay depends on the balance of two forces: the motivating effect of higher pay vs. the (possibly) demotivating effect of inequality, or social punishment from breaking a culture's norm of looking after others than themselves. Indeed, performance pay inevitably generates inequality between high and low performers, and this might decrease utility if individuals have culturally mediated preferences for conformity within the group or aversion to inequality (Fehr and Schmidt 1999; Ashraf and Bandiera 2018).

[1] Data from the World Management Survey in manufacturing 2004-2015 (Bloom and Van Reenen 2007).

These two forces map into a key dimension of culture, known as the individualism-collectivism spectrum (Hofstede 1980; 2001), where individualism is related to the degree of interdependence a society maintains among its members.[2] Simply put, individualistic cultures value individuals enjoying the fruits of their efforts and largely accept differences that derive from it (Alesina and Angeletos 2005). In collectivistic cultures, the self-image is defined as "we" rather than "I" and members of the in-group loyally take care of each other Hofstede Insights, 2020). In line with this, Figure 4.1b shows a positive correlation between individualism and the use of performance pay. Importantly, the correlation is not only due to differences in the overall quality of management practices. Even among the top three countries in management practices (see Figure 4a.4 in the Appendix),[3] incentives are more common in the most individualistic country (the U.S.) and less common in the least individualistic country (Japan) (see Figure 4.1b).

The challenge in establishing causality is that firms are, understandably, reluctant to offer contracts that they expect to be ineffective or detrimental (Bandiera, Barankay and Rasul 2011), thus field experiments are generally limited to interventions that firms are reasonably optimistic about. To address this challenge, we set up our own data entry firms to exogenously vary workers' exposure to performance pay in three countries that are in the same GDP per capita class (Lower Middle Income Countries, LMIC)[4], but are in the lowest, middle and highest part of the individualism scale among LMIC: Ghana, India and the Philippines.

In all three countries we implement three classic pay packages: our control group receives fixed wages, while treatment is divided into individual incentives (paid by keystrokes per hour) and group incentives (paid by the average keystrokes per hour in a group of on average four workers[5]). To shut down income effects under each scheme we set the piece

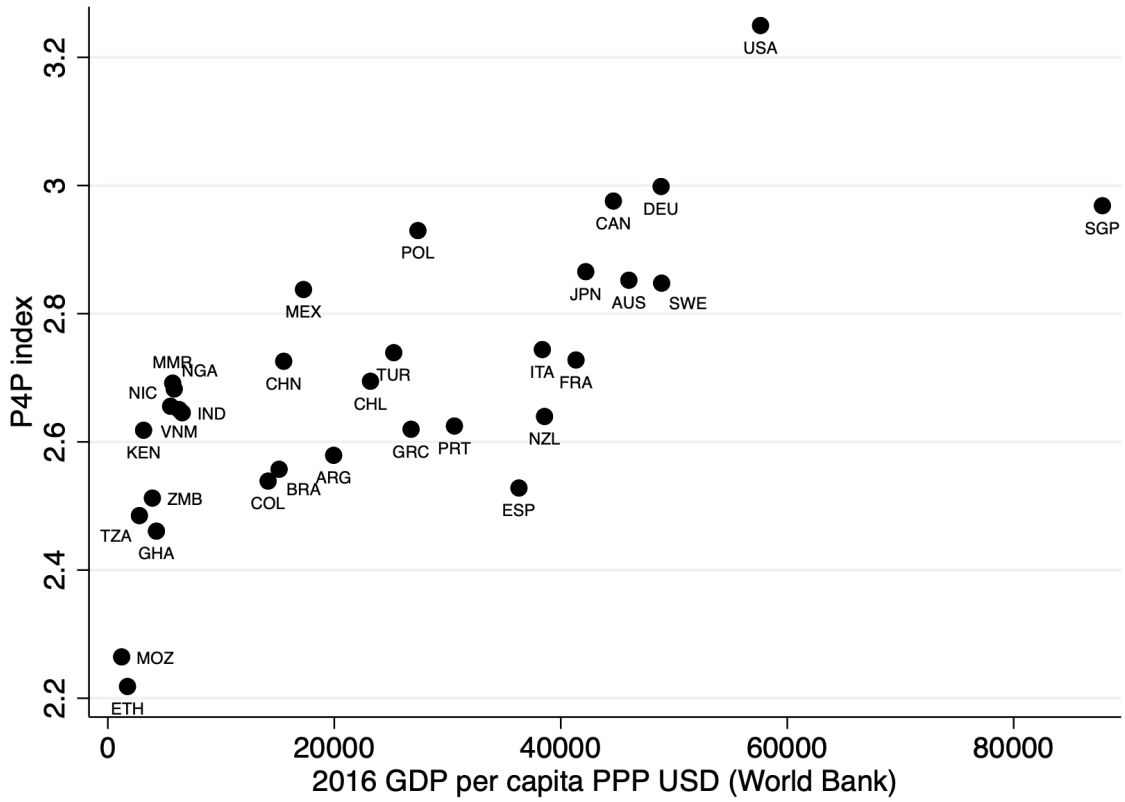
[2]We use the established measure of individualism from Hofstede's (2001) survey of 70k+ IBM employees in over 80 countries. Hofstede's measures are commonly used in economics (Bloom, Sadun and Van Reenen 2012; Gorodnichenko and Roland 2011; Hjort, Li and Sarsons 2020). Other Hofstede cultural dimensions such as *power distance* ("the extent to which the less powerful members of institutions and organisations within a country expect and accept that power is distributed unequally") or *masculinity* ("society is driven by competition, achievement and success") (Hofstede Insights, 2020) could also be partially related to the response to incentive pay. However, our view is that individualism is more closely related, as placing self-interest over collective interest would logically lead to working harder on incentive pay, while an emphasis on harmony within the group and an increased concern for others would lead to not wanting to perform only when there is material self-interest in working harder.

[3]Figure 4a.4 in the Appendix shows that the US, Japan and Germany have top management practices scores. These scores of manufacturing firms are from Bloom, Sadun and Van Reenen (2016) and include the Incentives questions used to create the Pay for Performance index used in Figure 1, as well as a host of other questions on Operations, Monitoring and Targeting.

[4]Ghana moved from the Low Income to the LMIC category in 2011 (World Bank, 2011) and in 2019 the three countries remain in this category (World Bank, 2019).

[5]The number of workers that 2-day contract were divided into 2 equal groups if possible. As the number of people working each day varied, groups had 2-5 participants, with the mean and median being 4 workers.

A. Pay for performance index by GDP



B. Pay for performance index by individualism

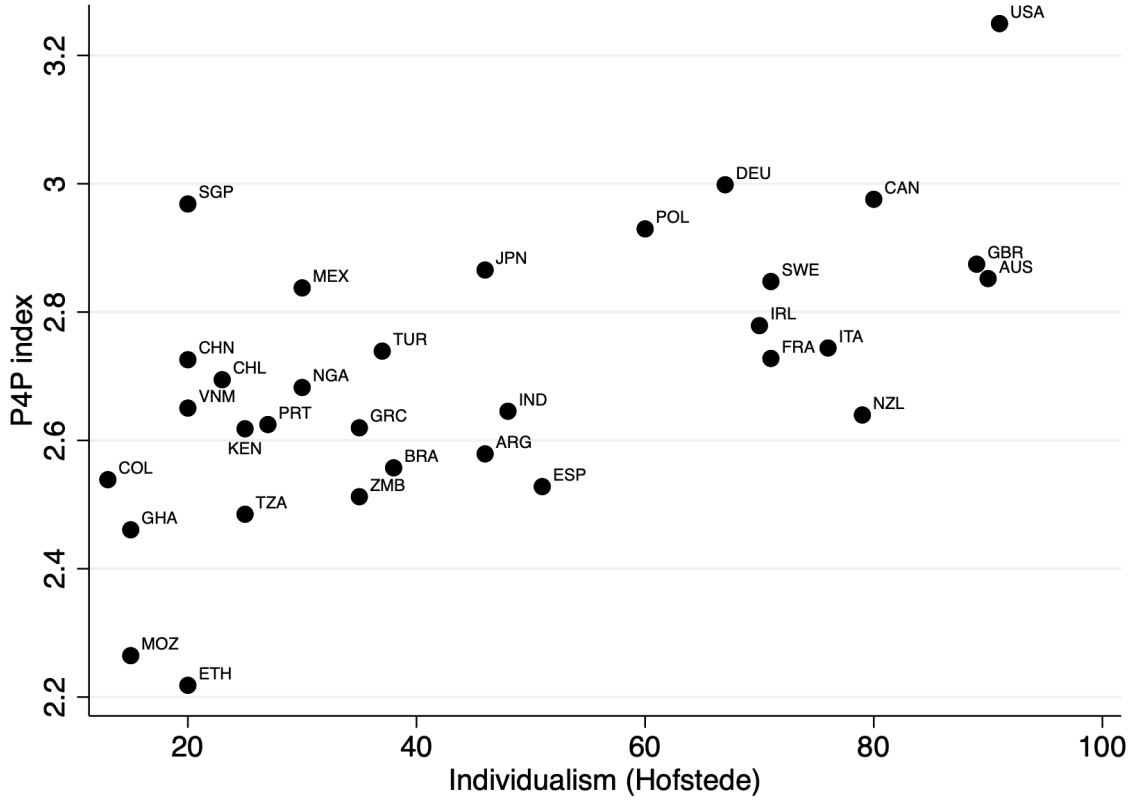


Figure 4.1: Pay for performance index against GDP and culture.

An index of the use of Incentives/Pay for performance (P4P) in manufacturing from the World Management Survey data 2004-2015 (Bloom and Van Reenen 2007), plotted against (a) per capita income data from the World Bank and (b) Hofstede's measure of individualism of each country. The Incentives index is calculated as a mean of the measures listed as Incentives in Appendix I.A in Bloom and Van Reenen (2007).

rates so that a worker with median productivity is expected to earn the same. To test whether culture is the mediating factor, we vary the visibility of individual performance and compare the effect of incentives when their consequences for inequality are made public (by publishing individual productivity ranks during the working day) and when they are not. To the extent that cultural norms are not fully internalised and rather depend on whether conformity is observable, this visibility should dampen the response to incentives in more collectivistic cultures.

Each worker is hired for one or two 2-day contracts. Short term contracts are common in the data entry industry, and this design of the experiment allows us to speak to the growing zero-hour contract sector, where interactions are short — meaning that reputation and career concerns are minimized. The three firms we set up were active for 8-11 months in each country, and in total the experiment was active every month from December 2010 to February 2012, a similar length as other data entry firm experiments (Kaur, Kremer and Mullainathan 2015). Experiments with data entry workers in the US also involve contracts of days rather than months (see, e.g., Hedblom, Hickman and List 2019). Moreover, the expected cultural differences in incentive effects are not dependent on a long-term assignment but could be expected to be present already in a one-off task, as shown in Gneezy et al (2019).

We have almost no evidence that performance pay schemes that have been shown to be effective in Anglo-Saxon countries would work in lower-income countries. The exceptions concern primarily the public sector, notably education and public health, e.g., Muralidharan and Sundararaman (2011) and Ashraf, Bandiera and Jack (2014), with results suggesting that there may be significant heterogeneity in the response to these incentive schemes across countries in these sectors. This study seeks to contribute with evidence from *firms* in lower-income countries, and to our knowledge we are the first to compare such identical experiments across culturally different countries.

