THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMICS OF SCALE AND COMPLEXITY IN HEALTHCARE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
MAYA LOZINSKI

CHICAGO, ILLINOIS
MARCH 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to David Meltzer, Dan Black, Tamara Konetzka, Zarek Brot-Goldberg and Joshua Gottlieb for advising, feedback, and support. Thank you to my family, without whom none of this would have been possible. Thank you to the program directors and staff of the University of Chicago Medical Scientist Training Program (MSTP) and the MD-PhD Program in Medicine, the Social Sciences, and Humanities (MeSH) for helping with resources, complicated logistics, and funding. Thank you to the Agency for Healthcare Research and Quality (R36HS028592, PI: Maya Lozinski), the NIH Medical Scientist Training Program (T32GM007281), the NIA Program in Medicine, the Social Sciences, and Aging (T32AG051146), and the Center for Research Informatics (2U54TR002389-06) for funding support.

ABSTRACT

This dissertation investigates the economics of scale and complexity in healthcare. Chapter one provides an overview of the dissertation. Chapter two measures trade and economies of scale in healthcare, and considers the implications for rural healthcare policy. Chapter three finds medical knowledge growth causes some oncologists to become more specialized, leading to growing economies of scale in healthcare. Chapter four explores the trade-off between accuracy and interpretability in the use of machine learning in formula driven healthcare policy.

CHAPTER 1

INTRODUCTION

Complexity and scale are unavoidable aspects of modern medical practice. This dissertation investigates the economics of scale and complexity in healthcare, motivated by observation of how these factors impact healthcare. In this introduction, I summarize the circumstances that motivated each chapter and the key findings.

1.1 Overview of Chapter Two

The seed for chapter two was planted in my first year of medical school when I went on a service trip to the Rosebud reservation in South Dakota. While there, I shadowed in a small local clinic. The town was too small to support even a single full-time doctor. Instead, a family physician visited the clinic one day a week. The clinic was staffed with nurse practitioners the rest of the time.

Many patients saw these nurse practitioners who would have seen a subspecialist in Chicago. One boy came in with a punctured eardrum. At the University of Chicago, he would likely have been referred to a pediatric ear-nose-throat specialist. Another kid had an extensive skin infection - likely impetigo. In Chicago, he likely would have been sent to a pediatric dermatologist. However, in Rosebud, there were no such specialists; there was barely even a doctor. The town was too small to support those kinds of medical professionals. Instead, the nurse practitioners managed both children.

This set of observations got me talking with my collaborators about how providers' tasks shift across space. In small places, generalists take on many tasks that would be the purview of specialists in larger places. The small places do not have the scale to support the specialists. This line of thinking got us to consider the role of scale in healthcare, which then snowballed into the research project in chapter two.

Chapter Two documents substantial interregional trade in medical services and investigates whether regional economies of scale explain it. In Medicare data, one-fifth of production involves a doctor treating a patient from another region. Larger regions produce greater quantity, quality, and variety of medical services, which they "export" to patients from smaller regions. These patterns reflect scale economies: greater demand enables larger regions to improve quality, so they attract patients from elsewhere. Most policies to support access to healthcare for people in rural areas relocate production to small regions. However, contrary to concerns that production is too concentrated, larger regions increase access more per subsidy dollar. Another viable policy approach is to lower travel costs rather than relocating production.

1.2 Overview of Chapter Three

The seeds of chapter three were also planted in my second and third years of medical school, as I met a dizzying array of astonishingly specialized physicians. I met a neurosurgeon who only operated on one particular type of brain cancer, glioblastoma. I rotated with a urologist who only removed prostate cancer and a urologist who only removed kidney stones. I encountered a gastrointestinal (GI) specialist who treated mainly inflammatory bowel disease and a different one who focused on liver disease. Medical care at the University of Chicago was unimaginably specialized to me. I was amazed that a doctor could and would want to treat one type of disease for most of their career.

When I asked physicians why they were so specialized, they almost uniformly had the same answer. It was the only way to keep up — with surgical skills, the literature, the dizzying pace of new treatments. As I progressed in medical school, I grew more sympathetic. The volume of knowledge in medicine was growing, and I was being asked to learn more than previous generations of medical students. First Aid for the USMLE Step 1 was around 400 pages in 2000. In 2020, the year I took my Step 1 exam, it was over 800 pages. In preparing

for that exam, I reached a point where my forgetting rate started to equal my learning rate. I felt I reached the physical limit of what could be held in my head.

These experiences spurred me to investigate how physicians manage the growth of medical knowledge empirically. However, this investigation faced a central hurdle. Everyone says that there is more to know in medicine — but how do you actually measure growth in knowledge? Growth in knowledge proved surprisingly tricky to plausibly quantify. I considered using new imaging technologies in radiology, new medical devices in orthopedics, new machinery in urology, and drug approvals in oncology. I finally settled on the length of clinical guidelines in oncology as my measure of knowledge growth from chapter three, summarized below.

Chapter three investigates the impact of knowledge growth on specialization. Many fields, such as computer science, molecular biology, and medicine, have a rapidly growing knowledge base. Do expert workers respond to growth in knowledge by becoming more specialized? We study this empirically in the context of oncology, which has experienced explosive growth in knowledge. Using a panel of Medicare claims data and historical cancer treatment guidelines, we test if oncologists exposed to greater knowledge growth become more specialized. We proxy knowledge growth in each cancer subfield using the increase in the length of clinical guidelines. Exposure to knowledge growth causes ex-ante specialized oncologists to become even more specialized but does not affect general oncologists. We find that the resulting growth in specialization among specialist oncologists occurs entirely in large markets. These trends lead to growing divergence in extreme specialization between large and small markets and suggest that knowledge growth increases economies of scale in expert work.

1.3 Overview of Chapter Four

The seeds of chapter four were planted in my first year of graduate school in conversation with one of my main advisors, Dr. Meltzer. He noted that risk adjustment tended to consistently underestimate costs for complex patients, a problem for the Comprehensive

Care Program, a program he had started. I thought about this problem and wondered if it could be driven by all the interactions between health conditions. For example, diabetes impairs the immune system, leading to worse infections. Many drugs are cleared by the liver or kidneys and, therefore, cannot be used in people with significant liver or kidney disease. There are innumerable such interactions and challenges for complex patients. However, standard risk adjustment models largely assumed these health conditions were additive. They did not account for these potential interactions.

Thinking about interactions led me to think about machine learning. After all, many machine learning estimators allow for more flexible functional forms than standard linear models. However, another question arose when talking to other advisors about using machine learning for risk adjustment. Could these complex machine-learning models ever actually be used in public policy? Or would they be too complex and uninterpretable for policymakers actually to use? Contemplating these questions led me to write chapter four.

Chapter four empirically investigates the trade-off between accuracy and interpretability in Medicare Advantage risk adjustment models. I introduce a formal metric for model complexity in payment policy, which equates complexity to the number of coefficients in a model, a factor central to stakeholder interpretation of payment rates. Machine learning models significantly improve prediction accuracy and robustness to upcoding but also dramatically increase complexity. Analyzing policymakers' preferences reveals that these models likely do not justify their additional complexity. Future research should explore aligning machine learning advances with payment policy constraints.

CHAPTER 2

MARKET SIZE AND TRADE IN MEDICAL SERVICES

2.1 Introduction

Rural Americans have worse health outcomes [Deryugina and Molitor, 2021, Finkelstein et al., 2021], but America's doctors are disproportionately located in big cities [Rosenblatt and Hart, 2000]. This contrast might suggest a spatial mismatch between consumers and producers of medical services, and arguments about whether physicians are geographically "maldistributed" go back decades [Newhouse et al., 1982a, Skinner et al., 2019]. To evaluate this concern, we must consider two economic mechanisms: economies of scale and patients' travel costs. We find that both are key to understanding spatial patterns of healthcare within the United States [1].

When medical services exhibit increasing returns to scale, there are benefits to geographically concentrating production. Indeed, medicine has long been suggested as an industry in which the division of labor is limited by the extent of the market [Arrow, 1963, Baumgardner, 1988a]. But if healthcare markets are geographically segmented, the only way to serve patients in smaller regions is to disperse production across space, foregoing the benefits of scale.[2] For time-sensitive emergency care, this assumption is plausible. But the vast majority of medical spending is not for such emergencies. For example, if patients with cancer can travel across regions in search of the ideal oncologist—one specialized in their particular type of cancer, one with a better reputation, or simply a better personal match—the economic

1. This chapter is co-authored with Jonathan Dingel, Joshua Gottlieb, and Pauline Mourot.

2. Many economists assume trade costs for medical services are prohibitively high. Hsieh and Rossi-Hansberg [2021]: "Producing many cups of coffee, retail services, or health services in the same location is of no value, since it is impractical to bring them to their final consumers." Jensen and Kletzer [2005]: "Outside of education and healthcare occupations, the typical 'white-collar' occupation involves a potentially tradable activity." Bartik and Erickcek [2007]: "An industry can bring in new dollars by selling its goods or services to persons or businesses from outside the local economy ('export-base production')... For health care institutions, demand for services tends to be more local."

geography of medical care may resemble other tradable industries. Society would face a proximity-concentration tradeoff: patients who import medical services produced elsewhere incur trade costs but benefit from higher quality generated by scale economies.

We quantify the roles of local increasing returns and trade costs in medical services. Using millions of Medicare claims, we find that "imported" medical procedures—defined as a patient's consumption of a service produced by a medical provider in a different region—constitute about one-fifth of US healthcare consumption. Imports are a larger share of consumption for patients in smaller markets. "Exported" medical services are disproportionately produced in large markets. Larger regions specialize in producing less common procedures, and these procedures are traded more. These patterns are attributable to local increasing returns to scale: larger regions produce higher-quality services because they serve more patients. We estimate a model and use it to quantify how production or travel subsidies would affect patients' access to care and the quality produced in each region. Spatially neutral policies affect regions differently depending on their size and trade patterns.

Section 2.2 develops a model of trade in medical services to guide our analysis. We adapt standard models of agglomeration and trade to a setting in which the government sets prices, so endogenous quality and travel patterns clear markets. If there are local increasing returns, larger markets produce higher-quality care and export it. When economies of scale are sufficiently strong relative to market size, the model predicts that larger markets will be net exporters of medical services. Market size matters more at smaller scales, so less common medical procedures respond more to differences in market size.

Section 2.3 describes our Medicare claims data. Medicare is the federal government's insurance program for the elderly and disabled and the largest insurer in the United States. Medical service providers submit claims that report the treatment location, where the patient lives, and distinguish among thousands of distinct medical procedures.

Section 2.4 begins our empirical investigation by examining how production and con-

sumption vary with market size. Production is geographically concentrated in larger markets, while consumption is much less so. This contrast implies that larger markets are net exporters of medical services to smaller markets. To test whether this pattern reflects a home-market effect—that is, larger demand causes larger regions to export medical services—we estimate a gravity model of bilateral gross trade flows [Costinot et al., 2019]. Controlling for the geographic distribution of demand and travel distances, regions with larger residential populations export more medical care. Local increasing returns to scale are so strong that greater demand induces a larger increase in exports than imports, making larger markets net exporters of medical care. We show that these scale effects cannot be attributed to larger markets having lower input costs or medical production raising population size.

