Market Size and the Returns to Surgeon Volume: Evidence from Joint Replacements

Dante Domenella*

[Click here for latest version]

November 13, 2025

Abstract

Why are people in larger geographic markets more productive? This paper investigates one potential mechanism in the health care sector–the relationship between a surgeon's procedural volume and patient health outcomes–that may generate benefits to market size. I study this topic in the context of hip and knee replacements, two of the most frequently performed medical procedures in the U.S. Using differential distance as an instrument, I find that surgeon volume significantly affects patient health outcomes. I then attribute 22% of the benefits of market size to this mechanism. This result highlights an externality, as a patient's choice of surgeon affects other patients' health outcomes through surgeon quality. Incorporating this externality into a demand model, I evaluate the welfare consequences of a first-best policy and three feasible policies: a minimum volume standard, transportation subsidies, and a policy that moves surgeons to shortage areas. Failing to incorporate the externality substantially understates the welfare effects of the feasible policies. Among the feasible policies, the minimum volume standard generates the largest welfare increase, yet it only achieves 7% of the gain from the first-best policy.

*Department of Economics, Stanford University. Email: ddomenel@stanford.edu. This paper benefitted from the incredible guidance of Mark Duggan, Neale Mahoney, Adam Sacarny, Isaac Sorkin, Doug Staiger, and Heidi Williams. I also thank Lea Bottmer, Matt Brown, Ian Calaway, Liran Einav, Nick Grasley, Harsh Gupta, Caroline Hoxby, Helena Roy, and seminar participants at Stanford for their valuable input. Data for this project were accessed using the Stanford Center for Population Health Sciences Data Core. The PHS Data Core is supported by a National Institutes of Health (NIH) National Center for Advancing Translational Science Clinical and Translational Science Award (UL1TR003142) and from internal Stanford funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This research was supported by the Leonard W. Ely and Shirley R. Ely Graduate Student Fellowship through a grant to the Stanford Institute for Economic Policy Research. All remaining errors are mine.

1 Introduction

People in larger geographic markets are more productive. Workers are paid higher wages, innovators produce more patents, and entrepreneurs are more likely to start firms (Glaeser and Maré, 2001; Bettencourt, Lobo, and Strumsky, 2007; Moretti, 2021; Rubinton, 2025). Distinguishing the mechanisms that generate these spatial differences–such as whether they reflect the selection of higher skill individuals into larger markets or some causal effect of being in a large market–is challenging due to endogeneity and data constraints (Glaeser and Gottlieb, 2009).

Understanding these mechanisms is crucial for designing public policy. This knowledge is especially relevant for designing place-based policy in health care, as the federal government spends billions of dollars each year in policies targeted at rural areas (MedPAC, 2022). The optimal policy to improve patient outcomes in small markets depends on the mechanism generating the health benefits to market size. If patients in larger markets have better health outcomes because higher quality physicians choose to practice there, then subsidizing physicians to practice in smaller markets may be optimal. On the other hand, if larger markets improve physician productivity through some agglomerative externality, then policy that leverages these benefits to market size by incorporating this externality, such as subsidies for patient transportation, may instead be optimal.

In this paper, I investigate the role of one potential externality–the returns to surgeon volume–in explaining the benefits of market size. Intuitively, if larger markets support more specialization, physicians can perform more of the same procedure, which can improve physician quality and thus generate differences in outcomes between patients in small and large markets. I investigate this idea in the context of hip and knee replacements, two of the most frequently performed medical procedures in the U.S. (Pfuntner, Wier, and Stocks, 2013). Using the differential distance between patients and surgeons as an instrument, I find that surgeon volume significantly affects patient outcomes, and I attribute 22% of the benefits of market size to this returns to surgeon volume mechanism. This result underscores a potentially policy-relevant externality–a patient's choice of surgeon affects other patients' outcomes through the surgeon's quality. Incorporating the externality into a demand model, I evaluate the welfare consequences of a first-best policy and three frequently discussed, feasible policies: implementing a minimum volume standard, subsidizing patient transportation, and moving surgeons to government-designated shortage areas.

The health care context, and specifically hip and knee replacement, is an economically meaningful setting to study this question, and it provides rich micro-data to capture the key elements for the analysis. In 2013, the U.S. spent $2 trillion, or 17% of its GDP, on health care, of which about $24 billion was on payments to hospitals for hip and knee replacements (Molloy et al., 2017). Annually, physicians perform 1.3 million of these procedures, and this number has been increasing in recent years (Sloan, Premkumar, and Sheth, 2018). Additionally, while hip and knee replacements are themselves important to understand, many of the features of this setting, especially the role of surgeon experience, are similar to other surgical procedures (Halm, Lee, and Chassin, 2002; Levaillant et al., 2021). To study this context, I use insurance claims data from Medicare, a public health insurance program covering virtually all U.S. individuals sixty-five years and older. Medicare covers

more than 50% of all hip and knee replacements and spends about $12 billion annually on these procedures (Molloy et al., 2017). Using this dataset, I measure a surgeon's hip and knee volume as the surgeon's number of Medicare hip and knee replacements in the 365 days prior to a patient's surgery date and define the outcome as a surgery-related complication.

To begin the analysis, I show three descriptive facts that motivate the returns to surgeon volume as a mechanism underlying the benefits of market size. I first document that patients who receive care in larger markets have better health outcomes, a pattern that speaks to other settings in which market size is associated with outcomes or productivity (Glaeser and Maré, 2001; Syverson, 2004; Berry and Waldfogel, 2010). Second, I demonstrate that surgeons who choose to practice in larger markets perform more hip and knee replacements. Finally, I show potentially large returns to surgeon volume, as patients who choose surgeons with higher hip and knee volume have better health outcomes (Arrow, 1962; Luft, Bunker, and Enthoven, 1979). This last relationship underscores the externality of patient choice on other patients that may drive the benefits of market size and thus may influence policy design.

Next, to structure the empirical analysis and evaluate counterfactual policies, I introduce a multinomial logit model of patient demand for surgeons that captures the potential externality implied by the returns to surgeon volume relationship. In the model, patients trade off a surgeon's quality with the distance they must travel for that surgeon. Surgeon quality depends both on an exogenous component, due to innate ability, for example, and an endogenous component through hip and knee volume. Allowing patient utility to depend on a surgeon's hip and knee volume embeds the returns to surgeon volume externality–a patient's choice affects the outcomes of other patients. In conjunction with the main estimating equation, the model helps me isolate quasi-random variation in a surgeon's hip and knee volume to estimate the returns to surgeon volume. I also use the model to understand how patients re-sort to surgeons under alternative policy regimes and to compute welfare.

To estimate the returns to surgeon volume, I then use this demand model to construct a control function that relies on differential distance as an instrument. The challenge with estimating the returns to surgeon volume is that patients select surgeons non-randomly, and surgeons select locations non-randomly. Thus, I use the demand model to isolate quasi-random variation in hip and knee volume based on the differential distance between patients and surgeons. Intuitively, some surgeons may have higher hip and knee volume based both on where patients live and other surgeons practice relative to where they practice. Since the first-stage is non-linear, I implement a control function approach that generalizes the Heckman selection technique to multiple choices (Dubin and McFadden, 1984; Abdulkadiroğlu et al., 2020; Einav, Finkelstein, and Mahoney, 2022; Mourot, 2024). This empirical strategy assumes that differential distance predicts a patient's choice of surgeon (the first stage) and that differential distance only affects outcomes through the choice of surgeon (the exclusion restriction). I probe the validity of these assumptions and find that the evidence supports them.

I find that surgeon volume significantly impacts patient outcomes and that this relationship explains 22% of the benefits of market size. I first demonstrate that doubling a surgeon's hip and knee volume reduces the probability of a complication by 9%. This highlights the externality–a patient's choice of surgeon affects other

patients' health outcomes through the quality of the surgeon. Using the control function approach, I also show that the ordinary least squares (OLS) estimate overstates the true effect in part because patients choose higher exogenous quality surgeons. This result suggests that re-allocating patients to higher volume surgeons may improve quality, even without a returns to surgeon volume relationship.

Using the demand model and the estimate of the returns to surgeon volume, I incorporate this externality into a welfare function and estimate the welfare effects of a first-best policy and three frequently discussed, feasible policies. For this analysis, I define social welfare as consumer surplus less fiscal costs and allow for both the returns to surgeon volume externality and an externality that reflects the fact that patients do not face full financial responsibility for their care. For the feasible policies, I first consider a minimum volume standard for surgeons to perform hip and knee replacements, which various states have implemented for hospitals, especially for stroke and cardiac care. I then consider a policy that subsidizes patient transportation, which is becoming increasingly common in the private Medicare Advantage market (Shen et al., 2024). Finally, I consider a policy in which I move surgeons to government-designated shortage areas. This policy mimics numerous other interventions that financially incentivize physicians to move to shortage areas, such as loan forgiveness programs. Because of the returns to surgeon volume externality, the welfare effects for all three feasible policies have an ambiguous sign and magnitude. Namely, welfare increases among the patients choosing surgeons who gain volume under these policies but decreases it for those choosing surgeons who lose volume.

To provide a relevant comparison for these three feasible policies and guide future policy design, I first compute the welfare gain from a first-best policy and show that it generates large welfare gains. I calculate the welfare gain from this first-best policy by allowing a social planner to set surgeon-specific Pigouvian taxes or subsidies to correct the inefficiency. Under this first-best policy, the social planner introduces large subsidies for high-quality surgeons. However, the relationship between the subsidy amount and the surgeon's quality is smaller when incorporating the returns to surgeon volume externality. This difference reflects a fundamental tradeoff–funneling patients to high-quality surgeons is welfare-increasing for patients who choose those high-quality surgeons but welfare-decreasing for those who choose other surgeons. Thus, while the optimal policy substantially increases the concentration of care, it does so less with the externality than without. Finally, I show that the welfare gain from the optimal policy is large, as it exceeds the average amount Medicare pays to the surgeon for the replacement.

Then, I find that incorporating the returns to surgeon volume externality substantially increases the welfare effects of each of the three feasible policies. Implementing a minimum volume standard mechanically centralizes care, which increases the welfare effect by 26% through the externality. Likewise, the transportation subsidy consolidates care among higher volume surgeons, although the change in concentration is much smaller than the minimum volume standard. Thus, without the externality, the welfare effect of this subsidy would actually have been negative, but it is positive with the externality. Finally, while moving surgeons to shortage areas decentralizes care, welfare increases sevenfold with the externality as the movers' hip and knee volume increases. This increase occurs because the increase in volume from nearby patients outweighs the decrease in volume from the now further patients.

Next, I find that the minimum volume standard generates the largest welfare increase among the feasible policies, yet it only achieves 7% of the first-best policy. The minimum volume standard achieves such large welfare effects because it generates large fiscal savings both through the returns to surgeon volume externality and as patients choose higher exogenous quality surgeons. Hence, the average welfare effect of the minimum volume standard is more than an order of magnitude larger than the effects of the other two policies. This difference largely reflects the fact that the other two policies do not substantially change patients' choice of surgeon, which is crucial to generate large welfare effects with this externality. Despite these large welfare gains under the minimum volume standard, however, consumer surplus still declines as patients lose choice and must travel farther.

Finally, exploring the distributional consequences of these policies, I find that the returns to surgeon volume externality substantially increases the welfare effects for both rural and urban patients for the three feasible policies. For both the minimum volume standard and the transportation subsidy, rural patients become more likely to choose surgeons in larger markets, so the externality improves welfare for both rural and urban patients. Meanwhile, when moving surgeons, the externality increases welfare because the moving surgeons' volume increases both in rural and urban areas, as the increase in volume from nearby patients outweighs the decrease in volume from the now further patients. Exploring differences in the levels of the welfare effects, I then show that the minimum volume standard generates similar welfare effects for rural and urban patients. However, rural patients experience larger declines in consumer surplus, as they must travel farther after the policy but urban patients do not. On the other hand, subsidizing transportation increases welfare for rural patients but decreases it for urban patients because rural patients travel to much higher exogenous quality providers but urban patients do not. Moving surgeons also increases welfare for rural patients but decreases it for rural patients because average surgeon quality improves in rural areas but declines in urban areas.

This paper provides key empirical insight into the mechanisms that drive the benefits of agglomeration. Understanding the drivers of these benefits is critical for formulating models of economic geography, such as quantitative urban models, and for evaluating and designing effective public policy, such as local taxation. (Greenstone, Hornbeck, and Moretti, 2010; Redding, 2023, 2025). Since Marshall (1890), urban economists have theorized that specialization and learning generate benefits to agglomeration (Duranton and Puga, 2004). Nevertheless, the endogeneity of location decisions and the lack of microdata on worker tasks have made it difficult to empirically demonstrate that any source, let alone specialization or learning, actually generates benefits to agglomeration (Rosenthal and Strange, 2004; Glaeser and Gottlieb, 2009). Several more recent papers have strongly suggested that specialization may play a crucial empirical role in facilitating benefits to market size, such as in the U.S. labor market or the Italian restaurant industry (Leonardi and Moretti, 2023; Moretti and Yi, 2024). This paper contributes by empirically showing that the returns to surgeon volume–a potential consequence of specialization–are an important mechanism underlying the benefits of agglomeration in health care.

This paper also shows that market size is an important driver of some of the substantial spatial variation in health outcomes. A wide literature in health economics documents large geographic variation in health outcomes (Chetty et al., 2016; Finkelstein, Gentzkow, and Williams, 2019). While the literature documents that

supply-side factors are important in explaining much of this variation, the sources that generate this supply-side variation still remain mostly unknown, despite the importance of understanding these sources to evaluate the welfare consequences of health interventions to decentralize or centralize the provision of care (Chandra and Staiger, 2007; Deryugina and Molitor, 2021). This paper provides evidence that market size is an important driver of these geographic differences. Perhaps most related to this project, Dingel et al. (2023) studies trade across markets for medical services and its relationship to market size. In exploring the mechanisms underlying the benefits of market size, however, this paper makes an important contribution to understanding what generates these benefits and thus how to design policy.

Finally, this paper contributes causal evidence to the "volume-outcomes" literature and considers what this relationship implies for policy. This literature has documented strong correlations between provider volumes and patient health outcomes, including in the context of orthopedic surgeries (Luft, Bunker, and Enthoven, 1979; Birkmeyer et al., 2002; Halm, Lee, and Chassin, 2002; Phibbs et al., 2007). While a much smaller number of these papers attempt to causally identify this relationship, they typically rely on quite strong identification assumptions and focus on hospitals, not physicians (Gaynor, Seider, and Vogt, 2005; Hentschker and Mennicken, 2018; Avdic, Lundborg, and Vikström, 2019; Kugler et al., 2022). Additionally, these papers do not consider this "volume-outcomes" relationship as an externality and thus do not incorporate it into a quantitative framework to analyze the policy implications.

2 Context, data, and measurement

I investigate the role of the returns to surgeon volume in explaining the benefits of market size in the context of hip and knee replacements. Hip and knee replacements are one of the most commonly performed medical procedures in the U.S., and hospitals receive $24 billion each year for these procedures. Since many patients who undergo a hip or knee replacement are elderly, this context lends itself well to the dataset I use, insurance claims from the Medicare program.

2.1 Context: Total hip and knee replacement

I study this question in the context of total hip and knee replacements, a surgical procedure in which an orthopedic surgeon removes a patient's hip or knee joint and replaces it with an artificial implant. Physicians perform over 1.3 million of these procedures annually, and hospital payments for them comprise 0.8% of health care spending. To provide intuition for the empirical analysis, I discuss the clinical background for total hip and knee replacements.

Hip and knee replacements are frequent medical procedures and comprise a large share of U.S. health care costs. About 1.3 million of these replacements are performed each year. In 2010, 2.3% of the U.S. population was living with a hip or knee replacement, and this number jumps to 13.7% among those 70 years or older (Kremers et al., 2015). The number of hip and knee replacements is also growing over time. From 2000 to 2014, the annual number of hip and knee replacements per 100,000 people increased by over 100% (Sloan, Premkumar, and Sheth, 2018). Hip and knee replacements are also costly to the U.S. health care system. Hospitals alone

receive about $24 billion each year for care, comprising 0.8% of health care expenditures (Molloy et al., 2017). In the Medicare context, roughly 7% of discharges and 8.5% of inpatient spending, or $12 billion annually, are for hip and knee replacements. Additionally, while complication rates from hip and knee replacement are low, the costs of complications are quite large. For instance, 8% of complications result in death, and Yi et al. (2015) estimates that Medicare pays double for a hip or knee replacement with an infection than what it pays without one.

Total hip and knee replacement, or more formally total hip and knee arthroplasty, is a surgical procedure in which an orthopedic surgeon removes a patient's hip or knee joint and replaces it with an artificial implant. Patients who have severe joint arthritis and who do not respond well to more conservative treatments often choose to undergo replacements. Thus, unlike other common conditions, such as a heart attack, they are almost always elective. In fact, in my sample, 94% of total hip and knee replacements are elective. The procedures are typically successful at eliminating pain and improving patient mobility so patients can enjoy a higher quality of life (Räsänen et al., 2007). Orthopedic surgeons generally perform these replacements, and hip and knee replacements are two of the most common surgeries they perform. Orthopedic surgeons are specialized surgeons, who have long training periods and earn higher income compared to physicians in other specialties (Gottlieb et al., 2023). Many orthopedic surgeons receive fellowship training or subspecialize in both hip and knee replacements, which is why I consider these two procedures together.

The typical patient chooses an orthopedic surgeon through referrals and word-of-mouth reviews. Primary care physicians or rheumatologists often refer their severely arthritic patients to orthopedic surgeons for more intensive treatments. In some cases, patients receive referrals from one orthopedic surgeon to another. At the same time, the typical patient learns about orthopedic surgeons through word-of-mouth reviews, such as from family members, friends, or even people at the same church. Importantly, the typical patient chooses an orthopedic surgeon, as opposed to a hospital or a physician practice. Because hip and knee replacement is an elective procedure and there are virtually no network restrictions in this context, patients can in theory travel quite far for a surgeon. However, the median patient only travels ten miles for a replacement.

Once a patient chooses a surgeon, the surgeon's skill and experience play an important role in determining patients' outcomes. First, the surgeon orders physical exams, X-rays, and lab tests for their patient, which the surgeon uses to prepare for any potential risk factors. Then, while operating, the surgeon's most salient concern is an infection, which is more likely to occur the longer the wound is open. As surgeons gain skill and experience, they can operate more quickly, reducing the probability of an infection. For example, Nairn et al. (2021) find that the operating time for a new hip replacement technique drops from 157 minutes on the first replacement to 93 minutes for the thirtieth replacement. The typical patient remains in the hospital for one to three days and then begins months of physical therapy.[1] Throughout this process, other providers, including anesthesiologists, nurses, and physical therapists, also assist with care. Thus, their skill and the team-specific human capital between them and the surgeon also affect the patient's outcomes (Chen, 2021). While hip and

[1]Many of these procedures now take place in the outpatient setting. For all but one year of my sample period, these procedures were on the "inpatient-only list," implying that Medicare would not reimburse providers if these replacements were billed in an outpatient setting. In 2018, the last year of my sample, knee replacements were moved off the inpatient-only list, and hip replacements followed in 2020.

knee replacements are themselves important to understand, many of the features of this context, especially the role of surgeon experience, are similar to other surgical procedures, such as cardiovascular or urological surgeries (Halm, Lee, and Chassin, 2002; Levaillant et al., 2021).

2.2 Data and sample construction

Another reason I study this question in the context of hip and knee replacement is because it lends itself well to my data, the Medicare claims data. This data includes a large number of hip and knee replacements, orthopedic surgeons, and geographic areas.

For my primary dataset, I use insurance claims from the Medicare FFS program. The dataset consists of claims, or billing statements, that providers submit to Medicare. The dataset contains claims for FFS patients, the traditional Medicare coverage that is publicly provided.[2] Specifically, I use the carrier files, which are professional claims typically submitted by physicians; outpatient claims; and institutional hospital and nursing home stays.[3] The data includes patient, physician, and hospital identifiers and procedural and diagnostic medical codes. I link these claims to other data from the Center for Medicare and Medicaid Services (CMS) that gives basic demographic information, such as age and sex, for both patients and surgeons.[4] The dataset spans the years starting in 2006 and ending in 2018. I use a 20% sample of the dataset based on patients, so I observe all claims for a given patient.

I use the Medicare FFS data because it includes a large share of hip and knee replacements, physicians, and geographies. The primary benefit of this data is its vast coverage compared to other insurance claims data–it includes about 40% of all hip and knee replacements in the U.S. Additionally, it includes almost 23,000 orthopedic surgeons, which is essentially all orthopedic surgeons in the U.S. Finally, unlike other claims data, this dataset provides nationwide geographic coverage. This dataset has such impressive coverage because there are minimal network or geographic restrictions for participation in the Medicare FFS program (both for surgeons or patients). In my context, the dataset's key limitations are that it lacks detailed clinical information and does not include non-Medicare FFS claims. Thus, I imperfectly measure patient risk and surgeon volume.

To complete the dataset, I supplement the Medicare claims with other data. First, I augment the claims data with annual, publicly available Provider of Service (POS) files, which give the ZIP code for all Medicare-certified institutional providers, such as hospitals, and thus the place of surgery. I also incorporate publicly available data on physicians' medical school from the CMS National Downloadable File, which I link to U.S. News and World Report medical school rankings.[5] Meanwhile, for the counterfactual policy analysis, I employ county-level wage data from the Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW) and data on Health Professional Shortage Area (HPSA) designations from the Health Resources and Services

[2] While the private version of Medicare, Medicare Advantage, has been rapidly growing in the past fifteen years, FFS still covers 73% of Medicare patients in the last year of my data.

[3] For institutional hospital and nursing home stays, I specifically use the Medicare Provider Analysis and Review (MedPAR) data. The Center for Medicare and Medicaid Services aggregates hospital and skilled nursing facility claims into the MedPAR files.

[4] Specifically, I use the Master Beneficiary Summary File and the Medicare Data on Provider Practice Practice and Specialty (MD-PPAS).

[5] Specifically, I use the CMS National Downloadable File for doctors and clinicians from December 2018, the last year of my primary dataset. I then manually collect data on U.S. News and World Report rankings of medical school for research in 2018.

Administration (HRSA).[6] Additionally, I use various geographic crosswalks to match patient and hospital ZIP codes to latitudes and longitudes, counties, and health care markets (Census, 2019; HUD, 2019; Dartmouth Atlas, 2025). Finally, I obtain inpatient, outpatient, and emergency department data files for all patients in the state of Florida in 2018. While I cannot link this data to my primary sample, I use it to better understand the relationship between a surgeons' Medicare FFS hip and knee volume and their total volume.

To adapt the Medicare FFS data to my setting, I define my sample and make two primary sample restrictions. My initial sample includes all hip or knee replacements during the sample period. I use the procedural codes from the claims together with the billing and coding guidelines from CMS to define total hip and knee replacement.[7] To classify the procedure as a hip or knee replacement, I also require the performing surgeon's self-reported specialty to be an orthopedic surgeon. In my first sample restriction, I require surgeons to perform ten or more procedures over the sample period (97.7% of hip and knee replacements) to avoid measurement error from improperly identifying the performing surgeon and to permit meaningful computations of quality metrics. Second, to simplify the empirical analysis, I restrict to patients who receive a hip or knee replacement for the first time (80.1% of hip and knee replacements).

2.3 Measurement and summary statistics

One advantage of studying this question in the health care context, and specifically for hip and knee replacements, is that there are well-defined measures of hip and knee volume and patient health outcomes. I measure a surgeon's hip and knee volume as the surgeon's number of Medicare FFS hip and knee replacements in the 365 days prior to a given procedure and rely on a previously established measure of patient health outcomes–surgery-related complications (CMS, 2022).

I measure volume as the surgeon's of Medicare FFS hip and knee replacements in the 365 days prior to a patient's surgery date and consider this measure as a proxy for a surgeon's recent experience. I exclude from the measure replacements where the surgeon assists but is not the primary surgeon. Even though my main sample excludes patients who receive more than one hip or knee replacement, I do include these patients in the volume measure, as surgeons may still learn from these procedures. I combine hip and knee volumes because hip and knee replacement together are a common orthopedic subspecialization and procedurally similar. I interpret annual volume as a proxy for a surgeon's "recent experience." As shown in Appendix Figure A1, correlating lagged measures of volume with patient health outcomes suggests that recent volume is more important than distant volume for patient outcomes, as the effect of volume on outcomes dies out after just one lag. The decision to measure hip and knee volume at the annual level also follows the "volume-outcomes" literature, in large part because the limited time frame of datasets prohibits the use of cumulative volume (Birkmeyer et al., 2002; Halm, Lee, and Chassin, 2002). Despite this limitation, annual Medicare FFS hip and knee volume proxies well for cumulative Medicare FFS hip and knee volume in my sample.[8]

[6] Using the BLS QCEW data, I measure hourly wages at the county-level using the average weekly county wage in 2012, the midpoint of the data, and I divide weekly county wages by 2,400, as I assume the average person works 40 hours, or 2,400 minutes, per week

[7] I exclude partial replacements and revisions from the sample, as they are procedurally distinct from and more complex than total hip or knee replacements.

[8] Among the surgeons for whom I can observe cumulative Medicare FFS hip and knee volume, the correlation between this measure

Using external data, I also show that a surgeon's Medicare FFS hip and knee volume is a strong proxy for total hip and knee volume. One concern with this volume measure is that surgeons specialize in certain insurers, such that increases in Medicare FFS hip and knee volume do not correspond to increases in total hip and knee volume but to a shift in the insurance status of patients on which the surgeon operates. Using the Florida data in 2018, I find that the cross-sectional correlation between a surgeon's Medicare FFS hip and knee volume and total hip and knee volume is 0.95. Additionally, a bivariate regression of log total hip and knee volume on log Medicare FFS hip and knee volume yields a coefficient of 1.06, implying that Medicare FFS hip and knee volume and total hip and knee volume co-move together at similar rates.

To measure patient health outcomes, I rely on a previously established measure of surgery-related complications. I directly employ this measure from the Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation, which developed it to help CMS publicly report hospital outcomes for hip and knee replacement. Specifically, this measure is a dichotomous indicator if a patient experiences one of eight surgery-related complications, such as an infection, within a specific time span following the replacement CMS (2022). While complications are infrequent, they are extremely serious adverse outcomes, as death occurs in 8% of complications. The three most common complications are periprosthetic joint infection, wound infection, or other wound complication (32.4%); pneumonia or other respiratory complications (19.4%); and pulmonary embolism (16.1%). All eight complication types and their frequency are in Appendix Table A1. Following CMS (2022), I adjust this complication measure for observable patient covariates, such as age, comorbid diseases, and indicators of patient frailty, which are clinically relevant and related to the outcome. I list all of these covariates in Appendix Table A2. Despite their severity, complications exhibit a strong negative correlation with self-reported quality of life, which is a key metric used to assess the surgery's success and a criterion on which patients select surgeons (Askari et al., 2024).

Finally, for geographic measurement, I employ two commonly used health care geographic units and formally define market size and distance. For geographic markets, I employ the frequently-used Hospital Referral Regions (HRRs) and Hospital Service Areas (HSAs), which are based on Medicare FFS patient travel for care (Dartmouth Atlas, 2025). HRRs are regional health care markets for tertiary medical care, and there are 306 in the U.S. Meanwhile, HSAs are local health care markets for hospital care, and there are 3,436 in the U.S. In my sample, 81% of patients receive a hip and knee replacement within their HRR, while 53% receive one within their HSA. Among markets with any hip and knee replacement in my sample, the median HRR contains thirteen hospitals in which a surgeon performs a hip and knee replacement at any point, while the median HSA contains only one such hospital. I measure market size using the number of Medicare FFS beneficiaries within an HSA. Finally, I measure distance as the distance between ZIP code centroids, such that the distance equals zero for a patient and surgeon located in the same ZIP code.

To provide a better picture of the patients and orthopedic surgeons in the sample, I discuss descriptive statistics for the patients and orthopedic surgeons in my sample. Examining these statistics for patients in Table 1 shows that the average patient is seventy-two years old. Additionally, 36% of the patients are male, 91%

and annual Medicare FFS hip and knee volume is 0.77.

are white, and 7% are dually eligible for Medicaid, a U.S. public health insurance program primarily for low-income individuals and thus a proxy for low income. Knee replacements comprise 66% of the replacements in the sample, and a complication occurs in 3.1% of replacements. The median patient travels ten miles for a replacement. Meanwhile, there are 11,243 surgeons in the sample, and the average surgeon is fifty-two years old and performs forty-one Medicare FFS hip and knee replacements within a 365 day span. As shown in Appendix Figure B1, surgeon hip and knee volume is right-skewed, such that the median surgeon performs twenty-seven hip and knee replacements annually. Using the Florida data, I calculate that the average physician's Medicare FFS share of total hip and knee replacements is 37% in 2018. Also, 98% of the orthopedic surgeons in the sample are male.

	Mean or median
Age	72.4
Fraction male	0.36
Fraction white	0.91
Fraction dual eligible	0.07
Fraction knee	0.66
Complication rate	0.031
Distance (median)	10.4
N	708,669

Table 1: Patient descriptive statistics

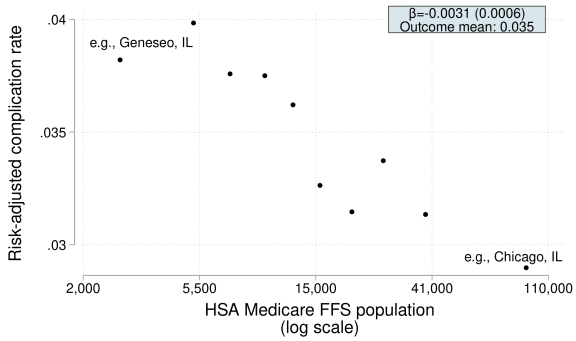
Notes: This table shows descriptive statistics for all patients in my sample. All statistics are averages except for distance, which is the median.