4.2 Experimental Design

The key feature of our design is that it allows us to isolate the effect of culture on the response to incentives by comparing the effect of the same performance pay packages on performance in the same job in firms that are identical but for the culture of the country where they are located.

This requires us to make choices on three dimensions: tasks, countries and contracts.

4.2.1 Task

Our choice of task is data entry. We made this choice for two reasons. First, because productivity (as well as quality-adjusted productivity) can be measured easily and accurately. Second, because it is representative of the type of employment available to individuals with a secondary education in many low-income countries, as data entry and similar tasks are being shifted there from high-income countries to take advantage of lower labor costs.

The workers involved in the experiment were people who would normally work in data entry, and were hired through normal channels and paid a typical wage. Each firm had at least two rooms, each with a large common table, five seats and laptops. The manager in each country was experienced in field work and data entry and additionally trained for two months to ensure that he or she could consistently implement the experiment protocol over the course of the experiment. Each country manager used the same materials to train workers, in order to minimize variations in training quality.

In addition to enabling consistency throughout the experiment, data entry is measurable and strongly or perfectly correlated with effort, productivity and work quality (Kaur, Kremer and Mullainathan 2015). Productivity is perfectly correlated with the number of key strokes per contract, work quality perfectly correlates with accuracy, and effort can be measured as a combination of productivity and accuracy. This allows us to precisely measure our key output variables.

4.2.2 Countries

To implement our design we relied on the existing administrative structures of Innovations for Poverty Action (IPA) and the Poverty Action Lab (J-PAL), both of which hire workers for data entry regularly. Among the countries with IPA or J-PAL offices, we chose three to maximise variation in individualism and minimize variation in income. These are Ghana (with an Individualism score of 15), the Philippines (with an Individualism score of 32), and India (with an Individualism score of 48)[6]. These span the entire Individualism-Collectivism scale for the 19 LMIC surveyed by Hofstede (2019).[7]. Identical firms were founded in Accra, Ghana in 2010; Kolkata, India in 2011 and Cagayan de Oro, Philippines in 2011, and all would digitize surveys on behalf of IPA and J-PAL. We called these firms

[6]See Appendix Figure 4a.5.

[7]The distribution of individualism scores among Lower middle income countries (LMIC) ranges from 14 to 52, with Ghana having the third lowest score and India the second highest among the LMIC listed at https://data.worldbank.org/income-level/lower-middle-income which also have a non-missing score at https://www.hofstede-insights.com/country-comparison/

IPA Data Services in Ghana and the Philippines, and *J-PAL Data Services* in India and advertised them to applicants as an entity dedicated to data entry.[8]

4.2.3 Contracts

Workers were hired for two-day contracts and carried out the work on computers in custom-made data entry interfaces, which we used to measure worker productivity in key strokes per hour. We randomized workers into contracts in a 2 by 2 design, wherein individual or group piece rate incentives were cross-randomised with public ranking of worker performance. The control group were paid a flat wage.[9] Randomization was stratified using three variables: gender, ethnicity (dominant vs. all others) and baseline performance (above vs. below median productivity). In Ghana, the ethnicity breakdown was Akan and non-Akan, in India General Caste and other, and in the Philippines Visaya and other.

The piece rate contracts paid a fixed amount per keystroke, whether correct or incorrect, defined as "a keyboard action that results in data capture in the used program." To shut down income effects we set the piece rates so that a worker with median productivity would earn the same under each scheme.[10] For the group incentives, the workers in each two-day contract were divided into two equally sized teams (if possible) with an average of four workers per team. Workers were paid a salary proportional to the total number of key strokes entered by all group members over two days.

In the contracts with public ranking of effort, managers ranked workers based on their productivity and publicly displayed rankings with full names three times a day on white boards at the front of each room, to provide public visibility of relative and absolute performance. For the group incentives with ranking of effort, it was still individual names which were ranked. See the Appendix for a more detailed description of the treatment groups.

Recruitment In all three countries, workers were recruited through advertisements on

[8] Table 4a.1 in the Appendix describes the experiment in terms of duration, number of workers and work done in the three identical firms.

[9] The fixed daily wage was set in each country to what was typical in that area and sector. This was 25 Ghana Cedis (GHC) = 2011 USD PPP 31.73 (using exchange rate 1 2011 USD PPP = 0.788 GHC, all from https://data.worldbank.org/indicator/PA.NUS.PRVT.PP), 450 Philippines Pesos (PhP) = 23.84 2011 USD PPP (using exchange rate 1 2011 USD PPP = 18.873 PhP), or 250 Indian Rupees (INR) = 2011 USD PPP 16.69 (using exchange rate 1 2011 USD PPP = 14.975 INR), respectively (paid in cash).

[10] In each country, we calculated the piece rate as the flat wage divided by median number of total key strokes entered during the 2-day first flat wage contract, which in Ghana was 50/46,729 = .00107 GHC per keystroke for the Pure Home Water Survey (PHWS), and 50/80,425 = .0006217 GHC for the Formal Savings (FS) survey, in India 500/31,083 = .0154421 INR for the Market survey and 500/70385 = .0071038 INR for the Village Welfare Society (WVS) survey, and in the Philippines 900/80,032 = .0112455 PhP for the CDO Fresh survey and 900/107136 = .0084 PhP for the Health and Financial Services Survey (HaFS) survey.

leading websites: JobsinGhana.com, Naukri.com in India and Jobstreet.com.ph in the Philippines. The advertisement required prospective employees to have a minimum of a secondary education, knowledge of computer applications, advanced English skills and to be at least eighteen years old. In the course of the recruitment period, IPA/J-PAL Data Services received 1560, 2085 and 1393 applications for employment in Ghana, India, and the Philippines, respectively. In addition to Internet postings, we also used street advertisements and contacts at local computer training centers.

In all three countries, the advertisement directed applicants to an form to provide their contact information. The benefits of using an advertisement included low cost, wide reach, the ability to collect electronically data on gender, age, education level, previous experience in data entry and previous general employment experience, as well as serving as a preliminary confirmation of prospective employees' level of computer knowledge. Applicants were interviewed and tested and answered survey questions on work preferences and attitudes. We invited all applicants who scored above a certain level of accuracy in the data entry test to at least one two-day contract and a maximum of two such contracts. Further details of the hiring process are provided in the Appendix.

Table 4.1: Balance table

Variable	(1) Fixed Wage	(2) Piece Rate	(3) Difference	(4) P-value
Ghana				
QA baseline prod.	2.000	2.035	0.035	(0.609)
Paid job before	1.000	0.961	-0.039	(0.154)
Data entry exper.	0.863	0.716	-0.147	(0.044)
Piece rate exper.	0.078	0.167	0.088	(0.137)
University	0.647	0.814	0.167	(0.023)
Male	0.725	0.667	-0.059	(0.463)
Age	27.928	28.051	0.123	(0.872)
Observations	51	102	153	
Philippines				
QA baseline prod.	2.122	2.095	-0.026	(0.716)
Paid job before	0.600	0.520	-0.080	(0.357)
Data entry exper	0.200	0.190	-0.010	(0.885)
Piece rate exper.	0.280	0.240	-0.040	(0.598)
University	0.820	0.800	-0.020	(0.772)
Male	0.320	0.410	0.090	(0.288)
Age	23.991	23.124	-0.866	(0.167)
Observations	50	100	150	
India				
QA baseline prod.	1.979	2.080	0.101	(0.100)
Paid job before	0.692	0.714	0.022	(0.778)
Data entry exper.	0.481	0.419	-0.062	(0.467)
Piece rate exper.	0.135	0.133	-0.001	(0.982)
University	0.885	0.790	-0.094	(0.149)
Male	0.750	0.762	0.012	(0.871)
Age	26.721	26.225	-0.496	(0.566)
Observations	52	105	157	

Balance table for all three countries with p-values in brackets. The first variable is log quality-adjusted baseline productivity: the natural lograrithm of the number of correct keystrokes per minute in the test conducted during interviews, using the same data interface as during actual production.

4.3 Identification strategy

We estimate :

$$y_{ic} = \sum_{c}^{3} \alpha_c T_i D_c + \sum_{c}^{3} \beta_c D_c + \sum_{c}^{3} \gamma_c y_{ic}^0 D_c + \epsilon_{ic} \tag{4.1}$$

where y_{ic} is log (quality-adjusted) average two-day productivity (key strokes per hour, adjusted for errors) or derived log firm profits of worker i in country c, T_i is an indicator variable for incentive treatment, D_c is an indicator for country, and y_{ic}^0 is baseline productivity. To measure baseline productivity we give all applicants the same 15 minute-data entry task at the interview stage. α_c measures the causal effect of incentives (compared to flat wage) on productivity in country c under the assumption that incentive treatments are orthogonal to ϵ_{ic}. The identifying assumption could fail because of either endogenous take-up differences or spillovers. However, neither appears to be relevant in this setting, as described below.

To minimize spillovers, we designed the experiment such that all workers within a given two-day work period received the same contract and worked on the same survey, and we invited workers to only one type of contract with each survey. Workers would naturally assume that this contract was the only one offered for entering this survey.

To eliminate selection effects in take-up of contracts, we did not reveal the contract type upon invitation. The manager invited workers to contracts one to three days in advance saying that the salary would depend on the contract type they received, but that the average worker should expect to receive 25 Ghana Cedis (GHC), 250 Rupees (INR) or 450 Philippines Pesos (PhP) per day,[11] and that the contract could be either flat wage or piece rate.

Randomization balance is shown in Table 4.1. In sum, that table shows that worker traits are balanced across treatments in each of the three countries. That table also shows evidence that baseline productivity[12] is similar *across countries*, indicating that different worker ability across countries does not explain our results. The baseline productivity was measured using the same data entry interface as during actual production, and workers

[11] In Table 4a.1 in the Appendix, we report the average total amount received over the course of the contract (almost two days on average) for all contracts except flat wage.

[12] Log quality-adjusted (QA) baseline productivity is the natural logarithm of the number of correct keystrokes per minute in the Epidata test conducted during interviews.

entered the same survey across all three countries in the test. Other worker traits differ across countries and this is the result of a deliberate design choice. When hiring we faced the choice between hiring the best workers, accepting that they might differ across countries because of local labour market conditions, or hiring to balance traits across countries. We chose the former both because our goal was to mimic real firms and because balancing on observables might worsen selection on unobservables. For instance, if data entry operators are normally high school graduates in Ghana and university graduates in India, going after university graduates in Ghana will result in negative selection or would require a much higher wage than is normally paid for data entry. In both cases the interpretation of the results would be confounded by these differences. To avoid that we chose to recruit as local firms do, and to later check for robustness by adding interactions between workers' traits and country dummies.

4.4 Results

4.4.1 Individual incentives

Workers can respond along two margins: effort (intensive) and labor supply (extensive). We measure the former by correct keystrokes per hour and the latter by hours worked across the two days.