Section 2.5 shows that trade and market size play a larger role in less common procedures. The imported share of consumption is 22% for above-median-frequency procedures and 35% for those below the median. Doctors performing rare procedures export their services more often and across a broader geographic scope, sometimes serving patients who reside thousands of kilometers away. For example, half of the patients having left ventricular assist devices (LVADs) inserted to restore their heart function come from outside the surgeon's region, while only 15% of screening colonoscopies are imported. Consistent with the model, the home-market effect is substantially stronger for less common procedures: a larger residential population drives a greater increase in net exports for rarer services.

Section 2.6 shows that larger markets produce higher-quality services thanks to economies of scale. We recover revealed-preference estimates of regional service quality by estimating patients' willingness to travel to each exporting region for medical services.[3] These estimates are positively related to external quality measures, such as hospital rankings published by *U.S. News and World Report*. Inferred quality rises considerably with the regional volume of production. We estimate the scale elasticity of production to be about 0.6: a region

3. Regional quality estimates and other results may be downloaded at `http://jdingel.com`.

producing 10% more because of greater demand produces about 6% higher quality.

A variety of mechanisms could generate these local increasing returns to scale: finer specialization among physicians, sharing of lumpy capital equipment, knowledge diffusion, learning by doing, and greater availability of complementary inputs [Marshall, 1890]. While we cannot test all these hypotheses, we find that physicians in larger markets are more specialized and more experienced in the procedures they perform. Trade enables patients from across regions to share in these benefits of scale: imports are more likely to be provided by a specialist—and the appropriate specialist—than locally produced services. Specialization and learning by doing likely contribute to the local increasing returns that produce higher-quality medical care in larger markets.

We use our estimates of scale economies and trade costs to quantitatively explore the proximity-concentration tradeoff. Section 2.7 shows that policies affect regions differently depending on their size and trade patterns. A nationwide increase in reimbursements raises local output quality more in smaller regions, but these regions experience smaller increases in patients' market access because fewer of their patients consume local services. We then examine the implications of increasing access to care in one region by either increasing reimbursements or reducing travel costs. Increasing reimbursements has a higher return in more populous regions: the nationwide improvement in patient market access is about 15% higher per dollar of spending when raising reimbursements in the largest regions instead of the smallest regions. Increasing reimbursements in one region reduces output quality in neighboring regions, while improving patients' market access to the extent they import from the treated region. Reducing travel costs for one region increases its import demand, which improves both output quality and market access in neighboring regions. The rich pattern of consequences when subsidizing patients in low-output regions highlights the importance of trade and agglomeration for the incidence of these policies on patients and producers.

The higher-quality care available in larger markets may not benefit all patients equally.

Patients of lower socioeconomic status are less likely to travel for better medical care. Gravity regressions show that patients from the lowest neighborhood-income decile exhibit a distance elasticity of -2.1, while those in the highest decile have a distance elasticity of -1.7. This finding is not driven by differences in the composition of care needed: these patients are more sensitive to distance even when we examine travel patterns within specific billing codes. Thus, the gains generated by local increasing returns do not benefit all patients equally.

This paper builds on research in urban, trade, and health economics. Urban economists have documented skill-biased agglomeration in production as knowledge workers have become more numerous and concentrated in skilled cities [Berry and Glaeser, 2005, Moretti, 2011, Diamond, 2016, Davis and Dingel, 2020, Eckert et al., 2020]. Connecting this to the production and trade of services has been more difficult. Most studies of the geography of services analyze restaurants and retailers [Davis et al., 2019, Agarwal et al., 2020, Allen et al., 2021, Miyauchi et al., 2021, Burstein et al., 2022]. We show that—even in a service-based economy—the sizes of both local and potential export markets influence production and quality. This suggests that healthcare can serve as an export base for large markets [Bartik and Erickcek, 2007].

The trade literature has examined market-size effects in manufacturing but investigated services much less. Davis and Weinstein [2003], Hanson and Xiang [2004], and Bartelme et al. [2019] link manufactures' market size to export patterns, in line with the home-market effect of Krugman [1980] and Helpman and Krugman [1985]. Dingel [2017] shows that market-size effects drive quality specialization across US cities. Market-size effects for pharmaceuticals have been estimated using demographic variation over time [Acemoglu and Linn, 2004a] and across countries [Costinot et al., 2019]. Services are much less studied, in part because of the paucity of reliable trade data [Lipsey, 2009, Muñoz, 2022]. We advance this literature using the detailed procedure and location information in medical claims data.

The importance of medical care for health, life expectancy, and welfare generates sub-

stantial public-policy interest. Rural locations have worse health outcomes but fewer doctors per capita. An important series of papers by Newhouse et al. [1982a,b,c], Newhouse [1990], and Rosenthal et al. [2005] considered this issue and argued against targeting a uniform geographic distribution of physicians. Building on these studies, we measure interregional trade in medical services, estimate the impact of geography on patient access, and connect this trade to economies of scale. Importantly, we use modern trade theory to guide our modeling, estimation strategy, and counterfactual policy analysis.

2.2 Theoretical framework

This section develops a model of trade in medical services tailored to our empirical analysis of US healthcare. Patients select quality-differentiated services and face trade costs. Regional increasing returns cause the quality-adjusted cost of producing a service to decline with scale. The distinction between lower costs and higher quality is important in our empirical context. The US government plays a unique role in healthcare, purchasing a large share of all output and imposing substantial regulations. We focus on Medicare, the large federal program that purchases healthcare for the elderly and disabled at regulated prices. In this context, prices do not play their traditional role in clearing markets. Instead, quality of care and patients' distance from care bring this market towards equilibrium.

For brevity, we present a competitive model, but the consequences of regional increasing returns for trade flows in a fixed-price environment do not hinge on this assumption. Appendix 2.10.1 shows that a monopolistic-competition model with one medical provider in each region delivers the same predictions. As in flexible-price models, many market structures can give rise to a home-market effect [Costinot et al., 2019].

Beyond healthcare, this model speaks to agglomeration effects in other markets subject to price controls. We show that such circumstances can be captured by a modest modification to conventional trade models. Our model continues to deliver a gravity equation for trade

flows and to predict home-market effects. This framework delivers testable predictions about spatial variation in services' quality and trade patterns when prices are fixed.

2.2.1 Demand

We use a logit model of individuals choosing providers for a given service. Providers and patients are in regions indexed by i or j, with $\mathcal{I}$ denoting the set of regions. Let N_j denote the number of patients residing in region j who make a choice.[4] All providers in a region are identical. Utility has a provider-region-specific component, a region-pair component, and an idiosyncratic component: patient k in region j choosing a provider in region i obtains utility

$$U_{ik} = \ln \delta_i + \ln \rho_{ij(k)} + \epsilon_{ik}.$$

The provider-region-specific component δ_i would usually include a product's characteristics and price. Since Medicare pays reimbursement rates that it sets administratively,[5] the δ_i relevant for the patient is the quality of the providers in region i. The region-pair component ρ_{ij} represents bilateral inverse trade costs (proximity). The idiosyncratic component ϵ_{ik} is independently and identically drawn from a standard Gumbel distribution, so the probability that patient k selects a provider in region i is

$$\Pr\big(U_{ik} > U_{i'k} \ \forall i' \neq i\big) = \frac{\exp\Big(\ln \delta_i + \ln \rho_{ij(k)}\Big)}{\sum_{i' \in 0 \cup \mathcal{I}} \exp\Big(\ln \delta_{i'} + \ln \rho_{i'j(k)}\Big)}.$$

4. Appendix 2.10.2 extends the model to have multiple patient types.

5. Patients pay a share of these reimbursements through copayments and deductibles. But note that these cost-sharing rules are constant nationally, and most Medicare patients have a supplemental insurance (Medigap or Medicaid) which covers most or all of this cost-sharing.

There is an outside option denoted by $i = 0$, which represents individuals choosing to forgo care, and we normalize its common component to zero, $\ln \delta_0 = \ln \rho_{0j(k)} = 0 \ \forall k$.[6]

This choice probability implies a gravity equation for the quantity of trade between any two regions when we aggregate patients' decisions. Let Q_{ij} denote the quantity of procedures supplied by providers in i to patients residing in j, and let Q_{0j} denote the number of patients in j selecting the outside option. Because each patient selects at most one provider, $N_j = \sum_{i \in \mathcal{I} \cup \{0\}} Q_{ij}$. The demand by patients in j for procedures performed in i is

$$Q_{ij} = \delta_i \frac{N_j}{\Phi_j} \rho_{ij}, \tag{2.1}$$

where $\Phi_j \equiv \sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'} \rho_{i'j}$ is the expected value of the choice set for patients in region j. We call this Φ_j "patient market access." Equation (2.1) is a gravity equation with an origin i component, a destination j component, and an ij pair component. Total demand for procedures produced in i is

$$Q_i = \delta_i \sum_j \frac{N_j}{\Phi_j} \rho_{ij}. \tag{2.2}$$

2.2.2 Production

We assume competitive production of services with free entry and local increasing returns that are external to the firm. That is, each price-taking provider chooses its output quality

6. This formulation of demand is familiar from the hospital competition literature, which has studied competition among hospitals on price and quality. The literature tends to assume competition occurs within a specified geographic radius [*e.g.,* Kessler and McClellan, 2000, Cooper et al., 2018] or within a metropolitan area or similar geographic unit [*e.g.,* Ho, 2009, Gowrisankaran et al., 2015, Clemens and Gottlieb, 2017, Lewis and Pflum, 2017, Ho and Lee, 2019, Dafny et al., 2019, Garthwaite et al., 2022]. Data in this literature are often limited to certain states [*e.g.,* Town and Vistnes, 2001, Gaynor and Vogt, 2003, Capps et al., 2003, Lewis and Pflum, 2015, Ericson and Starc, 2015, Ho and Lee, 2017]. Patients who are treated outside their home region may be dropped from the data or treated as choosing the outside option [as in Gaynor and Vogt, 2003]. These definitions may be appropriate for modeling competition within specified markets [though they have been questioned by Gaynor et al., 2013, Dranove and Ody, 2016] and are natural if one assumes healthcare demand is local—as has been standard (see footnote 2). We assume all regions are in each patient's choice set, so there are no "control" markets and modeling strategic interactions would be very computationally costly.

and quantity given total regional production, an exogenous factor price, and an exogenous productivity shifter. A provider in region i that employs L units of the composite input to produce service of quality δ produces the following output quantity:

$$A_i \frac{H(Q_i)}{K(\delta)} L.$$

Improving quality is costly so $K(\delta)$ is increasing. Regional increasing returns to scale are a weakly increasing, concave function $H(Q_i)$ of total regional production, Q_i, which competitive firms take as given [Chipman, 1970]. The regional productivity shifter A_i captures any other influences, such as historical investments. Provider size L is indeterminate (and unimportant) given the linear production function, external economies of scale, and price-taking behavior. The composite input is supplied to region i at factor price w_i.[7] Thus, the unit cost of producing quality δ in region i is

$$C(Q_i, \delta_i; w_i, A_i) \equiv \frac{w_i K(\delta_i)}{A_i H(Q_i)}.$$

In our institutional setting, output prices are not an equilibrium object determined solely by the intersection of supply and demand. Instead Medicare sets "reimbursement rates" largely independent of quality, quantity, or region,[8] which we denote $\overline{R}$. Each provider that produces output of the highest quality produced in region i earns revenue $\overline{R}$ per unit.