3 Descriptive evidence: The returns to surgeon volume as a mechanism

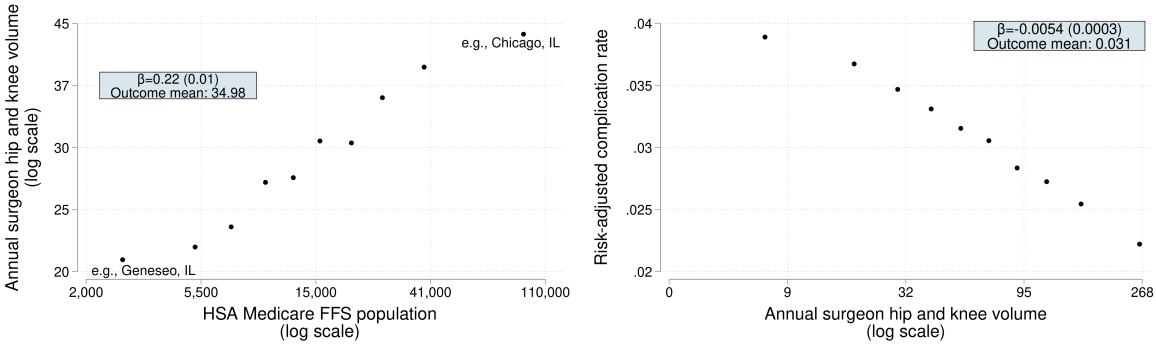
I show three descriptive facts that motivate the returns to surgeon volume as a potentially important mechanism underlying the benefits of market size. First, I demonstrate that patients who receive care in larger markets have better health outcomes. I then show that surgeons with higher hip and knee volume are in larger markets. Finally, I document that there appear to be large returns to surgeon volume, as patients who choose higher hip and knee volume surgeons have better health outcomes.

I first demonstrate that patients who receive care in larger markets have better health outcomes. Specifically, Figure 1a shows that doubling the size of a market corresponds to a 9% decline in the risk-adjusted complication rate. For example, in the first decile of market size in a market like Geneseo, Illinois, the probability of a complication is 3.8%. Nevertheless, it falls to 2.9% in the last decile, which includes the large market of Chicago, Illinois. While other work has found similar correlations for mortality or life expectancy, the magnitude and shape of this relationship is relatively unknown (Finkelstein, Gentzkow, and Williams, 2019; Deryugina and Molitor, 2021). This descriptive pattern also speaks to various other settings in which market size appears to drive outcomes or productivity, such as wages, restaurant quality, and plant-level productivity in the ready-mixed concrete industry (Glaeser and Maré, 2001; Syverson, 2004; Berry and Waldfogel, 2010).

While this paper does not decompose this relationship into the many factors that may cause it, such as



(a) Panel A: Relationship between risk-adjusted complication rate and market size



(b) Panel B: Relationship between annual surgeon hip and knee volume and market size

(c) Panel C: Relationship between risk-adjusted complication rate and annual surgeon hip and knee volume

Figure 1: Motivating the returns to surgeon volume as a mechanism underlying the benefits of market size

Notes: Figure 1a shows the relationship between the probability of a risk-adjusted complication and market size within an HSA. The probability of a risk-adjusted complication for an HSA is the HSA fixed effect from a linear probability model in which a binary indicator for a complication is regressed on all the patient risk covariates in Appendix Table A2 and HSA fixed effects for where the patient receives care. The figure excludes HSAs with five or fewer hip and knee replacements over the entire sample period and, to facilitate comparison with Figure 1b, HSAs that are never designated as a surgeon's primary practice location. For these exact same HSAs, Figure 1b shows the relationship between a surgeon's average hip and knee volume and the market size of their primary practice HSA. Surgeons are assigned to HSAs based on where they perform the plurality of hip and knee replacements. In Figures 1a and 1b, the market size is the HSA's average annual market size over the sample period. These figures then bins each HSA into ten equal-sized bins based on market size. Panel 1c shows the relationship between a binary indicator for a complication and a surgeon's annual hip and knee volume, controlling for the patient covariates in Appendix Table A2. Each hip and knee replacement is binned into 10 equal-sized bins based on the surgeon's hip and knee volume.

patient and surgeon selection or agglomerative externalities, the empirical evidence suggests that endogenous patient locations do not drive it. In Appendix Table B1, I first demonstrate that including patient risk covariates actually increases the magnitude of the slope in absolute value, indicating that, if anything, patients in smaller market are less risky. This result is consistent with Chandra and Staiger (2007), who argue that in areas that specialize more in a given procedure (in this context, the large markets), the patients who select into treatment will be less clinically appropriate for such a procedure and thus have worse health outcomes. Appendix Table B1 also shows that this magnitude is quantitatively similar when I add HRR fixed effects, restrict the sample to patients who receive care in their HSA of residence, and include only complications that occur prior to discharge from the hospital. Thus, differences in patient risk across HRRs, patient selection into travelling outside the market, and different behavior after hospital discharge do not appear to drive this relationship.

I then show that the average surgeon in a larger market performs more hip and knee replacements than the average surgeon in a smaller market. Namely, Figure 1b demonstrates that doubling the size of a market is associated with a 22% increase in the average surgeon's hip and knee volume. Using the same example as before, surgeons perform on average twenty-one hip and knee replacements annually in the first decile of market size, such as in Geneseo, Illinois, as compared to forty-three in the last decile, such as in Chicago, Illinois. While this relationship may capture both surgeon sorting and a causal effect of agglomeration, it suggests an important role for surgeon volume in explaining differences in health outcomes between small and large markets. This descriptive pattern relates to other work that has documented that physicians specialize more in larger markets due to a greater division of labor (Baumgardner, 1988; Dingel et al., 2023). This phenomenon even extends to other occupations and industries, such as lawyers and retail establishments, and is considered a potentially important microfoundation underlying the benefits of agglomeration (Duranton and Puga, 2004; Campbell and Hopenhayn, 2005; Garicano and Hubbard, 2007).

Finally, I show a suggestive returns to surgeon volume relationship, as patients who choose surgeons who perform more hip and knee replacements have better health outcomes. Specifically, Figure 1c shows that doubling a surgeon's hip and knee volume corresponds to an 17% reduction in the probability of a risk-adjusted complication. For example, when treated by surgeons in the first decile of hip and knee volume (fifteen or fewer replacements), patients have a complication in 3.9% of replacements, as compared to 2.3% when treated by surgeons in the top decile of hip and knee volume (200 replacements or more). This relationship echoes the results from the large "volume-outcomes" literature (Luft, Bunker, and Enthoven, 1979; Birkmeyer et al., 2002; Halm, Lee, and Chassin, 2002; Gaynor, Seider, and Vogt, 2005; Phibbs et al., 2007). It also touches on a broader literature that emphasizes the important role of learning and experience for firm and worker productivity (Arrow, 1962; Irwin and Klenow, 1994; Levitt, List, and Syverson, 2013; Chan, 2021; Caplin et al., 2022; Dinerstein, Megalokonomou, and Yannelis, 2022). From a policy standpoint, this relationship underscores an externality–patient choice of surgeon affects other patients' outcomes. Finally, this figure demonstrates that the returns to surgeon volume relationship is roughly linear, which motivates subsequent functional form choices and affirms the intuitive notion that performing one more replacement has a larger effect on patient outcomes when a surgeon's volume is low.

Taking the descriptive returns to surgeon volume as causal, these three figures together indicate that the returns to surgeon volume explain 42% of the benefits to market size. Namely, taking the spatial distribution of surgeons as given, the relationships from Figures 1b and 1c imply that doubling market size is associated with a 3.8% decline in the probability of a complication. Thus, using the estimate from Figure 1a, the returns to surgeon volume explain 42% of the benefits of market size, suggesting that policy design may benefit from incorporating the externality implied by the returns to surgeon volume. Such a conclusion, however, is premature given that the returns to surgeon volume estimated here is not necessarily causal. On one hand, this parameter may be biased upward, such that 42% would be an upper bound. In fact, Appendix Table B2 strongly suggests that it is, as adding patient covariates and hospital or surgeon fixed effects all decrease the magnitude of the returns to surgeon volume relationship. Nevertheless, 42% may also be a lower bound, if, for instance, the returns to surgeon volume estimated here are downward biased because surgeons with higher hip and knee volume operate on unobservably riskier patients. In the subsequent section, therefore, I describe the model and empirical strategy I use to obtain the causal effect of surgeon volume on patient outcomes.

4 Model and empirical strategy

To causally identify the returns to surgeon volume and assess the welfare consequences of various counterfactual policies, I introduce a model of patient demand for surgeons that allows for the externality implied by the returns to surgeon volume relationship. I then introduce a model of the returns to surgeon volume and show how the demand model allows me to identify the returns to surgeon volume using differential distance as an instrument for a surgeon's hip and knee volume.

4.1 Model of patient demand for surgeons

I introduce a multinomial logit model of patient demand for surgeons that allows for a given patient's choices to affect other patients' outcomes. In the model, patients trade off a surgeon's quality, which depends on other patients' choices through volume, with the distance they travel for the surgeon. The model initially acts as a "first-stage" to help me isolate quasi-random variation in hip and knee volume based on distance and later allows me to analyze the welfare effects of various counterfactual policies.

In this multinomial logit model, patients trade off surgeon quality with travel distance, and surgeon quality depends on hip and knee volume. I specify patient i's utility from surgeon j as follows:

$$u_{ij} = \delta_j(v_{ij}) - \tau \ln d_{ij} + \eta_{ij}, \tag{1}$$

where $\delta_j(v_{ij})$ is the average patient utility from surgeon j given surgeon j's hip and knee volume v_{ij} (plus one) in the 365 days prior to patient i's surgery date, d_{ij} is the distance (plus one) between patient i and surgeon j, and η_{ij} is an idiosyncratic demand shock drawn *i.i.d.* from a logit distribution.[9] Taking the log of distance

[9] Recall that distance is based on ZIP codes, so I add one to distance to include patients who live in the same ZIP code in which they receive care. I add one to volume because some surgeons have not performed any hip or knee replacements in the previous 365 days. I include these surgeons in the analysis because some of the policies I consider explicitly target these surgeons. In Appendix Tables D1 and D2, I show robustness of the relevant demand model estimates to not adding one to volume.

intuitively reflects the fact that patients bear less of a utility cost for one additional mile at larger distances, a frequent assumption in the health care literature that I show empirical support for in Appendix Figure C1 (Sivey, 2012; Cornell et al., 2019; Einav, Finkelstein, and Mahoney, 2022). The logit assumption on the error term aids the empirical strategy and allows for an easier evaluation of the policy counterfactuals. This utility function abstracts from the referral process, so patient i's choice can be interpreted as a joint decision between the patient and a primary care physician, a rheumatologist, or even another orthopedic surgeon. It also abstracts from the extensive margin decision to undergo hip and knee replacement, as the policies I consider do not directly target this margin.

To incorporate the returns to surgeon volume externality, I allow the average utility from a surgeon to depend on both exogenous quality and endogenous quality through volume. Specifically, I estimate demand using log hip and knee volume as a linear shifter of $\delta_j(v_{ij})$, such that:

$$\delta_j(v_{ij}) = \delta_{0j} + \delta_1 \ln v_{ij}, \tag{2}$$

where v_{ij} is again the hip and knee volume (plus one) of surgeon j in the 365 days prior to patient i's surgery date. The parameter δ_{0j} captures the demand for surgeon $j's$ exogenous quality, due to innate ability, for instance. Meanwhile, the parameter δ_1 reflects the returns to surgeon volume relationship, as it captures demand for endogenous quality due to surgeon j's hip and knee volume. I express $\delta_j(v_{ij})$ as a function of log volume to reflect diminishing utility gains from one additional procedure, which is empirically reflected in Figure 1c. Demand for exogenous surgeon quality, δ_{0j}, is identified based on how far patients travel for surgeon j, relative to other surgeons in the market. Demand for a surgeon's hip and knee volume, δ_1, is identified based on travel to surgeon j when the surgeon has different volume–if patients travel farther to the surgeon when the surgeon has higher hip and knee volume, then δ_1 is positive.

Due to latent choice constraints, such as information asymmetries, the demand parameters in this model may not capture true demand, so I interpret them as capturing both demand and these constraints. More precisely, δ_{0j} and v_{ij} may be correlated with other surgeon characteristics, such as available capacity or information asymmetries (Agarwal and Somaini, 2022). For example, δ_{0j} may capture demand under imperfect information, if patients do not have perfect information on surgeon quality. Meanwhile, δ_1 may partially reflect demand for surgical technologies, if hip and knee volume and surgeon investment in surgical technologies are correlated. I thus interpret these parameters as capturing these other characteristics, such that δ_1, for instance, reflects demand for both hip and knee volume and other characteristics correlated with surgeon volume.[10] For the empirical strategy in which I use these parameters to estimate the returns to surgeon volume, this interpretation does not present a serious issue because I primarily need to recover the parameters that capture the characteristics on which patients actually select surgeons, not true demand. Later, to address these latent choice constraints in the counterfactual policy analysis, I introduce a capacity constraint and assume that the policies I implement do not alter the other constraints.

[10] δ_1 is also endogenous because it is another measure of demand. While this endogeneity does not affect the subsequent empirical strategy, I more formally capture this interdependence when I discuss the counterfactual policy results.

Additionally, although measurement error from the inclusion of surgeon fixed effects may bias the demand parameters, this bias would only understate the subsequent empirical results and welfare analyses. Without sufficient within-surgeon variation in hip and knee volume, δ_{0j} would be biased upward and δ_1 biased downward because δ_{0j} would absorb some of the variation in hip and knee volume (Griliches and Hausman, 1986; Bound, Brown, and Mathiowetz, 2001). However, because the parameters of interest in this analysis are not these demand parameters but rather the returns to surgeon volume relationship, this bias also does not present serious issues. If anything, it would attribute more of the variation in demand for surgeons to their exogenous quality, rather than their volume. Thus, it would understate the returns to surgeon volume relationship I estimate in the subsequent section and the welfare effects of the counterfactual policies I evaluate.

4.2 Empirical strategy: Identifying the returns to surgeon volume

To causally identify the returns to surgeon volume, I introduce a model of the returns to surgeon volume and use the demand model to correct for selection. Intuitively, I construct a control function using the demand model to isolate quasi-random variation in hip and knee volume based on the differential distance between patients and the surgeons in their choice set.

To estimate the returns to surgeon volume, I relate patients' health outcomes to their surgeon's hip and knee volume with various controls. For patient i and surgeon j, I consider the following linear probability model:

$$\begin{aligned} \text{comp}_{ij} &= \beta_0 + \beta_1 \ln v_{ij} + X_i\alpha + \gamma_{t(i)} + \epsilon_{ij} \\ \epsilon_{ij} &= \xi_j + \psi_i + \lambda_{ij} + \omega_{ij} \end{aligned} \tag{3}$$

where comp_{ij} equals one if patient i has a complication, v_{ij} is surgeon j's hip and knee volume (plus one) in the 365 days prior to patient i's surgery date, X_i are the patient risk covariates listed in Appendix Table A2, and $\gamma_{t(i)}$ are fixed effects for the calendar year in which patient i receives the surgery. I also include four different unobserved error terms: surgeon-specific exogenous quality ξ_j, unobserved patient characteristics ψ_i, patient-surgeon match-specific characteristics λ_{ij}, and an idiosyncratic error ω_{ij}. For the functional form of this main specification, I use log volume based on Figure 1c, although I show robustness to other functional forms.[11] β_1 is the coefficient of interest, yet the three non-idiosyncratic error terms, ξ_j, ψ_i, and λ_{ij}, threaten causal identification.

Prior to discussing the empirical strategy, I highlight that the true estimate of the returns to surgeon volume, β_1, captures the bundled treatment effect from a patient choosing a higher volume surgeon. This treatment effect therefore includes both dynamic economies of scale, such as learning-by-doing, and static economies of scales, such as improved team-specific human capital. That is, the causal β_1 captures not just the "pure" effect of a surgeon's hip and knee volume on outcomes but other mechanisms that potentially accompany changes

[11] I also estimate Equation 3 using a constant relative risk aversion (CRRA) parameterization, such that I replace $\ln v_{ij}$ with $\frac{v_{ij}^{1-\upsilon}-1}{1-\upsilon}$. Recall the CRRA specification nests the log specification, as it approaches $\ln v_{ij}$ when $\upsilon = 1$. I estimate υ using non-linear least squares estimation and find that $\upsilon = 0.79$ with a 95% confidence interval of $[0.64, 0.93]$, suggesting that the log specification is a reasonable one in this setting. If anything, with the log specification, I understate the effect among higher volume surgeons.

in volume, such as higher team-specific human capital with nurses, additional human capital investments, or being overworked. In this setting, this effect is the policy-relevant parameter of interest because I consider policy that funnels patients to a specific surgeon, which may also coincide with changes in team-specific human capital, for instance. In the results section, however, I investigate some of these mechanisms underlying the returns to surgeon volume relationship.

To build intuition for the empirical strategy, I present examples of the types of selection that threaten causal identification for the three non-idiosyncratic error terms. First, if patients select surgeons with higher exogenous quality or higher exogenous quality surgeons practice in larger markets, β_1 would be biased upward due to selection on ξ_j. While surgeon fixed effects would address this selection, insufficient variation in hip and knee volume introduces substantial attenuation bias, which is only exacerbated with measurement error in volume (Griliches and Hausman, 1986; Bound, Brown, and Mathiowetz, 2001).[12] Additionally, since annual hip and knee volume is a proxy for a surgeon's experience, then a within-surgeon estimate is challenging to interpret. Second, patients may select surgeons based on unobservable patient characteristics, as captured in ψ_i, such as their underlying risk. For example, if unobservably less risky patients select surgeons with higher hip and knee volume, β_1 would be biased upward. Finally, patients may select surgeons based on an idiosyncratic match (Roy selection), as captured in λ_{ij}. For example, if higher hip and knee volume surgeons specialize in treating patients with heart conditions and patients with heart conditions choose those surgeons, then β_1 would be biased upward.

To address this non-random selection, I construct a control function that relies on the differential distance between patients and surgeons as an instrument. Intuitively, some surgeons may have higher hip and knee volume based both on where patients live and other surgeons practice relative to where they practice. To implement this approach, I use the demand model as a first-stage to construct a control function in which the vector of distances between a patient and each surgeon functions as the instrument. I employ a control function because the demand model is non-linear, implying that two-stage least squares (2SLS) implementation would yield inconsistent or biased estimates unless I make strong assumptions (Guo and Small, 2016; Peng, 2024). Following Dubin and McFadden (1984), this approach generalizes Heckman selection to to multinomial selection models with many alternatives (Heckman, 1976).

Formally, I construct a control function with three types of controls that explicitly address the three non-idiosyncratic error terms that may bias the returns to surgeon volume. Hence, I impose the following structure on the conditional expectation of the estimating equation error term, ϵ_{ij}, from Equation 3:

$$\mathbb{E}\left[\epsilon_{ij}|v_{ij}, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, \eta_{i1}, \cdots, \eta_{iJ}, D_i = j\right] = \kappa\widehat{\delta}_{0j} + \sum_k \phi_k \tilde{\eta}_{ik} + \varphi\tilde{\eta}_{ij}, \tag{4}$$

[12] In my sample, within-surgeon variation in hip and knee volume explains only 22% of the total variation in volume. Measurement error in hip and knee volume due to the fact that I only observe a 20% sample of Medicare FFS hip and knee volume only exacerbates this attenuation bias. This issue is reflected in the "volume-outcomes" literature, as many papers in the literature do not include provider fixed effects (Kim, Wolff, and Ho, 2017). Recall that in the demand model, I can include surgeon fixed effects because I do not seek to recover the true demand parameters. Here, though, I seek to recover the true returns to surgeon volume parameter, β_1, so I omit the surgeon fixed effects.

where $\widehat{\delta}_{0j}$ is the estimated demand for exogenous surgeon quality from the demand model in Equation 1, $\tilde{\eta}_{ij} = \eta_{ij} - \mu_\eta$ are the mean-zero logit shocks from the demand model, J is the total number of surgeons, and D_i denotes patient i's chosen surgeon. This functional form assumption closely follows the recent literature building on the Dubin and McFadden (1984) approach and can be thought of as a first-order approximation of a specification that allows for a richer relationship between choice and outcomes. (Abdulkadiroğlu et al., 2020; Einav, Finkelstein, and Mahoney, 2022; Mourot, 2024).

Intuitively, these three types of controls allow for correlations between patients' health outcomes and both observed surgeon quality and the unobserved patient or patient-surgeon characteristics on which patients may select surgeons. The first control, $\widehat{\delta}_{0j}$, addresses selection on exogenous surgeon quality, ξ_j. Namely, it controls for the relative distance that patients travel for the surgeon, independent of volume, which provides information on the surgeon's exogenous quality.[13] Meanwhile, the last two types of controls, the $\tilde{\eta}_{ik}$'s and $\tilde{\eta}_{ij}$, are the residuals from the first-stage. The $\tilde{\eta}_{ik}$'s address selection on unobserved patient factors, ψ_i. Intuitively, a patient's unobserved preference for every surgeon, as revealed through relative travel distance, provides information on this unobservable risk. Likewise, $\tilde{\eta}_{ij}$ address selection on match characteristics, λ_{ij}, as a patient's unobserved preference for the chosen surgeon, also as revealed through relative travel distance, provides information on this match.

Because the error terms are unobservable, I use the distributional assumptions to obtain closed-form selection correction controls and show that the vector of distances between a patient and each surgeon serves as the instrument. Namely, since I assume the error terms are drawn *i.i.d.* from a logit distribution, I can compute their expected values conditional on the choice, yielding the following conditional expectation function:

$$\mathbb{E}\left[\epsilon_{ij} | v_{ij}, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, d_i, D_i = j\right] = \kappa\widehat{\delta}_{0j} + \sum_k \phi_k \theta_{ik}\left(j, d_i\right) + \varphi\theta_{ij}\left(j, d_i\right), \tag{5}$$

where $d_i = \left(d_{i1}, \cdots, d_{iJ}\right)$ is the vector of distances between patient i and each surgeon that functions as the instrument.[14] Meanwhile, $\theta_{ik}(j, d_i)$ are functions of the logit choice probabilities from the demand model:

$$\theta_{ik}\left(j, d_i\right) = \begin{cases} -\ln \widehat{p}_{ik}\left(d_i\right), & k = j \\ \frac{\widehat{p}_{ik}(d_i)}{1-\widehat{p}_{ik}(d_i)} \ln \widehat{p}_{ik}\left(d_i\right), & \text{otherwise} \end{cases}, \tag{6}$$

where $\widehat{p}_{ik}$ is the predicted probability that patient i chooses surgeon j. I show these derivations in Appendix C.2. Note that the conditional expectation function in Equation 5 now conditions on the (excluded) instrument, the distance between a patient and every surgeon.

Finally, to recover the causal returns to surgeon volume, I incorporate these selection correction controls

[13] Following the procedure described in Appendix C.1, I use Empirical Bayes to shrink these parameters to the market-level mean to address measurement error. Additionally, insufficient variation in within surgeon volume coupled with measurement error in volume may bias downward δ_1 and thus attribute some of the variation in volume to exogenous quality $\widehat{\delta}_{0j}$. In this case, β_1 would be biased downward. Moreover, if there are latent choice constraints, such as capacity constraints, $\widehat{\delta}_{0j}$ does not reflect true demand, but for this empirical strategy, this interpretation does not present a serious issue, as I control for the characteristics on which patients actually select surgeons, not what they would have chosen absent the constraints.

[14] Note that I take the expectation of the true η's because the expectation of the predicted η's are by definition mean-zero. Intuitively, the mean of the predicted η's just captures noise, not selection on unobservable characteristics.

into the main specification. Combining Equations 3 and 5 yields the following equation for the conditional expectation of a complication:

$$\mathbb{E}\left[\text{comp}_{ij} | v_{ij}, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, d_i, D_i = j\right] = \beta_0 + \beta_1 \ln v_{ij} + X_i\alpha + \gamma_{t(i)} + \kappa\widehat{\delta}_{0j} + \sum_k \phi_k \theta_{ik}(j, d_i) + \varphi\theta_{ij}(j, d_i). \quad (7)$$

With this equation, I can recover an unbiased estimate of β_1, provided differential distance is a valid instrument. The coefficients κ, ϕ_k, and φ all provide insight on the types of selection that may bias the OLS estimate.

To provide more intuition for the approach, I discuss comparisons between this control function and a 2SLS implementation. Similar to a 2SLS implementation, the relevance assumption and exclusion restriction must hold. As opposed to the 2SLS setting, however, I microfound distance as an instrument using the demand model, rather than instrumenting for hip and knee volume with a more arbitrary function of distance, which implicitly imposes strong functional form assumptions (Hentschker and Mennicken, 2018). This approach thus allows me to use the demand model to address selection on exogenous surgeon quality, which the volume-outcomes literature generally does not address (Kim, Wolff, and Ho, 2017). Additionally, I obtain more precise estimates with the control function approach (Guo and Small, 2016).[15] Finally, this approach allows me to understand the types of selection that bias the OLS estimate. These advantages, though, come at the cost of a greater reliance on the functional form assumptions, for which I show support when discussing identification in Section 4.4. Unlike the 2SLS implementation, this approach also does not correct for measurement error in volume, so I employ an alternative measurement error correction strategy, which I describe in Appendix C.4.

4.3 Estimation

To ease the computational burden of estimation, I estimate the demand model by geographic market. I then discuss how I estimate the primary specification, which is challenging due to the large number of parameters.

To ease the computational burden of estimating the demand model, I restrict patients' choice sets to surgeons within their geographic market. Namely, I define a patient's choice set as all surgeons who perform a hip or knee replacement within the patient's HRR of residence starting at 365 days before a patient's surgery date and up to 365 days after. I assign surgeons to the ZIP code based on where and when they perform most of their hip and knee replacements within this time frame.[16] Since ZIP codes fall entirely within HRRs, a surgeon's HRR is simply the HRR of their assigned ZIP code. I define the outside option as choosing a surgeon outside of a patient's HRR, which occurs in 19% of all hip and knee replacements. For the outside option, I normalize utility to be zero. The median patient has forty-four surgeons in her choice set, and the median surgeon in her choice set is twenty-nine miles away.

Since the demand model is now estimated within a market, I modify the estimating equation slightly by adding fixed effects for a surgeon's practice location. Specifically, the δ_{0j}'s are now only relative to other

[15] Guo and Small (2016) argue that the control function estimate can be obtained using a 2SLS implementation with an augmented set of instrumental variables, but the efficiency of this estimate is less than that of the control function approach.

[16] For 7.9% of hip and knee replacements, surgeons do not perform a replacement in their assigned ZIP code. Thus, for patients who choose surgeons at these non-assigned ZIP codes, I assign the surgeon only to the ZIP code where the surgeon performs the replacement.

surgeons within the same geographic market because of the market-level demand estimation. Thus, denoting a surgeon's HRR as m, the main estimating equation becomes:

$$\text{comp}_{ijm} = \beta_0 + \beta_1 \ln v_{ij} + X_i\alpha + \gamma_{t(i)} + \kappa\widehat{\delta}_{0j} + \sum_{k \in S_i} \phi_k \theta_{ik} + \varphi\theta_{ij} + \mu_m + \epsilon_{ijm}, \tag{8}$$

where S_i denotes patient i's choice set and μ_m are fixed effects for the primary practice HRR of the surgeon.[17] Hence, now I only compare surgeons who practice within the same HRR, rendering the δ_{0j}'s interpretable. Note also that the ϕ_k parameters are only estimated for the surgeons within a patient's HRR.

Directly estimating this regression is computationally burdensome, so I leverage the Frisch-Waugh-Lovell theorem to estimate the parameters and bootstrap to obtain standard errors. Because Equation 8 has a parameter for each surgeon in the dataset, of which there are more than 11,000, I cannot estimate it quickly via OLS. I therefore exploit the fact that the θ_{ik}'s are zero for surgeons outside a patient's HRR and use the Frisch-Waugh-Lovell theorem to obtain the estimates. I describe this estimation strategy in Appendix C.3. Finally, to correct the standard errors for the variation from the first-stage demand model, I modify a two-score bootstrap approach to this setting in a procedure described in Appendix C.5 (Kline and Santos, 2012).

4.4 Identification

To recover a causal estimate of the returns to surgeon volume, distance must affect patient choice of surgeon but not affect the probability of a complication except through its effect on patient choice. I test these assumptions and find that evidence supports them.

To aid the discussion of identification, I first provide intuition for the identifying variation using a stylized Chicago as an example. In this market, shown in Figure 2, there are two surgeons, one of whom practices at Northwestern and one of whom practices at the University of Chicago, and three patients, numbered one through three. I show the distance between each patient and each surgeon with the numbers above or below the dashed lines. Note first that the differential distance the suburban patient 2 faces to each surgeon is identical to the differential distance the urban patient 1 faces. Hence, without the suburban patient 3, I could not identify the returns to surgeon volume because there is no variation in the differential distance these patients face. Nevertheless, with patient 3, identifying variation comes from comparing the outcomes of patients 1 and 2 to patient 3 (assuming all three patients do not choose the same surgeon). The two primary identification threats are thus the endogeneity of surgeon and patient locations. Namely, the surgeon at Northwestern might differ in exogenous quality from a surgeon at the University of Chicago, conditional on the demand for the surgeon's exogenous quality. Likewise, patient 3 may differ from patients 1 and 2, conditional on the risk covariates.

To further aid the identification discussion, I also provide intuition for the identification of the three types of control function coefficients. Broadly, these coefficients in Equation 8 are identified by comparing patients who travel short versus far distances for a hip or knee replacement. Identification of κ is standard. If patients travel

[17] In some cases, surgeons have primary practice locations in multiple HRRs. If the surgeon has a primary practice location that is the same as the patient's HRR of residence, I assign the surgeon to the $\widehat{\delta}_{0j}$ in that HRR. If the surgeon never has a primary practice location in the same HRR as the patient's HRR of residence, I then assign the surgeon to their practice location in which they perform the most hip and knee replacements, breaking ties randomly.