Intensive margin The response to incentives differs across countries and the differences line up with the individualism measure. Figure 4.2(a) shows that, relative to their colleagues paid fixed wages, workers on individual piece rate contracts (without ranking) are 20% more productive[13] in India, which is the most individualistic country of the three. By contrast, individual incentives do not increase productivity significantly in Ghana, which is the least individualistic country. In the Philippines, both the level of individualism and the productivity increase in response to incentives (12%) are between those in Ghana and India. The pattern is similar when restricting the data to working day 1 or working day 2 only, see Figure 4a.2 in the Appendix. Figure 4.2(c) shows similar results for the implied profits of the data entry firm under individual performance incentives (no rank) compared to fixed wage.

[13]Log quality-adjusted productivity is the natural logarithm of the number of correct keystrokes per hour over the two working days of the contract. The number of correct keystrokes per hour in the contract is calculated by total number of keystrokes minus wrong keystrokes during the two working days divided by the total number of working hours during the two working days.

Kernel density estimates of quality-adjusted productivity under fixed wages and piece rates displayed in Figure 4.3 show a shift of the distribution to the right in the Philippines, while in India the left tail disappears. In Ghana, there is no change to the distribution. We have verified this pattern with quantile treatment effect regressions, not shown here in the interest of space (see Appendix Figure 4a.1).

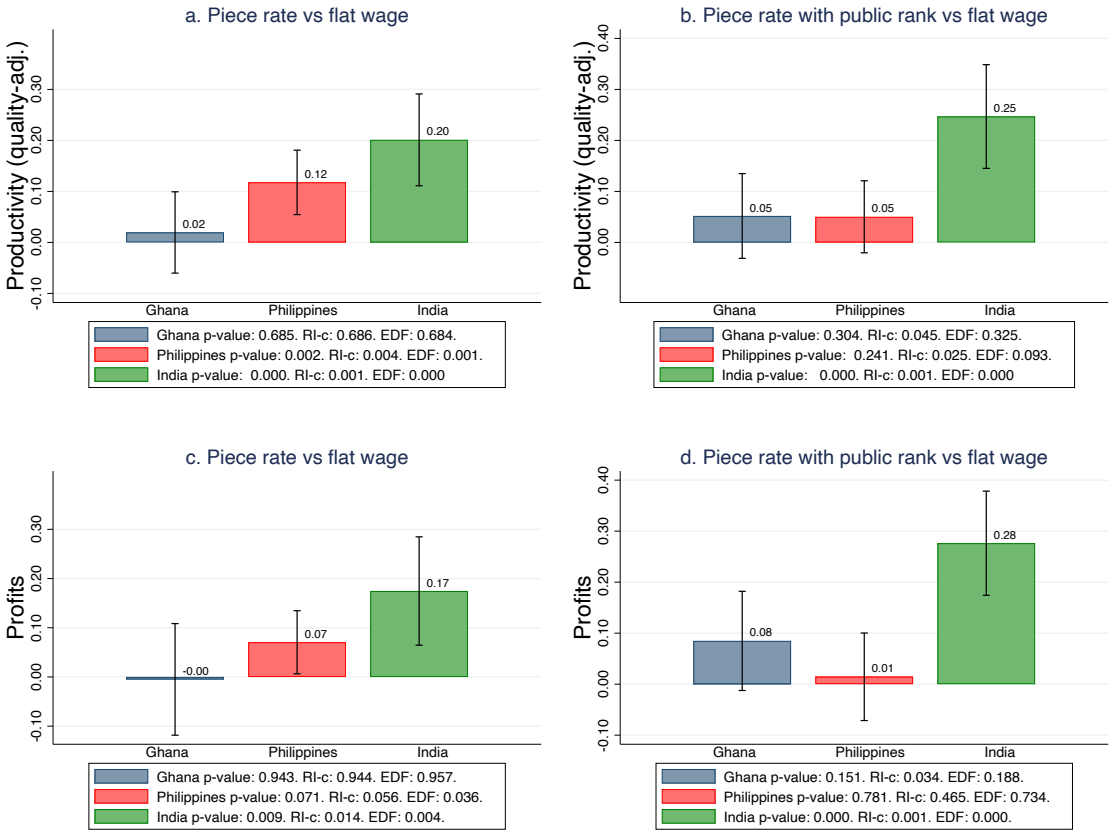
Extensive margin Figure 4.4(a) shows that in the Philippines, workers respond on the extensive margin by showing up less for work when they are randomly assigned individual piece rate, compared to those assigned to flat wage.[14] This can be understood similarly to the mechanism in DellaVigna, List and Malmendier (2012): people avoid facing social situations that they do not like. In India however, when individual piece rate is combined with publishing individual effort rankings (see further the section below), workers spend longer time at work (Figure 4.4(b)).

4.4.2 Individual incentives with public ranking

To corroborate our interpretation that the differences in responses are driven by cultural differences in the desirability of individual success, we vary the visibility of performance differences by posting individual productivity ranks. Figures 4.2(b) and 4.2(d) show that the effect of publicly ranking individual effort differs by country for both productivity and profits. In India, adding publication of individual performance has a positive effect on quality-adjusted productivity. However, in the Philippines, adding the publication of individual effort to the individual piece rate leads to a smaller productivity increase than just individual incentives without publication of ranks, when compared to flat wage. The effects in Ghana and the Philippines are similar and statistically indistinguishable. The difference-in-difference between incentives with vs. without public ranking between India and the Philippines for the outcome log quality-adjusted productivity has a p-value 0.10. This is (weakly) in line with the culture hypothesis: if individual effort is made more salient to other workers, the workers in the less individualistic country decrease their effort as the social punishment from non-conformity may outweigh the financial gain.

[14] The standard contract was for two days for a total of twelve working hours and two hours for lunch.

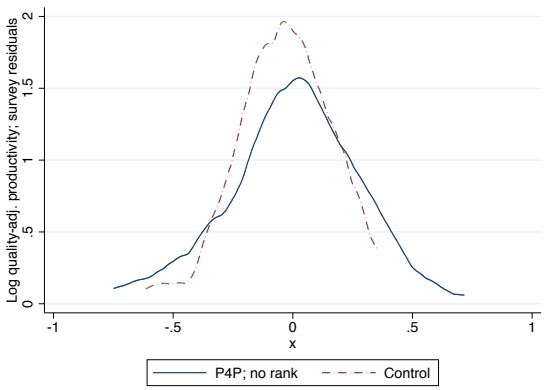
Figure 4.2: Response to Incentives: Productivity and Profits



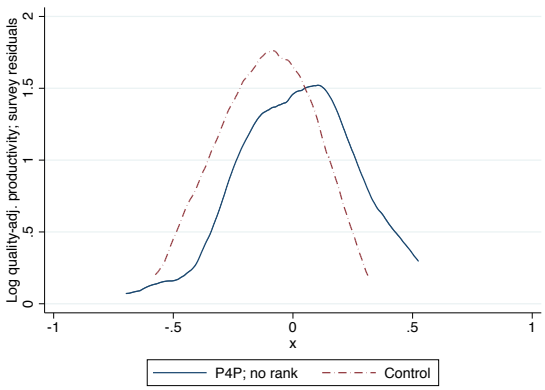
Coefficients from pooled OLS regressions with all independent variables (in natural logs) interacted with country. Control variables (which are also interacted with country) are log baseline productivity, gender, education, data entry experience, experience working with piece rate contracts, an indicator for the individual's first (of potentially 2) contract in this experiment, indicators for the dominant ethnicity/caste (Akan in Ghana, Visaya in the Philippines, and General Caste in India), and fixed effects for month and survey. The spikes represent 90% confidence intervals. The legend reports p-values for the effect of incentives differing from flat wage in each country. The randomization-c p-value is from randomization inference with individuals as unit of observation and country strata (Young, 2019). The EDF is the effective degrees of freedom correction by Young (2016).

Figure 4.3: Density plots of log quality adjusted productivity

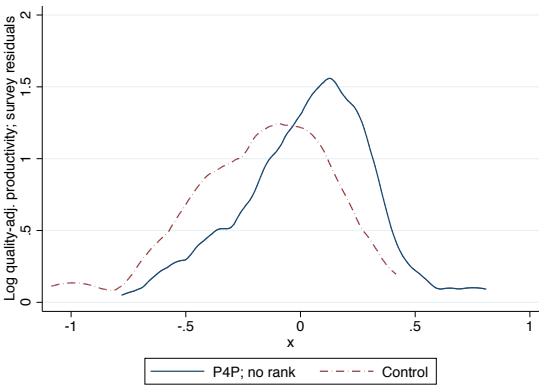
(a) Ghana



(b) The Philippines

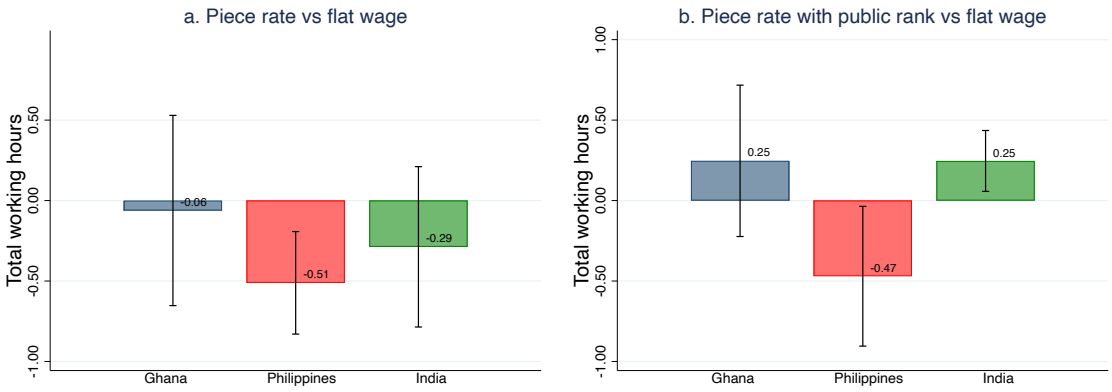


(c) India



Density plots of log quality adjusted productivity in the three countries, under IPR and flat wage. As there were 2 different surveys entered as production task in each country, the figures report the survey residuals to take out survey-specific means.

Figure 4.4: Extensive margin response



Coefficients from pooled OLS regressions with all independent variables interacted with country. Control variables (which are also interacted with country) are log baseline productivity, gender, education, data entry experience, experience working with piece rate contracts, an indicator for the individual's first (of potentially 2) contract in this experiment, indicators for the dominant ethnicity/caste (Akan in Ghana, Visaya in the Philippines, and General Caste in India), and fixed effects for month and survey. The spikes represent 90% confidence intervals.

4.4.3 Team incentives

To test the scope of conformity, that is whether it is more acceptable to respond to incentives when effort also benefits others in the group, we compare the response of individual incentives to that of team incentives across countries. The productivity effect of team incentives is slightly lower than that of individual incentives in each country (see Appendix Figure 4a.3), which is expected, as group incentives dampen the direct effect of individual effort on own pay. To test whether the effect of group incentives is the same as that of individual incentives, we divide the effect of individual incentives by team size (around 4 workers on average), and test whether that is equal to the effect of team incentives. The test does not reject equality for Ghana or the Philippines, suggesting that cultural pressures to conform operate at a higher level than the work team.