Provider optimization and free entry make the unit cost equal to the reimbursement rate

7. If the regional factor supply were upward-sloping rather than perfectly elastic, we would estimate increasing returns net of the cost of hiring additional inputs. That is, if the factor supply elasticity were β, our estimate of the scale elasticity α from equation (2.4) below would instead be an estimate of the effective scale elasticity $\tilde{\alpha} \equiv \alpha - \frac{\beta}{1+\beta}$.

8. While Medicare does have some quality incentive programs, the money at stake is a small share of Medicare's overall spending [Gupta, 2021]. Medicare has some spatial variation in physician reimbursements, but it is not very large and has diminished over time [Clemens and Gottlieb, 2014].

in each region. Given the factor price w_i and productivity shifter A_i, the free-entry condition

$$C(Q_i, \delta_i; w_i, A_i) = \overline{R} \tag{2.3}$$

defines a regional isocost curve: the set of quantity-quality combinations for which the average cost of production equals the reimbursement rate. This isocost curve is the set of potential equilibrium production outcomes in region i. Regional increasing returns make the isocost curve upward-sloping in (Q, δ) space. With free entry and fixed prices, the benefits of scale are realized as higher-quality services in higher-output regions.

While our assumptions thus far suffice for qualitative results, we later specify functional forms for additional predictions and empirical quantification; specifically, $K(\delta_i) = \delta_i$ and $H(Q_i) = Q_i^{\alpha}$, with a scale elasticity of $\alpha \in (0, 1)$. In this case, the free-entry condition (2.3) is

$$\overline{R} = \frac{w_i \delta_i}{A_i Q_i^{\alpha}}. \tag{2.4}$$

2.2.3 Equilibrium

Equilibrium equates supply and demand in each region, $Q_i = \sum_j Q_{ij}$. Given exogenous parameters $\overline{R}$, $\{w_i, A_i, N_i\}_{i \in \mathcal{I}}$, and $\{\rho_{ij}\}_{(i,j) \in (\mathcal{I}, \mathcal{I})}$, an equilibrium is a set of quantities and qualities $\{Q_i, \delta_i\}_{i \in \mathcal{I}}$ that simultaneously satisfy equations (2.2) and (2.3).

2.2.4 Scale effects in autarky

We first consider equilibrium in autarky: patients can choose whether to receive care, but they cannot travel between regions ($\rho_{ij} = 0$ for $i \notin \{0, j\}$). In this case, all demand is local and equation (2.2) simplifies to

$$Q_{jj} = \frac{\delta_j \rho_{jj}}{1 + \delta_j \rho_{jj}} N_j. \tag{2.5}$$

The autarkic equilibrium is at the intersection of the demand curve given by equation (2.5) and the free-entry isocost curve given by equation (2.3).[9] An increase in population size, $\Delta N_j > 0$, affects equilibrium outcomes by shifting the demand curve.

Figure 2.1 illustrates how greater demand affects quality in autarky. Panel 2.1a shows the role of increasing returns to scale. The vertical axis shows quality δ_i and the horizontal axis shows quantity Q_i (on logarithmic scales). Higher quality attracts more patients, so demand is upward-sloping.[10] We draw two cases of the free-entry isocost curve defined by equation (2.4): the horizontal line depicts constant returns ($\alpha = 0$) and the upward-sloping line depicts increasing returns ($\alpha > 0$). With constant returns, a rightward shift in demand ($\Delta N_j > 0$) causes a proportional increase in quantity produced and no change in output quality. With increasing returns, the demand shift elicits higher quality because producers move up the isocost curve and thus implies a more-than-proportional increase in quantity produced because the share of patients receiving care rises.

Panel 2.1b shows that an increase in demand raises quality more as the demand curve is increasingly elastic. The panel depicts two demand curves: the one on the left is more elastic, as we would expect for a less-common procedure.[11] Shifting each demand curve to the right raises the equilibrium quality of each procedure because of increasing returns to scale. This market-size effect is larger for the less common procedure with more elastic demand because the demand shift is amplified by a larger increase in quantity demanded.[12]

9. For the equilibrium to be Marshallian stable, the demand curve must be steeper than the isocost curve at the intersection. There is a stable equilibrium because equation (2.5) means $Q_{jj} \to N_j$ as $\delta_j \to \infty$.

10. For visual clarity, we draw a log-linear demand curve. The logit demand function (2.5) is in fact log-convex, which is consistent with all the comparative statics illustrated in Figure 2.1.

11. The demand function (2.5) is log-convex, so demand is indeed more elastic at lower quality. This is a fixed-price counterpart of Marshall's second law that demand is more elastic at higher prices.

12. Alternatively, one could obtain this prediction by assuming that demand is log-linear and the isocost curve is log-concave. A rightward shift in demand would cause a larger (log) difference in quality for the low-volume procedure on the steeper part of the isocost curve.

2.2.5 Market-size effects on trade flows

We now consider trade. With multiple regions and finite trade costs ($\rho_{ij} > 0$), some patients will engage in trade—*i.e.*, select a provider located in another region. This trade stems from two sources. First, in the logit demand system with finite trade costs, patients have idiosyncratic preferences that yield a strictly positive probability of choosing every region. Second, when quality varies, regions producing higher-quality services attract more patients.

Fixing the qualities produced in other regions, an increase in one region's demand affects its trade flows through three mechanisms. First, greater demand for services directly raises a region's demand for imports through the N_j term in equation (2.1). A larger population translates proportionally to a greater demand for imports. Second, with increasing returns, an increase in N_i elicits an increase in quality δ_i, which raises region i's *gross* exports to each region. Costinot et al. [2019] call this the "weak home-market effect." Third, if increasing returns are sufficiently strong, the increase in quality δ_i improves region i's patient market access Φ_i so much that $\ln \delta_i$ rises more than $\ln\left(\frac{N_i}{\Phi_i}\right)$ does. That is, the increase in region i's gross exports exceeds any increase in its gross imports. This is the "strong" home-market effect: an increase in local demand raises a region's *net* exports.

Figures 2.1c and 2.1d introduce trade and illustrate the distinction between weak and strong home-market effects.[13] Panel 2.1c depicts the quality and quantity produced in one region under two scale elasticities. Comparing points B and C, we see that a given increase in demand elicits a larger quality improvement when increasing returns are stronger. Panel 2.1d depicts equilibrium exports and imports as a function of the region's demand shifter N_j. The import curves are upward-sloping because an increase in local demand raises demand for imports. The export curves are upward-sloping because of increasing returns: an increase in local demand causes an increase in quality, which causes an increase in gross exports. This

13. These diagrams are fixed-price analogues of Figures II and III in Costinot et al. [2019]. See their discussion of the assumption that one region is large enough to affect its own quality but too small to affect the quality produced in other regions. This assumption is only made for this figure.

is the weak home-market effect. When the scale elasticity α is larger—the free-entry isocost curve in Figure 2.1c is steeper—greater demand elicits a larger increase in output quality, which steepens the export curve and flattens the import curve in Figure 2.1d. When the export curve is steeper than the import curve, there is a strong home-market effect: the increase in demand raises exports more than imports.

We predict larger effects of market size for less common procedures. When two procedures have the same production function and trade costs, demand is more elastic at the rare procedure's equilibrium quantity. As Figure 2.1b shows, an increase in demand raises quality more when the demand curve is more elastic, leading to a stronger home-market effect for the rarer procedure.

If rare procedures also have greater economies of scale (higher α)—for example, because they require specialized equipment—that would amplify this contrast. This result motivates a difference-in-differences research design: we compare the market-size effects of common and rare procedures.

These results continue to hold when an increase in demand in one region affects equilibrium outcomes in all other regions. To demonstrate this, we consider the isoelastic special case with scale elasticity $\alpha \in (0, 1)$ and examine the home-market effect in the neighborhood of a symmetric equilibrium. Suppose all regions are the same size, $N_i = \bar{N}\ \forall i$, and trade costs are symmetric: $\rho_{ii} = 1$ and $\rho_{ij} = \rho \in (0, 1)\ \forall i \notin \{0, j\}$. There is a symmetric equilibrium, which has quality $\bar{\delta}$ and patient market access $\bar{\Phi}$ in each region. As detailed in Appendix 2.10.3, we totally differentiate the system of equations in terms of $\{\mathrm{d}\delta_i, \mathrm{d}N_i\}_{i=1}^{\mathcal{I}}$ and evaluate this system with $\mathrm{d}N_1 > 0$ and $\mathrm{d}N_j = 0\ \forall j \neq 1$ at the symmetric equilibrium.

With increasing returns of any magnitude, there is a weak home-market effect; with sufficiently strong increasing returns, there is a strong home-market effect. When $\alpha > 0$, an increase in the population size of region 1 elicits an increase in the quality of service

produced in region 1 relative to the other regions:

$$\mathrm{d}\ln\delta_1 - \mathrm{d}\ln\delta_{j\neq 1} = \left[\frac{1-\alpha}{\alpha}\frac{(\bar{\Phi}-1)}{(1-\rho)\bar{\delta}} + \frac{(1-\rho)\bar{\delta}}{\bar{\Phi}}\right]^{-1} \mathrm{d}\ln N_1 > 0.$$

This higher quality causes region 1 to export more to every other region: $\frac{\mathrm{d}\ln Q_{1j}}{\mathrm{d}\ln N_1} > 0$. The effect on the region's net exports is

$$\mathrm{d}\ln Q_{1,j\neq 1} - \mathrm{d}\ln Q_{j\neq 1,1} = \left[\frac{1 - \frac{1-\alpha}{\alpha}\frac{1+(\mathcal{I}-1)\rho}{1-\rho}}{\frac{1-\alpha}{\alpha}\frac{(1+(\mathcal{I}-1)\rho)}{(1-\rho)} + \frac{(1-\rho)\bar{\delta}}{1+(1+(\mathcal{I}-1)\rho)\bar{\delta}}}\right] \mathrm{d}\ln N_1. \tag{2.6}$$

Net exports increase if and only if

$$\frac{\alpha}{1-\alpha} > \frac{1+(\mathcal{I}-1)\rho}{1-\rho}.$$

When this inequality holds, the larger population size of region 1 makes it a *net* exporter of the medical procedure; *i.e.*, the procedure exhibits a strong home-market effect around the symmetric equilibrium. This occurs if increasing returns are sufficiently strong (α is large enough) and trade costs are sufficiently large (ρ is small enough). Otherwise, there is a weak home-market effect but not a strong one. Given a strong home-market effect, the effect in equation (2.6) is diminishing in the number of potential patients $\bar{N}$, so we predict a stronger home-market effect for less common procedures.

While the existence of increasing returns seems likely—at least for some types of medical care—there is no guarantee they are sufficiently large to generate a strong home-market effect. When larger markets are net exporters, they produce care that smaller regions need. This trade can also support the larger markets' economies: rather than exporting manufactured goods, as in decades past, larger cities can reinvent themselves [Glaeser, 2005] and export medical services. Absent a strong effect, healthcare would be a net import, not an economic base, for larger regions.