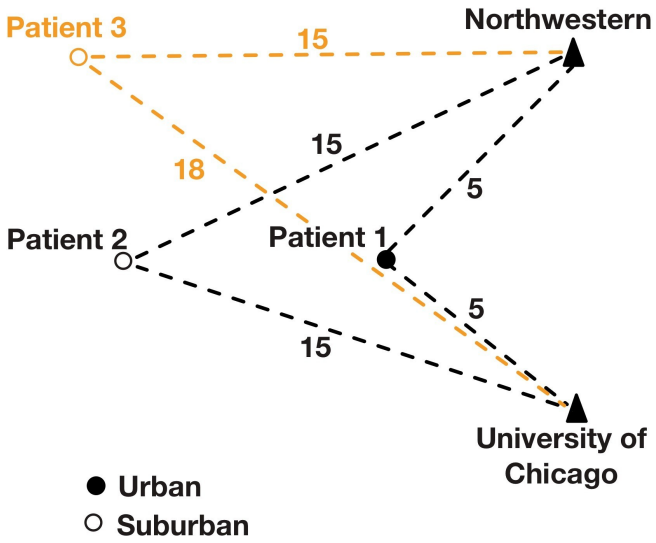


Figure 2: Conveying intuition for the identifying variation in a stylized Chicago market

Notes: This figure shows a stylized Chicago market with two surgeons and three patients to convey the intuition of the identifying variation. The numbers above and below the dashed lines show the distance between each patient and surgeon.

farther for j relative to other surgeons in the market and j's patients have a lower probability of a complication, $\kappa < 0$. Identification of ϕ_k and φ is slightly less conventional. If, for instance, surgeon k treats unobservably less risky patients, then $\phi_k < 0$, since the unobserved reasons for choosing surgeon k (as captured in η_i and revealed through distance) are negatively correlated with the probability of a complication. For the same reason, if patients select their chosen surgeon j for a match-specific reason, as captured in η_{ij}, then $\varphi < 0$. Critically, if the relative distance patients travel to a surgeon is correlated with the choice determinants, such as surgeon quality or a patient's unobserved risk, then it carries information about these factors. Hence, controlling for δ_{0j} and the η_i's (implicitly through the θ_{ij}'s) no longer fully capture these factors that matter for outcomes.

Causal identification of the returns to surgeon volume first requires a relevance assumption–distance must affect patient choice of surgeon. To assess this assumption, I show the predicted choice probability of a surgeon versus the distance to that surgeon for the Boston HRR in Appendix Figure C1. As evident in the figure, distance is a strong predictor of patient choice. The average choice probability falls by 71% when a patient is 0-5 miles away from a surgeon versus 5-10 miles away. This relationship generally holds across all markets as well, as the average coefficient on distance implies that doubling the distance to a surgeon decreases the odds of choosing that surgeon by 66%, as shown in Appendix Figure C2. In part for this reason, distance is a commonly used instrument in the health care literature (Grabowski et al., 2013; Einav, Finkelstein, and Williams, 2016; Einav, Finkelstein, and Mahoney, 2022; Card, Fenizia, and Silver, 2023; Karadakic et al., 2025). This figure also supports the primary control function functional form assumption, as the relationship between choice and distance is well-approximated with a logarithmic relationship.[18]

Causal identification also requires the exclusion restriction holds–distance can only affect health outcomes

[18]In Appendix Tables C1 and C2, I show more evidence that the estimates are robust to this parameterization. That is, for one HRR, I demonstrate that the estimates from the demand model used to construct the control function essentially do not change under alternative distance specifications.

through the choice of surgeon. Formally, this assumption implies that:

$$\begin{aligned} &\mathbb{E}\left[\epsilon_{ijm}|d_i, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, \mu_m, D_i = j\right] = 0 \\ \iff &\mathbb{E}\left[\xi_j + \psi_i + \lambda_{ij}|d_i, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, \mu_m, D_i = j\right] = 0, \end{aligned} \tag{9}$$

Here, I have re-expressed ϵ_{ijm} using the error term decomposition in Equation 3, as well as the fact that $\mathbb{E}\left[\omega_{ijm}|d_i, X_i, \gamma_{t(i)}, \delta_{0j}, \mu_m, D_i = j\right] = 0$. This formulation highlights the underlying exclusion restriction–knowing the relative distance patients travel produces no new information about the average value of the unobserved factors affecting outcomes, once the value of the controls are known. In other words, conditional on the controls, unobserved surgeon quality, ξ_j; unobserved patient characteristics, ψ_i; and unobserved patient-surgeon match characteristics, λ_{ij}, cannot be correlated with the relative distance patients travel. Hence, the major threats to identification occur if surgeons live closer to patients based on some component of ξ_j that isn't captured in $\widehat{\delta}_{0j}$, patients live closer to surgeons based on some component of ψ_i, or patients and surgeons co-locate based on their match, λ_{ij}.

Concretely, the primary threats to identification are the endogeneity of surgeon and patient location decisions. First, if higher exogenous quality surgeons sort into larger markets on characteristics other than patient demand, $\widehat{\delta}_{0j}$, the returns to surgeon volume would be biased upward. While much work points to labor demand as a key driver of skill-based geographic sorting (Berry and Glaeser, 2005; Moretti, 2013; Diamond, 2016; Baum-Snow and Hartley, 2020), other work highlights different causal forces, such as amenities or the rising value of time (Couture and Handbury, 2020; Su, 2022). Recall, however, that only skill-based sorting within an HRR is a threat, as the surgeon HRR fixed effects eliminate any bias induced by skill-based sorting across HRRs. Second, patient location decisions may be endogenous. For example, less risky patients may live closer to surgeons, which would bias upward the returns to surgeon volume. Finally, another threat is that distance has a mechanical effect on outcomes, as would be the case for an emergent procedure that requires immediate transport to the hospital. Nevertheless, this third threat is not quite as serious, since 94% of the hip and knee replacements in the sample are elective.

To address the threat from endogenous surgeon location decisions, I conduct two primary robustness exercises, both of which yield results that are consistent with the exclusion restriction. In the first one, I introduce several controls to the main estimating equation and show that these controls do not affect the main estimate. Specifically, I add a control for the rank of a surgeon's medical school several to address any selection on exogenous quality that is not captured with labor demand. I also add controls for more granular measures of market size, such that I compare surgeons who choose to practice in the same-sized markets, perhaps for unobservable reasons.[19] I further add controls for the log distance between surgeons' medical school and their practice locations and the amenity value of the surgeons' practice locations as estimated in Albouy (2016). Other work has demonstrated that both training locations and amenities are important drivers of geographical sorting for physicians and perhaps even differentially for higher exogenous quality physicians (Lee, 2010; Koehler et al.,

[19]Specifically, I include controls for both the market size of the HSA and the ZIP code where the hip or knee replacement is performed.

2016). In a second robustness exercise, I investigate whether high-quality surgeons move to larger markets, in the spirit of Chandra et al. (2016). Among 521 surgeons who move to a different HSA within an HRR, I find that higher quality surgeons do not move to larger markets, as shown in Appendix Figure C3.

To address the threat from endogenous patient location decisions, I conduct a covariate balance test and introduce a specification with patient ZIP code fixed effects and find evidence consistent with the exclusion restriction. Conducting robustness exercises with patients is simpler than it is with surgeons, as I have significantly more covariates and many more patients, allowing for more granular levels of fixed effects. Therefore, I first conduct a balance test in this setting to examine how the first-stage coefficient changes with the addition of observable covariates (Altonji, Elder, and Taber, 2005; Oster, 2017). Since a simple balance test is infeasible in this setting, as the instrument is the distance between a patient and every surgeon in the choice set, I implement the balance test by examining how the estimated distance coefficient from Equation 1 changes with the inclusion of observable covariates (Einav, Finkelstein, and Mahoney, 2022; Mourot, 2024).[20] In Appendix Figure C4, I show that estimated distance coefficients for the 306 HRRs are essentially identical with and without the patient covariates. This tight correlation indicates that the effect of distance on a patient's choice does not vary with observable patient covariates, lending credence to the exclusion restriction. To further bolster this point, I introduce an alternative specification with fixed effects for a patient's ZIP code and show that the main estimate does not change.[21]

5 Results: The returns to surgeon volume and the benefits of market size

Instrumenting for a surgeon's volume with differential distance, I find that an increase in a surgeon's hip and knee volume causally affects patients' health outcomes. This result suggests an externality–a patient's choice of surgeon affects other patients' health outcomes through the surgeon's quality. Using this estimate, I attribute 22% of the benefits of market size to the returns to surgeon volume.

5.1 The returns to surgeon volume

The main results indicate that doubling a surgeon's hip and knee volume decreases the probability of a complication by 9%. Differences in exogenous surgeon quality and patient selection on unobservable risk bias the non-causal estimate upward. Both static and dynamic economics of scale drive this relationship.

Addressing the potential selection issues using differential distance as an instrument, I show that doubling a surgeon's hip and knee volume reduces the probability of a complication by 9%, suggesting that a patient's choice of surgeon affects other patients' health outcomes. The coefficient on log volume in Table 2 shows these results. Column (1) shows the OLS estimate of the returns to surgeon volume. It indicates that doubling a surgeon's volume is associated with an 18% reduction in the probability of a complication. However, as discussed in the empirical strategy in section 4.2, this coefficient may be biased upward or downward for

[20] Specifically, I re-estimate demand in Equation 1, but I now let $\delta_{jt}(v_{jt})$ from Equation 2 depend on a composite risk index, $\widehat{X}_{it}$, such that $\delta_{jt}(v_{jt}) = \delta_{0j} + \delta_{1j}\widehat{X}_{it} + \delta_1 v_{jt}$. The composite risk index is simply the linear prediction of all the risk covariates on a complication. I use an index, as opposed to each individual risk characteristic, for computational tractability.

[21] Even though the finest geographic unit I observe is a ZIP code, I still have variation in differential distance because a patient's choice set also depends on when they undergo the replacement.

several reasons, such as selection on exogenous surgeon quality or unobservable patient risk. Thus, in column (2), I instrument for surgeon volume using differential distance and show that the effect falls in half. This coefficient yields the main result and indicates that doubling a surgeon's hip and knee volume decreases the probability of a complication by 9%. Because other papers that examine the "volume-outcomes" relationship for hip and knee replacement typically do not address all these types of selection, this effect is smaller than what has been estimated in other similar settings for hospitals (Hentschker and Mennicken, 2018; Kugler et al., 2022). Crucially, this result may have policy implications, as it suggests an externality–a patient's choice of surgeon affects other patients' health outcomes through the quality of the surgeon.

	(1)	(2)
Model	OLS	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0056	-0.0028
	(0.0003)	(0.0005)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$		-0.0047
		(0.0008)
Roy selection $\left(\theta_{ij}(j)\right)$		-0.0001
		(0.0004)
N	681,260	681,260
Mean complication	0.0314	0.0314
Median surgeon selection term $\left(\widehat{\phi}_k\right)$		-0.0033
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓

Table 2: Estimates of the returns to surgeon volume

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. The specification in column (1) provides the OLS estimate. The specification in column (2) instruments for surgeon volume using differential distance. Standard errors for the control function are calculated as described in Appendix C.5.

Using the control function implementation to explore the types of selection that bias the OLS estimate, I find that patients are more likely to choose higher exogenous quality surgeons. Note that in Table 2 column (2), the coefficient on the demand for exogenous surgeon quality, $\widehat{\delta}_{0j}$, is negative and statistically significant at the 10% level. Thus, an increase in the demand for exogenous surgeon quality is associated with a decline in the probability of a complication, an observation that is also supported in Appendix Figure D1. The magnitude implies that choosing a surgeon at the 75th percentile of $\widehat{\delta}_{0j}$, as compared to a surgeon at the 25th percentile, corresponds to a 12% decline in the probability of a complication. Hence, the OLS coefficient is biased upward because patients are more likely to choose higher exogenous quality surgeons, higher quality surgeons are more likely to locate in larger markets due to patient demand, or both. This result echoes other work that indicates that patient demand plays an important role in allocating patients to health care providers (Cutler, Huckman, and Landrum, 2004; Gaynor, Moreno-Serra, and Propper, 2013; Chandra et al., 2016). From a policy perspective, it suggests that, even without a causal effect of surgeon volume on outcomes, policies that reallocate patients to surgeons with higher hip and knee volume may generate welfare improvements through the surgeon's exogenous quality.

I also find suggestive evidence that unobservably healthier patients choose higher hip and knee volume surgeons. Returning to column (2) in Table 2, note that the median estimate of the surgeon-specific selection term coefficients, ϕ_k, is negative, which suggests that the median surgeon operates on unobservably healthier patients. However, as shown in Appendix D2, there is a distribution of positive and negative $\widehat{\phi}_k$'s, indicating that some surgeons appear to treat unobservably riskier patients while others appear to treat unobservably healthier patients. To better demonstrate how this selection affects the OLS estimate, I show in Appendix Figure D3 that surgeons with higher average annual hip and knee volume have lower values of $\widehat{\phi}_k$, suggesting that unobservably less risky patients choose higher hip and knee volume surgeons. This selection thus biases the OLS estimate upward, a result that is consistent with the decline in the coefficient on log volume from column (1) to column (2) in Appendix Table B2 when adding patient covariates. One potential explanation for this selection mechanism is that healthier individuals may have higher income and thus be less sensitive to traveling far or have better information (Dingel et al., 2023; Mourot, 2024).

Finally, I find that patients do not appear to select on an idiosyncratic match with a surgeon. Once again returning to column (3) in Table 2, note that the coefficient on the Roy selection control, $\theta_{ij}(j)$, is negative but quite small and statistically insignificant. While the negative coefficient would be suggestive of Roy selection, the small magnitude and lack of statistical significance indicate that such match-specific selection is likely unimportant in this setting. This type of selection therefore likely does not meaningfully bias the OLS estimate. Roy selection has also shown to be negligible in other health care settings (Einav, Finkelstein, and Mahoney, 2022; Mourot, 2024).

I partially disentangle the mechanisms underlying this returns to surgeon volume relationship and find that both static and dynamic economies of scale are important. To disentangle these mechanisms, I introduce lagged surgeon hip and knee volume into the estimating equation (Gaynor, Seider, and Vogt, 2005).[22] The coefficient on contemporaneous volume captures static differences in outcomes due to higher volume today, holding fixed last year's volume, such as greater team-specific human capital. Meanwhile, the coefficient on lagged volume captures dynamic differences, such as learning-by-doing, on-the-job experience, or forgetting, which have been shown to be important in other settings (Arrow, 1962; Irwin and Klenow, 1994; Benkard, 2000; Levitt, List, and Syverson, 2013; Caplin et al., 2022; Dinerstein, Megalokonomou, and Yannelis, 2022).[23] Table 3 shows that the coefficients on both contemporaneous and lagged volume are economically meaningful, statistically significant, and sum to slightly more than the total effect in column (1). Doubling contemporaneous hip and knee volume corresponds to a 7% decline in the probability of a complication, whereas the effect is about 4% for lagged volume. Because I cannot statistically reject that these estimates differ, this result indicates that both static and dynamic economies of scale drive the returns to surgeon volume relationship.

To address the potential bias from endogeneous surgeon and patient locations, I show that these results do

[22] To avoid the computational burden of re-estimating the demand model and then the main estimating equation, I omit lagged volume from the demand model. However, for one HRR, I show that the estimates from the demand model used in the main estimating equation essentially do not change when introducing lagged surgeon hip and knee volume into the demand model. Specifically, in Appendix Tables D1 and D2, I show that regressing the predicted choice probabilities and the estimated demand for exogenous surgeon quality, $\widehat{\delta}_{0j}$, from Equation 1 on the corresponding terms from a demand model with lagged hip and knee volume yield coefficients quite close to one.

[23] Motivated by the fact that the descriptive relationship between the probability of a complication and lagged measures of volume dies out after just one lag, as shown in Appendix Figure A1, I just include one lag.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0030	-0.0023
	(0.0005)	(0.0006)
Log lag volume $\left(\ln v_{ij}^{lag}\right)$		-0.0013
		(0.0005)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0049	-0.0044
	(0.0009)	(0.0009)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0002	-0.0002
	(0.0005)	(0.0005)
N	633,141	633,141
Mean complication	0.032	0.032
Median surgeon selection term $\left(\widehat{\phi}_k\right)$	-0.0036	-0.0036
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓

Table 3: Decomposing the returns to surgeon volume into static and dynamic economies of scale

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. Both specifications are estimated using a linear probability model. The specification in column (1) omits log lagged hip and knee volume, whereas the specification in column (2) adds it as a control. As compared to the main estimated in Table 2, the sample size falls because it only includes observations after 2008, such that both contemporaneous and lag volume can be measured. Standard errors are calculated as described in Appendix C.5.

not change when I add various surgeon and market-level controls or add fixed effects for the ZIP codes where patients live. Recall that if higher exogenous quality surgeons sort into larger markets based on characteristics other than patient demand, the returns to surgeon volume estimate would be biased upward. Thus, comparing column (1) to column (2) in Table 4, I show that the estimate of the returns to surgeon volume does not change when I control for the rank of a surgeon's medical school, HSA market size, ZIP code market size, the distance between the practice location and medical school as controls, and the amenity value of the Metropolitan Statistical Area (MSA) or non-metropolitan area.[24] Hence, these results thus lend support to the assumption that surgeon locations are conditionally exogenous (Kessler, Sage, and Becker, 2005). Additionally, in Appendix Table D3, I show that adding fixed effects for a patient's ZIP code also do not noticeably change the results, suggesting that patient locations are also conditionally exogenous.

In an additional specification, I introduce hospital and surgeon fixed effects into the main estimating equation. In Appendix Table D4, I show that including hospital or surgeon fixed effects decreases the returns to surgeon volume relationship noticeably but that these estimates do not fall outside the 95% confidence interval of the main estimate in column (4) of Table 2.[25] Even if they did fall outside the 95% confidence interval, however, interpreting this decline would be challenging. On one hand, these fixed effects directly address endogeneity concerns due to surgeons' endogenous location decisions. However, it is difficult to interpret the decline in the coefficient as simply a correction for endogeneity, as they also exacerbate the attenuation bias due

[24] Because HRRs are larger than MSAs and do not overlap entirely with MSAs, the median HRR has two different areas (either MSAs or non-metropolitan areas), with some having up to six.

[25] In practice, for ease of computation, I treat these fixed effects as local parameters, implying that I actually add hospital- and surgeon-by-patient market fixed effects.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0026	-0.0025
	(0.0006)	(0.0006)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0055	-0.0054
	(0.0011)	(0.0011)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0003	-0.0002
	(0.0008)	(0.0008)
N	495,821	495,821
Mean complication	0.0314	0.0314
Standard deviation surgeon selection $\left(\widehat{\phi}_k\right)$	-0.0030	-0.0032
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓
Surgeon covariates	X	✓
Market covariates	X	✓

Table 4: Estimate of the returns to surgeon volume with surgeon and market covariates

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. Both specifications are estimated using a linear probability model. The specification in column (1) omits surgeon and market-specific controls. Column (2) adds these controls, which include three dummies for the rank of a surgeon's medical school (1-25 rank; 26-50 rank; 51+ rank), the log distance between the surgeon's medical school and their practice locations, log HSA market size, log ZIP code market size, and the amenity value of the MSA or non-metropolitan area from Albouy (2016). As compared to the main estimated in Table 2, the sample size falls because it only includes observations when these controls are non-missing. Standard errors are calculated as described in Appendix C.5.

to measurement error in volume and provide further insight on the mechanisms underlying the relationship (Griliches and Hausman, 1986; Bound, Brown, and Mathiowetz, 2001). For instance, the fixed effects eliminate cross-hospital or cross-surgeon differences in nurse quality, which are potentially important drivers of the main relationship I estimate. Additionally, if annual hip and knee volume is considered a proxy for cumulative hip and knee volume, then it is difficult to interpret a within-surgeon estimate of the returns to surgeon volume relationship.

To probe the robustness of my results, I show that the returns to surgeon volume relationship is consistent across alternative specifications.[26] In Appendix Table D5, I show that the results are robust to eliminating surgeons with zero hip and knee volume and not adding one to volume. In a similar manner, I show the estimates of the returns to surgeon volume using volume quintiles, instead of log volume, in Appendix Table D6, which yields similar results. Additionally, I separately identify the effects for hip replacements and knee replacements in Appendix Table D7 and find a larger returns to surgeon volume relationship for knee replacements than for hip replacements. The estimate on each of these coefficients is smaller than the combined estimate, suggesting that skills may spillover between the two procedures. Finally, in Appendix Table D8, I examine how spillovers across other procedures affect the estimate by including log volume for other common orthopedic surgical procedures in the estimating equation but do not find that these controls affect the estimate.

[26] As with the specification that includes lag volume, I do not change the demand model under these alternative functional forms to avoid the computational burden of re-estimating the demand model. However, for one HRR, I show that the estimates from the demand model used in the main estimating equation essentially do not change under these alternative demand model functional forms. Specifically, in Appendix Tables D1 and D2, I show that regressing the predicted choice probabilities and the demand for exogenous surgeon quality, $\widehat{\delta}_{0j}$, from Equation 1 on the corresponding terms from a demand model with an alternative functional form yield coefficients quite close to one.

5.2 Quantification exercises

To provide insight into the mechanisms underlying the benefits of market size, I use the estimate of the causal returns to surgeon volume and show that the returns to surgeon volume explain 22% of the benefits of market size. I also find that the differences in average surgeon hip and knee volume only explain 3% of the geographic variation in outcomes and 2% of the variation in surgeon quality.

Using the causal returns to surgeon volume, I attribute 22% of the benefits of market size to the returns to surgeon volume. I arrive at this number by performing the same calculation as in Section 3 but with the causal returns to surgeon volume. Namely, doubling the size of the market decreases the probability of a complication by 2.0% through the returns to surgeon volume, since surgeon volume does not increase as much as market size. Thus, since Figure 3 shows that doubling the size of the market corresponds to a 9% decline in the probability of a complication, the returns to surgeon volume explain 22% of the benefits of market size. Because the returns to surgeon volume relationship also underscores an externality-a patient's choice affects other patients' outcomes through surgeon quality-it may be especially relevant for place-based policy. It also provides an empirical magnitude to previous theoretical contributions that have demonstrated that learning and specialization are microfoundations underlying the benefits of agglomeration (Duranton and Puga, 2004; Rosenthal and Strange, 2004).

I also show that differences in surgeon hip and knee volume explain only 3% of the geographic variation in health outcomes and 2% of the variation in surgeon quality. As a large literature has emphasized, there is substantial spatial variation in health outcomes (Chetty et al., 2016; Finkelstein, Gentzkow, and Williams, 2019). Yet, little is understood about what drives this variation, despite its importance for policy and welfare (Chandra and Staiger, 2007; Deryugina and Molitor, 2021). Using the returns to surgeon volume estimate, I find that differences in volume explain only 3% of the difference in the probability of a complication between HSAs in the 25th and 75th percentiles. Meanwhile, another strand of literature has demonstrated substantial heterogeneity in worker quality, such as for physicians, teachers, and managers, and noted the difficulty in identifying the drivers of these differences (Chetty, Friedman, and Rockoff, 2014; Lazear, Shaw, and Stanton, 2015; Ginja et al., 2022). Understanding this variation, as well as how much of a physician quality is endogenous, is critical for designing policy to potentially improve physician quality and better allocate physicians across space. Using the returns to surgeon volume estimate, I find that differences in volume explain only 2% of the difference in quality between surgeons at the 25th and 75th percentiles of quality. Thus, more work is needed to understand the primary drivers of this variation across space and surgeons.

6 Policy implications

The returns to surgeon volume relationship underscores an externality—patients' choice of surgeon affects other patients' outcomes. To understand the policy implications of these findings, I use the demand model and the estimate of the returns to surgeon volume to evaluate the welfare consequences of a first-best policy and three frequently discussed, feasible policies. I find that failing to incorporate the returns to surgeon volume

relationship substantially understates the welfare effects of the feasible policies. A minimum volume standard increases welfare by more than subsidizing transportation or moving surgeons to shortage areas, yet even this welfare effect falls far short of a first-best policy.

6.1 Measuring and interpreting welfare

To assess the consequences of various policies for patients, I incorporate the returns to surgeon volume externality into a welfare function. This welfare function allows me to evaluate how welfare changes under various counterfactual policies.

I define welfare as consumer surplus less fiscal costs from complications and policy implementation and incorporate the returns to surgeon volume externality into both consumer surplus and fiscal costs. More precisely, I define welfare in a given state of the world, $\mathcal{W}$, as the weighted average of welfare across HRRs, denoted by m:

$$\begin{aligned}\mathcal{W} &= \sum_m s_m \mathcal{W}_m \\ &= \sum_m s_m \left(CS_m - c \times N_{\text{comp,m}} - G_m \right), \end{aligned} \tag{10}$$

where s_m is the share of total patients in market m, CS_m is the consumer surplus in market m discussed more below, c is the dollar value of a complication, $N_{\text{comp},m}$ is the number of complications in HRR m, and G_m is the cost of implementing the policy in market m.

To expound on this welfare measure, I use the demand model to formally define consumer surplus, a key component in this welfare equation. Using Equation 1, for patient i and surgeon j, define $I_i = \ln\left(\sum_j \exp\left(\delta_j(v_{ij}) - \tau \ln d_{ij}\right)\right)$ as the inclusive value, or the expected utility. This expression measures patient i's expected utility from all the surgeons in the patient's choice set. I measure consumer surplus as the sum of inclusive values across patients, assuming all patients have equal welfare weights, and convert the change in utils to dollars. Thus, for patient i in market m, consumer surplus is:

$$CS_m = \sum_{i \in m} I_i \times \frac{(1 + \overline{d_i})\nu_c}{\tau_m}, \tag{11}$$

where $\overline{d_i}$ is the probability-weighted sum of the pre-policy distances patient i faces to each surgeon in her choice set, ν_c is the dollar value of distance constructed based on the average wage in patient i's county of residence c, and τ_m is the estimated distance coefficient in market m from Equation 1. Intuitively, $\frac{\nu_c}{\tau_m}$ acts as a conversion factor that converts utils to miles using $\frac{1+\overline{d_i}}{\tau_m}$, the marginal rate of substitution between utils and miles at patient i's expected travel distance, and then to changes in dollars using ν_c. In expressing consumer surplus in this way, I assume that patients fully internalize the cost of a complication to themselves, as they trade off a surgeon's quality with the distance they must travel.

This welfare function embeds two key externalities. The first is that a patient's choice of surgeon affects other patients' outcomes. I incorporate this externality both in CS_m, as patient utility depends on a surgeon's

hip and knee volume through the demand model, and in $N_{comp,m}$, as the number of complications depends on the quality of surgeons and thus on surgeon volume. I also incorporate the fiscal cost of complications because patients do not face full financial responsibility for their care with insurance, and the fiscal cost of a complication to the Medicare program is high (Yi et al., 2015).[27] This fiscal externality that has been shown to be important in other health care settings, such as in choosing managed care plans (Baicker, Chernew, and Robbins, 2013; Layton and Politzer, 2025). Intuitively, with these externalities, a policy will change welfare if it change patient choice. The new allocation of patients to surgeons not only affects those who switch surgeons but also the patients who do not switch, as the quality of their surgeon changes.

While this measure of welfare captures the key elements of patient and government welfare, it does not consider producer surplus, nor any general equilibrium effects. First, I assume that these policies do not change surgeon profits. While this assumption may not be realistic, it is beyond the scope of this paper to consider but a promising area for future work (Baker et al., 2014; Kleiner, White, and Lyons, 2015). This welfare measure also does not capture general equilibrium effects, such as if these policies affect cost-sharing or premiums or incentivize more patients to receive a hip or knee replacement or fewer doctors to specialize in orthopedics. This partial equilibrium analysis, however, would help form the basis for a general equilibrium analysis. Additionally, the primary goal of the counterfactual policy analysis is to better understand how incorporating the externality implied by the returns to surgeon volume relationship affects the policy implications, not necessarily to comprehensively analyze each policy.

Finally, since latent choice constraints, such as available capacity or imperfect information, may affect demand and hence welfare, I implement a capacity constraint and hold fixed all other constraints under alternative policy regimes. Recall from Section 4.1 that δ_{0j} and δ_1 may be correlated with other surgeon characteristics that affect patient demand, such as available capacity. If, for instance, surgeon j is capacity constrained, δ_{0j} would be understated among patients who have this surgeon in their choice set. I directly address the challenge this poses for welfare in two ways. First, I explicitly incorporate capacity constraints for hip and knee volume into the counterfactual policy implementation, which I discuss in more detail in the subsequent Subsection 6.2. Second, I assume all other latent choice constraints are fixed across alternative policy regimes.[28] For example, if patients harbor incorrect information about surgeon j's exogenous quality, δ_{0j}, I assume their information set does not change under different policy counterfactual policies. Similarly, if δ_1 captures both changes in volume and changes in information about a surgeon, I assume this relationship also holds under the counterfactual policies.

6.2 Implementing the counterfactual policies

To evaluate the counterfactual policies, I incorporate the externality from patient choice using feedback loops between patient demand and surgeon volume. I also implement the capacity constraint, choose a value to

[27] Other work suggests that higher volume providers impose additional fiscal externalities as they perform the same surgery for higher costs, perhaps due to greater bargaining power (Porter et al., 2025). In this setting, I do not observe this same relationship, which likely reflects the fact that in the Medicare FFS program, prices are administratively set.

[28] This assumption also implies that a surgeon's total "capacity" is held constant. That is, I assume that the total time that a surgeon spends working is constant across policy regimes. They implicitly substitute to and from other procedures or work activities.

convert distance into dollars, calibrate the cost of a complication, and discuss other assumptions related to implementation.

To help capture the key externality in the counterfactuals–the relationship between a patient's choice and other patients' outcomes–I use the demand model to introduce feedback loops between patient demand and surgeon volume. Recall from Equation 1 that surgeon demand depends on hip and knee volume: $\delta_{jt}(v_{jt}) = \delta_{0j} + \delta_1 v_{jt}$. Economically, this equation captures the externality from the returns to surgeon volume relationship–a given patient's choice affects other patients' outcomes through surgeon volume. Therefore, when I introduce a policy and a surgeon's volume changes, demand changes. The parameter δ_1 governs the magnitude of the change, so I show the distribution of these parameters in Appendix Figure .[29] In turn, this change in demand changes volume and so on. Thus, demand and volume change iteratively, so I update v_{jt} and $\delta_{jt}(v_{jt})$ until convergence.[30] I interpret this fixed point as the long-run equilibrium of the system following policy implementation.