4.4.4 Different culture or different workers?

As discussed above, our strategy was to replicate as closely as possible the operations of data entry firms in the three countries. As a consequence, workers have somewhat different traits because the profile of data entry operators is not the same across countries. To allay the concern that the observed differences in response to incentives are driven by worker heterogeneity we add interactions between country and worker baseline productivity, education, experience, gender, minority religion, and age. The results reported in Table 2 in the Appendix show that our key finding is quantitatively and qualitatively robust, that is, differences in response across countries are not driven by workers' heterogeneity.[15]

4.5 Conclusion

The principle that underpins performance pay is that it serves to align the interest of the agent to that of the principal by giving the agent a reward that they value (more pay) for behaviour that benefits the principal. Performance pay however also creates inequality in pay among agents of different ability who put in the same effort in the same task. We have shown that the same incentive scheme is less effective in societies where these differences are less culturally acceptable. This highlights the importance of understanding how social

[15]We also checked whether there is heterogeneity in treatment effects across individual motivations within a country (individual motivations were also measured with the Hofstede questionnaire). Despite the fact that individuals showed considerable differences in their motivations according to the questionnaire, we did not find any treatment effect heterogeneity along these lines (see Appendix Figure 4a.6 and Table 4a.3) – workers seem to conform to the culture in their country in their behaviour.

norms interact with individual motives when designing rewards systems in organizations.

4.6 Appendix

4.6.1 Details of Experimental Design

Firms

The experimental firms were established in the Osu neighborhood of Accra, Ghana in September 2010, in the Salt Lake neighborhood of Kolkata, India in January 2011 and in the Divisoria neighborhood of Cagayan de Oro, on the southern island of Mindinao, Philippines in May 2011. Each firm had at least two rooms, each with a large common table and five seats. We distributed workers evenly between these two rooms in all three offices. The offices contained ten identical laptops with disabled internet connectivity for minimal distractions. A Wireless Local Area Network (WLAN) connected these computers to a central server computer, which permitted a manager to monitor worker progress and electronically collect worker output and generate wages at the end of contracts based on productivity. The treatments/contracts were the following:

Production task

The production task required workers to enter coded data from paper surveys into the Epidata interface, a commonly used data entry software. In general, the surveys contained short numeric and string fields. At the end of the two-day contract, the on-site manager paid workers in cash. For workers under the flat wage contract, we did not carry out a salary calculation and paid each worker a cash wage of 50 GHC, 500 INR or 900 PhP. For the incentive contracts, the on-site manager collected all data entered by the workers and calculated their pay using based on the conditions of the workers' assigned contracts. Workers received cash along with a receipt stating the number of key strokes entered and the calculation used to define their two-day wage.

In Ghana, during the first contract, all workers entered the Pure Home Water Survey (PHWS) survey. During the second contract, they entered the Formal Savings survey (FS). In the Philippines, workers entered the CDO Fresh survey during the first round of contracts and the Health and Financial Services survey (HaFS) during the second round. In India, workers entered the Market survey during the first contract and the Village Welfare Society (VWS) survey in the second.

Table 4a.1: Summary statistics of the experiment.

	Ghana		Philippines		India	
	mean	sd	mean	sd	mean	sd
Duration (months)	10.3		8.2		11.2	
N unique workers	294		404		557	
Average man days per contract	2.0	0.16	2.0	0.11	2.0	0.19
Avg. man days in group per contract	7.8	1.34	9.1	0.97	7.4	1.76
Avg. number of workers in group 1	4.0	0.59	4.7	0.54	3.8	0.78
Avg. number of workers in group 2	3.9	0.72	4.5	0.50	3.6	0.87
Average minutes spent per survey	3.3	0.98	2.4	0.80	2.5	2.17
Test: Total fields entered	130.3	47.26	148.2	47.77	137.7	45.60
Test: Total correct fields entered	122.1	45.82	136.4	46.36	122.8	43.28
Total fields entered, 2 days	13327	3415	24342	13708	24566	10516
Total correct fields entered, 2 days	12625	3321	23714	13943	23235	10397
2-day payment (excl FW) 2011$PPP	64.0	14.53	65.3	11.47	28.4	8.25
Payment as share of FW (excl. FW)	1.0	0.23	1.1	0.19	1.1	0.31
Surveys	PHWS FS		Market VWS		CDO HaFS	
Observations	459		553		571	

Notes: The surveys referred to are the Pure Home Water Survey (PHWS), Formal Savings Survey (FS), CDO Fresh (CDO), Health and Financial Services Survey (HaFS), Market Survey (Market) and Village Welfare Society (VWS).

Contracts

1. **Flat wage** (FW): Workers received a fixed rate[16] for each day of work, regardless of the quantity or quality of data entered.

2. **Individual piece rate** (IPR): Workers received a fixed piece rate based on the total number of key strokes entered over the two days of the contract.[17]

3. **Individual piece rate with rank** (IPR rank): Same as above, but managers also ranked workers based on their productivity and publicly displayed rankings with full names three times a day to provide public visibility of relative and absolute performance.

4. **Group piece rate** (GPR): Workers were divided into two groups (of up to five workers each; with on average four workers each; see Table 1) and paid a salary

[16]The fixed daily wage was 25 Ghana Cedis (GHC) = 2011USD 16.53 (using exchange rate 1 2011USD = 1.512 GHC, all from https://data.worldbank.org/indicator/PA.NUS.FCRF), 450 Philippines Pesos (PhP) = 2011USD 10.43 (using exchange rate 1 2011USD = 43.31 PhP), or 250 Indian Rupees (INR) = 2011USD 5.36 (using exchange rate 1 2011USD = 46.67 INR), respectively (paid in cash).

[17]In each country, we calculated the piece rate as the flat wage divided by the median number of total key strokes entered during the 2-day first flat wage contract, which in Ghana was 50/46,729 = .00107 GHC per keystroke for the Pure Home Water Survey (PHWS), and 50/80,425 = .0006217 GHC for the Formal Savings (FS) survey, in India 500/31,083 = .0154421 INR for the Market survey and 500/70385 = .0071038 INR for the Village Welfare Society (WVS) survey, and in the Philippines 900/80,032 = .0112455 PhP for the CDO Fresh survey and 900/107136 = .0084 PhP for the Health and Financial Services Survey (HaFS) survey.

proportional to the total number of key strokes entered by all group members over two days.

5. **Group piece rate with rank** (GPR rank): Same as the GPR, but managers also ranked workers based on their productivity and publicly displayed rankings of individual performance with full names three times a day to provide public visibility of relative and absolute performance.

Interviewing and testing

Managerial staff called the applicants in the order they applied and invited them to in-person interviews in groups of up to ten applicants. The interviews took approximately one hour and consisted of two baseline typing tests and a demographic and work preferences survey. Prior to the start of the interview, the on-site manager would explain IPA/J-PAL Data Services' mission to applicants: to provide high-quality paper-to-computer data entry services. The on-site manager also detailed the logistics associated with the two-day contracts and explained that we would store all applicant information in a secure database.

Applicants took one baseline typing test in Excel, and the other in Epidata, a commonly used data entry interface software. The Excel test featured a recurring list of numbers that applicants entered as quickly and accurately as possible for five minutes. During the fifteen-minute Epidata test, applicants in all three contries entered the Pure Home Water Survey (PHWS) to enable direct comparisons of baseline productivity across countries. We then used this same survey in Ghana during the first experimental contract, but did not use it in either India or the Philippines, where we instead used local surveys. Following both tests, applicants took a 20-minute survey. In addition to demographic information, this survey collected information regarding work experience and preferences, including prior experience in data entry and prior wages received (flat wage or piece rate). The work preferences/motivations module utilized questions drawn from the 2008 Geert Hofstede Values Survey Module (see Motivation questions in Table 4). Responses to these questions allow us to measure our workers' ex ante preferences along individualism-collectivism and other cultural dimensions. We translated the survey into the relevant language in each country.

Hiring and Working

We invited all applicants who scored above a certain level of accuracy in the Epidata test to at least one contract and a maximum of two contracts. We eliminated from consideration individuals whose accuracy during the Epidata test was below 65% in Ghana. In

India, that level was 50% to account for recruitment difficulties, although an outlier scored 38.8% accuracy due to an error in the initial calculation. In the Philippines, the individual accuracy threshold was 62%. We then stratified individuals using three variables: gender, ethnicity (dominant vs. all others) and baseline Epidata performance (above vs. below median productivity). In Ghana, the ethnicity breakdown was Akan and non-Akan, in India General Caste and other, and in the Philippines Visaya and other. By using a Power Point presentation to train workers, we minimized variations in training quality that could have occurred over the course of the experiment.

We ran two 2-day contracts per week over the course of a year and a half, with an experimental duration of ten months, one year and nine months in Ghana, India and the Philippines, respectively. We executed 73 contracts in Ghana, 86 in India and 61 in the Philippines. Despite inviting ten workers to each contract, we found that approximately seven workers attended each contract on average. To limit this phenomenon, we invited workers to work on an alternative date in the event that they were unable to attend on their originally assigned date. If a worker did not answer his or her phone after at least two attempts to contact him or her, we eliminated them from consideration for future contracts.

In order to capture order effects, we offered each worker up to two contracts under either the same or different payment terms. To minimize selection bias, the on-site manager invited workers to contracts one to three days in advance without indicating which contract type the worker would receive. The on-site manager told workers that the actual salary would depend on the contract type they received, but that the average worker should expect to receive 25 Ghana Cedis (GHC), 250 Rupees (INR) or 450 Philippines Pesos (PhP) per day, and that the contract could be either flat wage or piece rate.[18] During the call, the on-site manager also reiterated that the contract was for two days. To avoid spillovers, all workers in a given two-day contract received the same contract (i.e., all workers in the office at a given time received flat wage or IPR).

[18]In each country, we calculated the piece rate as the flat wage divided by median number of total key strokes entered during the 2-day first flat wage contract, which in Ghana was 50/46,729 = .00107 GHC per keystroke for the PHWS survey and 50/80,425 = .0006217 GHC for the FS survey, in India 500/31,083 = .0154421 INR for the Market survey and 500/70385 = .0071038 INR for the WVS survey, and in the Philippines 900/80,032 = .0112455 PhP for the CDO Fresh survey and 900/107136 = .0084 PhP for the HaFS survey.

4.6.2 Additional evidence referenced in paper

Table 4a.2: Culture vs. worker traits in explaining log quality-adjusted productivity.