2.3 Data description

Our primary dataset is 2017 claims data from Medicare, the US federal government's insurance program for the elderly and disabled. Medicare is the largest health insurer in the United States. It does not directly employ physicians or run its own hospitals. Instead, it pays bills submitted by independent physicians, physician groups, hospitals, and other medical service providers. These bills—called "claims" in industry terminology—report the specific services provided using 5-digit codes from the Healthcare Common Procedure Coding System (HCPCS). There are over 12,000 distinct HCPCS codes, which identify individual procedures at a granular level.[14] Federal regulation determines the payment for each claim, rather than physicians' or hospitals' pricing decisions. In alternative analyses we use groupings of patient *diagnoses* to account for potential substitution between treatments.[15]

The claims data report the geographic location of both the physician providing the care and the patient receiving it, allowing us to construct a trade matrix for medical services. We study all medical care provided by physicians outside an emergency room, whether in an office or hospital facility.[16] Because Medicare rarely reimbursed telehealth in 2017, this trade involves traveling to receive a service delivered in-person.[17] We aggregate the ZIP-code-level information up to 306 hospital referral regions (HRRs), which are geographic units defined by the Dartmouth Atlas Project to represent regional health care markets for tertiary

14. For instance, there are distinct codes for providing flu vaccines based on patient age, whether the vaccine protects against three or four strains of flu, and whether administration is intramuscular or intranasal. There are distinct codes for chest X-rays based on whether the images are of ribs, the breastbone, or the full chest, both sides or one side of the body, and the number of images taken (1, 2, 3, or 4+).

15. We use the Clinical Classifications Software Refined (CCSR) diagnosis categories produced by the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project. CCSR aggregates over 70,000 ICD-10-CM diagnosis codes into "clinical categories," of which 482 have at least 20 patients each in our data. We split these categories at the median frequency to separate common from rare diagnoses.

16. Our results are robust to adding the value of hospital facility fees on top of physicians' professional fees.

17. In 2012, Medicare spent only $5 million—less than 0.001% of its expenditures—on telehealth services [Neufeld and Doarn, 2015], lagging other insurers [Dorsey and Topol, 2016].

medical care based on 1992–93 data. We construct HRR-to-HRR trade flows by interpreting the patient's residential HRR as the importing region and the service location's HRR as the exporting region.[18] The Dartmouth Atlas Project defines HRRs by aggregating residential areas based on where patients were referred for major cardiovascular surgical procedures and for neurosurgery and requires each HRR to have at least one city where both major cardiovascular surgical procedures and neurosurgery were performed. Thus, the construction of these geographic units should tend to minimize trade between different HRRs.[19]

Physicians, hospitals, pharmacies, and other healthcare providers submit different types of claims. We use a random 20% sample of all physician claims paid by Traditional (fee-for-service) Medicare in 2017, selected randomly by patient.[20,21] One year of data from this sample contains 229 million services, representing $19 billion in spending. The Medicare claims are not perfectly representative of all US healthcare, since Medicare beneficiaries are elderly or disabled. But the geographic distribution of Medicare beneficiaries is quite similar to the overall population, and Medicare alone finances one-fifth of medical spending. So it is likely to capture the key features of overall healthcare production and consumption.

Since we only see a sample of Medicare data—and hence an even smaller share of overall medical care—we might completely miss physicians or procedures so rare that a 20% sample includes none of them in a particular location. We use two other sources to address this concern. First, we use a less-detailed but more comprehensive extract of Medicare data

18. The Medicare claims are US patients receiving care at US service facilities. These data do not report any international transactions. Throughout this paper, "imports" and "exports" refer to domestic transactions between regions of the United States.

19. We have also used alternative geographies, including core-based statistical areas (CBSAs) and metropolitan statistical areas, a subset of CBSAs that excludes the smaller micropolitan areas. Because these yield consistent findings, we do not report all such estimates.

20. We also use data from 2011 to 2016 to investigate trade patterns over time in Appendix Figure 2.13.

21. One-third of Medicare patients opt out of the traditional version of Medicare, where care is paid directly by the government, in favor of a private insurance scheme ("Medicare Advantage"). In these private schemes, the government pays the insurer a fixed amount per patient and the insurers are responsible for the patient's care. Because Medicare does not pay claim-level bills in these private insurance schemes, the availability and quality of data for the privately insured patients is lower. We exclude these patients from our analysis.

(based on all Traditional Medicare patients) to replicate some of our analyses and obtain extremely similar findings.[22] Second, we use physician registry data to study the geographic patterns of production by specialty. These data provide the ZIP code and specialty of all physicians registered to practice in the United States. Physician specialty is conceptually distinct from medical service—and there is not a one-to-many mapping of specialties to services, since many services can be provided by physicians of different specialties—but we expect many of the same economic forces to apply at the level of physician specialties.

2.4 Is there a home market effect in medical services?

This section estimates how scale economies and trade costs shape the geography of aggregate healthcare production and consumption. Section 2.4.1 documents size-related spatial variation in both production and consumption. Section 2.4.2 shows that bilateral trade declines with distance. Section 2.4.3 describes our empirical strategy, which identifies the consequences of market size using gravity equations to model bilateral trade flows of medical services. Section 2.4.4 reports the empirical estimates, which demonstrate a strong home-market effect.

2.4.1 Spatial variation in production and consumption

Figure 2.2 shows maps of healthcare production and consumption across regions. The consumption map shows the substantial geographic variation that has been well-documented by the Dartmouth Atlas and related literature on geographic variation in healthcare [Fisher et al., 2003a,b, Finkelstein et al., 2016]. The production map shows even more pronounced variation: more production in large urban agglomerations and less in rural areas. There is

22. Appendix 2.11.1 explains why we must use the 20% sample and uses the 100% data to confirm some of our measures. It also shows that the relative frequencies of services purchased by private insurance are similar to those in Medicare.

substantial variation in production even between neighboring regions, while spatial variation in consumption is smoother.

The subsequent panels show patterns of trade, which constitutes the difference between production and consumption. Nationally, 22.4% of production is exported to a patient in another region.[23] Panel 2.2c shows the ratio of production to consumption; a value larger than one means an HRR is a net exporter. Net-exporting regions tend to be major urban agglomerations, plus places such as Rochester, Minn. and Hanover, N.H. that specialize in healthcare. Panel 2.2d shows gross exports as a share of local production for each HRR. Three-quarters of services produced in the Rochester metropolitan area, home to the top-rated Mayo Clinic, are provided to patients from other regions, who travel an average of 545 km to Rochester. As a major healthcare exporter with a population of merely 220,000, Rochester is an outlier: larger regions are responsible for a disproportionate share of medical services production.

Figure 2.3 plots the average production and consumption per capita across HRRs of different sizes. Both rise monotonically with population. Production rises about twice as steeply, with a population elasticity of 0.13 versus 0.06 for consumption. The difference between production and consumption is net trade: larger markets are net exporters and smaller markets are net importers. Gross trade flows exceed net trade flows, with imports comprising about one-third of consumption in the smallest regions. Exports per capita are approximately flat, which means total exports are increasing with local population. Imports per capita decline with an elasticity of -0.25 with respect to population.

23. This value is nearly identical whether measured across HRRs or across CBSAs. Appendix Figure 2.13 shows that the exported share rose steadily from 18.6% in 2011 to its 2017 level of 22.4%. For manufactured goods, the export share across CBSAs is about 68%.

2.4.2 Bilateral trade and bilateral distance

Despite the clear patterns in Figure 2.3, geographic variation in trade is far from entirely explained by market size. The four regions with the lowest export shares are Anchorage, Honolulu, and Yakima and Spokane, Wash., likely reflecting their isolated geographic locations. The highest export shares are in Rochester, Minn., Ridgewood, N.J. (just outside of New York City), and Hinsdale, Ill. (just west of Chicago). Other than Rochester—home to the Mayo Clinic—these exporting regions are all on the edge of major metropolitan areas and serve patients from those metros' hinterlands. To ensure our analysis captures these geographic patterns, we next examine bilateral trade flows.

Figure 2.4 depicts how trade varies with the distance between the patient and place of service. Figure 2.4a shows the distribution of distances patients travel for care, distinguishing between care provided in the patient's home region and other regions.[24] Within HRRs, there is a narrow distribution of distances that peaks around 10 km. When visiting providers in a different HRR, patients travel a great variety of distances. There is a local plateau between approximately 30–100 km, suggesting a fair amount of travel to nearby HRRs, perhaps indicating regional medical centers. There is another substantial peak at thousands of kilometers, demonstrating substantial long-distance travel for care.[25] Patients' willingness to travel these distance underpin our revealed-preference estimates of regional service quality.

Figure 2.4b shows that trade declines with distance. The blue curve depicts trade volume against distance (for pairs of HRRs with positive trade flows) after removing fixed effects for each exporter and each importer.[26] This intensive-margin relationship is roughly log-linear.

24. For travel within an HRR, we use the distance between the centroids of the patient's residential ZIP code and the ZIP code of the service location. We obtain the centroid coordinates from the Census Bureau's corresponding ZIP code tabulation areas (ZCTAs). For travel across HRRs, we use ZCTA-to-ZCTA distances when they are within 160 km, and (for computational ease) use HRR-to-HRR distances beyond 160 km.

25. The average patient travels 500 km to Chicago and 605 km to New York City, compared with less than 135 km to Urbana-Champaign, Ill. or Charlottesville, Va. An older literature cited in Dranove and Satterthwaite [2000] finds that patients who travel farther to hospitals tend to incur higher treatment costs.

26. This application of the Frisch-Waugh-Lovell theorem is only feasible for positive trade volumes.

The red curve shows the extensive margin: the share of pairs with positive trade as a function of distance. This is 100% for nearby pairs and under 60% for the most distant pairs. These patterns motivate the inclusion of distance covariates in our gravity-based analysis.

Patients may vary in their ability or willingness to travel, especially by socioeconomic status. We quantify it here, to the extent feasible in our data, for use in counterfactual scenarios and interpreting welfare implications. Figure 2.4c depicts distance elasticities estimated separately by neighborhood income decile.[27] We find a strong, nearly monotonic relationship between socioeconomic status and the distance elasticity: patients from the highest neighborhood-income decile exhibit a distance elasticity 25% smaller than those in the lowest decile.[28] This means patients from higher-income neighborhoods are more amenable to travel for medical care. Thus, the benefits of agglomeration—higher-quality rare care produced in major centers—may not be shared evenly. This is especially notable given the empirical setting: Medicare insures the near-universe of elderly and disabled Americans.

2.4.3 Gravity-based empirical strategy

We base our empirical examination of trade flows on a gravity equation that summarizes the geography of demand. We obtain this equation from the model by assuming the region-pair component in equation (2.1) satisfies $\ln \rho_{ij} = \gamma X_{ij} + \upsilon_{ij}$, where X_{ij} is a vector of observed trade-cost shifters and υ_{ij} is an orthogonal unobserved component. Taking expectations and then logs yields gross bilateral trade flows:

$$\ln \mathbb{E}\left(\overline{R}Q_{ij}\right) = \ln \delta_i + \ln\left(\frac{N_j}{\Phi_j}\right) + \gamma X_{ij}. \tag{2.7}$$

27. Our data do not contain patients' wealth or income, so we use their residential ZIP code. We split ZIP codes into deciles by median household income and estimate equation (2.12) separately by decile.