I introduce a capacity constraint on surgeons' hip and knee volume to reflect surgeons' time constraints and address a latent choice constraint. In particular, I impose a volume constraint at the 99th percentile of observed hip and knee volume and ration patients based on their surgery date. Namely, if a surgeon's volume in the 365 days prior to the patient's surgery date exceeds the constraint, I omit the surgeon from the patient's choice constraint. This constraint is "soft," as surgeon volume can still exceed the constraint. For example, suppose a surgeon exceeds the constraint for a given patient but no previous patients. Then, implementing this capacity constraint does not change the volume of the surgeon at the time of the patient's hip or knee replacement (although it does for future patients).

Evaluating the counterfactual policies requires a conversion factor between distance and dollars, which I obtain following the literature. I calculate the dollar cost of a mile in county c, ν_c, as the dollar value of driving plus the opportunity cost of time, which I provide more details on in Appendix E.1. Following the literature, I compute the dollar value of driving using the Internal Revenue Service (IRS) mileage reimbursement rate and the opportunity cost of time using the average wage in a patient's county (Einav, Finkelstein, and Williams, 2016; Dolfen et al., 2023; Rosenberg, 2025). This formulation imposes the assumption that the opportunity cost of time is homogeneous across patients within a county and that the opportunity cost of time for a Medicare hip and knee patient is the same as that of a working individual.

Analyzing the policy implications also requires an estimate of the fiscal cost of a complication, which I calibrate according to the literature. To calibrate this fiscal cost, c, I rely on Yi et al. (2015), who estimate the total inpatient, outpatient, skilled nursing facility, and home health Medicare payments following a periprosthetic joint infection from hip and knee replacement. They find that patients with a periprosthetic infection spend \$53,470 more in the four years following a replacement than those without an infection. I convert this value to 2012 dollars, implying $c = \$65,442$. While my measure of a complication includes other adverse health

[29] Recall also that δ_1 may be biased downward due to measurement error in volume. In this case, I would understate the welfare effects of the policies.

[30] To ensure convergence and limit the influence of noisier estimates of τ and δ_1 in smaller markets, I shrink both of these market-level parameters to the national mean using the Empirical Bayes procedure described in Appendix C.1.

consequences, infections are the most common. If anything, this value likely understates the true cost, given that death occurs in 8% of complications.

The counterfactual policy evaluation relies on three other assumptions that ease the computational burden and simplify the analysis. First, I do not consider how policy affects the choice of the outside option because it is difficult to measure and interpret changes in quality to what is, in practice, a large number of surgeons practicing outside the market. Hence, I consider only how policy changes within-market substitution across surgeons, a reasonable simplification in a setting in which the median patient travels only 10 miles for a hip or knee replacement. Second, because annual surgeon-level complication rates are noisy, I use the average surgeon complication rate across all years in the counterfactual evaluations.[31] Finally, to simplify the analysis, I do not allow the counterfactual policies to affect patient selection of surgeons based on unobservable choice determinants. Namely, if a policy increases a certain surgeon's hip and knee volume, I do not allow the policy to influence a patient's unobserved choice determinants, such as the patient's unobserved riskiness, for that surgeon.

6.3 Calculating a first-best policy

To provide a relevant comparison for the three feasible policies and guide future policy design, I first compute the welfare gain from a first-best policy in which a social planner optimally assigns patients to surgeons. I calculate the welfare gain from this first-best policy by allowing a social planner to set surgeon-specific Pigouvian taxes or subsidies. I show that under this first-best policy, the social planner re-allocates patients to higher-quality surgeons and thus generates large efficiency and distributional welfare gains.

To provide intuition for this exercise, I consider how the social planner would assign patients in two simplified cases. Suppose first that patients face no travel costs and surgeons cannot be capacity constrained. In this case, the social planner would assign all patients to the highest exogenous quality surgeon. The large increase in volume for this surgeon would result in a further welfare increase through the returns to surgeon volume externality. Now, suppose that patients face non-zero travel costs, but surgeons still do not face a capacity constraint. Even in this simplified setting, the solution is not obvious. In Appendix E.3, I simplify this example even further and show that the social planner's assignment depends on the relative importance of the externality versus the importance of travel costs. When the externality is small relative to the travel cost, the planner would just assign patients to the closest surgeon, as assigning a farther away patient to a surgeon does not generate sufficient public benefit to justify the larger travel cost. However, when the returns to surgeon volume relationship is large relative to the travel cost, allocating all patients to one surgeon outweighs the costs of having patients travel farther.

I generalize this intuition by introducing a surgeon-specific Pigouvian tax or subsidy into the demand model that the social planner manipulates to maximize welfare and achieve the first-best allocation. Formally,

[31] To address additional measurement error, I also shrink these estimates to the national mean using the Empirical Bayes procedure described in Appendix C.1.

I modify the demand model from Equation 1 as follows:

$$u_{ij} = \delta_j(v_{ij}) - \tau d_{ij} + \pi_j + \eta_{ij}, \tag{12}$$

where all else is defined as before and π_j denotes the tax (or subsidy) for surgeon j. Intuitively, the social planner manipulates these parameters to steer patients to certain surgeons based on the welfare function. These terms can be thought of as surgeon-specific co-payments (or subsidies) that a savvy private insurer might set. Because they function as pure transfers and exactly cancel out with fiscal spending or revenue, they are not used to calculate welfare.

Solving for the vector of tax parameters in the demand model allows me to recover welfare under the first-best policy. In particular, for each market, I solve for π_j for every surgeon in the market using a non-linear unconstrained optimization algorithm. In contrast to a combinatorial optimization algorithm, which could also be used to solve for the first-best policy, this method is more computationally tractable. With this method, solving for the tax parameters induces the equilibrium assignment. Specifically, to solve for the parameters, I use bound optimization by quadratic approximation, a local, derivative-free algorithm, which is designed to find the local maximum of reasonably well-behaved functions without mathematical derivatives. Because, however, this optimization is still quite computationally burdensome, I solve for the optimal tax parameters among HRRs with less than or equal to forty surgeons, which is 63% of HRRs.[32]

I show first that the social planner introduces large subsidies for high-quality surgeons, but the subsidy amounts are substantially lower and exhibit a weaker relationship with quality once I incorporate the returns to surgeon volume externality. In Figure 3, I demonstrate that the social planner's optimal tax has a strong negative correlation with a surgeon's risk-adjusted complication rate, both with and without incorporating the externality. That is, the social planner subsidizes surgeons who have a below average complication rate and taxes those above the average. Nevertheless, with the externality, the tax parameters are substantially smaller (in absolute value), reflecting the fact that the market amplifies the changes in surgeon volume that the taxes induce, as patients respond to these subsidies when surgeon volume changes. The differences in slopes, meanwhile, reflect a fundamental tradeoff once I incorporate the externality–funneling patients to high-quality surgeons is welfare-increasing for patients who choose those high-quality surgeons but welfare-decreasing for those who choose other surgeons. Notably, these tax parameters only weakly correlate with other surgeon characteristics, such as their average volume, their average distance to patients, or the demand for their exogenous quality. Thus, the social planner primarily targets quality.

With the first-best policy, care is significantly more concentrated among higher-quality surgeons, yet the concentration is 22% lower when incorporating the returns to surgeon volume externality. Because the social planner sets large tax parameters, patients substantially change their choice and concentrate care among a much smaller set of surgeons, as shown in Appendix Figure E3. Averaging over the results with and without the externality, concentration increases from 0.06 to 0.48 under the optimal policy, as measured by the

[32]The welfare effects across all HRRs for the other policies are similar to those among these smaller HRRs.

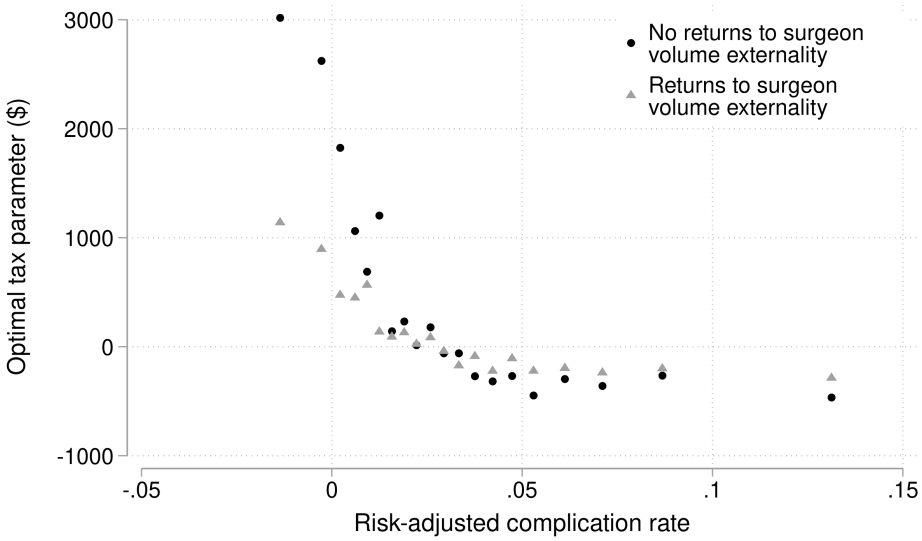


Figure 3: Relationship between optimal tax parameter and risk-adjusted complication rate

Notes: This figure shows the relationship between the optimal tax parameter for a surgeon, π_j, and the surgeon's risk-adjusted complication rate. The optimal tax parameter has been converted into dollars following the method in Appendix E.1 and using an average conversion factor across markets. The risk-adjusted complication rate has been shrunk to the mean using Empirical Bayes, as described in Appendix C.1. The relationship with the black dots shows this relationship without incorporating the returns to surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8). The relationship with grey triangles shows the relationship when incorporating the returns to surgeon volume externality. The underlying unit of observation is an HRR-surgeon that has been grouped into twenty equal-sized bins based on a surgeon's risk-adjusted complication rate after residualizing out HRR fixed effects.

Herfindahl-Hirschman Index (HHI). Nevertheless, when incorporating the externality, the HHI is 20% smaller, which once again reflects the fact that re-allocating patients to higher-quality surgeons decreases the welfare of those left with the lower-quality surgeons. Thus, under the optimal policy, the average patient travels 47% farther without the externality but only 34% farther with it. As shown in Appendix Figures E4 and E5, the social planner concentrates care among high-quality surgeons, not necessarily among high-volume surgeons. Thus, while greater concentration yields substantial welfare increases, the types of surgeons in which care is concentrated also matters.

Finally, I show that with and without returns to surgeon volume externality, the average and heterogeneous welfare results are quite similar. In Table 5, I show that the average welfare effects of the optimal policy are quite large, as they exceed the average amount paid to surgeons for the replacement and comprise about 7% of the total costs of a replacement, including the post-operative care (Dummit et al., 2016). Both with and without the externality, the welfare increase is quite similar, indicating that gains from concentrating care when incorporating the externality are offset by the losses from those who still choose lower volume surgeons. Increases in fiscal savings drive these effects, as the social planner re-allocates patients to higher exogenous quality surgeons. Consumer surplus actually declines, although the magnitude is trivial compared to the change in fiscal savings. When incorporating the externality, the decline in consumer surplus is much larger, reflecting the fact that most surgeons are now low-volume. In Appendix Figure E6, I show that the welfare effects for rural and urban patients are quite similar under this optimal policy both with and without the externality.

	No returns to surgeon volume externality	Returns to surgeon volume externality
$\Delta\mathcal{W}$ / person ($)	2150.74	2171.12
Δ (CS - G) / person ($)	-1.75	-21.30
Δ fiscal savings / person ($)	2152.49	2192.42

Table 5: Average welfare effect of optimal policy

Notes: This table shows the average welfare effect of implementing the optimal policy. The first row of each panel shows the average change in per-person welfare in dollars, as calculated using Equation 10. The second and third rows show the average change in the (per-person) sub-components of welfare: consumer surplus less fiscal spending on policy implementation and fiscal savings from reduced complications, respectively. Consumer surplus is defined in Equation 11. The fiscal costs of the policy are calculated as in Section 6.4 and are equal to zero. The fiscal savings per person are calculated as in Equation 10. The first column shows the effect without the returns surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8), while the second column incorporates the returns to surgeon volume externality.

6.4 Introducing and implementing the feasible counterfactual policies

I evaluate how incorporating the returns to surgeon volume externality affects the welfare implications of three frequently discussed, feasible policies: a minimum volume standard, subsidies for patient transportation, and moving surgeons to government-designated shortage areas. I discuss the tradeoffs of these policies, the policy context, and the implementation details. Because of the externality, the direction and magnitude of the welfare effect for each of the policies is theoretically ambiguous.

I first consider a minimum volume standard, which imposes a volume threshold for hip and knee volume such that surgeons below that threshold are not permitted to perform hip and knee replacements and thus trades off higher quality care with less choice and greater travel distance for patients. This policy effectively imposes an infinite tax on low-volume surgeons, forcing patients to re-allocate among the remaining surgeons. Theoretically, the welfare effect of this policy is unclear. On one hand, patients may receive higher quality care. First, as patients re-allocate among surgeons with higher hip and knee volume, they will be choosing among higher exogenous quality surgeons, as demonstrated previously in the results in Table 2. Additionally, in consolidating care, it may increase quality through the returns to surgeon volume relationship. On the other hand, this policy will reduce consumer surplus through both a reduction in the number of choices and an increase in travel distance. Thus, the welfare effect of this policy depends on these competing factors.

In the U.S., minimum volume standards are especially common at the state-level for stroke and cardiac care. For example, some states, such as New York, impose minimum volume requirements for hospitals to treat severe stroke patients.[33] Other states have passed similar laws for Coronary Artery Bypass Grafting (CABG). Other prominent interest groups, such as the Leapfrog Group, advocate for minimum volume standards for many surgical procedures. Minimum volume standards are also quite similar to restricted networks that introduce strong financial incentives for patients to choose high volume surgeons. While Medicare FFS patients have no network restrictions, these restrictions are quite common in the Medicare Advantage setting, and some Medicare Advantage insurers even require networks to contain high volume specialists.[34]

[33]The policy details are here: https://health.ny.gov/diseases/cardiovascular/stroke/designation/docs/stroke_center_guidance.pdf.

[34]For a specific example, please see https://www.blueshieldca.com/content/dam/bsca/en/provider/docs/2024/February/PRV_1-24-A11504-HG-Manual.pdf

To implement the minimum volume standard, I remove any surgeon from a patient's choice set who performs less than or equal to twenty replacements in the 365 days prior to the patient's surgery date. This policy corresponds to a large change in the number of surgeons but a small change relative to the total number of replacements performed. Namely, since the median surgeon in my sample performs twenty-three replacements, this policy reduces the number of choices in the average patient's choice set by 45%. However, since these surgeons have low hip and knee volume, they only perform 12% of hip and knee replacements. After removing these surgeons from patients' choice sets, I then allow patients to re-optimize over the remaining surgeons. I set G_m, the cost of implementing the policy, to be zero, as the government would pay no direct cost of imposing such a policy.

Second, I consider a policy that subsidizes transportation for patients and thus trades off higher quality care with the costs of the subsidy. In reducing patients' sensitivity to travel, this policy may better allocate patients to surgeons. On one hand, therefore, this policy may improve welfare if the subsidy incentivizes patients to travel to higher exogenous quality surgeons and to further centralize care. On the other hand, without the returns to surgeon volume externality and the fiscal externality in the welfare function, this subsidy is purely distortionary, since patients will only travel farther if their expected utility gain does not exceed the value of the subsidy. Hence, with a small returns to surgeon volume relationship, the fiscal cost from the subsidy may exceed the benefit of the policy. The sign and magnitude of the welfare change therefore depend on the relative importance of the welfare decrease from distorting patient choices versus the welfare increase from greater surgeon volume and fiscal savings.

Subsidies for patient transportation are quite common within the Medicare program. While they do not currently exist within the Medicare FFS program, they are prevalent as supplemental benefits in Medicare Advantage plans, which cover more than 50% of Medicare beneficiaries. In 2024, 44% of Medicare Advantage plans offered some type of coverage for non-emergency medical transportation, an increase of four percentage points even from just four years prior (Shen et al., 2024). This coverage often reimburses patients for transportation within certain geographic bounds, and it may restrict the number of trips in a policy year or require some cost-sharing. While I do not consider these more complicated policy designs, the welfare analysis still provides insight into the effects of these supplemental insurance benefits.

I implement a transportation subsidy by changing the coefficient on distance in the demand model, setting the cost of the policy equal to the average wage in a patient's county plus the dollar cost of driving, and computing compensated changes in consumer surplus. Specifically, to implement this counterfactual, I change the coefficient on distance in the demand model, τ, and use the model to examine how patients re-sort. Specifically, I implement a policy that subsidizes 30% of patients' transportation, so I cut τ by 30%. Meanwhile, I calibrate G_m, the fiscal cost of implementing the policy, in the same way that I calibrate ν_c, the dollar value of a mile, as described in Appendix E.1. That is, the cost of the policy is the wage in the patient's county plus the dollar cost of driving. Intuitively, this value might reflect the distance and time cost to patients if family members or friends drive them. I assume that the cost of raising taxes to finance this policy is zero, as the government can raise revenues with a non-distortionary lump sum tax. Finally, to isolate welfare changes resulting from the

reallocation of patients to surgeons, I compute the welfare change using a compensated change in consumer surplus, as described in Appendix E.2. That is, I compensate patients for the mechanical increase in utility from surgeons being effectively closer, had they not been allowed to change their behavior.

Finally, I consider a policy in which I move surgeons to government-designated shortage areas, which trades off decreased travel distance for patients in these shortage areas with theoretically ambiguous changes in quality for the moving surgeons and non-moving surgeons in both the non-shortage and shortage areas. On one hand, for the destination locations, the presence of more surgeons decreases travel distance for patients in shortage areas. On the other hand, because many shortage areas are also small markets and may not be able to support large hip and knee volume, the moving surgeons may lose hip and knee volume and thus have lower quality. Additionally, decentralizing care in the shortage area may decrease the quality of the other surgeons in the shortage area if they lose hip and knee volume. Meanwhile, in the origin locations, welfare will decrease as patients must travel greater distances, but this decrease may be offset by an increase in quality as care becomes more centralized. Complicating this discussion is the fact that these welfare results depend on which surgeons move and the fiscal cost of subsidizing surgeons to move, considerations that I mostly abstract from in this paper.

One reason to consider this policy is that policymakers commonly advocate for and implement similar policies. Specifically, the Health Resources and Services Administration (HRSA) designates Health Professional Shortage Areas (HPSAs), and Congress has enacted various policies to incentivize surgeons to move to these shortage areas. For example, under the National Health Service Corps, physicians are eligible for educational loan repayments or scholarships if they practice in HPSAs. Likewise, immigrant physicians are eligible for J-1 visa waivers if they practice in HPSAs. Finally, Medicare FFS pays more to physicians when they perform care in these areas. Thus, this counterfactual exercise serves to provide insight into some of these previously implemented policies, as well as to guide future policy design.

I implement this policy by moving surgeons whose primary practice location falls outside a shortage area into hospitals located in shortage areas within the same HRR. Specifically, to implement the policy, I first identify all ZIP codes in my sample in which a hip or knee replacement is performed during my sample and is located in a geographic primary care HPSA county.[35] I define eligible surgeon movers as those whose primary practice locations falls outside of one of these HPSA ZIP codes.[36] I then assign one eligible surgeon to each HPSA ZIP code within the same HRR randomly, such that surgeons can only be moved within the same HRR.[37] Moving surgeons within the same HRR perhaps more accurately reflects how such a policy would actually affect the spatial distribution of surgeons, as surgeons may not move cross-county but rather within the same broader health care market. Additionally, this implementation facilitates the analysis, as the demand

[35] I focus on primary care HPSAs because there are no shortage designations for specialty care. HRSA designates primary care physician (PCP) shortages areas based on four statistics: the population-to-PCP ratio, the percent of the population below the federal poverty level, the infant health index (based on the infant mortality rate or the low birth weight rate), and the travel time to the nearest source of care outside the HPSA designation area. The geographic area ranges in size and can be a county, a group of counties, or part of a city.

[36] As before, I define surgeons' primary practice location as the ZIP code where they perform the plurality of their hip and knee replacements. This time, though, I define the primary practice location over the entire sample period.

[37] For HRRs where there are more destinations than movers, some destinations do not receive a mover. For destinations with more movers than destinations, I assign surgeons randomly.

parameters δ_{0j} are only interpretable within an HRR. Finally, I set the cost of implementing the policy $G_m = 0$, so this welfare analysis can be considered an upper bound on the true effect. In total, I move 1,329 surgeons, which is about 12% of all the surgeons in my sample, to the same number of shortage areas within 68% of the HRRs.[38] In general, surgeons move to smaller markets, as 60% of the shortage areas are rural.

6.5 The effect on the concentration of care

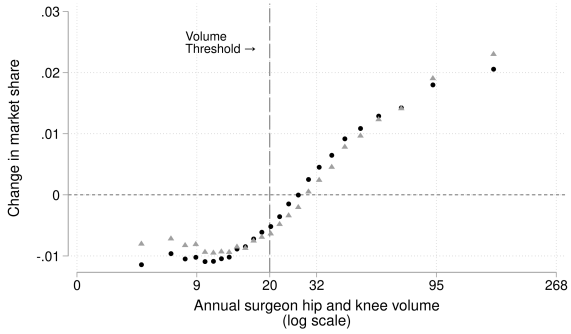
Since the welfare implications of these policies depend on the extent to which they change patient choice, I assess the effect of each feasible policy on the concentration of care. For all three policies, the externality amplifies the changes in patient choice. The minimum volume standard generates the largest changes in patient choice, as the market share of surgeons with higher hip and knee volume substantially increases.

I first show that the minimum volume standard substantially increases the market share of surgeons with high hip and knee volume, and this increase is 16% greater once I incorporate the returns to surgeon volume externality. Figure 4 shows the change in the market share of surgeons versus the surgeon's average log hip and knee volume both with and without incorporating the externality for each of the three policies. The first panel, Figure 4a, demonstrates that the minimum volume standard significantly increases the concentration of care. Regardless of whether I incorporate the externality, surgeons who perform fewer hip and knee replacements than the threshold (denoted by the vertical dashed line) lose virtually all of their market share, as these patients are forced to change their choice.[39] This volume is then concentrated among surgeons with much larger average hip and knee volumes. Focusing on the surgeons with especially high average annual volume indicates that incorporating the externality generates larger increases in market share than without incorporating it. For instance, among surgeons who perform greater than ninety-five hip and knee replacements in a 365 day time period increase their market share by 80% on average following the policy. This increase is 16% greater when incorporating the returns to surgeon volume externality.

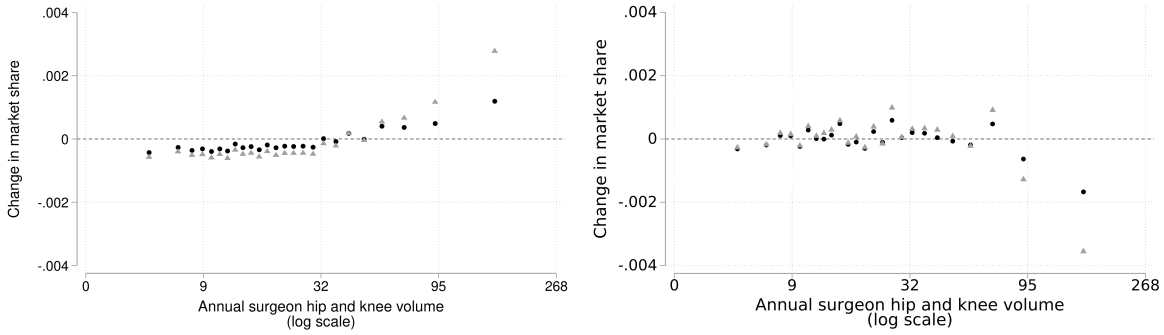
Next, I show that although subsidizing patient transportation increases the market share of higher hip and knee volume surgeons, especially with the returns to surgeon volume externality, the change is substantially smaller than with the minimum volume standard. In Figure 4b, the two upward sloping relationships once again indicate that care is concentrated among surgeons with higher hip and knee volume when I implement the policy, regardless of whether I incorporate the externality. Thus, the policy induces patients to choose surgeons with higher hip and knee volume. Once again, focusing on surgeons with high hip and knee volume indicates that the subsidy increases the market share of these surgeons by substantially more when incorporating the externality. Using the same example as before, among surgeons who perform greater than ninety-five hip and knee replacements within a 365 day time frame on average, the increase in market share is 141% larger when incorporating the externality. Nevertheless, as compared to the minimum volume standard, this policy does not substantially change patient choice, as these changes in market share are almost an order of magnitude smaller. For instance, even when incorporating the externality, the increase in market share among surgeons

[38] I therefore only consider the welfare effects among these HRRs who have a moving surgeons.

[39] The grey triangles lie slightly above the black dots because when I incorporate the returns to surgeon volume, the baseline concentration of care is already greater prior to the implementation of the policy. Thus, these surgeons with low hip and knee volume have less market share to lose.



(a) Panel A: Minimum volume standard



(b) Panel B: Subsidizing transportation

(c) Panel C: Moving surgeons

• No returns to surgeon volume externality
▲ Returns to surgeon volume externality

Figure 4: Relationship between change in market share and surgeon hip and knee volume

Notes: Each of these figures shows the relationship between the change in a surgeon's market share and the surgeon's average log volume after implementing a minimum volume standard (Figure 4a), a 30% subsidy for patient transportation (Figure 4b) and moving surgeons to government-designated shortage areas (Figure 4c). The relationship with the black dots shows this relationship without incorporating the returns to surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1). The relationship with grey triangles shows the relationship when incorporating the returns to surgeon volume externality. The underlying unit of observation is a surgeon-HRR that have been binned into twenty-five equal-sized bins based on a surgeon's average log hip and knee volume after residualizing out HRR fixed effects. For better visualization, the top figure is on a different scale than the bottom two figures.

who perform greater than ninety-five hip and knee replacements within a 365 day time frame on average is only 10%.

Finally, I show that, unlike the previous two policies, moving surgeons to government-designated shortage areas decreases the market share of surgeons with high hip and knee volume, especially once incorporating the returns to surgeon volume relationship, but that this effect is small relative to the minimum volume standard. Figure 4c shows a slightly downward sloping relationship between the change in the market share and a surgeon's hip and knee volume, implying that surgeons with relatively lower hip and knee volume gain market share, while those with high hip and knee volume lose market share. This result accords with the intention of policymakers to decentralize care so as to provide better access to patients in shortage areas. Incorporating the returns to surgeon volume relationship once again dramatically changes the results. Among surgeons who perform greater than ninety-five hip and knee replacements within a 365 day time frame on average, the decrease in market share is 111% larger when incorporating the externality. Once again, though, the magnitude of the change from the policy is small, as patients do not substantially change their choice. Even when incorporating the externality, the decline in market share among surgeons who perform greater than ninety-five hip and knee replacements within a 365 day time frame on average is only 15%.

6.6 Efficiency: Evaluating the average welfare effects

I now investigate the change in efficiency from each of these feasible policies. I find that each policy generates a welfare increase and that incorporating the returns to surgeon volume substantially increases these welfare effects. Among the three policies, the minimum volume standard generates the largest welfare increases, yet it achieves only 7% of the gain from a first-best policy.

I first show that implementing the minimum volume standard increases welfare on average even without incorporating the returns to surgeon volume externality, as patients choose higher exogenous quality surgeons. Namely, Table 6a demonstrates that the increase in per-patient welfare from this policy is $123 without incorporating the externality. Moving down the rows in this first column further reveals that fiscal savings due to a decline in the complication rate drive this large increase in welfare. That is, when surgeons with low hip and knee volume can no longer operate, patients re-sort to surgeons with higher hip and knee volume, who are also higher exogenous quality surgeons. This welfare estimate echoes the results from the control function estimation shown earlier in Table 2, indicating that patients are more likely to choose higher exogenous quality surgeons. These fiscal savings, in turn, vastly outweigh the (mechanical) decline in consumer surplus resulting from less choice and longer travel distances. For example, prior to the minimum volume standard, 27% of patients live in an HSA with an orthopedic surgeon, but after the policy only 15% do, increasing patients' expected travel distance by 6%.

Once I incorporate the returns to surgeon volume externality, the welfare effect of the minimum volume standard increases by 26%, yet this effect is only 7% of the first-best policy. In the second column of Table 6a, I show this result, which indicates that the welfare increase is $154 per-patient once I incorporate the externality. Investigating the sub-components of welfare in the subsequent rows of the table demonstrates that both

	No returns to surgeon volume externality	Returns to surgeon volume externality
Panel A: Minimum volume standard		
$\Delta\mathcal{W}$ / person ($)	122.64	154.20
Δ (CS - G) / person ($)	-19.06	-11.97
Δ fiscal savings / person ($)	141.70	166.16
Panel B: Subsidizing transportation		
$\Delta\mathcal{W}$ / person ($)	-7.03	3.60
Δ (CS - G) / person ($)	-20.62	-20.63
Δ fiscal savings / person ($)	13.59	24.23
Panel C: Moving surgeons		
$\Delta\mathcal{W}$ / person ($)	1.67	14.07
Δ (CS - G) / person ($)	8.25	9.25
Δ fiscal savings / person ($)	-6.58	4.82

Table 6: Average welfare effects of policies

Notes: This table shows the average welfare effect of implementing a minimum volume standard (Panel A), a 30% transportation subsidy (Panel B), and moving surgeons to government-designated shortage areas (Panel C). The first row of each panel shows the average change in per-person welfare in dollars, as calculated using Equation 10. The second and third rows show the average change in the (per-person) sub-components of welfare: consumer surplus less fiscal spending on policy implementation and fiscal savings from reduced complications, respectively. Consumer surplus is defined in Equation 11. The fiscal costs of the policy are calculated as in Section 6.4 and are equal to zero for Panel A and Panel C. The fiscal savings per person are calculated as in Equation 10. The first column shows the effect without the returns surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8), while the second column incorporates the returns to surgeon volume externality.

decreased reductions in consumer surplus and increased fiscal savings due to a lower complication rate drive this result. Namely, while the contribution of consumer surplus to welfare in this policy is small relative to fiscal savings, the decline in consumer surplus almost 40% smaller when incorporating the returns to surgeon volume. This result reflects the fact that consolidating care increases surgeons' hip and knee volume, which patients demand. Second, fiscal savings increase as surgeon quality improves through the returns to surgeon volume relationship. Nevertheless, this effect is only 7% of the first-best benchmark, which is perhaps unsurprising given that the minimum volume standard imposes an infinite tax on some surgeons and no tax on all others.