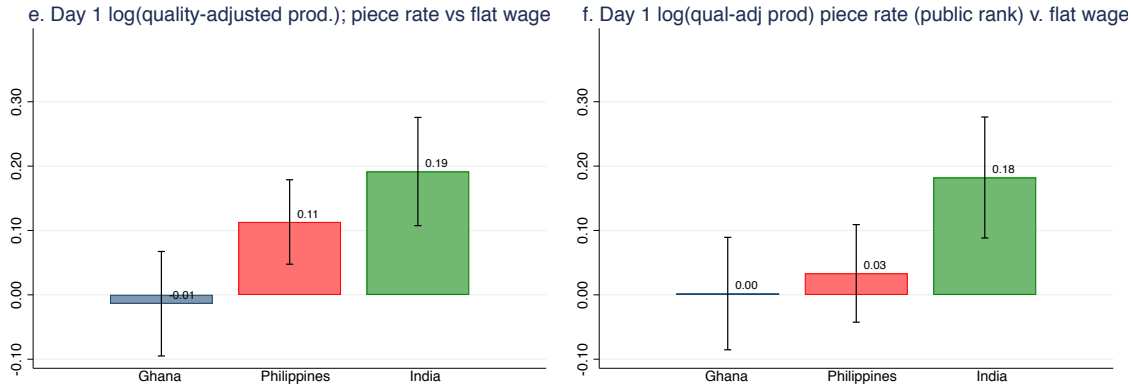
	(1) No control	(2) Main	(3) Ability	(4) Education	(5) Experience	(6) Gender	(7) Religion	(8) Age
Incentives (Ghana)	0.046	0.065	0.065	0.072	0.061	0.063	0.067	0.065
	(0.031)	(0.030)	(0.030)	(0.031)	(0.030)	(0.030)	(0.031)	(0.030)
India flat wage	0.774	0.789	0.830	0.778	0.694	0.712	0.804	0.802
	(0.060)	(0.064)	(0.123)	(0.072)	(0.087)	(0.068)	(0.065)	(0.147)
Phil. flat wage	-0.344	-0.371	-0.417	-0.391	-0.415	-0.398	-0.363	-0.197
	(0.045)	(0.060)	(0.119)	(0.063)	(0.081)	(0.062)	(0.061)	(0.139)
India incentive	0.164	0.136	0.137	0.129	0.139	0.137	0.131	0.136
	(0.057)	(0.055)	(0.055)	(0.055)	(0.054)	(0.055)	(0.056)	(0.055)
Phil. incentive	0.049	0.025	0.026	0.018	0.027	0.027	0.024	0.015
	(0.044)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.041)
Log basel. prod.		0.257	0.257	0.257	0.255	0.259	0.260	0.249
		(0.019)	(0.035)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Log basel. prod. India			-0.020					
			(0.050)					
Log basel. prod. Phil.			0.020					
			(0.045)					
Interaction variable				-0.051	-0.042	-0.052	0.039	-0.003
				(0.025)	(0.054)	(0.024)	(0.039)	(0.004)
Interaction var. India				0.030	0.104	0.097	-0.084	-0.000
				(0.043)	(0.060)	(0.037)	(0.061)	(0.005)
Interaction var. Phil.				0.040	0.025	0.025	-0.074	-0.008
				(0.034)	(0.057)	(0.031)	(0.048)	(0.005)
N	1379	1379	1379	1379	1379	1379	1378	1376

Incentives are all 4 incentive treatments pooled. In columns 4-8, the interaction variable is the column header.
Standard errors in brackets. Controls in all columns but (1): ethnicgroup, ever worked in data entry, piece rate experience, whether this was their first contract, survey and month FE.
Experience is measured by an indicator of ever having had paid employment. Education is an indicator for university.
Gender is an indicator for being male. Religion is an indicator for belonging to a minority religion in each country, i.e., not Christian in Ghana and the Philippines and not Hindu in India. Age is a continuous variable for age based on birth date.

Figure 4a.1: Quantile treatment effects

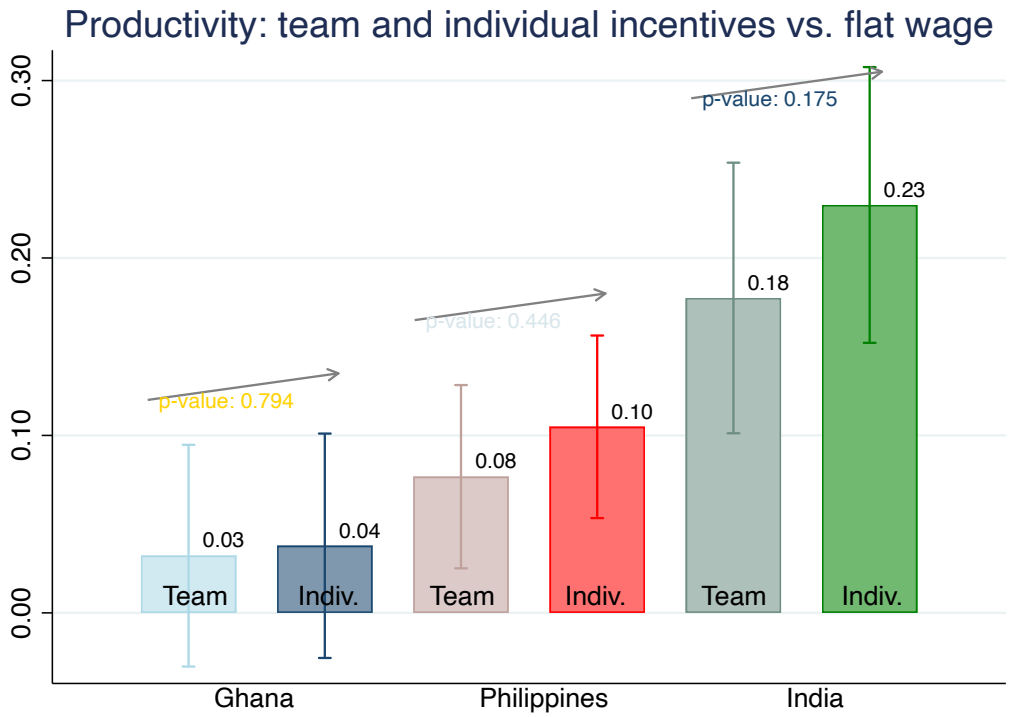
Notes: Individual Piece Rate Treatment: Coefficients from separate QTE regressions for each country with 95% confidence intervals from bootstrapped standard errors with 1000 draws. Control variables are log baseline productivity, gender, education, data entry experience, experience working with piece rate contracts, an indicator for the individual's first (of potentially 2) contract in this experiment, indicators for the dominant ethnicity/caste (Akan in Ghana, Visaya in the Philippines, and General Caste in India), and fixed effects for month and survey.

Figure 4a.2: Working day one effects only.



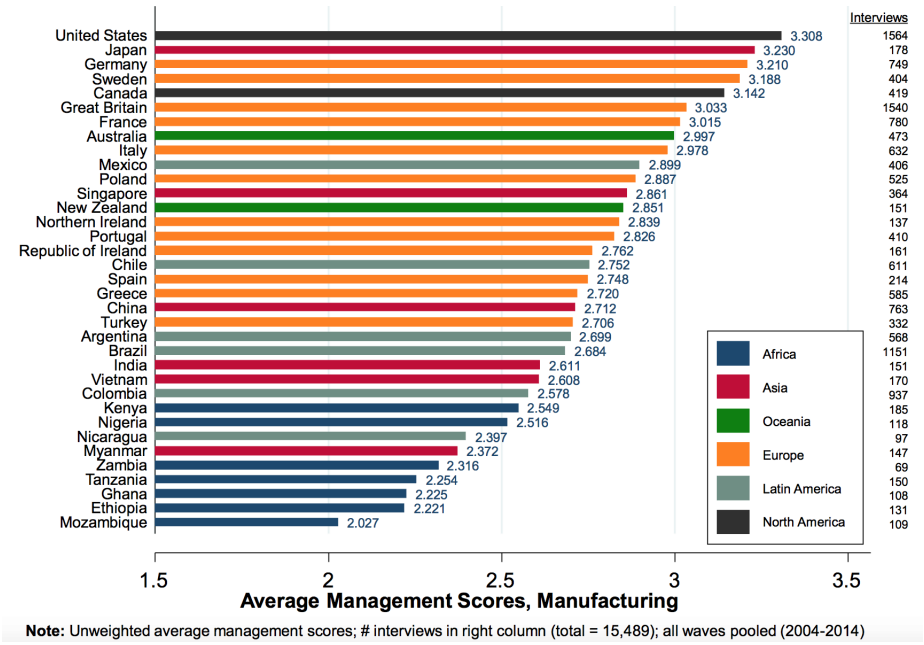
Notes: Day 1 only: Coefficients from pooled OLS regressions with all independent variables interacted with country. Control variables (which are also interacted with country) are log baseline productivity, gender, education, data entry experience, experience working with piece rate contracts, an indicator for the individual's first (of potentially 2) contract in this experiment, indicators for the dominant ethnicity/caste (Akan in Ghana, Visaya in the Philippines, and General Caste in India), and fixed effects for month and survey. The spikes represent 90% confidence intervals.

Figure 4a.3: Team and individual incentives.



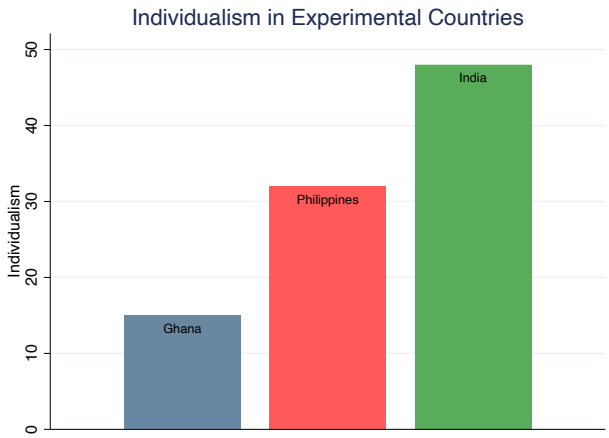
Notes: Comparison of effects on log quality-adjusted productivity from group and individual incentives, compared to flat wage. Note: ranked and not ranked treatments are pooled. Spikes represent 90% confidence intervals.

Figure 4a.4: Variation in management practices across countries



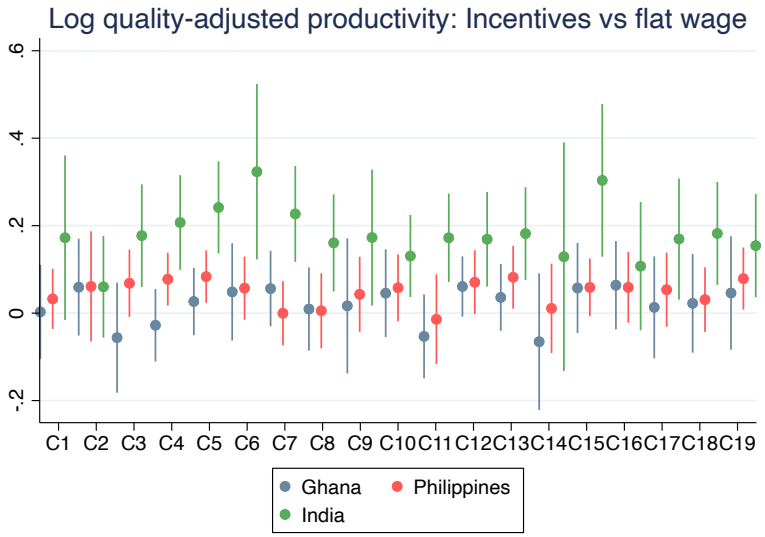
Notes: Evidence of large variation in management practices across countries. Source: Bloom et al 2016 (7 questions of 18 about incentives).

Figure 4a.5: Hofstede's individualism measures for the three countries.