28. These estimates are consistent with the interaction that Silver and Zhang [2022] estimate between income and distance to care. These differences in distance elasticities are not driven by differences in the composition of procedures. When we estimate elasticities separately for rare and common services—or even for individual procedures (see Appendix Table 2.7)—the income gradient of distance elasticities persists.

The left side of (2.7) is the value of procedures exported from region i to patients residing in j. We specify the first two right-side regressors as either observable demand shifters or fixed effects in different specifications described below. We generally parameterize observed trade-cost shifters as containing log distance and a same-region dummy, so that $\gamma X_{ij} = \gamma_1 \ln \text{distance}_{ij} + \gamma_0 \mathbf{1}(i = j)$. Alternative specifications include $(\ln \text{distance}_{ij})^2$ or replace these continuous distance covariates with indicators for distance deciles.

When using the total value of bilateral exports as the dependent variable in (2.7), we aggregate quantities across thousands of distinct medical procedures using the average national Medicare reimbursement rate for each procedure. This produces an expenditure measure independent of any spatial variation in reimbursement rates.[29] We also estimate procedure-level versions of (2.7) for selected procedures, such as LVAD insertion and screening colonoscopy. The dependent variable in this case is the procedure count and no aggregation is required. Since observed bilateral trade is zero for many pairs of regions, especially when looking at trade in individual procedures, we estimate (2.7) using Poisson pseudo-maximum-likelihood [PPML; Santos Silva and Tenreyro, 2006].

We test for a home-market effect in medical services using population as an observed demand shifter. Following Costinot et al. [2019], we differentiate the system of equations (2.2) and (2.3). around the symmetric equilibrium. This delivers the local relationship between trade and population, independent of market access Φ_j. The estimating equation is

$$\ln \mathbb{E}\left[\overline{R}Q_{ij}\right] = \lambda_X \ln \text{population}_i + \lambda_M \ln \text{population}_j + \gamma X_{ij}. \tag{2.8}$$

Relative to (2.7), equation (2.8) replaces $\ln \delta_i$ and $\ln\left(\frac{N_j}{\Phi_j}\right)$ by log population in the producing and consuming regions, respectively. A positive coefficient $\lambda_X > 0$ implies a weak home-market effect as defined in Costinot et al. [2019]: *gross* exports increase with market size. If

29. Mechanically, we multiply the quantity of each procedure by the national average price for that procedure and denote the sum across all procedures by $\overline{R}Q_{ij}$.

$\lambda_X > \lambda_M > 0$, the home-market effect is strong: *net* exports increase with market size.

One potential concern with estimating (2.8) directly is reverse causality. Suppose that success in exporting medical services serves as an employment base that raises current population size, as epitomized by "anchor institutions." For example, William Worrall Mayo settling in Rochester, Minn. in the 1860s, and subsequent investment in medical care and reputation, helps explain Rochester's current population [Clapesattle, 1969].

We use two instrumental variables to address this concern. First, we use historical population. Medicine was a far smaller industry in 1940, and it is implausible that it could have driven local population in the way it might today. Since population is persistent over time, population in 1940 predicts contemporary population, and we are interested in capturing any effects of historical population that operate through current population. We therefore instrument for both the exporting region's and importing region's contemporaneous log populations with the respective log populations in 1940.

Our second instrument goes farther back than 1940 and uses local geology to predict population. Rosenthal and Strange [2008] and Levy and Moscona [2020] show that shallower subterranean bedrock makes construction easier, leading to higher population density. Bedrock depth also predicts population size, so we use this as a second instrument for local demand, again for both the importing and exporting regions.[30]

2.4.4 A strong home-market effect in medical services

Table 2.1 reports the results of estimating (2.8). The first column shows significant, positive coefficients on both patient and provider market population. The coefficient on provider-market population is two-thirds greater than that on patient-market population. This demonstrates what Costinot et al. [2019] term a *strong home-market effect*. Not only does a

30. This instrument is currently only available for CBSAs, but not for HRRs. We demonstrate that our main results are robust to defining markets based on CBSAs and to using both instruments at this level. Levy and Moscona [2020] show that the instrument has ample first-stage power for predicting population density; the same is true for our endogenous variables (population levels).

larger population increase gross exports, but it does so more than it increases gross imports by local patients. The distance elasticity of medical services trade between hospital referral regions is -1.7. This is substantially larger than the distance elasticity of -0.95 estimated for trade in manufactures between CBSAs [Dingel, 2017].[31] This suggests that trade in personal services incurs greater distance-related costs, relative to the degree of product differentiation across regions, than trade in manufactured goods. The most obvious difference is that patients themselves must travel to the provider.

The next two columns of Table 2.1 demonstrate that more flexible distance-covariate specifications do not alter the result. Column 2 introduces the square of log distance as an additional covariate. Column 3 replaces the parametric distance controls with dummies for deciles of distance. The result is stable across the columns: gross and net exports both increase with market size. The magnitudes are stable in columns 2 and 3, and the magnitude of gross (though not net) exports increases when excluding zeros.

The last column of Table 2.1 uses the historical population instrument to address concerns about reverse causality. We obtain similar home-market-effect estimates to our baseline results. Appendix Table 2.8 reports similar results estimated using CBSAs rather than HRRs as our geographic unit. It also shows the CBSA-based results are robust to instrumenting with either historical population or bedrock depth. Appendix Table 2.9 reports similar results when adding facility payments on top of physician fees.

The primary competing explanation for these results is other factors that reduce the cost of production w_i in larger markets. If doctors prefer to live in big cities [Lee, 2010], as college graduates generally do [Diamond, 2016], they could accept lower nominal wages and thus reduce healthcare production costs in such cities.

We investigate whether this mechanism is sufficiently large quantitatively to drive a net

31. We find a distance elasticity of medical services trade between CBSAs of -2.3. The analogous elasticity of health care and social assistance services trade between Canadian provinces is -1.42 [Anderson et al., 2014]. The distance elasticity of international trade is typically near -0.9 [Disdier and Head, 2008].

cost reduction in larger markets. We use data from Gottlieb et al. [2020] to measure the population elasticities of doctors' earnings and the American Community Survey [Ruggles et al., 2022] to examine other healthcare workers' earnings and real estate costs.[32] We confirm that doctors are cheaper in larger markets [Gottlieb et al., 2020], but other costs rise with population size. Appendix Figure 2.14 shows that the population elasticity of doctors' earnings is -0.01, but that for non-physicians is 0.045. To compute the population elasticity of labor costs, we use ACS data to estimate that non-physician labor's share of healthcare production is three times as much as physician labor's share. The population elasticity of labor costs is thus positive. The higher cost of real estate in larger markets reinforces these higher labor costs. This spatial variation in costs undercuts the idea that amenities make production cheaper in larger markets.

A number of related phenomena do not threaten our results. If doctors accept lower wages because they prefer the sort of work available in healthcare agglomerations, this is not a confound. Rather, it is a mechanism increasing profitability in healthcare agglomerations: greater scale lowers the cost of an input. Similarly, teaching hospitals are not a confounder. Teaching hospitals tend to be large, suggesting an agglomeration benefit of combining training with treatment at scale. Indeed, medical training exposes trainees to a large volume of patients so that they learn clinical skills by practicing them. The most salient example is Cornell University: after an abortive attempt to have medical training in both Ithaca and New York City, the Cornell Trustees quickly closed down the Ithaca location and centered the medical school in New York—where the patients and doctors were more abundant—in the early 20th century [Flexner, 1910, Gotto and Moon, 2016]. As this history illustrates, the potential local demand for care can drive the location of medical training.[33] If academic hospitals attract doctors, and their location is driven by market size, they are

32. Appendix 2.11.2 discusses subtleties of the income data.

33. In general education, in contrast, university placement induces economic growth [Moretti, 2004].

part of the agglomeration mechanism, not a confounder.

One final concern is measurement error in Medicare's records of patients' residences. To address this, Appendix 2.11.4 first demonstrates our results' robustness to excluding states with large seasonal populations. Second, we examine how far dialysis patients appear to travel. We find that residential measurement error is limited and does not drive our results.

2.5 Comparing rare and common services

Because our model predicts larger home-market effects for rarer procedures, comparing market-size effects by service frequency is a finer test of our theory. Section 2.5.1 examines how spatial variation in the production and consumption of each procedure relates to market size. Section 2.5.2 generalizes our gravity-based regression analysis to estimate home-market effects separately for rare and common procedures.

2.5.1 Spatial variation in production and consumption by frequency

We estimate the population elasticity of production and consumption per Medicare beneficiary for each procedure.[34] We find that production rises with market size more than consumption, especially for less common procedures.

Method

We first estimate the population elasticity of production per Medicare beneficiary for each procedure. Let Q_{pi} denote the count of procedure p produced in region i and its national volume be $Q_p = \sum_i Q_{pi}$. Let M_i denote the number of Medicare beneficiaries residing in i.

34. Davis and Dingel [2020] relate population elasticities to other measures of geographic concentration, such as location quotients, and estimate population elasticities of employment for various skills and sectors.

For each procedure p, we estimate the following relationship across regions:

$$\ln \mathbb{E}\left[\frac{Q_{pi}}{M_i}\right] = \zeta_p + \beta_p \ln \text{population}_i. \tag{2.9}$$

The estimated population elasticity of production per beneficiary, $\hat{\beta}_p$, describes how production varies with market size, and we estimate it using Poisson pseudo-maximum-likelihood.[35] If the quantity produced were simply proportional to population, β_p would be zero.

Our model suggests that scale effects play a larger role for rarer procedures. It predicts less common services will have higher population elasticities of production. We therefore estimate a linear regression relating $\hat{\beta}_p$ to the total national volume of service p, $\ln Q_p$.

To summarize size-linked variation in consumption patterns, we separately estimate the population elasticity of *consumption* per beneficiary for each procedure. That is, we estimate a Poisson model in which the outcome variable is the count of procedure p consumed by patients residing in region i, G_{pi}, per Medicare beneficiary residing there:

$$\ln \mathbb{E}\left[\frac{G_{pi}}{M_i}\right] = \zeta_p^C + \beta_p^C \ln \text{population}_i. \tag{2.10}$$

If $\beta_p^C \neq \beta_p$, there is size-predicted net trade in procedure p. Our model predicts that procedure frequency influences the pattern of trade, a prediction we test in Section 2.5.2.

Results

Production per beneficiary rises with market size, especially for less common procedures. Figure 2.5a relates the population elasticity of production per beneficiary $\hat{\beta}_p$ for each procedure to its national volume $\ln Q_p$. Across all volumes, procedure output per beneficiary increases with market size. Less common procedures have higher elasticities, consistent with

35. In a robustness check, we have also estimated a zero-inflated Poisson model, to account for the possibility that fixed costs are especially important for the decision of whether to provide the first instance of a service in a region. These results (not reported here) are quite similar.

economies of scale that decline with quantity.

This finding raises questions about patients' access to care. What happens to patients who live in smaller markets but need rare services? To investigate this question, we estimate equation (2.10), the population elasticity of *consumption* per beneficiary of each procedure.