Next, I show that implementing a 30% transportation subsidy decreases welfare without incorporating the returns to surgeon volume externality. This result, shown in the first column of Table 6b, indicates that welfare per patient falls by $7 following the subsidy. Moving down the rows shows that the main driver of these welfare losses is a decline in consumer surplus less the fiscal costs of policy implementation. This decline is mechanical, as without the externality, the policy only distorts decisions and lowers utility. The small decline generates a small increase in utility as patients choose higher exogenous quality surgeons. Nevertheless, the resulting fiscal savings do not generate sufficient fiscal savings to offset this difference.

Incorporating the returns to surgeon volume externality actually reverses the direction of the welfare effect for this 30% transportation subsidy, yet the magnitude is minuscule compared to both the minimum volume standard and the first-best policy. Namely, in the second column of Table 6b, I show that incorporating the externality actually generates a positive welfare change. As evident in the table, the primary driver of this difference is fiscal savings due to an increase in fiscal savings as care becomes more consolidated and surgeon quality improves. In fact, complication-related fiscal savings almost double when I incorporate the externality, thereby offsetting the decline in consumer surplus.[40] Despite this difference, these effects are still quite small. This welfare effect is only 2.3% of the welfare effect from implementing the minimum volume standard and less than 1% of the first-best benchmark. Since the welfare implications in this setting depend on the extent to which the policy changes choice and this policy does not substantially change choice, the welfare effect is small.

I now show that without incorporating the returns to surgeon volume externality, moving surgeons to shortage areas generates a small increase in welfare. In the first column of Table 6c, I demonstrate that moving surgeons to shortage areas increases per-patient welfare by $1.67 on average. Increases in consumer surplus barely offset increases in fiscal spending due to a higher complication rate. Intuitively, patients on average become better off as they travel less but worse off as their probability of a complication increases. Notably, however, this policy is the only one that generates increases in consumer surplus, as patients now travel less distance on average.

Once I incorporate the returns to surgeon volume externality, I find that moving surgeons to shortage areas increases welfare by more than seven times the amount without it. Namely, in the second column of Table 6c, I show that moving surgeons to shortage areas generates a $14 increase in per-patient welfare on average.

[40] Consumer surplus also increases when I incorporate the externality, but this increase is essentially entirely offset by greater fiscal costs of the subsidy.

As evident in the table, once incorporating the externality, fiscal savings actually become positive, as surgeons become higher quality. This surprising result counters the notion that moving surgeons to shortage areas would decrease their quality as they would not be able to maintain their hip and knee volume. In fact, it does the opposite because the increase in the probability that nearby patients choose the mover offsets the decline in probability that now farther patients choose the mover. Hence, moving surgeons experience a 61% increase in their hip and knee volume on average. Once I incorporate the externality, this increase in volume improves quality and thus generates fiscal savings. Since moving surgeons are chosen randomly, this interpretation is consistent with Figure 4c–surgeons with high hip and knee volume experience a decline in hip and knee volume as moving surgeons, who are more likely to be surgeons with low or medium hip and knee volume, "steal" some of their patients.[41]

As with subsidizing transportation, the welfare effect from moving surgeons is quite small compared to the minimum volume standard and the first-best policy. Specifically, moving surgeons achieves only 7% of the minimum volume standard and less than 1% of the first-best benchmark. These effects are about $10 greater per-person if I consider only HRRs with rural shortage areas, but overall they are still quite small because, as with subsidizing transportation, this policy does not substantially change patient choice. Recall, however, that this welfare effect is an upper bound, as it does not incorporate the fiscal cost of incentivizing surgeons to move, implying that the welfare change from this policy would likely be even smaller. For this policy to continue to have a positive welfare effect, the cost of moving an individual surgeon must not exceed $3,100 per surgeon per year. Given that the average annual income for orthopedic surgeons is almost $800,000, this welfare effect would actually likely be negative (Gottlieb et al., 2023).

6.7 Distributional effects: Evaluating the welfare effects for rural versus urban patients

Finally, since these policies often explicitly target patients in rural areas, I investigate how welfare changes differentially for rural and urban patients. I first find that, regardless of the policy or the type of patient, the returns to surgeon volume externality substantially increases the welfare effects. Second, while I find little heterogeneity in the welfare effect of the minimum volume standard, as compared to the other two feasible policies, even the largest heterogeneous welfare effects of the other two policies are still relatively small.

I show first that the minimum volume standard generates similar welfare effects for rural and urban patients both with and without the returns to surgeon volume externality, although rural patients bear more of the costs of the policy. Figure 5 first demonstrates that the welfare effects are roughly the same for rural and urban patients both with and without the externality. The externality has similar effects because rural patients are now more likely to choose surgeons in urban markets, increasing welfare for both rural and urban patients. In fact, the share of hip and knee replacements performed in urban markets increases by 6%. Nevertheless, despite these similar welfare increases, rural patients bear more of the cost of the policy. Namely, as shown in Appendix Figure E7, rural patients experience declines in consumer surplus that are about double what urban patients experience. This greater decline in consumer surplus occurs as rural patients' expected travel distance

[41] In fact, the surgeons at moving surgeons' destination ZIP codes lose 20% of their hip and knee volume on average.

increases by 7% while urban patients actually experience a slight decrease in their expected travel distance. The greater decline in consumer surplus for rural patients is offset by a larger increase in fiscal savings, as these patients now choose higher exogenous quality surgeons.

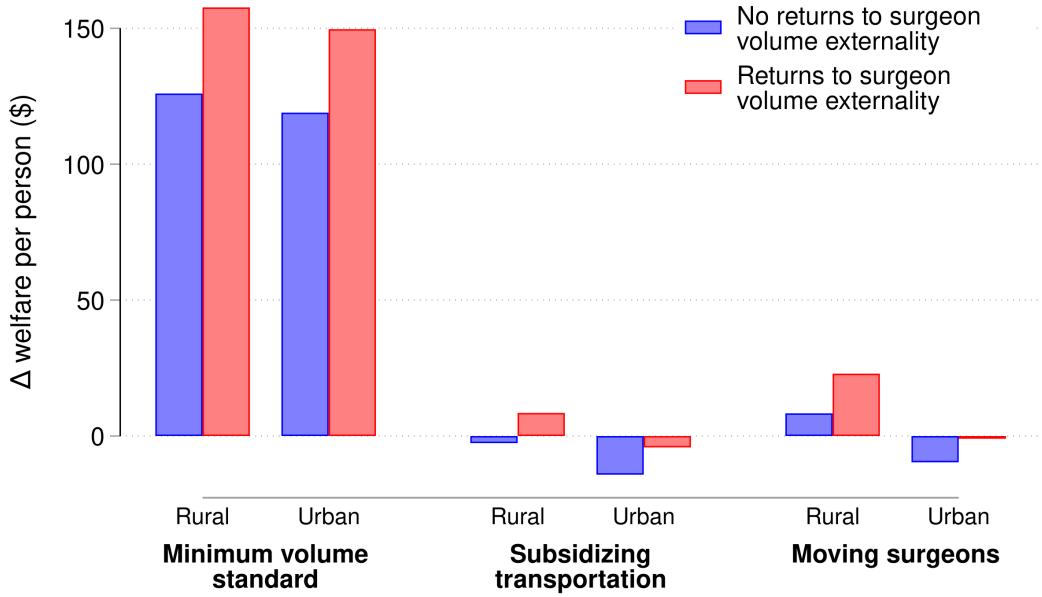


Figure 5: Welfare effects of minimum volume standard for rural versus urban patients

Notes: This table shows the welfare effect of implementing a minimum volume standard, a 30% transportation subsidy, and moving surgeons to government-designated shortage areas for both rural and urban patients. Welfare is calculated in dollars using Equation 10. Rural patients are defined as those having ZIP codes of residence outside a "city center," where a "city center" is the HSA with the largest market size within an HRR. Urban patients are defined as those having ZIP codes of residence within a "city center." In 3 of the 306 HRRs, patients only live in one of the HSAs within the HRR, so these HRRs are omitted from this analysis. The blue bars show the effect without incorporating the returns surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8), while the red bars incorporate this externality.

I also show that the returns to surgeon volume externality increase the welfare effect of the 30% transportation subsidy for both rural and urban patients, but the resulting effect is negative for urban patients and positive for rural patients. Figure 5 first demonstrates that incorporating the externality increases the welfare effect for both urban and rural patients. The externality has this effect because rural patients use the subsidy to travel to larger markets, which improves outcomes for both rural and urban patients. In fact, the share of hip and knee replacements performed in urban markets increases by 3%. The figure also shows that the welfare effect for rural patients is positive for rural patients and negative for urban patients. This difference occurs because rural patients use the subsidy to travel to surgeons with higher hip and knee volume and thus higher exogenous quality. In fact, the average hip and knee volume of the surgeons that operate on rural patients increases by 8% with the subsidy, which is double that of the urban patients. For urban patients, however, the welfare effect is negative, since the gain in fiscal savings is substantially smaller, as shown in Appendix Figure E8, because these patients generate welfare increases primarily through the externality. Nevertheless, even the welfare effect for rural patients is more than an order of magnitude smaller than that of the minimum volume standard or the first-best policy.

Finally, I demonstrate that, as with the transportation subsidy, the returns to surgeon volume externality increase the welfare effect for both rural and urban patients when moving surgeons, but the final effect is negative for urban patients and positive for rural patients. Figure 5 demonstrates that incorporating the externality increases the welfare effect for both rural and urban patients because moving surgeons to both rural and urban areas increases the movers' volume and thus welfare. The figure also shows that the welfare effects are positive for rural patients and negative for urban patients. This difference occurs because the policy generates both larger increases in consumer surplus and larger increases in fiscal savings for rural patients, as shown in Appendix Figure E9. In particular, while moving surgeons increases the consumer surplus of the nearby patients regardless of if they are rural or urban, moving surgeons to rural shortage areas increases fiscal savings for rural patients, but moving surgeons to urban shortage areas decreases fiscal savings for urban patients. This result reflects the fact that moving surgeons are average quality, which is relatively better in rural markets but relatively worse in urban markets. Once again, it is important to highlight that this policy is the only one to increase consumer surplus, and especially so for rural patients, which may be an important objective for policymakers. Nevertheless, as with subsidizing transportation, even the welfare effect for rural patients is only 14% of that for the minimum volume standard and less than 2% of that for the first-best policy.

7 Conclusion

While people in larger markets are more productive, relatively little is known about what drives these patterns, despite the importance of understanding these mechanisms for designing policy. In this paper, I first show that market size matters in health care. Doubling the size of the market corresponds to a 9% reduction in the probability of a complication for patients following a hip or knee replacement. This descriptive pattern speaks to other settings where market sizes appears to drive outcomes or productivity, such as wages, innovation, mortality, and even restaurant quality (Glaeser and Maré, 2001; Bettencourt, Lobo, and Strumsky, 2007; Berry and Waldfogel, 2010; Deryugina and Molitor, 2021). Understanding whether selection or some agglomerative externality drives this pattern is crucial for designing policy.

Therefore, I investigate the role of one potential mechanism–the returns to surgeon volume–in generating these benefits to market size. Using the differential distance between patients and surgeons as an instrument, I causally identify the returns to surgeon volume. I attribute 22% of the benefits to market size to this mechanism. This returns to surgeon volume relationship highlights a policy-relevant externality–patients' choice of surgeon affects other patients' outcomes through the quality of the surgeon. It also yields new insight into the mechanisms underlying the benefits of agglomeration and thus provides a better understanding of cities (Duranton and Puga, 2004; Rosenthal and Strange, 2004; Moretti and Yi, 2024).

Incorporating this externality into a demand model, I evaluate the welfare consequences of a first-best policy and three feasible policies: implementing a minimum volume standard, subsidizing transportation, and moving surgeons to government-designated shortage areas. Failing to incorporate the returns to surgeon volume relationship substantially understates the welfare effects of the feasible policies. The minimum volume standard and the transportation subsidy concentrate care among surgeons with higher hip and knee volume,

thereby increasing increasing welfare through the externality. Meanwhile, moving surgeons dramatically increases the hip and knee volume of the moving surgeons, which thus increases welfare through the externality. Moving surgeons is the only policy to increase consumer surplus, and the increases are even larger for rural patients. A minimum volume standard, though, generates larger welfare increases than either of the other two feasible policies, both on average and for rural patients. Even the minimum volume standard, however, achieves only 7% of the gains from the first-best policy.

All together, these results suggest that more targeted policy with large price differences across surgeons may generate large welfare gains. In this setting, because of the returns to surgeon volume relationship, the welfare results depend on the extent to which the policy changes patient choice. Subsidizing patient transportation and moving surgeons does not induce sufficient changes in patient choice to substantially increase welfare. However, the minimum volume standard, which sets an infinite price for surgeons with low hip and knee volume, generates a large welfare effect, suggesting that large price changes, such as regulating entry, may be necessary to substantially change patient choice and thus welfare. Because the welfare gain from a first-best policy is substantially larger than that of the minimum volume standard, more targeted policies may yield substantial gains. For instance, since the social planner's optimal taxes target quality, insurers may want to target surgeon quality with large differences in in- and out-of-network prices.

More work is still needed in this area to better understand the mechanisms underlying the benefits of market size. One limitation of this paper is that I do not consider how these policies affect the supply-side. Thus, research examining how physicians make decisions on where to locate and what to specialize in is crucial to better evaluate the welfare consequences. Further research on other mechanisms, such as knowledge spillovers, that may generate benefits to market size may also provide better understanding of the benefits of large markets and of how to design policy. Likewise, empirical work documenting the drivers of differences in surgeon quality and the spatial variation in health outcomes, including those unrelated to market size, may be beneficial for designing policy to improve outcomes. Nevertheless, this paper provides insight into why people in larger markets are more productive and how to design policy in light of this result.

References

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters (2020). Do parents value school effectiveness? *American Economic Review*, 110(5):1502–39.

Agarwal, Nikhil and Paulo J Somaini (2022). Demand analysis under latent choice constraints. Working Paper 29993, National Bureau of Economic Research.

Albouy, David (2016). What are cities worth? land rents, local productivity, and the total value of amenities. *The Review of Economics and Statistics*, 98(3):477–487.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184.

Arrow, Kenneth J. (1962). The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173.

Askari, Alireza, Mehdi Mohammadpour, Mahmoud Jabalameli, Niloofar Naeimipoor, Babak Goodarzy, Behnam Jafari, Heeva Rashidi, Fatemeh Mousazadeh, Maziar Rajei, Amir Khazanchin, Mansour Bahardoust, and Mohammad Hassanzadeh (2024). Predictors of health-related quality of life after total knee arthroplasty: a case–control study. *Scientific Reports*, 14.

Avdic, Daniel, Petter Lundborg, and Johan Vikström (2019). Estimating returns to hospital volume: Evidence from advanced cancer surgery. *Journal of Health Economics*, 63:81–99.

Baicker, Katherine, Michael E. Chernew, and Jacob A. Robbins (2013). The spillover effects of medicare managed care: Medicare advantage and hospital utilization. *Journal of Health Economics*, 32(6):1289–1300.

Baker, Laurence C., M. Kate Bundorf, Anne B. Royalty, and Zachary Levin (2014). Physician practice competition and prices paid by private insurers for office visits. *JAMA*, 312(16):1653–1662.

Baum-Snow, Nathaniel and Daniel Hartley (2020). Accounting for central neighborhood change, 1980–2010. *Journal of Urban Economics*, 117:103228.

Baumgardner, James R. (1988). Physicians' services and the division of labor across local markets. *Journal of Political Economy*, 96(5):948–982.

Benkard, C. Lanier (2000). Learning and forgetting: The dynamics of aircraft production. *American Economic Review*, 90(4):1034–1054.

Berry, Christopher R. and Edward L. Glaeser (2005). The divergence of human capital levels across cities. *Papers in Regional Science*, 84(3):407–444.

Berry, Steven and Joel Waldfogel (2010). Product quality and market size. *The Journal of Industrial Economics*, 58(1):1–31.

Bettencourt, Luis M.A., José Lobo, and Deborah Strumsky (2007). Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy*, 36(1):107–120.

Birkmeyer, John D., Andrea E. Siewers, Emily V.A. Finlayson, Therese A. Stukel, F. Lee Lucas, Ida Batista, H. Gilbert Welch, and David E. Wennberg (2002). Hospital volume and surgical mortality in the united states. *New England Journal of Medicine*, 346(15):1128–1137.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001). Chapter 59 - measurement error in survey data. volume 5 of *Handbook of Econometrics*, pages 3705–3843. Elsevier.

Campbell, Jeffrey R. and Hugo A. Hopenhayn (2005). Market size matters. *The Journal of Industrial Economics*, 53(1):1–25.

Caplin, Andrew, Minjoon Lee, Søren Leth-Petersen, Johan Saeverud, and Matthew D Shapiro (2022). How worker productivity and wages grow with tenure and experience: The firm perspective. Working Paper 30342, National Bureau of Economic Research.

Card, David, Alessandra Fenizia, and David Silver (2023). The health impacts of hospital delivery practices. *American Economic Journal: Economic Policy*, 15(2):42–81.

Census (2019). Us zip codes to longitude and latitude (version 1.1).

Chan, David C. (2021). Influence and information in team decisions: Evidence from medical residency. *American Economic Journal: Economic Policy*, 13(1):106–37.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson (2016). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review*, 106(8):2110–44.

Chandra, Amitabh and Douglas O Staiger (2007). Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of political Economy*, 115(1):103–140.

Chen, Yiqun (2021). Team-specific human capital and team performance: Evidence from doctors. *American Economic Review*, 111:3923–3962.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

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

CMS (2022). 2022 procedure-specific complication measure updates and specifications report. Technical report, Centers for Medicare and Medicaid Services. Version 11.0, prepared by Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (YNHHSC/CORE).

Cornell, Portia Y., David C. Grabowski, Edward C. Norton, and Momotazur Rahman (2019). Do report cards predict future quality? the case of skilled nursing facilities. *Journal of Health Economics*, 66:208–221.

Couture, Victor and Jessie Handbury (2020). Urban revival in america. *Journal of Urban Economics*, 119:103267.

Cutler, David M., Robert S. Huckman, and Mary Beth Landrum (2004). The role of information in medical markets: An analysis of publicly reported outcomes in cardiac surgery. *American Economic Review*, 94(2):342–346.

Dartmouth Atlas (2025). Zip code crosswalks.

Deryugina, Tatyana and David Molitor (2021). The causal effects of place on health and longevity. *Journal of Economic Perspectives*, 35(4):147–70.

Diamond, Rebecca (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3):479–524.

Dinerstein, Michael, Rigissa Megalokonomou, and Constantine Yannelis (2022). Human capital depreciation and returns to experience. *American Economic Review*, 112(11):3725–62.

Ding, Peng (2021). The frisch–waugh–lovell theorem for standard errors. *Statistics Probability Letters*, 168:108945.

Dingel, Jonathan I, Joshua D Gottlieb, Maya Lozinski, and Pauline Mourot (2023). Market size and trade in medical services. *NBER Working Paper*, 31030.

Dolfen, Paul, Liran Einav, Peter J. Klenow, Benjamin Klopack, Jonathan D. Levin, Larry Levin, and Wayne Best (2023). Assessing the gains from e-commerce. *American Economic Journal: Macroeconomics*, 15(1):342–70.

Dubin, Jeffrey A. and Daniel L. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362.

Dummit, Laura A., Daver Kahvecioglu, Grecia Marrufo, Rahul Rajkumar, Jaclyn Marshall, Eleonora Tan, Matthew J. Press, Shannon Flood, L. Daniel Muldoon, Qian Gu, Andrea Hassol, David M. Bott, Amy Bassano, and Patrick H. Conway (2016). Association between hospital participation in a medicare bundled payment initiative and payments and quality outcomes for lower extremity joint replacement episodes. *JAMA*, 316(12):1267–1278.

Duranton, Gilles and Diego Puga (2004). Micro-foundations of urban agglomeration economies. In Henderson, J. V. and J. F. Thisse, editors, *Handbook of Regional and Urban Economics*, volume 4 of *Handbook of Regional and Urban Economics*, chapter 48, pages 2063–2117. Elsevier.

Einav, Liran, Amy Finkelstein, and Neale Mahoney (2022). Producing health: Measuring value added of nursing homes. Working Paper 30228, National Bureau of Economic Research.

Einav, Liran, Amy Finkelstein, and Heidi Williams (2016). Paying on the margin for medical care: Evidence from breast cancer treatments. *American Economic Journal: Economic Policy*, 8(1):52–79.

Finkelstein, Amy, Matthew Gentzkow, and Heidi L Williams (2019). Place-based drivers of mortality: Evidence from migration. Working Paper 25975, National Bureau of Economic Research.

Garicano, Luis and Thomas N. Hubbard (2007). Managerial leverage is limited by the extent of the market: Hierarchies, specialization, and the utilization of lawyers' human capital. *The Journal of Law and Economics*, 50(1):1–43.

Gaynor, Martin, Rodrigo Moreno-Serra, and Carol Propper (2013). Death by market power: Reform, competition, and patient outcomes in the national health service. *American Economic Journal: Economic Policy*, 5(4):134–66.

Gaynor, Martin, Harald Seider, and William B. Vogt (2005). The volume-outcome effect, scale economies, and learning-by-doing. *American Economic Review*, 95(2):243–247.

Ginja, Rita, Julie Riise, Barton Willage, and Alexander Willén (2022). Does Your Doctor Matter? Doctor Quality and Patient Outcomes. Working Papers 2022-016, Human Capital and Economic Opportunity Working Group.

Glaeser, Edward and David Maré (2001). Cities and skills. *Journal of Labor Economics*, 19(2):316–42.

Glaeser, Edward L. and Joshua D. Gottlieb (2009). The wealth of cities: Agglomeration economies and spatial equilibrium in the united states. *Journal of Economic Literature*, 47(4):983–1028.

Gottlieb, Joshua D, Maria Polyakova, Kevin Rinz, Hugh Shiplett, and Victoria Udalova (2023). Who values human capitalists' human capital? the earnings and labor supply of u.s. physicians. Working Paper 31469, National Bureau of Economic Research.

Grabowski, David C., Zhanlian Feng, Richard Hirth, Momotazur Rahman, and Vincent Mor (2013). Effect of nursing home ownership on the quality of post-acute care: An instrumental variables approach. *Journal of Health Economics*, 32(1):12–21.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Griliches, Zvi and Jerry A. Hausman (1986). Errors in variables in panel data. *Journal of Econometrics*, 31(1):93–118.

Guo, Zijian and Dylan Small (2016). Control function instrumental variable estimation of nonlinear causal effect models.

Halm, Ethan A., Clara Lee, and Mark R. Chassin (2002). Is volume related to outcome in health care? a systematic review and methodologic critique of the literature. *Annals of Internal Medicine*, 137(6):511–20.

Heckman, James J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, NBER Chapters, pages 475–492. National Bureau of Economic Research, Inc.

Hentschker, Corinna and Roman Mennicken (2018). The volume–outcome relationship revisited: Practice indeed makes perfect. *Health Services Research*, 53(1):15–34.

Irwin, Douglas A. and Peter J. Klenow (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political Economy*, 102(6):1200–1227.

Jacob, Brian A. and Lars Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101–136.

Karadakic, Rene, David C. Chan, Nancy L. Keating, Bruce E. Landon, and Michael L. Barnett (2025). The value of (sub) specialization: Evidence from oncology. Technical report.

Kessler, Daniel P., William M. Sage, and David J. Becker (2005). Impact of malpractice reforms on the supply of physician services. *JAMA*, 293(21):2618–2625.

Kim, Woojin, Sven Wolff, and Vivian Ho (2017). Measuring the volume–outcome relation for complex hospital surgery. *Applied Health Economics and Health Policy*, 15(4):537–538.

Kleiner, Samuel A., William D. White, and Sean Lyons (2015). Market power and provider consolidation in physician markets. *International Journal of Health Economics and Management*, 15(1):99–126.

Kline, Patrick and Andres Santos (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41.

Koehler, Tracy J., Jaclyn Goodfellow, Alan T. Davis, John E. vanSchagen, and Lori Schuh (2016). Physician retention in the same state as residency training: Data from 1 michigan gme institution. *Journal of Graduate Medical Education*, 8(4):518–522.

Kremers, Hilal Maradit, Dick R Larson, Cynthia S. Crowson, Walter K Kremers, Raynard E Washington, Claudia A Steiner, William A Jiranek, and Daniel J Berry (2015). Prevalence of total hip and knee replacement in the united states. *The Journal of Bone and Joint Surgery*, 97(17):1386–97.

Kugler, C. M., K. Goossen, T. Rombey, K. K. De Santis, T. Mathes, J. Breuing, S. Hess, R. Burchard, and D. Pieper (2022). Hospital volume–outcome relationship in total knee arthroplasty: a systematic review and dose–response meta-analysis. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(8):1795.

Layton, Timothy J. and Eran Politzer (2025). The dynamic fiscal costs of outsourcing health insurance - evidence from medicaid. *Journal of Public Economics*, 247:105417.

Lazear, Edward P., Kathryn L. Shaw, and Christopher T. Stanton (2015). The value of bosses. *Journal of Labor Economics*, 33(4):823–861.

Lee, Sanghoon (2010). Ability sorting and consumer city. *Journal of Urban Economics*, 68(1):20–33.

Leonardi, Marco and Enrico Moretti (2023). The agglomeration of urban amenities: Evidence from milan restaurants. *American Economic Review: Insights*, 5(2):141–57.

Levaillant, Mathieu, Romaric Marcilly, Lucie Levaillant, Philippe Michel, Jean-François Hamel-Broza, Benoit Vallet, and Antoine Lamer (2021). Assessing the hospital volume-outcome relationship in surgery: a scoping review. *BMC Medical Research Methodology*, 21:204.

Levitt, Steven D., John A. List, and Chad Syverson (2013). Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *Journal of Political Economy*, 121(4):643–681.

Luft, Harold S., John P. Bunker, and Alain C. Enthoven (1979). Should operations be regionalized? *New England Journal of Medicine*, 301(25):1364–1369. PMID: 503167.

Marshall, Alfred (1890). *The Principles of Economics*. McMaster University Archive for the History of Economic Thought.

(MedPAC), Medicare Payment Advisory Commission (2022). Chapter 3: Supporting safety-net providers.

Molloy, Ilda, Brook Martin, Wayne Moschetti, and David Jevsevar (2017). Effects of the length of stay on the cost of total knee and total hip arthroplasty from 2002 to 2013. *The Journal of Bone and Joint Surgery*, 99:402–407.

Moretti, Enrico (2013). Real wage inequality. *American Economic Journal: Applied Economics*, 5(1):65–103.

Moretti, Enrico (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–75.

Moretti, Enrico and Moises Yi (2024). Size matters: Matching externalities and the advantages of large labor markets. Working Paper 32250, National Bureau of Economic Research.

Morris, Carl N. (1983). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Mourot, Pauline (2024). Should top surgeons practice at top hospitals? sorting and complementarities in healthcare.

Nairn, Leah, Lauren Gyemi, Kyle Gouveia, Seper Ekhtiari, and Vickas Khanna (2021). The learning curve for the direct anterior total hip arthroplasty: a systematic review. *International Orthopaedics*, 45.

Oster, Emily (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business ; Economic Statistics*, 37(2):187–204.

Peng, Jing (2024). Machine learning for causal inference: Is a nonlinear first stage really forbidden in 2sls?

Pfuntner, A., L. M. Wier, and C. Stocks (2013). Most frequent procedures performed in u.s. hospitals, 2010. HCUP Statistical Brief 149, Agency for Healthcare Research and Quality (US), Rockville (MD).

Phibbs, Ciaran S., Laurence C. Baker, Aaron B. Caughey, Beate Danielsen, Susan K. Schmitt, and Roderic H. Phibbs (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *New England Journal of Medicine*, 356(21):2165–2175. PMID: 17522400.

Pischke, Jörn-Steffen (2007). Lecture notes on measurement error. Lecture notes for EC524: Empirical Methods in Applied Economics, London School of Economics.

Porter, Giselle M., Jeffrey Balian, Ayesha P. Ng, Hugo Mannings, Devon M. Jeffcoat, and Peyman Benharash (2025). Cost-volume analysis of primary total knee and hip arthroplasty in the united states. *The Journal of Arthroplasty*, 40(9):2259–2267.e1.

Redding, Stephen J. (2023). Quantitative urban models: From theory to data. *Journal of Economic Perspectives*, 37(2):75–98.

Redding, Stephen J. (2025). Chapter 2 - quantitative urban economics. In Donaldson, Dave and Stephen J. Redding, editors, *Handbook of Regional and Urban Economics*, volume 6 of *Handbook of Regional and Urban Economics*, pages 73–141. Elsevier.

Rosenberg, Adam (2025). Regulating firearm markets: Evidence from california.

Rosenthal, Stuart S. and William C. Strange (2004). Chapter 49 - evidence on the nature and sources of agglomeration economies. In Henderson, J. Vernon and Jacques-François Thisse, editors, *Cities and Geography*, volume 4 of *Handbook of Regional and Urban Economics*, pages 2119–2171. Elsevier.