Notes: Source: www.hofstede-insights.com/country-comparison/

Figure 4a.6: Heterogeneous treatment effects on workers' motivations



Notes: Coefficients from 19 regressions of log quality-adjusted productivity on incentives (all 4 incentive treatments pooled) vs flat wage in each country, one for each subsample which has the highest importance (=1) on the motivation question c1-19 (see Table 4a.3 for a description of the motivations). The spikes represent 90% confidence intervals.

Table 4a.3: Summary of results on motivations

Question	Motivation	Ghana	Philippines	India
c1	Do Challenging Work			
c2	Live in Desirable Area			
c3	Opportunity for High Earnings			X
c4	Work with People who Cooperate Well		X	X
c5	Have Training Opportunities		X	X
c6	Good Fringe Benefits			X
c7	Get Recognition			X
c8	Good Physical Working Conditions			X
c9	Freedom for Own Approach to Job			X
c10	Have Job Security			X
c11	Opportunity for Advancement			X
c12	Have Good Relationship with Manager		X	X
c13	Fully use skills		X	X
c14	Have Sufficient Personal Time			
c15	Boss You Respect			X
c16	Pleasant Colleagues		X	
c17	Interesting Work			X
c18	Consulted by Boss			X
c19	Job Respected by Family/Friends		X	X

Notes: The motivations that are checked are those for which workers who rate that of utmost importance have a positive and significant (on 10% level) effect on log quality-adjusted productivity of incentives (all 4 incentive treatments pooled) compared to flat wage.

The motivations are not mutually exclusive. The variables are defined as answering "1 = of utmost importance" on the question "Please think of an ideal job, disregarding your present job, if you have one. In choosing an ideal job, how important would it be to you to ... (please circle one answer in each line across): 1 = of utmost importance 2 = very important 3 = of moderate importance 4 = of little importance 5 = of very little or no importance", for e.g. the following statements: "Do work that is interesting", "Work with People who Cooperate Well", "Have opportunity for high earnings" and "Have opportunity for advancement" (see further Table 4a.3 listing all the motivations) In each country, there are people who do not answer the highest importance (=1) to any of these - they are the control groups in these regressions (183 people in Ghana, 184 in India, and 139 in the Philippines).

Chapter 5

Conclusion

The three core chapters making up this thesis have studied service provision and work in various subfields of economics, but with a common theme of emphasising the spatial perspective. One chapter has considered the different matches of doctors and patients across geographic locations that become possible with digital technology. Another chapter has used geospatial data and a spatial regression discontinuity design to study the effects of early infrastructure investment. The final chapter takes seriously that the response to incentives may vary across geographic locations which have different cultural norms, and studies this with three randomized controlled trials in different sites.

A common theme in the conclusions is that there are vast differences across locations, which can be attributed either to policy, to geograpic sorting, or to norms. In Chapter 3, neighborhoods have developed differently over 40 years depending on whether they were covered by the policy of minimal early infrastructure investment. This is partly because this policy crowded in private investment, but also because of sorting. In Chapter 2, local healthcare provision before digital care varied more than expected due to sorting, but could be equalized with digital technology. In Chapter 4, norms differ across countries which creates the need for adaptations of, e.g., multinationals' management and personnel policies. In sum, economics depends on spatial forces.

Chapter 6

Bibliography

6.1 Chapter 2 References

Abaluck, Jason, Agha, Leila, Kabrhel, Chris, Raja, Ali, and Venkatesh, Arjun. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care". The American Economic Review 106.12 (2016): 3730-764.

Abdulkadiroglu, Atila, Yeon-Koo Che, Parag A. Pathak, Alvin E. Roth, and Olivier Tercieux, "Efficiency, Justiffied Envy, and Incentives in Priority-Based Matching". American Economic Review: Insights, Vol. 2, No. 4, (Pp. 425-42).

Agency for Healthcare Research and Quality. 2020. "2019 National Healthcare Quality and Disparities Report". Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Pub. No. 20(21)-0045-EF. Link. Accessed 4 October 2021.

Alsan, Marcella, Owen Garrick, and Grant Graziani. 2019. "Does Diversity Matter for Health? Experimental Evidence from Oakland". American Economic Review 109.12: 4071-4111.

Andrews, Isaiah, Toru Kitagawa and Adam McCloskey. 2021. "Inference on Winners." Unpublished.

Aucejo, Esteban M., Patrick Coate, Jane Cooley Fruehwirth, Sean Kelly and Zachary Mozenter. 2019. "Teacher Effectiveness and Classroom Composition". CEP Discussion Paper No 1.

Bau, Natalie. 2021. "Estimating An Equilibrium Model of Horizontal Competition in

Education". Unpublished. Accessed 14 September 2021. Link.

Baumhäkel M, Müller U, Böhm M. 2009. "Influence of gender of physicians and patients on guideline-recommended treatment of chronic heart failure in a cross-sectional study". Eur J Heart Fail, 11(3):299-303.

Becker, Gary S. 1973. "A Theory of Marriage: Part I." *Journal of Political Economy*. Volume 81 Number 4.

Bendz, Anna (2011) "Vårdvalet i Västsverige" in Annika Bergström (ed.) *Västsvensk vardag*. Göteborgs universitet: SOM-institutet.

Bergeron A, Bessone P, Kabeya JK, and Tourek G. "Quality and Optimal Matching of Bureaucrats: Evidence from Tax Collection in the DRC." Unpublished. Link. Accessed August 21, 2021.

Berthold HK, Gouni-Berthold I, Bestehorn KP, Böhm M, Krone W. 2008. "Physician gender is associated with the quality of type 2 diabetes care." J Intern Med. 264(4):340-350.

Bhattacharya, Debopam, and Pascaline Dupas. 2011. "Inferring welfare maximizing treatment assignment under budget constraints." Journal of Econometrics 167 (2012) 168–196. Journal of the American Statistical Association 104 (486), 486–500.

Bhattacharya, Debopam. 2009. "Inferring optimal peer allocation using experimental data". Journal of the American Statistical Association 104 (486), 486–500.

Bond, Philip and James Dow. 2021. "Failing to forecast rare events." *Journal of Financial Economics*, 2021-06.

Bonhomme, S., Lamadon, T. and Manresa, E. 2019. "A Distributional Framework for Matched Employer Employee Data." Econometrica, 87: 699-739.

Cabral, Marika and Marcus Dillender. 2021. "Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctors." Manuscript. Link. Accessed on August 3, 2021.

Centers for Disease Control and Prevention. 2019. "Antibiotic Resistance Threats in the. United States." Link. Accessed on September 1, 2021.

Chen, Yiqun, Petra Persson and Maria Polyakova. 2021. "The Roots of Health Inequality

and The Value of Intra-Family Expertise." Mimeo. https://web.stanford.edu/~perssonp/FamilyDoctors.pdf

Chen, Yiqun, Raj Chetty, Luke Chu, David Cutler, Lorena Di Bono, Andreas Haller, Katja Hofmann, Jonas Minet Kinge, Claus Thustrup Kreiner, Hsien-Ming Lien, Kevin Milligan, Benjamin Milner, Thomas Minten, Torben Heien Nielsen, Petra Persson, Maria Polyakova, Tammy Schirle, Benjamin Ly Serena, Johannes Spinnewijn, Stefan Staubli, Michael Stepner, Tzu-Ting Yang, Yuting Zhang and Josef Zweimüller. 2021. "Health Inequality Around the World: Comparing the Relationship Between Income and Life Expectancy in Ten High-Income Countries." Unpublished.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." American Economic Review, 104 (9): 2593-2632.

Condie, Scott, Lars Lefgren, David Sims. 2014. Teacher heterogeneity, value-added and education policy. Economics of Education Review, Volume 40, 76-92.

Currie, Janet, and W Bentley MacLeod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." Journal of Labor Economics, 35(1): 18977.

Cutler, David, Jonathan S. Skinner, Ariel Dora Stern, and David Wennberg. 2019. "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending." American Economic Journal: Economic Policy, 11(1): 192-221.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err". Journal of Experimental Psychology: General, 144: 114–126.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. 1998. "Comorbidity measures for use with administrative data". Medical Care, 36(1), 8–27.

Fenizia, Alessandra. 2020. "Managers and Productivity in the Public Sector." Manuscript. Link. Accessed 24 July 2021.

Fernemark, Hanna, Janna Skagerström, Ida Seing, Carin Ericsson and Per Nilsen. 2020. "Digital consultations in Swedish primary health care: a qualitative study of physicians' job control, demand and support." BMC Family Practice, 21:241.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2021. "Place-Based Drivers of Mortality: Evidence from Migration." American Economic Review, 111 (8): 2697-2735.

Finkelstein, Amy, Petra Persson, Maria Polyakova and Jesse Shapiro. 2021. "A Taste of Their Own Medicine: Guideline Adherence and Access to Expertise." Unpublished. Link. Accessed 24 July 2021.

Fitzmaurice, Garrett M., Nan M. Laird, and James H. Ware. 2004. "Applied Longitudinal Analysis." Hoboken, NJ: Wiley, ISBN: 0-471-21487-6. xix + 506 pp.

Folkhälsomyndigheten. 2013. "Samhällsekonomiska konsekvenser av antibiotikaresistens." Link. Accessed on September 13, 2021.

Frakes, Michael, Jonathan Gruber, and Anupam Jena. 2021. "Is great information good enough? Evidence from physicians as patients." Journal of Health Economics, 75, 102406.

Gabrielsson-Järhult, Felicia, Kristina Areskoug-Josefsson, and Peter Kammerlind. 2019. "Digitala vårdmöten med läkare: Rapport av kvantitativ och kvalitativ studie." Manuscript.

Gale, D., and Shapley, L. (1962). "College Admissions and the Stability of Marriage." The American Mathematical Monthly, 69(1), 9-15.

Glenngård, Anna H. 2020. "International Health Care System Profiles: Sweden." In Tikkanen, Roosa, Robin Osborn, Elias Mossialos, Ana Djordjevic, and George Wharton, eds. 2020. *International Health Care System Profiles* Link.

Goldfarb, Avi and Catherine Tucker. 2019. "Digital Economics." Journal of Economic Literature, 57(1): 3–43.

Göransson M, Persson A-C, Abelsson A. 2020. "Triage in primary healthcare." Nordic Journal of Nursing Research: 40(4):213-220.

Graham, Bryan S. 2011. "Chapter 19 - Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers." in Benhabib, Jess, Alberto Bisin, Matthew O. Jackson, eds.: *Handbook of Social Economics* North-Holland, Volume 1, 965-1052.

Graham, Bryan S., Guido W. Imbens and Geert Ridder. 2020. "Identification and Efficiency Bounds for the Average Match Function Under Conditionally Exogenous Matching." Journal of Business and Economic Statistics, 38:2, 303-316.

Graham, Bryan S., Geert Ridder, Petra Thiemann and Gema Zamarro. 2021. "Teacher-to-classroom assignment and student achievement." Manuscript. Retrieved on 2 June, 2021. Link.

Guyton, John, Patrick Langetieg, Daniel Reck, Max Risch and Gabriel Zucman. 2021. "Tax Evasion at the Top of the Income Distribution: Theory and Evidence." Mimeo. Link. Accessed 22 June 2021.

Hill, Andrew, Daniel Jones and Lindsey Woodworth. 2018. "Physician-patient race-match reduces patient mortality." Unpublished.