The population elasticity of consumption per beneficiary is smaller for the vast majority of procedures and less steeply related to a procedure's national frequency. Figure 2.5a also plots the population elasticity of consumption per beneficiary $\hat{\beta}_p^C$ for each procedure against its national volume $\ln Q_p$. While the relationship is negative, the slope for consumption is only one third that for production. Appendix Table 2.13 reports the production, consumption and trade patterns for two exemplar procedures: screening colonoscopy and LVAD implantation. Colonoscopies are common and geographically dispersed, while LVAD procedures are rare, geographically concentrated, and traded over longer distances.

We have thus far modeled patients as demanding (and providers as producing) specific service codes. An alternative view is that patients have a particular medical condition that requires treatment, but the patients may not know what particular care they need; they simply know they require care. As physicians might use different treatments across regions for the same condition, our estimates thus far could reflect substitution among procedures. We address this by conducting a similar analysis at the level of clinical condition.

Figure 2.5b shows production and consumption elasticities by diagnosis, rather than by procedure. The key patterns remain similar: production elasticities are higher than consumption and decline more rapidly with national patient volume. Both consumption and production elasticities have less steep relationships with national volume than for procedures. This could reflect measurement error within each category: the 482 diagnosis categories we use are far coarser than the 8,253 procedures in Figure 2.5a. Alternatively, it could indicate true substitution among procedures within a condition that varies with location.

The contrasting population elasticities of production and consumption summarized in

Figure 2.5 imply trade in medical services between markets of different sizes. Just as theories of trade with scale effects would predict, larger markets export rare procedures and smaller markets import them. For almost all procedures, production increases more than proportionately with market size. Consumption also increases more than proportionately with market size, but much less so than production. The differences between these elasticities mean net exports vary with market size. The implied net trade between markets of different sizes is particularly large for procedures that have small national volumes.

2.5.2 Market-size effects are stronger for less common procedures

Procedure-level variation in bilateral trade provides a finer test of how market-size effects depend on a procedure's frequency. Appendix Figure 2.15a shows a wide distribution of imports as a share of consumption by procedure.[36] We divide procedures into two equal-sized groups, common and rare, based on the quantity produced nationally and show each group's distribution of import shares across regions in Panel 2.15b. The difference is dramatic: rare procedures (those with national frequency below the median) have much higher import shares, while the common procedures are overwhelmingly lower.[37] To formally test for differences in home-market effects, we again employ gravity models.

Empirical strategy

To test the model's difference-in-differences prediction for trade volumes, we estimate market-size effects separately for common and rare services. We compute trade flows between each HRR pair $\overline{R}Q_{ijc}$ separately for these two categories of care, $c \in \{\text{common}, \text{rare}\}$. We thus

36. This kernel density plot exhibits a spike at just above 20%, indicating that trade is, quite common in most procedures. There is a long tail reaching all the way to 1 and also many procedures with few or even zero imports.

37. Nationally, the imported share of consumption is 22% for below-median-frequency procedures and 35% for those above the median. Within both groups of procedures, there is substantial variation in import shares across hospital referral regions.

have two observations for each ij pair, allowing us to estimate:

$$\ln \mathbb{E}\left[\overline{R}Q_{ijc}\right] = \lambda_X \ln \text{population}_i + \lambda_M \ln \text{population}_j + \gamma X_{ij} + \left(\mu_X \ln \text{population}_i + \mu_M \ln \text{population}_j + \psi X_{ij}\right) \cdot \mathbf{1}(c = \text{rare}). \quad (2.11)$$

An alternate specification introduces ij-pair fixed effects, which absorb all the covariates not interacted with $\mathbf{1}(c = \text{rare})$. The theory from Section 2.2.5 predicts stronger market-size effects for rare procedures, $\mu_X > 0$.

Results

Table 2.2 reports estimates for a gravity regression in which each pair of location has two observations: one for rare services and one for common. Column 1 repeats our baseline regression from Table 2.1 but with this new structure and obtains identical results. Column 2 limits the sample to pairs of location that have positive trade in at least one of the two procedure groups, which is the estimation sample used in the remainder of the table. In columns 3 and following, we interact both provider-market and patient-market population with an indicator for rare services. We find significant and robust evidence that the home-market effect is stronger for rare services. The coefficient on provider-market population increases by about 50% relative to common services. The coefficient on patient-market population shrinks by nearly half. Column 4 introduces location-pair fixed effects. Columns 5 and 6 are analogues of the previous two, but add a quadratic distance control. These results are statistically indistinguishable from the previous columns.

Table 2.3 shows that these results are robust to instrumenting for market size with either historical population or depth to bedrock. Columns 1 and 2 show estimates for common and rare services, respectively, when instrumenting for population in each region by its 1940 population. Columns 3 and 4 repeat the exercise using CBSAs rather than HRRs,

and columns 5 and 6 switch to the bedrock-depth instrument. The results are consistent regardless of geographic unit or instrument. The estimates' stability suggests that neither the aggregate result nor the variation with procedure frequency is driven by anchor institutions or similar omitted variables.

The finding that less common procedures exhibit stronger home-market effects is robust to different ways of defining rare and common care. Table 2.4 demonstrates that our result holds when we look across diagnoses rather than procedures, and Appendix Table 2.14 shows the same when including facility spending. As with the production and consumption elasticities in Figure 2.5b, the magnitude of the difference between rare and common care shrinks. This could reflect substitution across care within a diagnosis or a less precise classification of diagnoses than of procedures. But the qualitative pattern holds and remains significant, consistent with the model's difference-in-difference prediction.

These findings reflect each procedure's national frequency, not how often an individual patient receives the same procedure. We call the latter concept the procedure's "engagement". If patients are less willing to travel for high-engagement services and these services are more common, higher engagement could drive the stronger home-market effect we observe for rare procedures. In fact, the national frequency of a service has a very low correlation with various measures of engagement for that service, so it does not confound this result.[38] While the distance elasticity is more negative for high-engagement procedures, Appendix Table 2.15 shows that separating high- from low-engagement procedures does not meaningfully alter the estimated differential impacts of population size for rare procedures.

Figure 2.6 returns to categorizing services by frequency, reporting estimates of (2.8) separately for each national frequency decile. The blue circles show estimated provider-market population elasticities, which decline monotonically from the least common to most common procedures. The red squares show patient-market population elasticities, which

38. For example, the correlation between the share of patients who had more than one claim for the procedure in a given year and the procedure's national frequency is 0.14.

increase across the frequency distribution. The difference between the respective coefficients demonstrates a strong home-market effect for all deciles. This effect is stronger the less common the procedure. Appendix Table 2.16 shows the same pattern among illustrative procedures.[39]

The potential concern about omitted cost shifters from Section 2.4.4 has an analogue here: Do the doctors who provide rare services benefit more from urban amenities than those providing common ones, lowering the cost of producing rare services in larger markets? This has facial plausibility if rare services are produced by elite specialists, who are higher-earning and more willing to pay for urban amenities through lower compensation.

Examining the population elasticities of physician earnings for each specialty alleviates this concern. If urban amenities drive specialists' locations, earnings elasticities should be negative, especially for rare specialties. But Appendix Figure 2.16 shows that the income elasticities are close to zero on average and uncorrelated with the specialty's national abundance. However urban amenities affect physicians' choices, they do not exhibit the compensating differentials necessary to explain the relationship between market size and specialization.

2.6 Estimating the scale elasticity of quality

To estimate the scale elasticity of regional medical services production, we first estimate each region's quality in Section 2.6.1. Section 2.6.2 describes our empirical strategy for estimating the scale elasticity, which Section 2.6.3 reports to be around 0.6 for aggregate medical services. Section 2.6.4 documents one mechanism linking these increasing returns and interregional trade: larger markets support a finer division of labor, and traded services

39. We show two common procedures—screening colonoscopy and cataract surgery—along with four rare ones: two treatments for brain cancer, implantation of a left ventricular assist device (LVAD), and total colectomy. All six procedures exhibit strong home-market effects, but the differences between $\hat{\lambda}_X$ and $\hat{\lambda}_M$ are smaller for the common procedures than the rare ones.

are performed by more specialized and more experienced physicians.

2.6.1 Quality estimates

We use a two-step procedure, which begins by estimating a fixed-effects version of the gravity equation. In equation (2.7), the exporting region i component of the bilateral trade flow is its perceived service quality. We can thus estimate $\ln \delta_i$ as the origin fixed effect in this gravity equation. Similarly, $\ln\left(\frac{N_j}{\Phi_j}\right)$ can be estimated as a destination fixed effect, denoted $\ln \theta_j$. This implies the following estimating equation:

$$\ln \mathbb{E}\left(\overline{R}Q_{ij}\right) = \underbrace{\ln \delta_i}_{\text{exporter FE}} + \underbrace{\ln \theta_j}_{\text{importer FE}} + \gamma X_{ij}. \tag{2.12}$$

We interpret the exporter fixed effects $\widehat{\ln \delta_i}$ as a revealed-preference measure of quality, an interpretation we validate using hospital rankings and measures of physician specialization. The importer fixed effects $\widehat{\ln \theta_j}$, plus an assumption about potential market size, enable us to compute $\Phi_j = N_j/\hat{\theta}_j$, a measure of patient market access for those who reside in location j. We also estimate (2.12) separately by service frequency, yielding $\widehat{\ln \delta_i}^{\text{rare}}$ and $\widehat{\ln \delta_i}^{\text{common}}$.

To test whether $\widehat{\ln \delta_i}$ reflects quality, the first three panels of Figure 2.7 compare the estimated exporter fixed effects to external measures of regional hospital quality. We count the number of times each region's hospitals appear on *U.S. News* Best Hospitals.[40] We also obtain Hospital Safety Grades from the Leapfrog Group and average them by HRR. The significant positive slopes in both Figures 2.7a and 2.7b show that patients prefer to obtain care from HRRs with better *U.S. News* rankings. There is also a positive relationship with Hospital Safety Grades, shown in Figure 2.7c.[41] The positive relationships with both measures suggest that our estimates capture a meaningful measure of hospital quality.

40. Appendix 2.11.2 explains how we use these rankings.

41. The distance elasticity does not meaningfully vary with procedure frequency. This suggests that patients' preference for a particular region loads onto the region fixed effects, consistent with our interpretation.

The *U.S. News* rankings are intended to capture the "Best Hospitals," a concept associated with providing highly specialized care. So it is natural that there is a stronger relationship between the *U.S. News* rankings and exporter fixed effects for rare services; the slope in Figure 2.7b is twice as large as that for common services in Figure 2.7a.[42]

2.6.2 Empirical approach

We use the estimated exporter fixed effects $\widehat{\ln \delta_i}$ to examine the determinants of regional service quality, in particular the scale elasticity, α. In the free-entry condition (2.4), service quality in region i is an isoelastic function of the quantity produced, conditional on revenue, cost, and productivity shifters. Taking the log of (2.4) and rearranging terms yields an estimating equation for the quality-quantity relationship across locations:

$$\ln \delta_i = \alpha \ln Q_i + \ln \overline{R} - \ln w_i + \ln A_i. \tag{2.13}$$

Replacing $\ln \delta_i$ with its estimate $\widehat{\ln \delta_i}$ from (2.12) yields an estimating equation for $\hat{\alpha}$.[43]

One potential concern with estimating equation (2.13) by ordinary least squares is reverse causality. Shifts of the isocost curve would cause movements along the upward-sloping demand curve, biasing the estimated scale elasticity upwards. We address this with three instruments, starting with current population. Population is relevant for healthcare output and is valid if not correlated with healthcare quality other than by driving local demand. The "anchor institutions" concern discussed in Section 2.4.3 could violate this exclusion restriction, so we also use the historical population and bedrock-depth instruments.