Rubinton, Hannah (2025). The geography of business dynamism and skill-biased technical change. *The Review of Economic Studies*, page rdaf063.

Räsänen, Pirjo, Pekka Paavolainen, Harri Sintonen, Anna-Maija Koivisto, Marja Blom, Olli-Pekka Ryynänen, and Risto P Roine (2007). Effectiveness of hip or knee replacement surgery in terms of quality-adjusted life years and costs. *Acta Orthopaedica*, 78(1):108–115.

Shen, Yunrong, Xin Hu, Ryan D. Nipp, K. Robin Yabroff, Arthur S. Hong, Joshua M. Liao, and Changchuan Jiang (2024). Nonemergency medical transportation benefit in traditional medicare advantage and value-based plans. *JAMA Network Open*, 7(12):e2449038–e2449038.

Sivey, Peter (2012). The effect of waiting time and distance on hospital choice for english cataract patients. *Health Economics*, 21(4):444–456.

Sloan, Matthew, Ajay Premkumar, and Neil P Sheth (2018). Projected volume of primary total joint arthroplasty in the u.s., 2014 to 2030. *The Journal of Bone and Joint Surgery*, 100(17):1455–1460.

Small, Kenneth A. and Harvey S. Rosen (1981). Applied welfare economics with discrete choice models. *Econometrica*, 49(1):105–130.

Stanford Center for Population Health Sciences (2021a). Medicare 20% [2006-2018] carrier, carrier base claim file.

Stanford Center for Population Health Sciences (2021b). Medicare 20% [2006-2018] medpar, medpar.

Stanford Center for Population Health Sciences (2021c). Medicare 20% [2006-2018] outpatient, outpatient base claim file.

Su, Yichen (2022). The rising value of time and the origin of urban gentrification. *American Economic Journal: Economic Policy*, 14(1):402–39.

Syverson, Chad (2004). Market structure and productivity: A concrete example. *Journal of Political Economy*, 112(6):1181–1222.

United States Department of Housing and Urban Development (2019). Us zip codes to county (version 1.0).

Yi, Sarah H., James Baggs, Steven D. Culler, Sandra I. Berríos-Torres, and John A. Jernigan (2015). Medicare reimbursement attributable to periprosthetic joint infection following primary hip and knee arthroplasty. *The Journal of Arthroplasty*, 30(6):931–938.e2.

Appendices

Table of contents

A Context, data, and measurement appendix

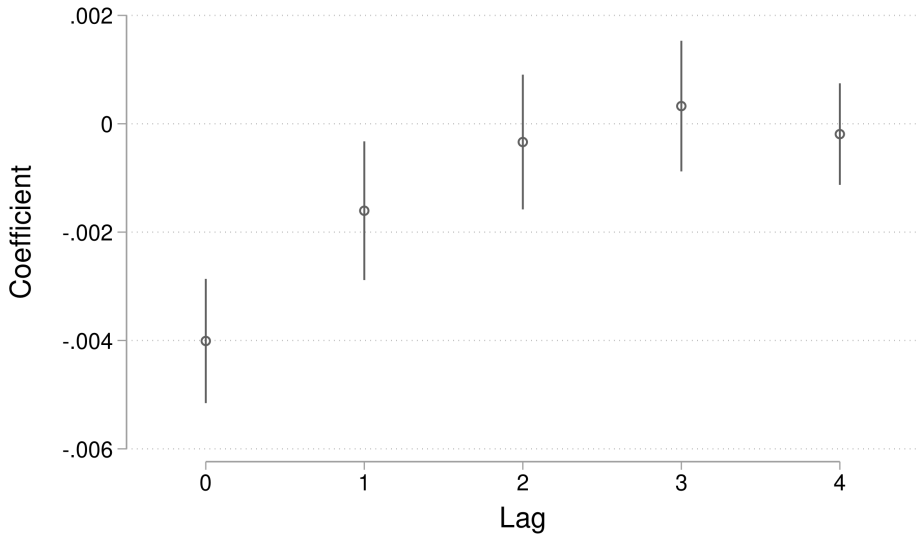


Figure A1: Relationship between probability of a risk-adjusted complication and lags of surgeon hip and knee volume

Notes: This figure shows the relationship between the probability of a risk-adjusted complication and lagged measures of a surgeon's hip and knee volume. Complications are adjusted for year fixed effects and the patient covariates listed in Appendix Table A2. The lags are defined in 365 day intervals, such that the zero lag corresponds the surgeon's hip and knee volume up to 365 days prior to the surgery date, for instance. The unit of observation is a patient.

Complication	Measurement period (days from surgery)	% complications
Periprosthetic joint infection/wound infection or other wound complication	90	32.4
Pneumonia or other acute respiratory complication	7	19.4
Pulmonary embolism	30	16.1
Mechanical complication	90	12.4
Death	30	7.7
Acute myocardial infarction	7	6.4
Sepsis/septicemia/shock	7	3.4
Surgical site bleeding or other surgical site complication	30	2.3

Table A1: Complication types for hip and knee replacement

Notes: This table shows each of the eight complication types used to define a complication in my sample, following CMS (2022). The second column shows the period over which this complication is measured, relative to the surgery date. The third column shows the share of total complications that each complication type comprises in my sample.

- Age
- Joint (hip or knee)
- Dually eligible for Medicaid
- Other congenital deformity of hip
- Metastatic cancer and acute leukemia
- Respiratory / heart / digestive / urinary / other neoplasms
- Protein-calorie malnutrition
- Bone / joint / muscle infections / necrosis
- Osteoarthritis of hip or knee
- Dementia or other specified brain disorders
- Hemiplegia, paraplegia, paralysis, functional disability
- Coronary atherosclerosis or angina
- Vascular or circulatory disease
- Pneumonia
- Dialysis status
- Decubitus ulcer or chronic skin ulcer
- Vertebral fractures without spinal cord injury
- Major complications of medical care and trauma
- Gender
- White race
- Multiple procedures
- Post-traumatic osteoarthritis
- Other major cancers
- Diabetes mellitus (DM) or DM complications
- Morbid obesity
- Rheumatoid arthritis and inflammatory connective tissue disease
- Osteoporosis and other bone / cartilage disorders
- Major psychiatric disorders
- Cardio-respiratory failure and shock
- Stroke
- Chronic obstructive pulmonary disease
- Pleural effusion / pneumothorax
- Renal failure
- Trauma
- Other injuries

Table A2: List of patient characteristics and comorbidities

Notes: This table shows all of the patient covariates used to adjust for complications, following CMS (2022). These patient risk covariates are obtained from Medicare claims extending up to twelve months prior to the index admission and include the index admission itself. A binary indicator for whether a patient is white and a binary indicator for if a patient is dually eligible for Medicaid have been added to the covariate list from CMS (2022) specifically for this analysis.

B Descriptive evidence appendix

Outcome	(1) Complication	(2) Complication	(3) Complication	(4) Complication	(5) Complication
Log market size	-0.0023	-0.0031	-0.0033	-0.0024	-0.0011
	(0.0007)	(0.0006)	(0.0007)	(0.0009)	(0.0004)
N	1,967	1,967	1,967	1,961	1,967
Mean comp.	0.035	0.035	0.035	0.035	0.012
Risk covariates	X	✓	✓	✓	✓
HRR FEs	X	X	✓	X	X
Home market procedure	X	X	X	✓	X
Primary stay complication	X	X	X	X	✓

Table B1: Relationship between complication rate and market size

Notes: This table shows the results from a regression of the complication rate in an HSA on the market size in that HSA averaged over the sample period. The unit of observation is an HSA. Column (1) does not include patient risk covariates. Column (2) adds these risk covariates, as shown in Appendix Table A2. Column (3) adds HRR fixed effects. Column (4) only includes patients who receive care in the HSA in which they reside. The observation count falls here because there are a small number of HSAs where patients receive care but no patients live. Column (5) includes only complications that are diagnosed on the same stay-level record as the hip or knee replacement, implying that they occur within the same inpatient stay. Standard errors are clustered at the HSA level.

Outcome	(1) Complication	(2) Complication	(3) Complication	(4) Complication
Log volume	-0.0066	-0.0054	-0.0044	-0.0018
	(0.0003)	(0.0003)	(0.0003)	(0.0005)
N	688,121	688,121	688,121	688,121
Mean comp.	0.031	0.031	0.031	0.031
Year FEs	✓	✓	✓	✓
Risk covariates	X	✓	✓	✓
Hospital FEs	X	X	✓	X
Surgeon FEs	X	X	X	✓

Table B2: Relationship between probability of a complication and surgeon log hip and knee volume

Notes: This table shows the results from a estimating a linear probability model, where the outcome is a binary indicator equal to one if a patient experiences a complication following the hip or knee replacement and the main independent variable is a surgeon's log hip and knee volume in the 365 days prior to the patient's surgery. Each column includes year fixed effects. Column (2) introduces all the patient covariates shown in Appendix Table A2. Column (3) introduces hospital fixed effects. Column (4) includes surgeon fixed effects but not hospital fixed effects. The units of observation is a patient. Standard errors are clustered at the surgeon level.

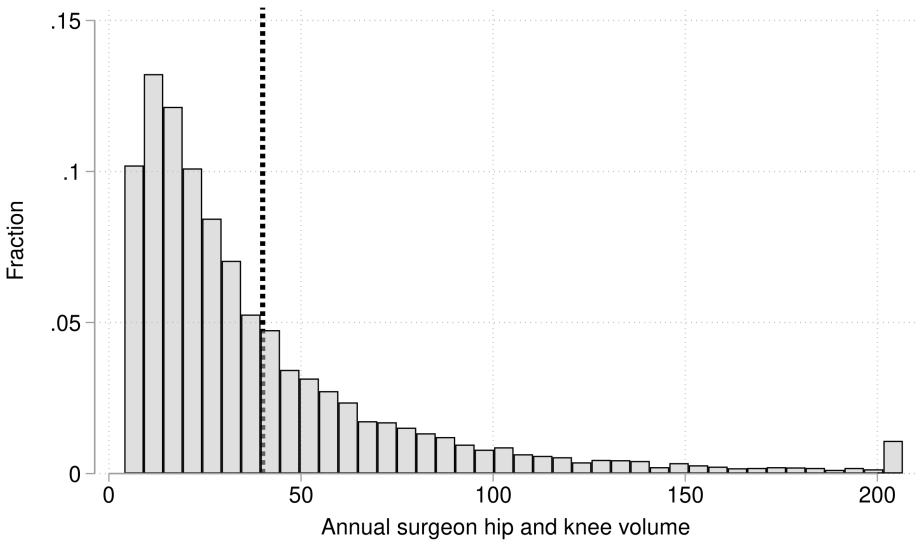


Figure B1: Histogram of surgeon hip and knee volume

Notes: This figure shows the histogram of average surgeon hip and knee volume in the main sample. The unit of observation is a surgeon, and volume is winsorized at the 1% level. The dotted vertical line denotes the median.

C Model and empirical strategy appendix

C.1 Empirical Bayes procedure

Measurement error in the estimated demand parameters poses challenges for both the empirical strategy and the counterfactual policy evaluations, so I use Empirical Bayes shrinkage to address this issue. Recall the demand model in Equation 1. Measurement error in the demand for exogenous surgeon quality, δ_{0j}, may attenuate the coefficient κ in the second-stage regression specification in Equation 8, which in turn could cause an upward bias in β_1 if δ_{0j} is improperly controlling for the demand for exogenous surgeon quality. Additionally, because I estimate the demand coefficients, τ and δ_1, within a geographic market, a smaller number of patients within a market may generate more extreme values for these parameters. Finally, the estimated surgeon fixed effects for complications that I use to conduct the counterfactual evaluations suffer from these same measurement error issues. To address these challenges, I adjust these parameters using Empirical Bayes shrinkage. This procedure reduces the mean-squared error and, crucially for the empirical strategy, eliminates attenuation bias in models using $\widehat{\delta}_{0j}$ as a regressor (Morris, 1983; Jacob and Lefgren, 2008).

To shrink the estimates to the mean, I adjust the parameters according to their standard errors following Morris (1983). For ease of exposition, I discuss the Empirical Bayes shrinkage procedure for estimating δ_{0j}, but the same logic applies to the coefficients, τ and δ_1, as well as to the surgeon fixed effects for complications. Formally, I assume that the estimated fixed effects are estimated with error:

$$\widehat{\delta}_{0j} = \delta_{0j} + e_j,$$

where δ_{0j} is the "true" fixed effect and e_j is the measurement error of the estimated fixed effect. Assume that the errors are independent and drawn from a normal distribution with variance χ_j^2, such that $e_j \sim N\left(0, \chi_j^2\right)$. Hence, the distribution of the estimated fixed effect conditional on the true effect and the measurement error variance is:

$$\widehat{\delta}_{0j}|\delta_{0j}, \chi_j^2 \sim N\left(\delta_{0j}, \chi_j^2\right)$$

Now, if we assume a prior for the true effect and use Bayes rules, we can derive the shrunk estimate for the parameters as the mean of the posterior. Namely, assume a prior distribution for the true effect such that:

$$\delta_{0j}|\sigma_j^2 \sim N\left(\rho, \sigma^2\right),$$

where ρ is the underlying mean and σ^2 is the variance of the true fixed effect. From Bayes' rule, we have that:

$$\delta_{0j}|\rho, \sigma^2, \chi_j^2, \widehat{\delta}_{0j}\ N\left(b_j\widehat{\delta}_{0j} + (1-b_j)\rho, b_j\chi_j^2\right),$$

where $b_j = \frac{\sigma^2}{\chi_j^2+\sigma^2}$. The empirical Bayes-adjusted fixed effects correspond to the mean of the posterior, implying

that:

$$\delta_{0j}^{EB} = b_j \widehat{\delta}_{0j} + \left(1 - b_j\right) \rho.$$

This formulation highlight how the procedure works–the larger the variance of the measurement error, χ_j^2, the more weight is given to the underlying mean, ρ. More simply, noisier estimates are shrunk toward the mean.

Empirically, I use moments from the data and plug them into this formula to shrink these Empirical Bayes parameters to the mean. From estimating the logit, I have obtained $\widehat{\delta}_{0j}$, as well as χ_j^2, which is simply the square of the standard error on the fixed effect. I cluster these standard errors at the patient-level to allow for non-independent errors within a patient's different alternatives. I shrink these estimates to the within-market mean, so ρ is simply the mean of the $\widehat{\delta}_{0j}$'s. Finally, by the law of total variance, I calculate the variance of the true fixed effect as $\sigma^2 = Var\left(\widehat{\delta}_{0j}\right) - \mathbb{E}\left[\chi_j^2\right]$, where $Var\left(\widehat{\delta}_{0j}\right)$ is the variance of the fixed effects and $\mathbb{E}\left[\chi_j^2\right]$ is the mean of the individual fixed effect variances. I perform a similar calculation for τ and for δ_1, except across markets. Similarly, for the surgeon fixed effects for complications, I perform the same computation except nationally, as opposed to within a geographic market.

C.2 Deriving the control function

Recall the conditional expectation of the estimating equation error term, ϵ_{ij} from Equation 3 is:

$$\mathbb{E}\left[\epsilon_{ij}|v_{ij}, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, \eta_{i1}, \cdots, \eta_{iJ}, D_i = j\right] = \kappa\widehat{\delta}_{0j} + \sum_k \phi_k \tilde{\eta}_{ik} + \varphi\tilde{\eta}_{ij},$$

where v_{ij} is surgeon j's volume (plus one) in the 365 days prior to patient i's surgery date, X_i are the patient risk covariates listed in Appendix Table A2, $\gamma_{t(i)}$ are year fixed effects, $\widehat{\delta}_{0j}$ is the estimated demand for exogenous surgeon quality from the demand model in Equation 1, $\tilde{\eta}_{ij} = \eta_{ij} - \mu_\eta$ are the mean-zero logit shocks from the demand model, J is the total number of surgeons, and D_i denotes patient i's chosen surgeon. We want to show that the expected values of the unobserved logit shocks, $\eta_{i1}, \cdots, \eta_{iJ}$ conditional on the choice, yields the following conditional expectation function:

$$\mathbb{E}\left[\epsilon_{ij}|v_{ij}, X_i, \gamma_{t(i)}, \widehat{\delta}_{0j}, d_i, D_i = j\right] = \kappa\widehat{\delta}_{0j} + \sum_k \phi_k \theta_{ik}\left(j, d_i\right) + \varphi\theta_{ij}\left(j, d_i\right),$$

where $d_i = \left(d_{i1}, \cdots, d_{iJ}\right)$ is the vector of distances between patient i and each surgeon that functions as the excluded instrument and $\theta_{ik}(j, d_i)$ are simply functions of the logit choice probabilities from the demand model:

$$\theta_{ik}(j) = \begin{cases} -\ln \widehat{p}_{ik}\left(d_i\right), & k = j \\ \frac{\widehat{p}_{ik}(d_i)}{1-\widehat{p}_{ik}(d_i)} \ln \widehat{p}_{ik}\left(d_i\right), & \text{otherwise} \end{cases},$$

Here, $\widehat{p}_{ij}(d_i)$ is the predicted probability that patient i chooses surgeon j as a function of the differential distance instrument. Because $\widehat{\delta}_{0j}$ is not a random variable, I focus on deriving the latter two terms. I derive the cases when $k = j$ and when $k \neq j$ separately.

I first begin this derivation by showing two crucial relationships. First, note that:

$$\mathbb{E}\left[\tilde{\eta}_{ij}|D_i = j\right] = \mathbb{E}\left[u_{ij}|D_i = j\right] - \delta_{0j} - \delta_1 v_{ij} + \tau \ln d_{ij} - \mu_\eta \tag{C1}$$

Additionally, using Small and Rosen (1981), we have:

$$\mathbb{E}\left[u_{ij}|D_i = j\right] = \ln\left[\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right)\right] + \mu_\eta \tag{C2}$$

Now, we can solve for the control function for the chosen alternative when $k = j$. Substituting Equation

C2 into Equation C1 yields:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\eta}_{ij}|D_i = j\right] &= \ln\left[\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)\right] + \mu_\eta - \delta_{0j} - \delta_1 v_{ij} + \tau\ln d_{ij} - \mu_\eta \\
&= \ln\left[\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)\right] - \delta_{0j} - \delta_1 v_{ij} + \tau\ln d_{ij} \\
&= \ln\left[\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)\right] - \ln\left[\exp\left(\delta_{0j} - \delta_1 v_{ij} + \tau\ln d_{ij}\right)\right] \\
&= \ln\left[\frac{\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)}{\exp\left(\delta_{0j} - \delta_1 v_{ij} + \tau\ln d_{ij}\right)}\right] \\
&= -\ln\left[\frac{\exp\left(\delta_{0j} - \delta_1 v_{ij} + \tau\ln d_{ij}\right)}{\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)}\right] \\
&= -\ln\widehat{p}_{ij}\left(d_i\right)
\end{aligned}$$

We can also solve for the control function for a non-chosen alternative, such that $k \neq j$. Re-writing Equation C1 for a non-chosen alternative yields:

$$\mathbb{E}\left[\tilde{\eta}_{ik}|D_i = j\right] = \mathbb{E}\left[u_{ik}|D_i = j\right] - \delta_{0k} - \delta_1 v_{ik} + \tau\ln d_{ik} - \mu_\eta \tag{C3}$$

Additionally, by the properties of conditional expectation, we have that:

$$\mathbb{E}\left[u_{ik}\right] = \mathbb{P}\left(D_i = k\right)\mathbb{E}\left[u_{ik}|D_i = k\right] + \mathbb{P}\left(c_i \neq k\right)\mathbb{E}\left[u_{ik}|D_i \neq k\right]$$

Then, substituting the relationship in Equation C2 into this relationship yields:

$$\delta_{0k} + \delta_1 v_{ik} - \tau\ln d_{ik} + \mu_\eta = \mathbb{P}\left(D_i = k\right)\left[\ln\left[\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)\right] + \mu_\eta\right] + \mathbb{P}\left(D_i \neq k\right)\mathbb{E}\left[u_{ik}|D_i \neq k\right].$$

Therefore, we have that:

$$\begin{aligned}
\mathbb{E}\left[u_{ik}|D_i \neq k\right] = \frac{1}{1 - \mathbb{P}\left(D_i = k\right)}\Bigg(&\delta_{0k} + \delta_1 v_{ik} - \tau\ln d_{ik} + \mu_\eta \\
&- \mathbb{P}\left(D_i = k\right)\left[\ln\left[\sum_{s=1}^{J}\exp\left(\delta_{0s} + \delta_1 v_{is} - \tau\ln d_{is}\right)\right] + \mu_\eta\right]\Bigg)
\end{aligned} \tag{C4}$$

We can then substitute in these two expressions to solve for the control function term for a non-chosen alternative. Specifically, from Equation , we know the conditional expectation of u_{ik} conditional on the choice

not being k. Supposing that the choice is j, we can then plug Equation C.2 into Equation C3:

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\eta}_{ik} | D_i = j\right] = & \frac{1}{1 - \mathbb{P}\left(D_i = k\right)} \Bigg(\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik} + \mu_\eta \\
& - \mathbb{P}\left(D_i = k\right) \left[\ln \left[\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right) \right] + \mu_\eta \right] \Bigg) - \delta_{0k} - \delta_1 v_{ik} + \tau \ln d_{ik} - \mu_\eta \\
= & \frac{1}{1 - \mathbb{P}\left(D_i = k\right)} \Bigg(\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik} + \mu_\eta - \left(1 - \mathbb{P}\left(D_i = k\right)\right) \left[\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik} + \mu_\eta\right] \\
& - \mathbb{P}\left(D_i = k\right) \left[\ln \left[\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right) \right] + \mu_\eta \right] \Bigg) \\
= & \frac{\mathbb{P}\left(D_i = k\right)}{1 - \mathbb{P}\left(D - i = k\right)} \left(\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik} - \ln \left[\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right) \right] \right) \\
= & \frac{\mathbb{P}\left(D_i = k\right)}{1 - \mathbb{P}\left(D - i = k\right)} \left(\ln \left[\exp\left(\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik}\right)\right] - \ln \left[\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right) \right] \right) \\
= & \frac{\mathbb{P}\left(D_i = k\right)}{1 - \mathbb{P}\left(D - i = k\right)} \left(\ln \frac{\exp\left(\delta_{0k} + \delta_1 v_{ik} - \tau \ln d_{ik}\right)}{\sum_{s=1}^{J} \exp\left(\delta_{0s} + \delta_1 v_{is} - \tau \ln d_{is}\right)} \right) \\
= & \frac{\widehat{p}_{ik}\left(d_i\right)}{1 - \widehat{p}_{ik}\left(d_i\right)} \ln \widehat{p}_{ik}\left(d_i\right),
\end{aligned}
$$

where $\mathbb{P}\left(D_i = k\right)\left(d_i\right) = \widehat{p}_{ik}(d_i)$ by definition.

C.3 Estimation procedure

Directly estimating the main empirical specification in Equation 8 is computationally burdensome, so I leverage the Frisch-Waugh-Lovell theorem to estimate the parameters and bootstrap to obtain standard errors. Because Equation 8 has a parameter for each surgeon in the dataset, of which there are more than 11,000, I cannot estimate it quickly via OLS. I therefore exploit the fact that the θ_{ik}'s are zero for surgeons outside a patient's HRR and use the Frisch-Waugh-Lovell theorem to obtain unbiased estimates of the parameters. Namely, I partition the parameters into the local parameters, which are the θ_{ik}'s, and the national parameters, which are all the other parameters. I first residualize out the local parameters from the national parameters within a market. I then aggregate these residualized national parameters to run one national-level regression and obtain coefficients on all national parameters. To recover the local parameters, I estimate the full estimating equation at the market-level holding fixed the national parameters at their true values.

C.4 Measurement error

Because there is measurement error in the key explanatory variable, volume, the estimate of the returns to surgeon volume may be biased. Measurement error comes from two sources. The first is classical measurement error and reflects the fact that I only observe a 20% random sample of Medicare FFS volume. The second may be classical or non-classical and reflects the fact that I only observe Medicare FFS volume, not all-payer volume, including Medicare Advantage volume. Using properties of random sampling and a validation dataset from the state of Florida, I show that this measurement error attenuates the estimated returns to surgeon volume parameter, such that the parameter I estimate is only 59% of the true value.

C.4.1 Measurement error due to random sampling

Because I only observe a random 20% sampling of Medicare FFS claims, the estimated returns to surgeon volume parameter is attenuated. Since this measurement error is classical, I compute a correction factor to purge the estimate of this attenuation using a classical errors-in-variable model and the properties of random sampling (Pischke, 2007).

Using a classic error-in-variables formula, I compute the correction factor needed to adjust for the attenuation bias in measured volume due to random sampling. Let $v_{obs} = \ln(N_{obs} + 1)$ denote the log Medicare FFS volume (plus one) that I observe, $v_{FFS} = \ln(N_{FFS} + 1)$ denote true log Medicare FFS volume (plus one), and e denote the measurement error from the random sampling. Following the classical errors-in-variables model, I assume that the error enter additively (in logs), implying $v_{obs} = v_{FFS} + e$. Then, following the derivations in Pischke (2007) with classical measurement error, the correction factor, or reliability ratio, from random sampling c_{samp} is given by:

$$c_{samp} = \frac{\sigma^2_{v_{FFS}}}{\sigma^2_{v_{FFS}} + \sigma^2_e}$$

where $\sigma^2_{v_{FFS}}$ is the variance of v_{FFS} and σ^2_e is the variance of e. This correction factor is the amount by which the estimate is attenuated and thus can be used to recover the unattenuated estimate. Theoretically, c_{samp} would correct for the attenuation bias. However, since v_{FFS} is not observed, I cannot compute $\sigma^2_{v_{FFS}}$. Thus, I use properties of the measurement error to reformulate it to something I can compute. Because the error term enters additively in logs and since the covariance between the true variable and the measurement error is zero due to random sampling, the variance of the measured variable, v_{obs}, is just the sum of the variance of the true variable v_{FFS} and the variance of the error term, e: $\sigma^2_{v_{obs}} = \sigma^2_{v_{FFS}} + \sigma^2_e$. Hence, the correction factor due to random sampling, c_{samp} can be re-expressed as:

$$c_{samp} = 1 - \frac{\sigma^2_e}{\sigma^2_{v_{obs}}}.$$

I compute the variance components of this correction factor using properties of the random sampling and the data and show that the returns to surgeon volume is 69% of the true value due to random sampling. Computing $\sigma^2_{v_{obs}}$ is simple, as it is just the variance of the residualized log-volume. To compute σ^2_e, note that by the law of total variance $\sigma^2_e = \mathbb{E}\left[Var\left[e|v_{FFS}\right]\right] + Var\left[\mathbb{E}\left[e|v_{FFS}\right]\right] = \mathbb{E}\left[Var\left[e|v_{FFS}\right]\right]$, since $\mathbb{E}\left[e|v_{FFS}\right] = 0$. Then, using the definition of the classical measurement error, note that $Var\left[e|v_{FFS}\right] = Var\left[v_{obs} - v_{FFS}|v_{FFS}\right] = Var\left[v_{obs}|v_{FFS}\right]$. Next, because the sample I use is a 20% random sample, I can model the distribution of N conditional on N_{FFS} as a binomial distribution with probability of success $p = 0.20$, which implies that $N_{obs}|N_{FFS} \sim binomial(N_{FFS}, p)$. Recalling that v_{obs} is the log of volume plus one, we can use the known variance of a binomial distribution and the delta method to calculate the expected value of this variance:

$$\begin{aligned}
\mathbb{E}\left[Var\left[v_{obs}|v_{FFS}\right]\right] &= \mathbb{E}\left[Var\left[\ln(1+N_{obs})|N_{FFS}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_{FFS}\times p+1}\right)^2 \times \left[N_{FFS}\times p\times(1-p)\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\frac{N_{obs}}{p}\times p+1}\right)^2 \times \left[\frac{N_{obs}}{p}\times p\times(1-p)\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_{obs}+1}\right)^2 \times \left[N_{obs}\times(1-p)\right]\right].
\end{aligned}$$

The second equality follows from both using the delta method where $f(N_{FFS}) = \ln(N_{FFS}+1)$ and the known variance of a binomial distribution. Meanwhile, the third equality follows from approximating N_{FFS} with $\frac{N_{obs}}{p}$. Computing this final expression is relatively easy, since I can just take the mean of this function of the observed volume in my sample. Using this formula to calculate σ^2_e, combined with the calculation of $\sigma^2_{v_{obs}}$, implies that $c_{samp} = 0.69$, such that the returns to surgeon volume estimate is 69% of the true value due to attenuation bias from random sampling.

C.4.2 Measurement error due to payer sampling

Because I only observe a surgeon's Medicare FFS volume and not their other payer volume, there may be additional bias in the estimated returns to surgeon volume parameter. Since this bias may be classical or non-classical, I solve for the resulting bias allowing for correlations between the measurement error and the true value using a validation dataset from the state of Florida.

To calculate the correction factor from this measurement error, I introduce and interpret the formula for non-classical measurement error. Once again, let v_{FFS} denote a surgeon's Medicare FFS log volume (plus one). Note, however, that now I treat v_{FFS} as observed, since it is in the validation dataset. Additionally, let v^* denote log all-payer volume (plus one), and let u denote the measurement error from only observing Medicare FFS as a payer. Following the derivations in Pischke (2007) for non-classical measurement error, we have that the

correction factor due to only observing Medicare as a payer, c_{payer} is:

$$c_{payer} = \frac{\sigma^2_{v^*} + \sigma_{v^*u}}{\sigma^2_{v_{FFS}}},$$

where $\sigma^2_{v^*}$ is the variance of the true volume, σ_{v^*u} is the covariance between the true volume and the measurement error, and $\sigma^2_{v_{FFS}}$ is the variance of the Medicare FFS volume. This formula is simply an alternate version of the omitted variable bias, where non-Medicare FFS volume is the omitted variable that may potentially be correlated with levels of the Medicare FFS volume. Thus, another way to compute this correction factor is by regressing the error, u, on the Medicare FFS volume, v_{FFS}. This formula also nests the formula with classical measurement error in which $\sigma_{v^*u} = 0$. Thus, the magnitude of $\sigma_{v^*u} = 0$ governs the influence of non-classical measurement error on the bias.