Kasy, Maximilian. 2016. "Partial identification, distributional preferences, and the welfare ranking of policies." The Review of Economics and Statistics, Vol.98 (1), p.111-131.

Kasy, Maximilian and Alexander Teytelboym. 2021. "Learning by Matching." Unpublished. Link. Accessed 21 Sep 2021.

Kim C, McEwen LN, Gerzoff RB, et al. 2005. "Is physician gender associated with the quality of diabetes care?" Diabetes Care, 28(7): 1594-1598.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics*, 133(1): 237–293.

Koopmans, T.C., Beckmann, M., 1957. "Assignment problems and the location of economic activities". Econometrica 25 (1), 53–76.

Lindenlaub, Ilse. 2017 "Sorting Multidimensional Types: Theory and Application." The Review of Economic Studies, Volume 84, Issue 2, Pages 718–789.

Lockwood, J. R. and Daniel F. McCaffrey 2009. "Exploring Student-Teacher Interactions in Longitudinal Achievement Data." Education Finance and Policy, 4(4); 439-467.

Manski, C., 2004. "Statistical treatment rules for heterogeneous populations". Econometrica 72 (4), 1221–1246.

Manski, CF. 2018. "Reasonable patient care under uncertainty. Health Economics." 27: 1397– 1421. Link.

Marx, Benjamin, Vincent Pons and Tavneet Suri. 2021. "Diversity and team performance in a Kenyan organization." *Journal of Public Economics*, Volume 197, 104332.

Miller, Amalia R. and Catherine E. Tucker. 2011. "Can Health Care Information Technology Save Babies?" *Journal of Political Economy*, 119(2): 289-324.

Molitor, David. 2018. "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration". American Economic Journal: Economic Policy 2018, 10(1): 326–356. Link.

Muñoz, Mathilde. 2021. Trading Non-Tradables: The Implications of Europe's Job Posting Policy. Unpublished.

National Board of Health and Welfare / Socialstyrelsen. 2018. "Tillgänglighet i hälso- och sjukvården." 2018-2-16.

OECD. 2017. "Health at a glance 2017: OECD indicators." Link.

O'Donnell, Catherine A. 2000. "Variation in GP referral rates: what can we learn from the literature?" Family Practice, 17(6): 462–471.

Page, Anthea, Sarah Ambrose, John Glover and Diana Hetzel. 2007. "Atlas of Avoidable Hospitalisations in Australia: ambulatory care-sensitive conditions." Adelaide: PHIDU, University of Adelaide.

Papageorgiou, T. 2014. "Learning Your Comparative Advantages." Review of Economic Studies, 81, 1263–1295

Reck, Daniel, Max Risch and Gabriel Zucman. 2021. "Response to a Comment by Auten and Splinter on 'Tax Evasion at the Top of the Income Distribution: Theory and Evidence'." Link. Accessed August 8, 2021.

Rocha, J.V.M., Marques, A.P., Moita, B. and R. Santana. 2020. "Direct and lost productivity costs associated with avoidable hospital admissions." BMC Health Serv Res 20(210). Link.

Roghman, K. J. and Zastowny, T. R. 1979. "Proximity as a factor in the selection of health care providers: emergency room visits compared to obstetric admissions and abortions." Social Science and Medicine 13, 61–9.

Scholle, S. H., Roski, J., Dunn, D. L., Adams, J. L., Dugan, D. P., Pawlson, L. G., and Kerr, E. A. 2009. "Availability of data for measuring physician quality performance". The American journal of managed care, 15(1), 67–72.

Shapley, L.S., Shubik, M. The assignment game I: The core. Int J Game Theory 1, 111–130 (1971). Link.

Skinner, Lucy, Douglas Staiger, David Auerbach and Peter Buerhaus. 2019."Implications of an Aging Rural Physician Workforce." New England Journal of Medicine, 381(4):299-301.

Smith, Adam. 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. McMaster University Archive for the History of Economic Thought

Socialstyrelsen och Sveriges Kommuner och Landsting. 2011. "Öppna jämförelser av hälso- och sjukvårdens kvalitet och effektivitet. Jämförelser mellan landsting." Accessed 13 September 2021.

Socialstyrelsen 2017. "Konsekvensutredning – förslag till föreskrift om belopp för vård av utskrivningsklara patienter." Link. Accessed 13 September 2021.

Statistics Sweden Statistikdatabasen, 2021. Link.

Strama. 2019. "Rekommendationer för kvalitetsindikatorer vid digitala vårdmöten." Link. Accessed 20 August 2021.

Strama. 2019. "Rekommendationer för kvalitetsindikatorer vid digitala vårdmöten." Link. Accessed 20 August 2021.

Studnicki, J. 1975. "The minimization of travel efforts as a delineating influence for urban hospital service area". International Journal of Health Service 5, 679–93.

Theis, R. P., Stanford, J. C., Goodman, J. R., Duke, L. L., and Shenkman, E. A. 2017. "Defining 'quality' from the patient's perspective: findings from focus groups with Medicaid beneficiaries and implications for public reporting." Health expectations : an international journal of public participation in health care and health policy, 20(3), 395–406.

Tinbergen, Jan. 1962. "An Analysis of World Trade Flows" in Shaping the World Economy, edited by Jan Tinbergen. New York, NY: Twentieth Century Fund.

Tsugawa, Yusuke, Joseph P Newhouse, Alan M Zaslavsky, Daniel M Blumenthal, Anupam B Jena. 2017a. "Physician age and outcomes in elderly patients in hospital in the US: observational study." BMJ; 357.

Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM, Jha AK. 2017. "Compar-

ison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians." JAMA Intern Med, 177(2):206–213.

Tucker, Catherine E. and Shuyi Yu. 2019. "Does IT Lead to More Equal or More Unequal Treatment? An Empirical Study of the Effect of Smartphone Use on Social Inequality in Employee-Customer interactions". Link. Accessed 2 August 2021.

Vasilik, Ashley. 2021. "Triage Process in Primary Care Clinics". DNP Scholarly Projects. 51. Link. Accessed 4 October 2021.

Willis, R. J. and Rosen, S. 1979. "Education and Self-selection." Journal of Political Economy, 87, S7–36.

Yeomans, M, Shah, A, Mullainathan, S, Kleinberg, J. 2019. "Making sense of recommendations." J Behav Dec Making, 32: 403–414.

Zeltzer, Dan, Liran Einav, Joseph Rashba, and Ran D. Balicer. 2021. "The Impact of Increased Access to Telemedicine." SIEPR Working Paper No. 21-038. Link. Accessed 20 August 2021.

6.2 Chapter 3 References

Ali, Daniel Ayalew, Matthew Collin, Klaus Deininger, Stefan Dercon, Justin Sandefur, and Andrew Zeitlin. 2016. "Small Price Incentives Increase Women's Access to Land Titles in Tanzania." J. Development Econ. 123:107–22.

Angel, Shlomo. 2012. Planet of Cities. Cambridge, MA: Lincoln Inst. Land Policy.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." American Econ. J. Appl. Econ. 7 (1): 22–53.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." J.P.E. 115 (4): 588–638.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen. 2011. "Inference with Dependent Data Using Cluster Covariance Estimators." J. Econometrics 165 (2): 137–51.

Bleakley, Hoyt, and Jeffrey Lin. 2012. "Portage and Path Dependence." Q.J.E. 127 (2): 587–644.

Brinkhoff, Thomas. 2017. “City Population.” https://www.citypopulation.de.

Buckley, Robert M., and Jerry Kalarickal. 2006. Thirty Years of World Bank Shelter Lending: What Have We Learned? Washington, DC: World Bank.

Castells-Quintana, David. 2017. “Malthus Living in a Slum: Urban Concentration, Infrastructure and Economic Growth.” J. Urban Econ. 98:158–73.

Cohen, Michael, Callisto E. Madavo, and Harold Dunkerley. 1983. “Learning by Doing: World Bank Lending for Urban Development 1972–82.” World Bank, Washington, DC.

Collin, Matthew, Justin Sandefur, and Andrew Zeitlin. 2015. “Falling off the Map: The Impact of Formalizing (Some) Informal Settlements in Tanzania.” Working paper, Centre for the Study of African Economies, Univ. Oxford.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2016. “The Production Function for Housing: Evidence from France.” IZA Discussion Paper no. 1037, Institute of Labor Economics, Bonn, Germany.

Conley, Timothy G. 1999. “GMM Estimation with Cross Sectional Dependence.” J. Econometrics 92 (1): 1–45.

Coville, Aidan, and Yu-hsuan Su. 2014. “From the Ground Up: An Impact Evaluation of the Community Infrastructure Upgrading Programme in Dar-es-Salaam.” Working paper, World Bank, Washington, DC.

Dar Ramani Huria. 2016. “Community-Based Mapping Project in Dar-es-Salaam.” ramanihuria.org/data/.

Dell, Melissa. 2010. “The Persistent Effects of Peru’s Mining Mita.” Econometrica 78 (6): 1863–903.

Diewert, W. Erwin, Jan de Haan, and Rens Hendriks. 2015. “Hedonic Regressions and the Decomposition of a House Price Index into Land and Structure Components.” Econometric Rev. 34 (1–2): 106–26.

DigitalGlobe. 2016. “WorldView Satellite Imagery.” http://worldview3.digitalglobe.com.

Duranton, Gilles, and Diego Puga. 2015. “Urban Land Use.” In Handbook of Regional and Urban Economics, vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 467–560. Amsterdam: North-Holland.

Duranton, Gilles, and Anthony J. Venables. 2020. "Place-Based Policies for Development." In Handbook of Regional Science, edited by Manfred Fisher and Peter Nijkamp. Berlin: Springer.

Field, Erica. 2005. "Property Rights and Investment in Urban Slums." J. European Econ. Assoc. 3 (2–3): 279–90.

Franklin, Simon. 2020. "Enabled to Work: The Impact of Government Housing on Slum Dwellers in South Africa." J. Urban Econ. 118:103265.

Freire, Maria E., Somik Lall, and Danny Leipziger. 2014. "Africa's Urbanization: Challenges and Opportunities." Working Paper no. 7, Growth Dialogue. http:// growthdialogue.org/africas urbanization-challenges-and-opportunities/.

Galiani, Sebastian, Paul J. Gertler, Raimundo Undurraga, Ryan Cooper, Sebastian Martinez, and Adam Ross. 2013. "Publisher's Note on Shelter from the Storm: Upgrading Housing Infrastructure in Latin American Slums." J. Urban Econ. 96:166–94.

Galiani, Sebastian, and Ernesto Schargrodsky. 2010. "Property Rights for the Poor: Effects of Land Titling." J. Public Econ. 94 (9): 700–729.

Gelman, A., and G. Imbens. 2017. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." J. Bus. and Econ. Statis. 37 (4): 447–56.

Giglio, Stefano, Matteo Maggiori, and Johannes Stroebel. 2014. "Very Long-Run Discount Rates." Q.J.E. 130 (1): 1–53.

Gollin, Douglas, Remi Jedwab, and Dietrich Vollrath. 2016. "Urbanization with and without Industrialization." J. Econ. Growth 21 (1): 35–70.