Despite our instruments, other channels related to population size could generate the

42. In contrast, safety grades are not differentially relevant for rare services: Appendix Figures 2.18a and 2.18b show virtually identical slopes.

43. Appendix 2.11.5 quantifies the potential bias resulting from our observing only the quantity produced for Traditional Medicare beneficiaries, rather than the total quantity produced for all patients. It shows that the bias is small: the estimates in Table 2.5 should be deflated by about 5%.

same relationship as the market-size effect we estimate. Most significantly, physicians might prefer to live in cities [Lee, 2010], regardless of patient demand. This could drive up quality in large markets, but through a different mechanism than the one we emphasize.

Before we address this problem, first note what is *not* a problem: physicians preferring to work in larger regions for job-related reasons. A larger population of patients allows physicians to specialize, conduct research, and train medical students. As discussed in Section 2.4.4, these forces operate through the scale of healthcare production in the region. Academic medical centers are often an important part of a region's medical industry. If their scale attracts workers, this is an agglomeration benefit α ought to capture.

The challenge to our interpretation arises if physicians prefer larger markets for non-professional reasons, and this labor supply shift increases quality. If urban amenities attract physicians—and higher-quality physicians in particular—this would represent variation in w_i or A_i that is correlated with population size and hence local output in equation (2.13). The analysis of local costs in Section 2.4.4 and Appendix Figure 2.14 mitigates this concern.

2.6.3 Scale improves quality

Estimated service quality $\widehat{\ln \delta_i}$ rises substantially with the regional volume of production $\ln Q_i$. Figure 2.7d depicts this relationship and Table 2.5 reports regression estimates. The estimated scale elasticity is around 0.6 and stable under various estimation approaches. The first row uses OLS, while subsequent rows instrument for output using contemporaneous or historical population. The first and third columns omit the diagonal Q_{ii} observations when estimating the gravity equation (2.12), to avoid any bias from having a region's own local consumption influence both the quality measures and output. The third and fourth columns control for spatial variation in reimbursements. Across twelve estimates, the lowest elasticity is 0.53 and the highest is 0.97. Instrumenting for output tends to reduce the estimated scale elasticity. Excluding the diagonal of the trade matrix when estimating quality

tends to raise it. The results for CBSAs, reported in Appendix Table 2.17, are also stable across specifications and when using the alternative bedrock instrument. While the existence of home-market effects implied local increasing returns, these estimates quantify their magnitude.[44] These estimates are central to our counterfactual calculations in Section 2.7.

The second panel of Table 2.5 estimates the scale elasticity for rare services. Section 2.2.5 shows that market-size effects are larger for rarer procedures even if all procedures have the same scale elasticity. These differences are amplified if rarer procedures have a larger scale elasticity than more common procedures. The scale elasticity is indeed substantially larger for rare services, with estimates centered around 0.9.

2.6.4 Scale facilitates the division of labor

One source of increasing returns—though certainly not the only one—could be division of labor among physicians. In particular, the specialized labor required to produce rare services could drive the patterns we found in Section 2.5 across treatments and diagnoses. Specialized services may require physicians with specific training, whom low demand in smaller HRRs may not support [Dranove et al., 1992].

Specialization as a source of local increasing returns

To study this mechanism, we estimate the population elasticity of physicians per capita for each specialization and relate it to the number of physicians in the specialization. Let Y_{si}

44. These estimates lie in the middle of other estimated agglomeration elasticities, Kline and Moretti [2013] estimate an elasticity of 0.4–0.47 from the Tennessee Valley Authority's investments. In manufacturing, Greenstone et al. [2010] report an analogous elasticity above 1 (a 12% increase in total factor productivity caused by adding a plant representing 8.6% of the county's prior output).

be the number of doctors of specialty s in location i.[45] We estimate a Poisson model,

$$\ln \mathbb{E}\left[\frac{Y_{si}}{\text{population}_i}\right] = \zeta_s^S + \beta_s^S \ln \text{population}_i, \tag{2.14}$$

for each specialty s by maximum likelihood.

Figure 2.8a shows a clear negative relationship between a specialty's per capita population elasticity $\hat{\beta}_s^S$ and the national number of physicians in that specialization.[46] A natural explanation for rare procedures and rare specializations both being geographically concentrated in larger regions is that the size of the market limits the division of labor. To the extent that producing rare procedures requires specialized physicians, a larger volume of patients makes production economically viable.

Consistent with this idea, Appendix Figure 2.17 shows the number of distinct procedures produced as a function of market size by procedure type. We group procedures into seven categories, count the number of procedures produced in each region, and project these onto regional population.[47] Larger regions produce a greater variety of procedures in all seven categories. If physicians specialize in particular procedures, this makes sense: larger markets have more specialties of physicians and thus a greater ability to provide rare procedures.

This evidence on specialization does not preclude other agglomeration mechanisms from also playing a role. Lumpy capital, knowledge diffusion [Baicker and Chandra, 2010], and thicker input markets could also be important productivity benefits of scale. We focus on

45. Data come from the National Plan and Provider Enumeration System (NPPES) data, which cover all physicians, not just those serving Medicare patients. These data only report the number of doctors/specialists and their location, but contain no further information about procedures performed. We restrict attention to the 223 specializations within Allopathic & Osteopathic Physicians. We restrict attention to national provider identifiers of the "individual" entity type (as opposed to "organization"). We consider each physician's primary specialty, as indicated in the NPPES file. Results (unreported) are similar when we allow for multiple specialties per physician, a common occurrence in the NPPES data.

46. This pattern is not attributable to spatial sorting driven by rare specialties commanding higher earnings. In fact, a specialty's number of physicians and mean earnings are uncorrelated. Appendix Table 2.19 shows that controlling for a specialty's earnings has no effect on the negative relationship between population elasticity and number of physicians across specialties.

47. Appendix Table 2.20 reports regression estimates for these relationships.

specialization and physician experience because of their close link to the procedure-level agglomeration we analyze and we can observe them in claims data.

Imports are specialist-intensive

We next ask whether the distribution of specialties helps explain trade. Figure 2.8b shows the share of imports and of local consumption that are provided by specialists as a function of regional population.[48] Imports are significantly more specialist-intensive than local production. This difference is especially pronounced in the smallest regions, and it remains true throughout the population distribution.

Does trade match patients with the appropriate specialist? Among all specialty care, we determine the two most common specialties to provide each unique service and label these the "standard" specialties for that care. We then determine whether each instance of the treatment was provided by a standard or non-standard specialty.

Figure 2.8c shows the share of imports and of local care provided by the standard specialties. Imports are more likely to come from the standard specialist than local care, and the distinction is especially pronounced in the smallest regions. The difference is substantial: Local care in the smallest regions is 40% more likely to be provided by a non-standard specialist than in the largest regions (7.0% vs. 5%). When importing medical services, this share falls to 5%—indistinguishable from the largest regions' local care.

We conduct a similar analysis based on provider experience. Using the public Medicare provider data (based on all Traditional Medicare patients), we count the number of times the physician billed for the specific service in the previous year. We divide this experience measure by the procedure's national mean and average it across all procedures provided to patients in an HRR. Figure 2.8d shows that, at all population sizes, care imported from other

48. We define "specialist" to mean all physicians except those whose primary specialty is internal medicine, general practice, or family practice.

regions is produced by more experienced providers than locally produced care.[49] Patients in larger regions see more experienced providers for both imported and locally produced care.

Specialists are disproportionately located in larger markets, as are physicians with more experience in any given procedure. Since imported care is predominantly specialty care, and provides patients access to this higher experience, we conclude that visiting the appropriate specialist based on training or experience is part of the value proposition for trade in medical care. This provides a second validation of our interpretation that trade reflects quality variation. Patients travel to regions with highly-ranked hospitals, which larger markets tend to have—along with the ability to provide rare services. This market-size effect strongly predicts gross and net exports. Together, this suggests that economies of scale play an important role in increasing the quality of care, and trade between regions enables patients from many regions to share the benefits of this agglomeration.

2.7 Tradeoffs and counterfactual scenarios

Given the estimated strength of local increasing returns, geographically concentrating health-care production has substantial benefits. Larger regions support specialists, house experienced physicians, and produce more specialized procedures. But this geographic concentration implies that patients in smaller regions may suffer from limited access to care. We use observed trade flows and our estimates of the scale elasticity α and region-specific qualities δ_i to quantify how various counterfactual policy scenarios would change each region's patient market access. Our results underline the importance of distinguishing between the quality of locally produced services and the quality of services to which local residents have access.

We compute counterfactual equilibrium outcomes relative to the baseline equilibrium. For the baseline equilibrium, define export shares $x_{ij} \equiv \frac{Q_{ij}}{\sum_{j'} Q_{ij'}}$ and import shares $m_{ij} \equiv \frac{Q_{ij}}{N_j}$.

49. This comparison restricts attention to procedures that are performed in all hospital referral regions (143 procedures). Thus, regional variation does not reflect the fact that larger markets produce a greater number of distinct codes (Appendix Figure 2.17).

For every variable or parameter y, denote the ratio of its counterfactual value y' to its baseline value y by $\hat{y} \equiv \frac{y'}{y}$. Appendix 2.12.1 shows how we solve for the relative counterfactual endogenous qualities ($\hat{\delta}$) using baseline equilibrium shares (x_{ij}, m_{ij}), the scale elasticity (α), and relative counterfactual exogenous parameters ($\hat{A}, \hat{R}, \hat{w}, \hat{\rho}, \hat{N}$). In particular, counterfactual qualities are given by a system of $\mathcal{I}$ equations with unknowns $\{\hat{\delta}_i\}_{=1}^{\mathcal{I}}$:

$$\hat{\delta}_i = \left(\hat{R}_i \hat{A}_i / \hat{w}_i\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in \mathcal{I}} \frac{x_{ij} \hat{\rho}_{ij} \hat{N}_j}{m_{0j} + \sum_{i' \in \mathcal{I}} m_{i'j} \hat{\delta}_{i'} \hat{\rho}_{i'j}}\right)^{\frac{\alpha}{1-\alpha}}.$$

The first term of this expression, $\left(\hat{R}_i \hat{A}_i / \hat{w}_i\right)^{\frac{1}{1-\alpha}}$, shows that the scale elasticity α governs the effect of exogenous supplier shifters, including reimbursements $\hat{R}_i$, on quality produced in a region. Reimbursement rates shift the scale of production, and stronger scale economies (higher α) amplify these shifts. The second term shows how changes in other regions influence local outcomes through trade, combined with scale. Thus, our counterfactual scenarios rely on both our estimates of the scale elasticity α and observed trade patterns.[50]

We first consider the impact of a nationwide change in reimbursements. Increasing reimbursements uniformly by 10% has heterogeneous effects. Figure 2.9a depicts the change in output quality in each region. Remote, rural areas tend to experience the largest increases in output quality δ_i. Large regions such as Boston, New York, Atlanta, and Florida have the smallest increases, because they produce more care at baseline.