I compute the correction term due to only observing Medicare FFS as a payer using a validation dataset from the state of Florida. Each of these objects is straightforward to calculate in the data. While this validation exercise provides significant benefits for this calculation, there are two important caveats with this exercise. First, because I do not observe all the controls in this validation dataset, I cannot residualize them out, so I assume the unconditional moments are good proxies for the conditional moments. Second, I assume that the relationship between Medicare FFS and all-payer volume in Florida generalizes to other states. The final calculation yields $c_{payer} = 0.85$, indicating that the returns to surgeon volume estimate is 85% of the true value due to bias from only measuring Medicare FFS volume. Had I assumed the measurement error was instead classical and omitted the σ_{v^*u} term, c_{payer} would be 0.87, indicating that the non-classical measurement error introduces further attenuation, as the measurement error is larger for higher (all-payer) volume surgeons. However, the magnitude of this non-classical measurement error is quite small, as evidenced in the small difference between the two correction factors with and without σ_{v^*u}.

C.4.3 Combining both measurement errors together

Since the measurement error from random sampling and from only observing Medicare FFS volume are independent, the bias is multiplicative in the correction terms. That is, the final correction term is $c = c_{samp} \times c_{payer} = 0.59$, implying that the estimated returns to surgeon volume parameter is 59% of the underlying structural parameter.

C.5 Computing standard errors

Calculating standard errors in this setting is difficult since the first-stage is a logit model that is computationally burdensome to estimate. To address this challenge, I use a two-step score bootstrap procedure originally developed in Kline and Santos (2012) and used in other similar settings (Abdulkadiroğlu et al., 2020; Einav, Finkelstein, and Mahoney, 2022). This procedure adjusts inference for the extra uncertainty introduced by the demand model estimates without recalculating the first-stage estimates or analytically deriving the influence of the first-stage estimates on the second-stage estimates. Intuitively, it yields perturbed values of the estimates by randomly weighting observations.

I first introduce simplified notation to help ease the exposition of this procedure. Using the demand model in Equation 1, I define Δ as the vector of market-specific parameters estimated via maximum likelihood from this first-stage:

$$\Delta = \left(\delta_{0j}, \cdots, \delta_{0|S_i|}, \delta_1, \tau\right),$$

where $|S_i|$ denotes the cardinality of the set of surgeons within a market, Similarly, for the second-stage, define Γ as the vector of parameters from Equation 8:

$$\Gamma = \left(\beta_0, \beta_1, \alpha, \gamma_{t(i)}, \kappa, \phi_1, \cdots, \phi_{|S_i|}, \varphi, \mu_m\right)$$

To generate bootstrap values of these parameters, I first perturb the first-stage estimates using random weights. Specifically, I generate a bootstrap distribution by taking repeated Newton-Raphson steps from the full-sample estimates and randomly reweighting each observation's score contribution. The first-stage bootstrap estimate of Γ in bootstrap replication b is:

$$\widehat{\Delta}^b = \widehat{\Delta} - H_1^{-1}\left(\widehat{\Delta}\right) \times \sum_{i=1}^{N} w_i^b s_1\left(i, \widehat{\Delta}\right),$$

where $H_1\left(\widehat{\Delta}\right)$ is the first-stage Hessian, N is the number of patients, $s_1\left(i, \widehat{\Delta}\right)$ is the first-stage score contribution of patient i at the optimal parameter vector $\widehat{\Delta}$, and $\{w_i^b\} \sim N(0,1)$ is a set of *i.i.d.* standard normal weights.

Then, using the same random weights and the perturbed first-stage values, I generate perturbed values for the second-stage bootstrap estimates. Namely, using an additional set of Newton-Raphson steps, I define a second-stage bootstrap estimate Γ in bootstrap replication b as:

$$\widehat{\Gamma}^b = \widehat{\Gamma} - H_2^{-1}\left(\widehat{\Gamma}; \widehat{\Delta}\right) \times \sum_{i=1}^{N} \left[w_i^b s_2\left(i, \widehat{\Gamma}; \widehat{\Delta}\right) - s_2\left(i, \widehat{\Gamma}; \widehat{\Delta}^b\right)\right], \tag{C5}$$

where where $H_2\left(\widehat{\Gamma}; \widehat{\Delta}\right)$ is the second-stage Hessian and $s_2\left(i, \widehat{\Gamma}; \widehat{\Delta}\right)$ is the second-stage score contribution of patient i at the two-stage optima $\left(\widehat{\Gamma}; \widehat{\Delta}\right)$. Intuitively, the term $s_2\left(i, \widehat{\Gamma}; \widehat{\Delta}^b\right)$ plus in the bootstrap estimate $\widehat{\Gamma}^b$ from Equation C5 into the second-stage score contribution function $s_2\left(i, \widehat{\Gamma}, \cdot\right)$ to account for the additional variability

in the second-stage score due to the first-stage estimate $\widehat{\Gamma}$.

To estimate these objects, I use standard econometric identities to exploit the fact that most of these objects are already estimated. Namely, I compute $H_1\left(\widehat{\Delta}\right)$ as the variance-covariance matrix from the first-stage logit estimation for a given market. Meanwhile, I compute the first-stage score contribution as the value of the covariate multiplied by the residual. For the second-stage, I compute $H_2^{-1}\left(\widehat{\Gamma};\widehat{\Delta}\right)$ as the variance-covariance matrix multiplied by a degrees of freedom correction after partialling out the market-specific parameters, as detailed in my estimation strategy in Appendix C.3. I obtain the degrees of freedom correction from Ding (2021), who derives the relationship between the partialled-out variance-covariance matrix and the full variance-covariance matrix under the Frisch-Waugh-Lovell theorem.[42] Finally, I construct the score vectors from the second-stage by multiplying the residuals from the partialled out regression by the value of either the covariate (for $s_2\left(i,\widehat{\Gamma};\widehat{\Delta}\right)$) or the perturbed covariate (for $s_2\left(i,\widehat{\Gamma};\widehat{\Delta}^b\right)$).

Finally, I employ one last correction to the standard errors to account for the adjustment due to measurement error described in Appendix C.4. Namely, I treat the measurement error correction coefficient as a constant. I thus multiply the standard errors from the two-step score bootstrap procedure by this constant to obtain the final standard errors. While this assumption ignores the uncertainty from computing the coefficient, it is computationally much simpler.

[42] Specifically, the degrees of freedom correction is $\frac{N-L}{N-K-L}$, where N is the number of patients, L is the number of national parameters and K is the number of market-level parameters.

C.6 Tables

	(1) $\widehat{p}_{ij}$	(2) $\widehat{p}_{ij}$	(3) $\widehat{p}_{ij}$
Predicted probability ($\widehat{p}_{ij}$)	0.9067	0.9537	0.8935
	(0.0041)	(0.0029)	(0.0035)
N	38,770	38,770	38,770
Distance functional form	Square	Square root	Quintiles

Table C1: Relationship between predicted probabilities with alternative demand model functional forms for distance in Springfield, Massachusetts

Notes: This table shows the relationship between the predicted choice probabilities, $\widehat{p}_{ij}$, in the demand model from Equation 1 and the predicted choice probabilities under alternative functional form specifications for distance in the demand model for the Springfield, Massachusetts HRR. The outcome in these regression is $\widehat{p}_{ij}$ under the main demand model, while the independent variables are the $\widehat{p}_{ij}$'s from the alternative functional forms. In column (1), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau_1 d_{ij} + \tau_2 d_{ij}^2 + \eta_{ij}$. In column (2), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau \sqrt{d_{ij}} + \eta_{ij}$. In column (3), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau_1 d_{ij}^{\text{bin } 2} - \tau_2 d_{ij}^{\text{bin } 3} - \tau_3 d_{ij}^{\text{bin } 4} - \tau_4 d_{ij}^{\text{bin } 5} + \eta_{ij}$, where v_{ij}^k denotes the kth quintile of a surgeon's hip and knee volume and the first quintile is the reference group. Standard errors are clustered at the patient level.

	(1) $\widehat{\delta}_{0j}$	(2) $\widehat{\delta}_{0j}$	(3) $\widehat{\delta}_{0j}$
Demand for exogenous quality $\widehat{\delta}_{0j}$	1.1629	1.0526	1.1587
	(0.0358)	(0.0140)	(0.0516)
N	32	32	32
Distance functional form	Square	Square root	Quintiles

Table C2: Relationship between demand for exogenous surgeon quality with alternative demand model functional forms for distance in Springfield, Massachusetts

Notes: This table shows the relationship between the demand for exogenous surgeon quality, $\widehat{\delta}_{0j}$, in the demand model from Equation 1 and the demand for exogenous surgeon quality under alternative functional form specifications for volume the demand model for the Springfield, Massachusetts HRR. The outcome in these regression is $\widehat{\delta}_{0j}$ under the main demand model, while the independent variables are the $\widehat{\delta}_{0j}$'s from the alternative functional forms. In column (1), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau_1 d_{ij} + \tau_2 d_{ij}^2 + \eta_{ij}$. In column (2), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau \sqrt{d_{ij}} + \eta_{ij}$. In column (3), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau_1 d_{ij}^{\text{bin } 2} - \tau_2 d_{ij}^{\text{bin } 3} - \tau_3 d_{ij}^{\text{bin } 4} - \tau_4 d_{ij}^{\text{bin } 5} + \eta_{ij}$, where v_{ij}^k denotes the kth quintile of a surgeon's hip and knee volume and the first quintile is the reference group. Standard errors are clustered at the patient level.

C.7 Figures

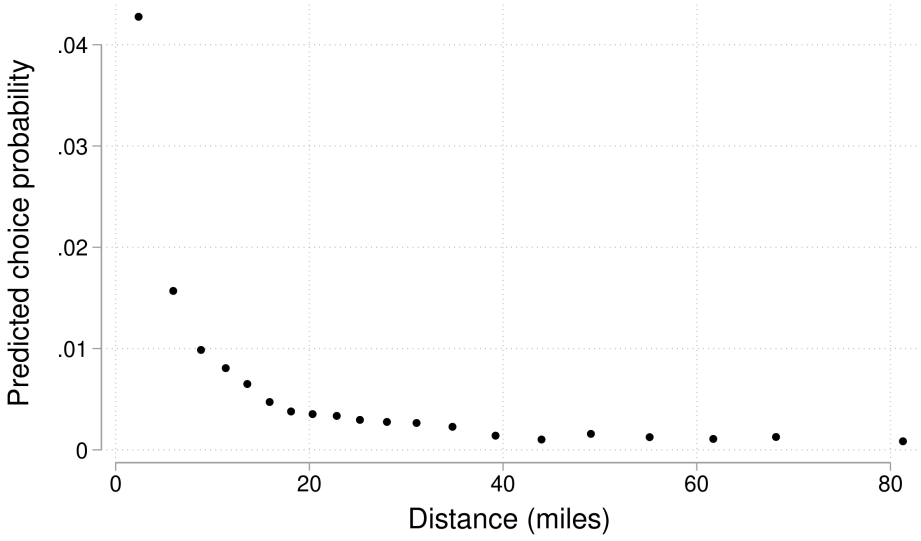


Figure C1: Relationship between predicted probability of choosing surgeon and distance in Boston HRR

Notes: This figure shows the relationship between the predicted probability of choosing a surgeon as estimated from Equation 1 and the distance to the surgeon in the Boston HRR. The unit of observation is a patient-surgeon.

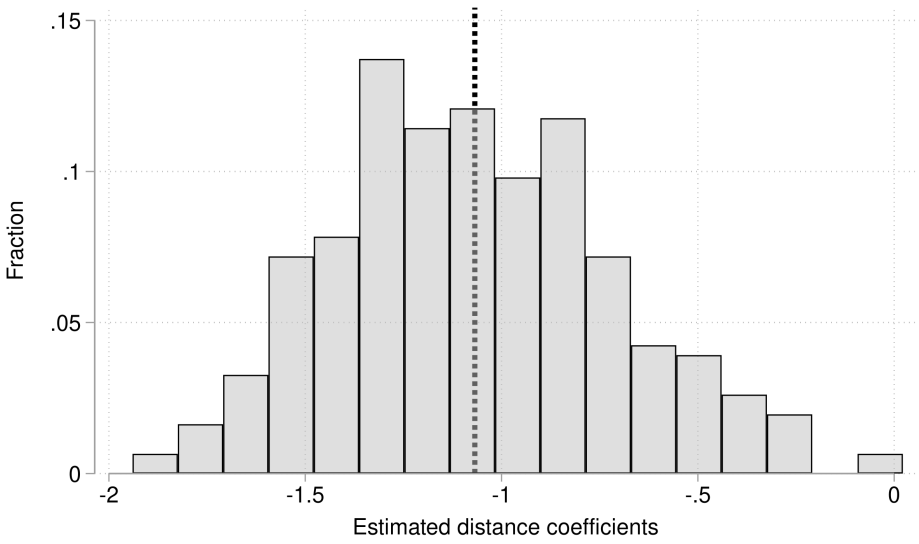


Figure C2: Histogram of estimated distance coefficients across markets

Notes: This figure shows the histogram of estimated distance coefficients for each HRR, as estimated from Equation 1. The coefficients have been shrunk to the national mean using the Empirical Bayes procedure described in Appendix Section C.1. The dotted vertical line denotes the median.

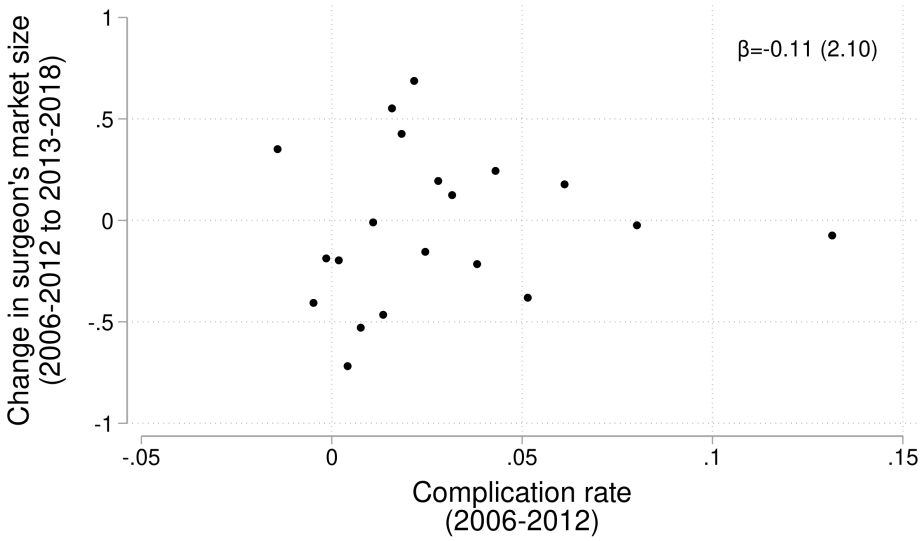


Figure C3: Relationship between change in HSA market size and surgeon's complication rate

Notes: This figure shows the relationship between the change in a surgeon's primary practice HSA from 2006-2012 to 2013-2018 versus the surgeon's risk-adjusted complication rate from 2006-2012 for 521 surgeons who move HSAs within the same HRR. A surgeon's primary practice HSA is the HSA where the surgeon performs a plurality of hip and knee replacements over the time period. Market size is the average market size of the HSA over the whole time period. A surgeon's risk-adjusted complication is a surgeon fixed effect from a linear probability model, where the outcome is a binary indicator equal to one if a patient experiences a complication following the hip or knee replacement and the independent variables are the covariates shown in Appendix Table A2, as well as year fixed effects. The unit of observation is a surgeon.

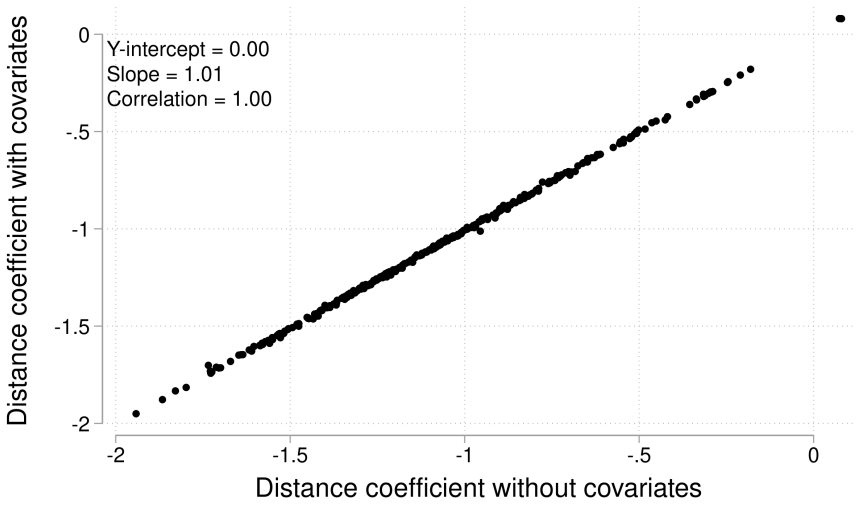


Figure C4: Estimated distance coefficient with vs. without patient covariates for HRRs

Notes: This figure shows the relationship between the estimated distance coefficients with and without patient covariates for the 306 HRRs. The estimated distance coefficient, τ, from the demand model in Equation 1 is plotted on the x-axis. The estimated distance coefficient, τ, from a demand model in which $\delta_{jt}(v_{jt}) = \delta_{0j} + \delta_{1j}\tilde{X}_{it} + \delta_1 v_{jt}$ in Equation 1 is plotted on the y-axis. The composite risk index, X_i, is the linear prediction of all the risk covariates on a complication.

D Results: The returns to surgeon volume and the benefits of market size appendix

	(1) $\widehat{p}_{ij}$	(2) $\widehat{p}_{ij}$	(3) $\widehat{p}_{ij}$	(4) $\widehat{p}_{ij}$	(5) $\widehat{p}_{ij}$
Predicted probability ($\widehat{p}_{ij}$)	0.9899	1.0002	0.9993	0.9970	0.9999
	(0.0007)	(0.0001)	(0.0005)	(0.0005)	(0.0001)
N	36,844	35,346	38,770	38,770	38,770
Volume functional form	Lag	No zeros	Quintiles	Hip and knee	Non-hip and knee

Table D1: Relationship between predicted probabilities with alternative demand model functional forms for volume in Springfield, Massachusetts

Notes: This table shows the relationship between the predicted choice probabilities, $\widehat{p}_{ij}$, in the demand model from Equation 1 and the predicted choice probabilities under alternative functional form specifications for volume in the demand model for the Springfield, Massachusetts HRR. The outcome in these regression is $\widehat{p}_{ij}$ under the main demand model, while the independent variables are the $\widehat{p}_{ij}$'s from the alternative functional forms. In column (1), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} + \delta_2 \ln v_{ij}^{lag} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{lag} is a surgeon's lagged log hip and knee volume (the surgeon's hip and knee volume from 720 days prior to the surgery date to 365 days prior). In column (2), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij} is no longer one plus hip and knee volume but just hip and knee volume (for the outside option, utility is still normalized to zero). In column (3), the demand model is $u_{ij} = \delta_{0j} + \delta_1 v_{ij}^{\text{bin }2} + \delta_2 v_{ij}^{\text{bin }3} + \delta_3 v_{ij}^{\text{bin }4} + \delta_4 v_{ij}^{\text{bin }5} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{k} denotes the kth quintile of a surgeon's hip and knee volume and the first quintile is the reference group. In column (4), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij}^{hip} + \delta_2 \ln v_{ij}^{knee} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{hip} and v_{ij}^{knee} are surgeon j's hip and knee volume plus one, respectively, in the 365 days prior to patient i's surgery date. In column (5), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} + \delta_2 \ln v_{ij}^{revision} + \delta_3 \ln v_{ij}^{arthroplasty} - \tau \ln d_{ij} + \eta_{ij}$, where $v_{ij}^{revision}$ and $v_{ij}^{arthroplasty}$ denote surgeon j's hip and knee revision and knee arthroplasty volume plus one, respectively. Standard errors are clustered at the patient level.

	(1) $\widehat{\delta}_{0j}$	(2) $\widehat{\delta}_{0j}$	(3) $\widehat{\delta}_{0j}$	(4) $\widehat{\delta}_{0j}$	(5) $\widehat{\delta}_{0j}$
Demand for exogenous quality $\widehat{\delta}_{0j}$	0.9205	1.0454	0.9379	0.9853	1.0132
	(0.0054)	(0.0037)	(0.0138)	(0.0171)	(0.0034)
N	32	32	32	32	32
Volume functional form	Lag	No zeros	Quintiles	Hip and knee	Non-hip and knee

Table D2: Relationship between demand for exogenous surgeon quality with alternative demand model functional forms for volume in Springfield, Massachusetts

Notes: This table shows the relationship between the demand for exogenous surgeon quality, $\widehat{\delta}_{0j}$, in the demand model from Equation 1 and the demand for exogenous surgeon quality under alternative functional form specifications for volume the demand model for the Springfield, Massachusetts HRR. The outcome in these regression is $\widehat{\delta}_{0j}$ under the main demand model, while the independent variables are the $\widehat{\delta}_{0j}$'s from the alternative functional forms. In column (1), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} + \delta_2 \ln v_{ij}^{lag} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{lag} is a surgeon's lagged log hip and knee volume (the surgeon's hip and knee volume from 720 days prior to the surgery date to 365 days prior). In column (2), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij} is no longer one plus hip and knee volume but just hip and knee volume (for the outside option, utility is still normalized to zero). In column (3), the demand model is $u_{ij} = \delta_{0j} + \delta_1 v_{ij}^{\text{bin }2} + \delta_2 v_{ij}^{\text{bin }3} + \delta_3 v_{ij}^{\text{bin }4} + \delta_4 v_{ij}^{\text{bin }5} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{k} denotes the kth quintile of a surgeon's hip and knee volume and the first quintile is the reference group. In column (4), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij}^{hip} + \delta_2 \ln v_{ij}^{knee} - \tau \ln d_{ij} + \eta_{ij}$, where v_{ij}^{hip} and v_{ij}^{knee} are surgeon j's hip and knee volume plus one, respectively, in the 365 days prior to patient i's surgery date. In column (5), the demand model is $u_{ij} = \delta_{0j} + \delta_1 \ln v_{ij} + \delta_2 \ln v_{ij}^{revision} + \delta_3 \ln v_{ij}^{arthroplasty} - \tau \ln d_{ij} + \eta_{ij}$, where $v_{ij}^{revision}$ and $v_{ij}^{arthroplasty}$ denote surgeon j's hip and knee revision and knee arthroplasty volume plus one, respectively.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0028	-0.0027
	(0.0005)	(0.0005)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0047	-0.0051
	(0.0008)	(0.0010)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0001	-0.0009
	(0.0004)	(0.0012)
N	677,024	677,024
Mean complication		
Median surgeon selection term $\left(\widehat{\phi}_k\right)$	-0.0033	-0.0017
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓
Patient ZIP code fixed effects	X	✓

Table D3: Estimate of the returns to surgeon volume with fixed effects for patient ZIP code of residence

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. Column (2) adds fixed effects for a patient's ZIP code of residence. Standard errors are calculated as described in Appendix C.5.

	(1)	(2)	(3)
Model	Control function	Control function	Control function
Outcome	Complication	Complication	Complication
Log volume $(\ln v_{ij})$	-0.0026	-0.0020	-0.0017
	(0.0005)	(0.0005)	(0.0006)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0044	-0.0060	
	(0.0008)	(0.0014)	
Roy selection $(\theta_{ij}(j))$	0.0001		
	(0.0004)		
N	659,868	659,868	659,868
Mean complication	0.032	0.032	0.032
Median surgeon selection term $(\widehat{\phi}_k)$	-0.0035	-0.0019	0.0015
Risk covariates	✓	✓	✓
Year fixed effects	✓	✓	✓
Surgeon HRR fixed effects	✓	✓	✓
Hospital fixed effects	X	✓	X
Surgeon fixed effects	X	X	✓

Table D4: Estimates of the returns to surgeon volume with hospital and surgeon fixed effects

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. Column (2) adds hospital fixed effects, and column (3) adds surgeon fixed effects. Standard errors are calculated as described in Appendix C.5.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0028	-0.0027
	(0.0005)	(0.0004)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0050	-0.0051
	(0.0008)	(0.0008)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0001	-0.0002
	(0.0004)	(0.0004)
N	689,565	674,496
Mean complication	0.032	0.032
Median surgeon selection term $\left(\widehat{\phi}_k\right)$	-0.0037	-0.0040
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓

Table D5: Robustness to adding one to hip and knee volume

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. Column (2) replaces log (plus one) hip and knee volume with log (plus one) hip and knee volume, resulting in a decline in the observation count. Standard errors are calculated as described in Appendix C.5.

	(1)	(2)
Model	OLS	Control function
Outcome	Complication	Complication
2nd volume quintile [6-9]	-0.0043	-0.0018
	(0.0007)	(0.0008)
3rd volume quintile [10-16]	-0.0071	-0.0025
	(0.0007)	(0.0009)
4th volume quintile [17-27]	-0.0102	-0.0047
	(0.0007)	(0.0010)
5th volume quintile [28-169]	-0.0147	-0.0070
	(0.0007)	(0.0012)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$		-0.0052
		(0.0008)
Roy selection $\left(\theta_{ij}(j)\right)$		-0.0001
		(0.0004)
N	689,565	689,565
Mean complication	0.0316	0.0316
Median surgeon selection term $\left(\widehat{\phi}_k\right)$		-0.0035
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓

Table D6: Robustness to using volume quintiles

Notes: This table shows the estimates from estimating Equation 8 except volume quintiles are used instead of log hip and knee volume. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. The specification in column (1) is the OLS specifications, whereas column (2) instruments for volume using differential distance using the control function approach. Standard errors for the control function are calculated as described in Appendix C.5.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0028	
	(0.0005)	
Log knee volume		-0.0018
		(0.0004)
Log hip volume		-0.0013
		(0.0004)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0050	-0.0050
	(0.0008)	(0.0008)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0001	-0.0001
	(0.0004)	(0.0004)
N	689,565	689,565
Mean knee complication	0.025	0.025
Mean hip complication	0.044	0.044
Median surgeon selection term $\left(\widehat{\phi}_k\right)$	-0.0037	-0.0036
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓

Table D7: Separating out hip and knee volume

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. Column (2) replaces log hip and knee volume with log hip volume and log knee volume separately. Standard errors are calculated as described in Appendix C.5.

	(1)	(2)
Model	Control function	Control function
Outcome	Complication	Complication
Log volume $\left(\ln v_{ij}\right)$	-0.0028	-0.0026
	(0.0005)	(0.0005)
Demand for exogenous quality $\left(\widehat{\delta}_{0j}\right)$	-0.0047	-0.0045
	(0.0008)	(0.0008)
Roy selection $\left(\theta_{ij}(j)\right)$	-0.0001	-0.0001
	(0.0004)	(0.0005)
N	681,260	681,260
Mean complication	0.0314	0.0314
Median surgeon selection term $\left(\widehat{\phi}_k\right)$	-0.0033	-0.0032
Risk covariates	✓	✓
Year fixed effects	✓	✓
Surgeon HRR fixed effects	✓	✓
Other volume	X	✓

Table D8: Controlling for hip and knee revisions and knee arthroplasties

Notes: This table shows the estimates from estimating Equation 8. The outcome is a binary indicator equal to one if a patient has a complication. All specifications are estimated using a linear probability model. Column (2) adds controls for a surgeon's log hip and knee arthroplasty revision volume, as well as their knee arthroscopy volume. Standard errors are calculated as described in Appendix C.5.

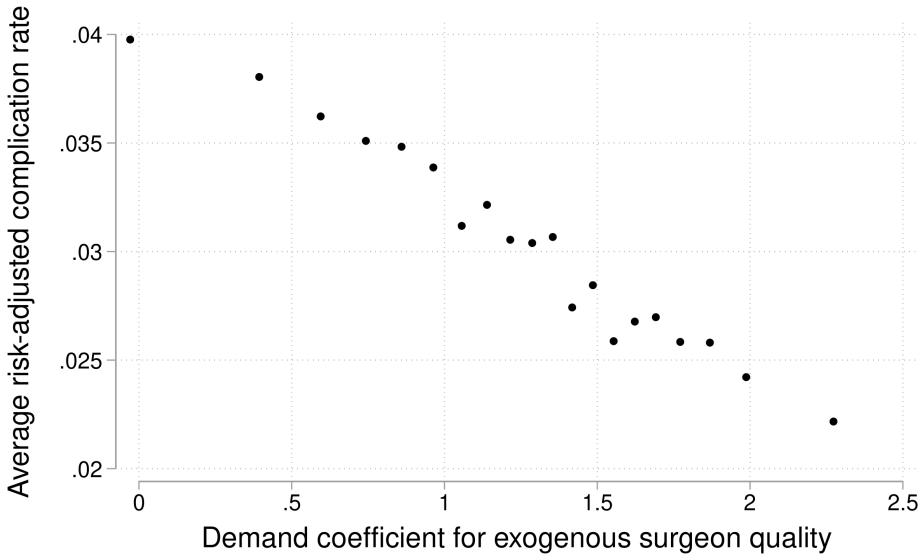


Figure D1: Relationship between risk-adjusted complication rate and demand for exogenous surgeon quality

Notes: This figure shows the relationship between the risk-adjusted complication rate and the demand for a surgeon's exogenous quality, $\widehat{\delta}_{0j}$, as estimated in Equation 1. The risk-adjusted complication rate is shrunk to the national mean using the Empirical Bayes procedure described in Appendix Section C.1. Meanwhile, δ_{0j} is shrunk to the market-level mean using the same procedure.