Gollin, Douglas, Martina Kirchberger, and David Lagakos. 2017. "In Search of a Spatial Equilibrium in the Developing World." Working Paper no. 23916, NBER, Cambridge, MA.

Harari, Mariaflavia, and Maisy Wong. 2017. "Long-Term Impacts of Slum Upgrading: Evidence from the Kampung Improvement Program in Indonesia." Working paper, Univ. Pennsylvania.

Henderson, J. Vernon, Tanner Regan, and Anthony J. Venables. 2017. "Building the City: Urban Transition and Institutional Frictions." Discussion Paper no. 11211, CEPR, London.

Henderson, J. Vernon, Anthony J. Venables, Tanner Regan, and Ilia Samsonov. 2016. "Building Functional Cities." Science 352 (6288): 946–47.

Hornbeck, Richard, and Daniel Keniston. 2017. "Creative Destruction: Barriers to Urban Growth and the Great Boston Fire of 1872." A.E.R. 107 (6): 1365– 98.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." Rev. Econ. Studies 79 (3): 933–59.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." Science 353 (6301): 790–94.

Kironde, J. M. Lusugga. 1991. "Sites-and-Services in Tanzania: The Case of Sinza, Kijitonyama and Mikocheni Areas in Dar-es-Salaam." Habitat Internat. 15 (1–2): 27–38.

———. 1992. "Creations in Dar-es-Salaam and Extensions in Nairobi: The Defiance of Inappropriate Planning Standards." Cities 9 (3): 220–31.

———. 1994. "The Evolution of the Land Use Structure of Dar es Salaam 1890– 1990: A Study in the Effects of Land Policy." PhD diss., Univ. Nairobi.

———. 2015. "Good Governance, Efficiency and the Provision of Planned Land for Orderly Development in African Cities: The Case of the 20,000 Planned Land Plots Project in Dar es Salaam, Tanzania." Current Urban Studies 3:348–67.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." Econometrica 75 (1): 83–119.

Komu, Felician. 2013. "Rental Housing in Tanzania: Power and Dialects of Misidentification." J. Building and Land Development, special issue: 29–41.

Laquian, Aprodicio A. 1983. "Sites, Services and Shelter: An Evaluation." Habitat Internat. 7 (5–6): 211–25.

Libecap, Gary D., and Dean Lueck. 2011. "The Demarcation of Land and the Role of Coordinating Property Institutions." J.P.E. 119 (3): 426–67.

Manara, Martina, and Tanner Regan. 2019. "Eliciting Demand for Title Deeds: Lab-in-the-Field Evidence from Urban Tanzania." Discussion Paper no. 19, LSE Geography and Environment, London.

Marx, Benjamin, Thomas Stoker, and Tavneet Suri. 2013. "The Economics of Slums in the Developing World." J. Econ. Perspectives 27 (4): 187–210.

———. 2019. "There Is No Free House: Ethnic Patronage in a Kenyan Slum." American Econ. J. Applied Econ. 11, no. 4 (2019): 36–70.

Mayo, Stephen K., and David J. Gross. 1987. "Sites and Services and Subsidies: The Economics of Low-Cost Housing in Developing Countries." World Bank Econ. Rev. 1 (2): 301–35.

McMillen, Daniel P., and Christian L. Redfearn. 2010. "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions." J. Regional Sci. 50 (3): 712–33.

MMM (Marshall Macklin Monaghan). 1979. "Main Report: Five Year Development Programme." In The Dar-es-Salaam Master Plan. MMM, Toronto.

NOAA (National Oceanic and Atmospheric Administration). 2012. "Version 4 DMSP-OLS Nighttime Lights Time Series." https://ngdc.noaa.gov/eog/dmsp /downloadV4composites.html.

Owens, Kathryn. 2012. "Enabling Sustainable Markets? The Redevelopment of Dar-es-Salaam." Paper prepared for the World Bank's Sixth Research and Knowledge Symposium, Barcelona, October 8–10.

Painter, K., and D. P. Farrington. 1997. "The Crime Reducing Effect of Improved Street Lighting: The Dudley Project." In Situational Crime Prevention: Successful Case Studies, 2nd ed., edited by R. V. Clarke, 209–26. Guilderland, NY: Harrow Heston.

Picarelli, Nathalie, Pascal Jaupart, and Ying Chen. 2017. "Cholera in Times of Floods: Weather Shocks Health Impacts in Dar es Salaam." Working Paper no. C-40404-TZA-1, Internat. Growth Centre, London.

Redding, Stephen J., and Daniel M. Sturm. 2016. "Estimating Neighborhood Effects: Evidence from War-Time Destruction in London." Working paper, Princeton Univ.

Romer, Paul. 2010. "Technologies, Rules, and Progress: The Case for Charter Cities." Center for Global Development.

———. 2012. "Urbanization as Opportunity." OECD.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Raymond Owens III. 2010 "Housing

Externalities." J.P.E. 118 (3): 485–535.

Tanzania Ministry of Lands. 2012. "Rates Land Value Mikoa (Regions) 10 2012." Ministry of Lands, Housing and Human Settlements Development, Dar es Salaam.

Tanzania National Bureau of Statistics. 2011. "Tanzania Population and Housing Census 2002." Tanzania National Bureau of Statistics, Dar es Salaam.

———. 2014. "Tanzania Population and Housing Census 2012." Tanzania National Bureau of Statistics, Dar es Salaam.

———. 2017. "2012 Census Shapefiles (machine readable data files)." Tanzania National Bureau of Statistics, Dar es Salaam.

Theodory, T. F., and M. M. Malipula. 2012. "Supplying Domestic Water Services to Informal Settlements in Manzese, Dar es Salaam: Challenges and Way Forward." J. Rural Planning Assoc. 14 (2): 60–72.

Tiba, A. D., G. Mwarabu, W. H. Sikamkono, et al. 2005. "The Implication of 20,000 Plots Project on the Emerging Form of Dar es Salaam City." MSc semester project, Dar es Salaam: Dept. Urban Planning and Management, UCLAS, UDSM.

Turner, Matthew A., Andrew Haughwout, and Wilbert Van Der Klaauw. 2014. "Land Use Regulation and Welfare." Econometrica 82 (4): 1341–403.

UN (United Nations). 2015. "World Population Prospects." Population Division, Dept. Econ. and Social Affairs, New York.

UN-Habitat. 2012. "State of the World's Cities 2012/2013: Prosperity of Cities."

World Bank. 1974a. "Appraisal of National Sites and Services Project." Report no. 337a-TA, Urban Projects Dept., World Bank, Washington, DC.

———. 1974b. "Development Credit Agreement between United Republic of Tanzania and International Development Association (Conformed Copy)." Credit no. 495 TA, World Bank, Washington, DC.

———. 1977a. "Development Credit Agreement between United Republic of Tanzania and International Development Association (Conformed Copy)." Credit no. 732 TA, World Bank, Washington, DC.

———. 1977b. "Tanzania: The Second National Sites and Services Project." Report no.

1518a-TA, Urban Projects Dept., World Bank, Washington, DC.

———. 1984. "Completion Report: Tanzania—First National Sites and Services Project." Report no. 4941, Eastern Africa Regional Office, World Bank, Washington, DC.

———. 1987. "Tanzania: The Second National Sites and Services Project." Report no. 6828, Operations Evaluation Dept., World Bank, Washington, DC.

———. 2010. "Project Appraisal Document for a Tanzania Strategic Cities Project." Report no. 51881-TZ, Urban and Water Dept., World Bank, Washington, DC.

———. 2013. "Tanzania Strategic Cities Project Housing Survey." Produced by President's Office, Regional Administration and Local Government, World Bank, Washington, DC.

6.3 Chapter 4 References

Ashraf, Nava, Oriana Bandiera and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics*, 120: 1-17.

Alesina, Alberto, and George-Marios Angeletos. 2005. "Fairness and Redistribution: US vs. Europe." *American Economic Review*, 95: 913-35.

Ashraf, Nava and Oriana Bandiera. 2018. "Social Incentives in Organizations". *Annual Review of Economics*, 10:1, 439-463.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2011. "Field Experiments with Firms." Journal of Economic Perspectives, 25 (3): 63-82.

Bandiera, Oriana, Amanda Dahlstrand-Rudin and Greg Fischer. 2019. "Incentives and Culture: Evidence from a Multi-Country Field Experiment." AEA RCT Registry. September 17. https://doi.org/10.1257/rct.4685-1.1.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2016. "Management as a Technology?" Harvard Business School Working Paper 16-133.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2012. "The Organization of Firms Across Countries." *Quarterly Journal of Economics*, 127(4): 1663-1705.

Bloom, Nicholas and John Van Reenen. 2007. "Measuring and Explaining Management Practices Across Firms and Countries." *The Quarterly Journal of Economics*, 122(4): 1351–1408.

DellaVigna, Stefano, John A. List and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1–56.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics,* 114(3): 817–868.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Saddoff, Xiangdong Qin and Yang Xu. 2019. "Measuring Success in Education: The Role of Effort on the Test Itself." *American Economic Review: Insights*. 1(3): 291-308.

Gorodnichenko, Yuriy, and Gerard Roland. 2011. "Which Dimensions of Culture Matter for Long-Run Growth?" *American Economic Review*, 101 (3): 492-98.

Hedblom, Daniel, Brent R. Hickman and John A. List. 2019. "Toward an Understanding of Corporate Social Responsibility: Theory and Field Experimental Evidence." NBER Working Paper No. 26222.

Hjort, Jonas, Xuan Li and Heather Sarsons. 2020. "Across-Country Wage Compression in Multinationals." https://drive.google.com/file/d/1cJ78hUrWMVuQIaaqEMmpah-taXUXOmZk/view. Accessed on February 6, 2020.

Hofstede, Geert. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage, Beverly Hills, CA.

Hofstede, Geert. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations.* 2nd ed., Sage, Thousand Oaks, CA.

Hofstede, Geert. 2008. "Values Survey Module." Updated version available at: https://geerthofstede.com/research-and-vsm/vsm-2013/. Accessed June 6, 2018.

Hofstede, Geert. 2011. "Dimensionalizing Cultures: The Hofstede Model in Context." *Online Readings in Psychology and Culture*, 2(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, Geert. "National Culture." https://geerthofstede.com. Accessed on May 23rd, 2018.

Hofstede Insights. 2019. https://www.hofstede-insights.com/country-comparison/ghana,india,the-philippines/. Accessed on September 13, 2019.

Hofstede Insights. 2020. https://www.hofstede-insights.com/models/national-culture/. Accessed on January 27, 2020.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. "Self-Control at Work." *Journal of Political Economy*, 123:6, 1227-1277.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India". *Journal of Political Economy*, 119(1): 39-77.

The World Management Survey. 2018. http://worldmanagementsurvey.org. Accessed on May 23, 2018.

World Bank. 2019. https://data.worldbank.org/income-level/lower-middle-income. Accessed September 13, 2019.

World Bank. 2011. https://blogs.worldbank.org/opendata/changes-country-classifications. Accessed on September 13, 2019.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results". *The Quarterly Journal of Economics*, 134(2): 557–598.

Young, Alwyn. 2016. "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections." http://personal.lse.ac.uk/YoungA/Improved.pdf. Accessed on February 6, 2020.