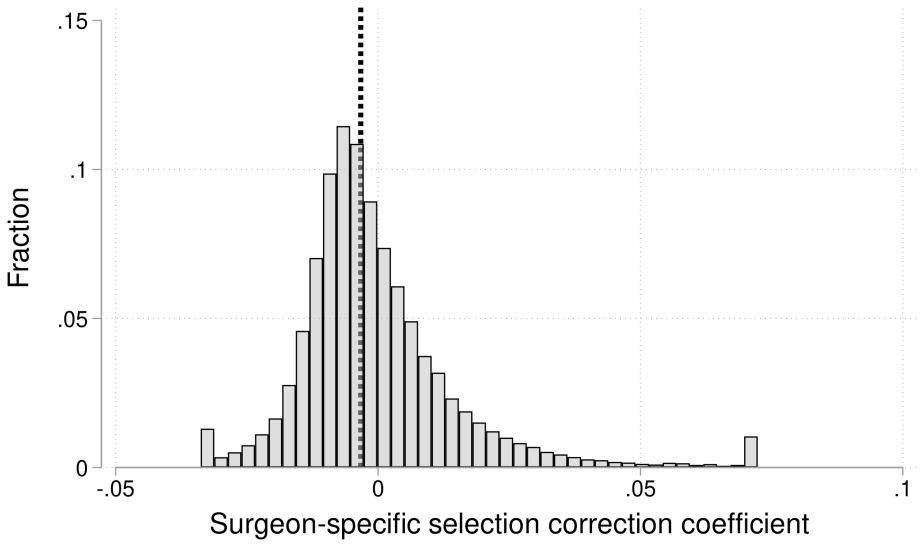


Figure D2: Distribution of surgeon-specific selection correction coefficients ($\widehat{\phi}_k$)

Notes: This figure shows the distribution of surgeon-specific selection correction terms, $\widehat{\phi}_k$, as estimated from Equation 8. These surgeon-specific selection correction terms have been winsorized at the 1% level. The vertical line denotes the median surgeon-specific selection correction term.

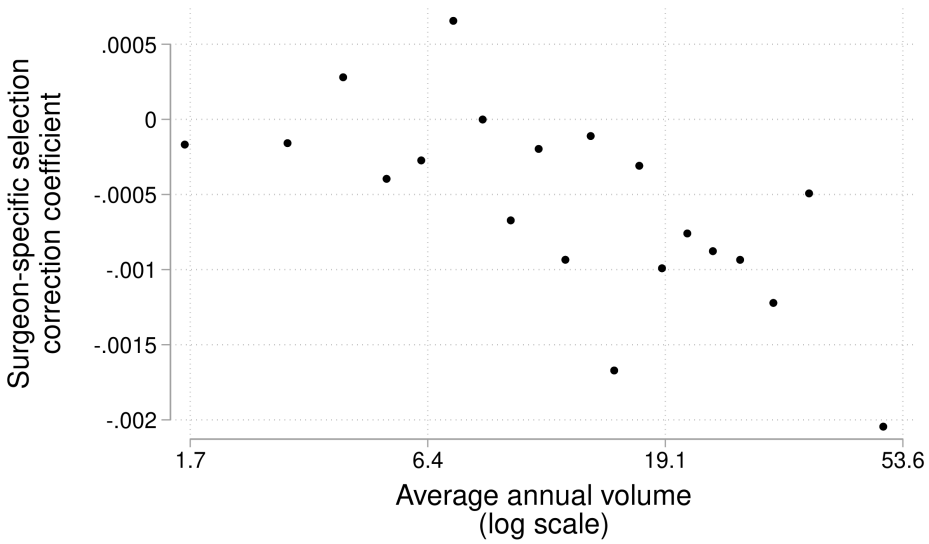


Figure D3: Relationship between surgeon-specific selection correction term ($\widehat{\phi}_k$) and average surgeon hip and knee volume

Notes: This figure shows the relationship between the average surgeon-specific selection correction term, $\widehat{\phi}_k$, as estimated from Equation 8, and a surgeon's average hip and knee volume. The surgeon-specific selection correction terms have been winsorized at the 1% level. The unit of observation is a surgeon-HRR, and each observation is weighted by cumulative hip and knee volume. HRR fixed effects have been residualized out.

E Policy implications appendix

E.1 Converting miles into dollars

I calculate the dollar cost of a mile in county c, ν_c, as the dollar value of driving plus the opportunity cost of time. Following the literature, I compute the dollar value of driving using the Internal Revenue Service (IRS) mileage reimbursement rate and the opportunity cost of time using the average wage in a patient's county:

$$\nu_c = \underbrace{2}_{\text{round trip}} \times \left(\underbrace{\frac{10.9 \text{ minutes}}{5.3 \text{ miles}}}_{\substack{\text{minutes/} \\ \text{straight-line mile}}} \times \underbrace{\text{Wage}_c}_{\$/\text{minute}} + \underbrace{\frac{7.9 \text{ drive miles}}{5.3 \text{ miles}}}_{\substack{\text{drive miles/} \\ \text{straight-line mile}}} \times \underbrace{\frac{0.55 \text{ dollars}}{\text{drive miles}}}_{\substack{\text{IRS mileage} \\ \text{reimbursement rate}}} \right). \tag{E1}$$

Here, the number outside the parentheses accounts for the round trip that patients must travel for their hip or knee replacement. Inside the parentheses is the dollar cost of a mile, where the first term before the sum is the lost wages from travel and the second is the cost of fuel and depreciation. I obtain the two conversion factors–between straight-line miles and minutes and between straight-line miles and drive miles–from Einav, Finkelstein, and Williams (2016). This factor ν_c is therefore common across all patients who reside within the same county.

E.2 Computing compensated changes in consumer surplus

For the policy counterfactual in which I subsidize patient transportation, I compute the welfare change using a compensated change in consumer surplus to isolate welfare changes resulting from the reallocation of patients to surgeons. Under this subsidy, patients may have higher utility for two reasons: they are now effectively closer to all surgeons (an "income effect") and they may choose different surgeons (a "substitution effect").[43] In this counterfactual analysis, I compute only the substitution effect, since it captures the welfare change only from the reallocation of patients to surgeons and thus how the policy may address the externality of patient choice. To isolate this effect, I hold fixed the baseline utility, such that the consumer surplus from the choice set with the subsidy is the same as that without. Formally, with some vector of surgeon volumes $\vec{v}$, I compute the change in consumer surplus, CS, under some subsidy s in market m as:

$$\begin{aligned}\Delta CS_m(s,\vec{v}) &= \underbrace{[CS_m(s=s,\vec{v}=\vec{v}_s) - CS_m(s=0,\vec{v}=\vec{v}_0)]}_{\text{full change in utility}} - \underbrace{[CS_m(s=s,\vec{v}=\vec{v}_0) - CS_m(s=0,\vec{v}=\vec{v}_0)]}_{\text{"income effect": patients closer to all surgeons}} \\ &= \underbrace{CS_m(s=s,\vec{v}=\vec{v}_s) - CS_m(s=s,\vec{v}=\vec{v}_0)}_{\text{"substitution effect": patients reallocate}} \end{aligned} \tag{E2}$$

where $\vec{v}_0$ is the vector of surgeon hip and knee volumes when $s = 0$ (i.e., there is no subsidy) and $\vec{v}_s$ is the vector of surgeon hip and knee volumes when $s = s$. Intuitively, therefore, I compensate patients for the mechanical increase in utility from surgeons being effectively closer, had they not been allowed to change their behavior.

[43] The income effect is non-zero because distance enters the utility function non-linearly, but the government values distance linearly. Computing the substitution effect therefore nets out this income effect.

E.3 Simplified first-best policy: Two surgeons at the ends of a Hotelling line

To provide intuition for the calculation of the first-best policy, I calculate the first-best policy in a stylized example. Consider two surgeons, surgeon A and surgeon B, with the same exogenous quality on the two endpoints of a Hotelling line, as depicted in Figure E1. Because it is a Hotelling line, patients are uniformly distributed between the two surgeons. The social planner's objective is to assign patients to surgeons to maximize welfare. In this case, this objective corresponds to choosing the share of patients who surgeon A operates on, as this share pins down exactly what share surgeon B operates on, since there are only two surgeons and a uniform mass of patients. To keep this example illustrative, I define welfare in this stylized example as consumer surplus, and I remove the logs from the demand model such that $u_{ij} = \delta_{0j} + \delta_1 v_{ij} - \tau d_{ij} + \eta_{ij}$. I also assume surgeons have no capacity constraints. These assumptions substantially simplify the calculations but still convey the main message.

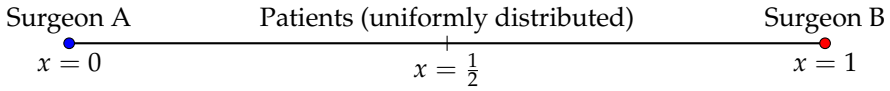


Figure E1: Patients uniformly distributed between two surgeons on a Hotelling line

I define this problem formally so as to solve for the solution mathemtically and build intuition. Let u_i^A denote the utility of patient i from surgeon A and u_i^B denote the utility of that same patient from surgeon B. Since the η's are mean-zero in expectation, the social planner's objective is:

$$\begin{aligned}\max_{s\in[0,1]} W(s) &= \int_0^s u_i^A + \int_s^1 u_i^B \\ &= \int_0^s (\delta_0 + \delta_1 s - \tau x)\partial x + \int_s^1 (\delta_0 + \delta_1(1-s) - \tau(1-x))\partial x, \end{aligned} \tag{E3}$$

where s is the share of patients assigned to surgeon A and x is the distance between a patient and surgeon A. These definitions imply that the share of patients assigned to surgeon B is $1-s$, and the distance from a patient to surgeon B is $1-x$.

I now solve for the optimal allocation of patients to surgeons in this stylized example under two separate conditions. First, consider the example where there are no returns to surgeon volume, such that $\delta_1 = 0$. Then $s^* = \frac{1}{2}$, since the surgeons are identical, so the social planner should just assign patients to the closest surgeon. Now, suppose that $\delta_1 > 0$. In this case, the solution is less obvious, as there is now a tradeoff between sending a patient farther and boosting (or lowering) a surgeon's volume, which increases (or decreases) everyone else's utility who is assigned to that surgeon. To proceed, I re-express the equation for welfare above:

$$\begin{aligned} W(s) &= \int_0^s (\delta_0 + \delta_1 s - \tau x)\partial x + \int_s^1 (\delta_0 + \delta_1(1-s) - \tau(1-x))\partial x \\ &= s\left[\delta_0 + \delta_1 s - \frac{\tau}{2}s\right] + (1-s)(\delta_0 + \delta_1(1-s)) - \frac{\tau}{2}(1-s)^2 \\ &= \delta_0 + \left(\delta_1 - \frac{\tau}{2}\right)(2s^2 - 2s + 1) \end{aligned}$$

Solving for the first-order condition (FOC) then implies:

$$\left(\delta_1 - \frac{\tau}{2}\right)(4s - 2) = 0$$

Thus, solving for the optimal s^* yields:

$$s^* = \begin{cases} \text{any } s \in [0,1], & \text{if } \delta_1 = \frac{\tau}{2} \\ \frac{1}{2}, & \text{if } \delta_1 < \frac{\tau}{2} \\ 0 \text{ or } 1, & \text{if } \delta_1 > \frac{\tau}{2} \end{cases}. \tag{E4}$$

I arrive at the first condition since any $s \in [0,1]$ satisfies the FOC if $\delta_1 = \frac{\tau}{2}$. I arrive at the next two, assuming $4s - 2 = 0$. s^* is a local maximum if $\delta_1 - \frac{\tau}{2} < 0$. However, it is a local minimum if $\delta_1 - \frac{\tau}{2} > 0$, implying that a border solution with $s = 0$ or $s = 1$ solves the maximization problem.

This solution lends insight into the conditions that determine how the social planner optimally assigns patients. The optimal assignment depends on the relative importance of the returns to surgeon volume relationship, as captured by δ_1, and the cost of travel, as captured by τ. When the returns to surgeon volume relationship roughly offsets the distance cost, any allocation of patients is optimal. When the returns to surgeon volume relationship is small relative to the distance cost, the planner should just assign patients to the closest surgeon. In this case, the externality is small, so assigning a farther away patient to a surgeon does not generate sufficient public benefit to justify the larger travel cost. Finally, when the returns to surgeon volume relationship is large relative to the distance cost, allocating all patients to one surgeon outweighs the costs of having patients travel farther. In this case, the externality is large.

E.4 Figures

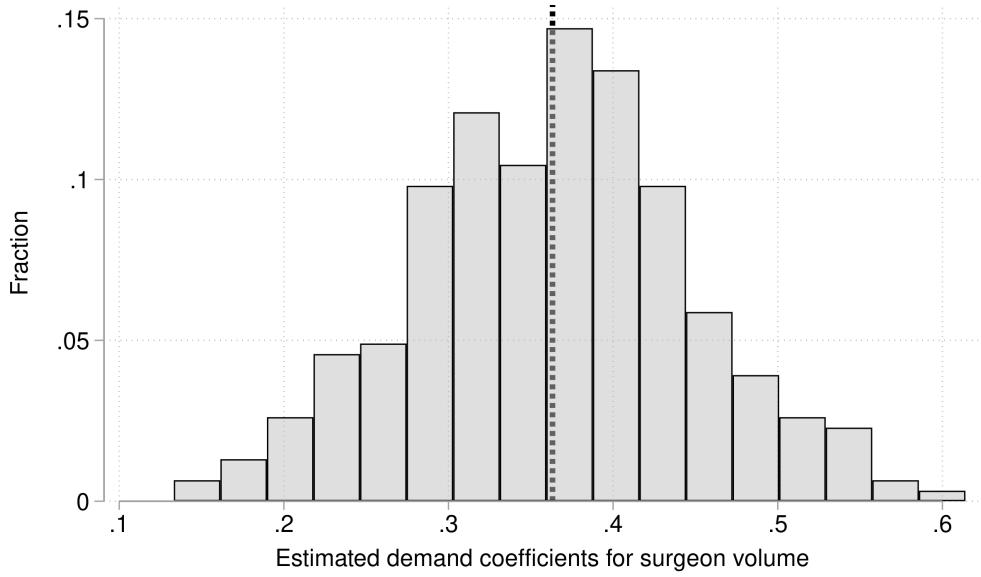


Figure E2: Histogram of estimated demand for surgeon volume across markets $\left(\widehat{\delta}_1\right)$

Notes: This figure shows the histogram of estimated distance coefficients for each HRR, as estimated from Equation 1. The coefficients have been shrunk to the national mean using the Empirical Bayes procedure described in Appendix Section C.1. The dotted vertical line denotes the median.

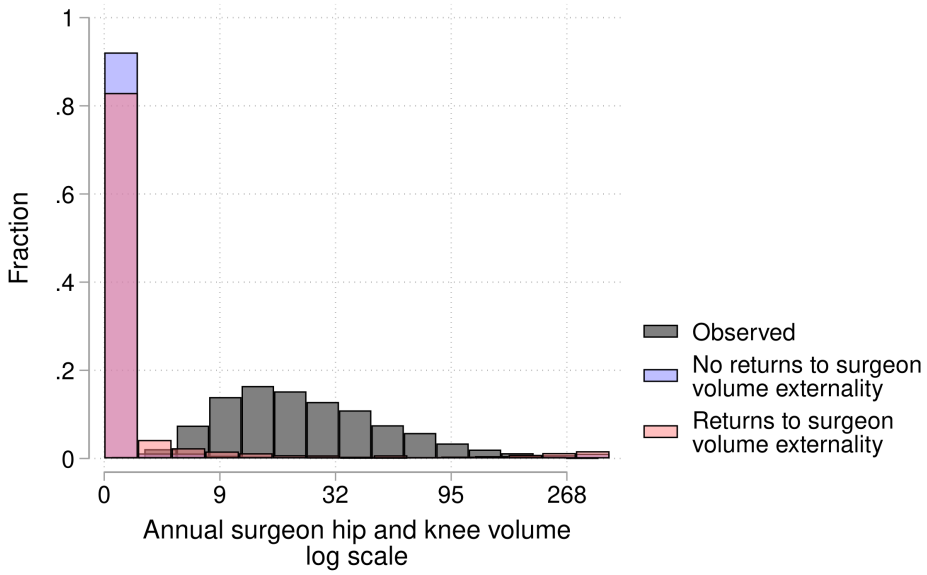


Figure E3: Histogram of observed and optimal annual hip and knee volume

Notes: This figure shows the histogram of observed volume, optimal volume without the returns to surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8), and optimal volume with the externality. The underlying unit of observation is a surgeon-HRR that has been binned into twenty-five equal-sized bins based on a surgeon's average log hip and knee volume after residualizing out HRR fixed effects. Volume above the capacity constraint has been trimmed from the histogram.

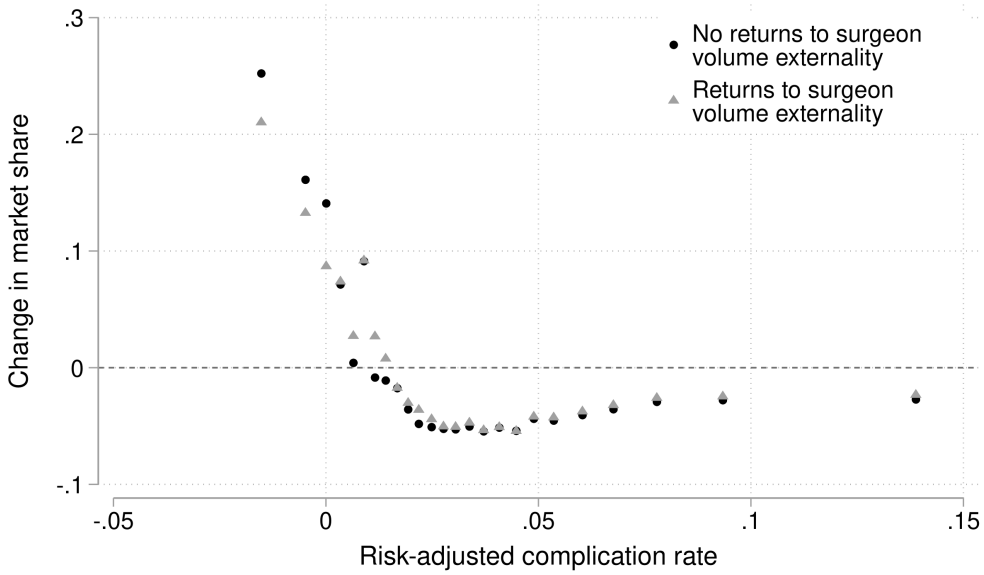


Figure E4: Relationship between change in market share and complication rate

Notes: This figure shows the relationship between the change in a surgeon's market share under the first-best policy and the surgeon's average risk-adjusted complication rate. The relationship with the black dots shows this relationship without incorporating the returns to surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1). The relationship with grey triangles shows the relationship when incorporating the returns to surgeon volume externality. The underlying unit of observation is a surgeon-HRR that has been binned into twenty-five equal-sized bins based on a surgeon's average risk-adjusted complication rate after shrinking them to the meaning using the Empirical Bayes procedure in C.1 and after residualizing out HRR fixed effects.

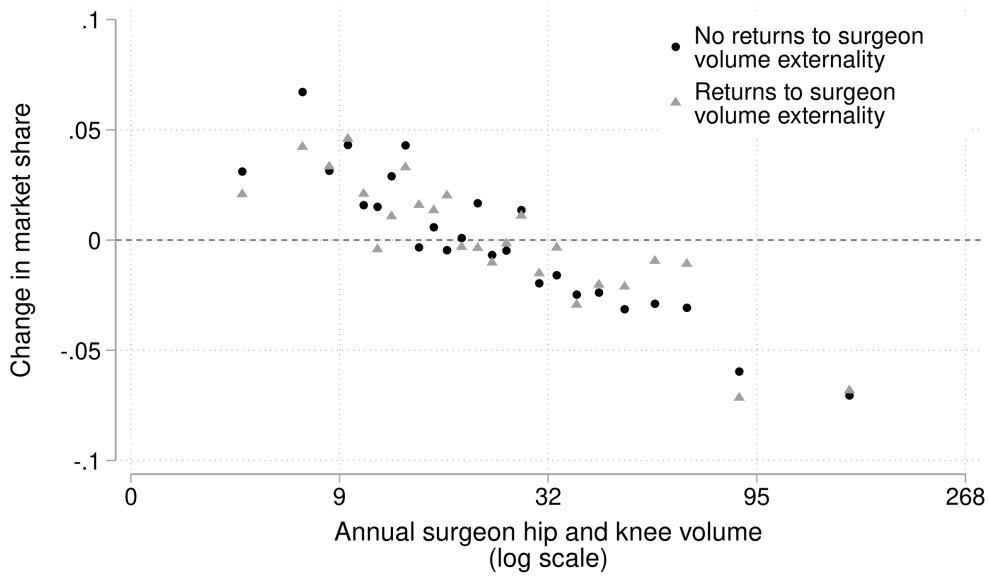


Figure E5: Relationship between change in market share and surgeon hip and knee volume

Notes: This figure shows the relationship between the change in a surgeon's market share under the first-best policy and the surgeon's average hip and knee volume across choice sets. The relationship with the black dots shows this relationship without incorporating the returns to surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1). The relationship with grey triangles shows the relationship when incorporating the returns to surgeon volume externality. The underlying unit of observation is a surgeon-HRR that has been binned into twenty-five equal-sized bins based on a surgeon's average log hip and knee volume after residualizing out HRR fixed effects.

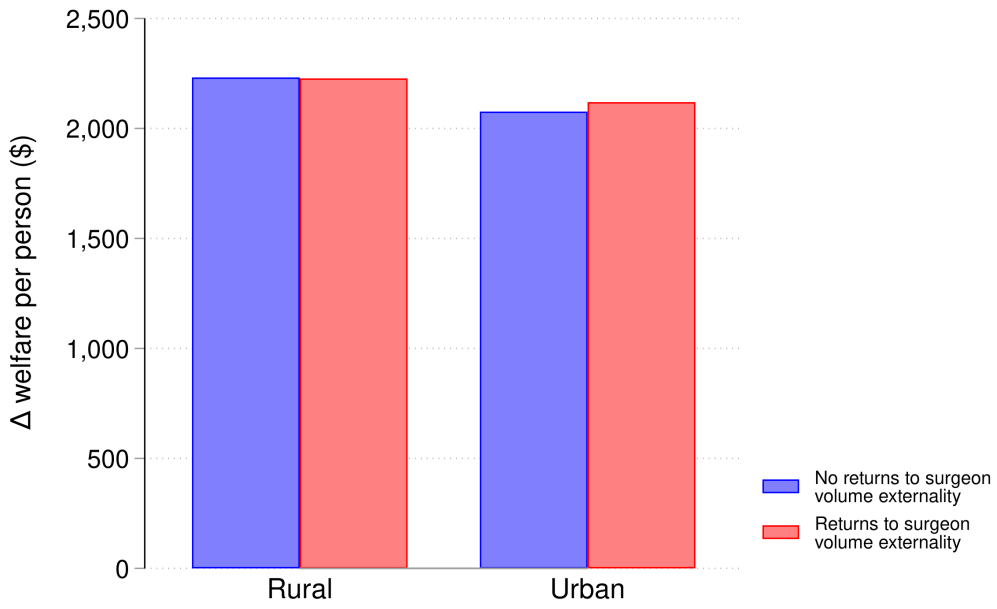


Figure E6: Welfare effects of optimal policy for rural versus urban patients

Notes: This table shows the welfare effect of implementing the optimal policy for both rural and urban patients. Welfare is calculated in dollars using Equation 10. Rural patients are defined as those having ZIP codes of residence outside a "city center," where a "city center" is the HSA with the largest market size within an HRR. Urban patients are defined as those having ZIP codes of residence within a "city center." The blue bars show the effect without incorporating the returns surgeon volume externality (i.e., setting $\delta_1 = 0$ from Equation 1 and $\beta_1 = 0$ from Equation 8), while the red bars incorporate this externality.

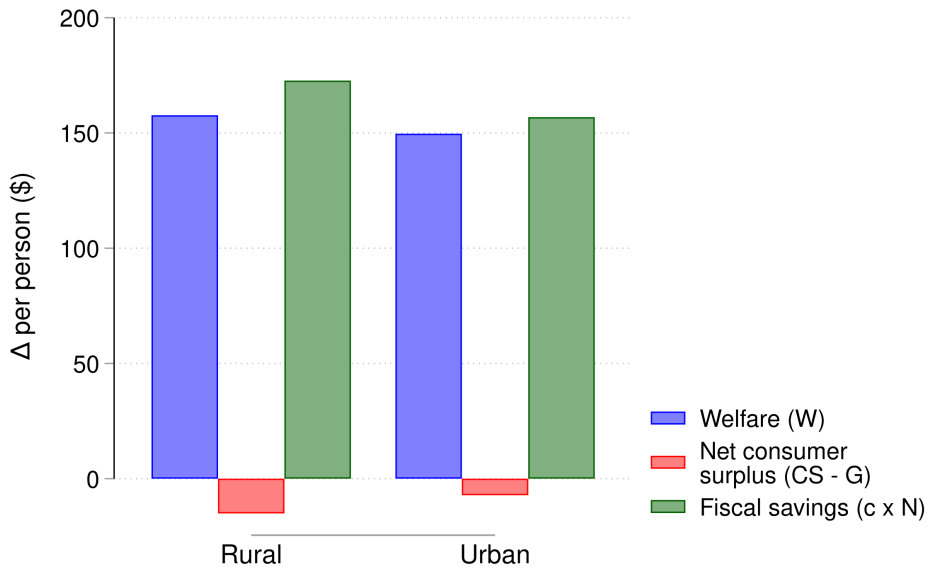


Figure E7: Effect of minimum volume standard on welfare, net consumer surplus, and fiscal savings for rural versus urban patients

Notes: This table shows the effect of implementing a minimum volume standard on welfare, consumer surplus less the fiscal costs of policy implementation (net consumer surplus), and complication-related fiscal savings for both rural and urban patients. Welfare is calculated in dollars using Equation 10. Consumer surplus is defined in Equation 11. The fiscal costs of the policy are calculated as in Section 6.4 and are equal to zero for this counterfactual. The fiscal savings per person are calculated as in Equation 10. Rural patients are defined as those having ZIP codes of residence outside a "city center," where a "city center" is the HSA with the largest market size within an HRR. Urban patients are defined as those having ZIP codes of residence within a "city center." In 3 of the 306 HRRs, patients only live in one of the HSAs within the HRR, so these HRRs are omitted from this analysis. These effects incorporate the returns surgeon volume externality.

Figure E8: Effect of subsidizing transportation on welfare, net consumer surplus, and fiscal savings for rural versus urban patients

Notes: This table shows the effect of subsidizing 30% transportation costs on welfare, consumer surplus less the fiscal costs of policy implementation (net consumer surplus), and complication-related fiscal savings for both rural and urban patients. Welfare is calculated in dollars using Equation 10. Consumer surplus is defined in Equation 11. The fiscal costs of the policy are calculated as in Section 6.4. The fiscal savings per person are calculated as in Equation 10. Rural patients are defined as those having ZIP codes of residence outside a "city center," where a "city center" is the HSA with the largest market size within an HRR. Urban patients are defined as those having ZIP codes of residence within a "city center." In 3 of the 306 HRRs, patients only live in one of the HSAs within the HRR, so these HRRs are omitted from this analysis. These effects incorporate the returns surgeon volume externality.

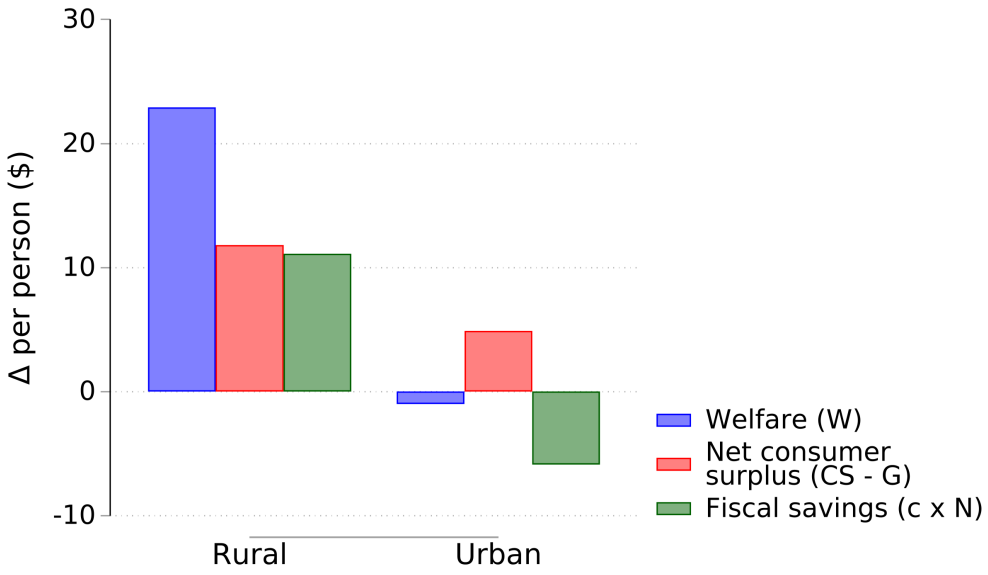


Figure E9: Effect of moving surgeons on welfare, net consumer surplus, and fiscal savings for rural versus urban patients

Notes: This table shows the effect of moving surgeons to government-designated shortage areas on welfare, consumer surplus less the fiscal costs of policy implementation (net consumer surplus), and complication-related fiscal savings for both rural and urban patients. Welfare is calculated in dollars using Equation 10. Consumer surplus is defined in Equation 11. The fiscal costs of the policy are calculated as in Section 6.4 and are equal to zero for this counterfactual. The fiscal savings per person are calculated as in Equation 10. Rural patients are defined as those having ZIP codes of residence outside a "city center," where a "city center" is the HSA with the largest market size within an HRR. Urban patients are defined as those having ZIP codes of residence within a "city center." In 3 of the 306 HRRs, patients only live in one of the HSAs within the HRR, so these HRRs are omitted from this analysis. These effects incorporate the returns surgeon volume externality.