Figure 2.9b shows the impact on patient market access is nearly opposite: regions with the largest increase in output quality have the smallest improvements in market access. Their residents already had high import shares, so the least reliance on local production. The increase in local quality thus has limited impact on their overall market access. For patients

50. In order to compute import shares, we assume that the number of potential patients is proportional to the number of enrolled Traditional Medicare beneficiaries. See Appendix 2.12.2 for details. The qualitative and spatial patterns of counterfactual outcomes do not depend on what share of potential patients we assume choose the outside option. Appendices 2.12.3 and 2.12.4 generalize this method of computing counterfactual outcomes to the model with multiple types of patients introduced in Appendix 2.10.2.

who switch to consuming local care, the gains are modest as local production is still lower-quality than the care they otherwise import. In contrast, patients in Houston, Dallas, or Florida had limited reason to travel. The increase in δ_i due to higher local reimbursements, even if modest, improves their access relatively more.

Figure 2.9c summarizes these contrasting changes in output quality and patient access. Regions with the lowest initial patient market access Φ_i have the biggest increase in local production quantity and quality, $\hat{Q}_i$ and $\hat{\delta}_i$, but the smallest increase in patient market access $\hat{\Phi}_i$. Appendix Figure 2.19 conducts this exercise separately for rare and common services. The patterns are qualitatively similar, but the impacts on quality are much larger for rare services because of their larger scale elasticity ($\alpha = 0.9$ rather than $\alpha = 0.6$), more concentrated baseline production, and higher baseline trade shares.

These results help reconcile two notable aspects of US healthcare policy. First, a range of recent studies find medical outcomes that match our predictions: patients who travel farther for care in larger markets tend to have better outcomes [Battaglia, 2022, Fischer et al., 2022, Petek, 2022]. Second, there is nevertheless a major political and policy effort to subsidize production in rural areas.[51] Our contrasting results for output and access rationalize this pattern: *producers* in rural areas are especially dependent on high reimbursements. This naturally leads to political pressure to subsidize production in such places. But *patients* do not necessarily benefit. They would often benefit from traveling to larger markets for better care, suggesting that the emphasis on local production may not be efficient—even from the perspective of rural patients.

We next consider the impact of this nationwide reimbursement increase on different income groups, indexed by κ. We compute changes in market access for each region and income group, and rescale them into percentage changes, $100(\hat{\Phi}_{j\kappa} - 1)$. At the region-by-

51. These policies include Critical Access Hospitals, Health Professional Shortage Areas, rural-biased adjustments to Medicare's Geographic Practice Cost Index for physician work, hospital geographic reclassification for Medicare reimbursements, increasing residency slots in rural areas, and more federal and state programs. The effectiveness of these policies is not always clear [Khoury et al., 2022, Falcettoni, 2021].

income-tercile level, Table 2.6 regresses these changes on income-group dummies (columns 1–3) and HRR fixed effects (columns 2–3). Column 1 shows that the market access gain for the highest income tercile is nearly 20% larger than for the lowest tercile. The lowest tercile experiences an 8.8% increase in patient market access (the constant in the regression). The highest tercile gains this plus an additional 1.6 percentage points. The difference is explained by differences in the groups' outside option shares, $m_{0j\kappa}$, as column 3 shows. Patients in the highest tercile are more likely to seek care, so benefit more from quality improvements.

Policies often target specific regions so we now examine how targeted production subsidies affect output quality and patient access. Figure 2.10 contrasts the consequences of raising reimbursements by 30% in Boston and in Paducah, Ky. Figure 2.10a depicts the impact on quality of care in each region relative to its baseline value in the Boston scenario. Free entry means that higher reimbursements translate to higher-quality care produced in Boston. Quality declines in the rest of New England as patients substitute away and scale economies translate lower volumes into lower quality [an "agglomeration shadow", as in Fujita and Krugman, 1995]. These effects diminish with distance to Boston.

Regions that experience larger declines in output quality due to Boston's expansion simultaneously experience larger improvements in patient market access. Figure 2.10b depicts the change in the value of patients' market access, $\hat{\Phi}_i$. Patients in Boston benefit the most from the higher reimbursement of their local production. Outside Boston, regional changes in patient market access are nearly opposite the changes in local output quality. The nearest regions import sufficient volumes that the benefits of improved quality in Boston exceed the declines in the quality of local production, causing their patient market access to improve. Regions closer to Boston experience larger declines in the quality of local production precisely because their residents' choice sets improve more, spurring more substitution. In more distant regions, the welfare impacts are neutral to ever-so-slightly negative.

We again see disproportionate gains for patients who live in higher-income neighborhoods,

shown in columns 4–6 of Table 2.6. The value of market access increases by 0.098% for first-tercile patients nationwide; this is orders of magnitude lower than in columns 1–3 because only one region's reimbursement is increasing. Gains are 70% larger for third-tercile patients. Once again this can be explained by baseline trade shares, as column 3 shows.

The consequences of higher reimbursement rates in Paducah, Ky. exhibit very different spatial patterns than in Boston. Figures 2.10c and 2.10d depict the regional changes in output quality and patient market access, respectively, caused by a 30% reimbursement increase in Paducah. Unlike Boston, Paducah is a net importer: its consumption of medical services exceeds local production by more than one-third. Higher reimbursements that improve output quality in Paducah cause Paducahans to reduce their imports from neighboring regions. This reduces the quantity produced in neighboring regions, lowering their output quality, similar to the regional spillovers in the Boston scenario. But Figure 2.10d shows that those regions where output quality declines more are the regions where patient market access declines more, contrary to the pattern of outcomes in the Boston scenario.

The contrasting outcomes reflect trade flows in the baseline equilibrium: Boston is a net exporter of medical services and Paducah is a net importer. Paducah imports one-third of its consumption, and Boston imports only six percent. Higher reimbursements in Boston cause output quality declines in nearby regions—largely because residents of those regions import more when Boston's quality improves. In contrast, higher reimbursements in Paducah reduce neighboring regions' output quality largely because Paducah residents demand fewer exports from these regions when Paducah's quality improves. Nearby regions import little from Paducah, so they benefit little from its improved quality. Appendix Figure 2.20 shows that the lessons from Boston and Paducah generalize: the pattern of spillovers from increasing reimbursements in one region is driven by that region's net trade in medical care. To summarize, the spillover consequences of subsidizing production in one region depend on the pattern of trade; changes in regional output quality need not align

with changes in regional patient market access.

The distributional consequences of region-specific subsidies depend on which region is subsidized. We compute the nationwide gains in market access from subsidizing production in each region, one at a time. Figure 2.11 shows this gain, scaled by the increase in total spending, as a function of region size. The aggregate gain in market access per dollar spent is higher in larger markets: further concentration of production has larger benefits. The graph also shows the gains per dollar separately by income tercile. Unlike the Boston scenario, in which benefits accrue more to higher-income ZIP codes, subsidizing production in less populous regions benefits lower-income ZIP codes more. These contrasts reflect geographic divides in incomes: lower-income patients are more likely to live in and near smaller regions.

Rather than subsidizing local production, policies might improve patient market access in a particular region by facilitating trade. We examine the consequences of a policy that reduces travel costs for Paducahans obtaining care elsewhere (specifically, $\hat{\rho}_{i,\text{Paducah}} = 1.3$ when $i \neq$ Paducah).[52] Figure 2.12 shows that, unlike an increase in Paducah reimbursements, this policy has positive spillovers on neighboring regions. These regions increase their exports to Paducah, and thus their own scale and quality. This improves their residents' market access.[53] Because lower-income patients are more sensitive to distance and are less likely to import care from other regions, a larger travel subsidy is necessary to achieve the same percentage improvement in their patient market access. To increase each income tercile's patient market access in Paducah by 7%, one would need to reduce trade costs by 40% for the first income tercile and by 37% for the third income tercile.

So this policy benefits both Paducah and its neighbors—though we do not estimate the costs of this travel subsidy. But facilitating travel reduces the quantity—and thus the

52. The impact of this change on Paducah residents' market access $\hat{\Phi}_{\text{Paducah}}$ is similar to an 8% increase in reimbursements in Paducah.

53. Recall that our model assumes elastic supply. The short-run impact on exporters may be more complex if there are short-term diseconomies of scale due to crowding or queuing.

quality—produced in Paducah. Analysts looking at the impact of travel subsidies on the quantity or quality of care provided in Paducah itself would reach very different conclusions than those looking at the impact on patient market access.

These counterfactual scenarios are subject to significant caveats, and we have not attempted to identify the optimal policy. Even so, this simple model rationalizes important aspects of the economics and politics of US healthcare policy. The counterfactual scenarios highlight our main findings: Healthcare production has substantial local increasing returns, and patient travel plays a meaningful role in enabling access to higher-quality care. Given these economic mechanisms, regional spillovers are larger when economies of scale are stronger, depend on the pattern of trade flows, and differ depending on whether policies subsidize production or travel. This shows the importance of distinguishing between regional output quality and regional patient access when evaluating healthcare policies.

2.8 Conclusion

Smaller markets have fewer specialized physicians, produce less medical care per capita, and have worse health outcomes than larger markets. Thanks to trade in medical services, less production does not translate one for one into less consumption of medical services. Instead, trade affords patients who live in smaller markets access to higher-quality care. This higher quality comes in part from consuming services that would otherwise be unavailable, visiting appropriate specialists, and accessing experienced physicians.

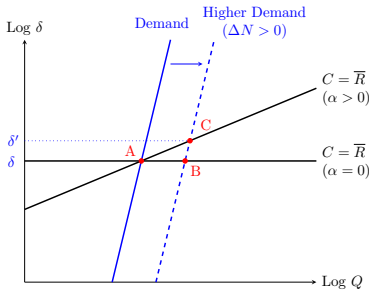
This trade amplifies the scale advantages of large markets and hence the quality of care they produce. This means the healthcare industry can serve as an export base for large cities. Substantial scale economies also imply that policies to reallocate care across regions may impact the quality of care available. We simulate policies that aim to improve care access in "under-served" markets. The rich and varied patterns of welfare consequences when subsidizing production or travel highlight the importance of trade and agglomeration

for the incidence of these policies on patients and producers.

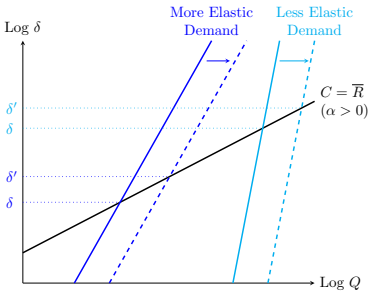
2.9 Exhibits

Figure 2.1: Illustrative model diagrams

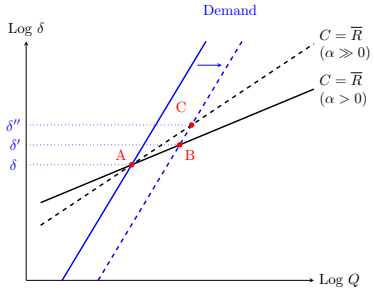
(a) Autarky: Constant vs. increasing returns



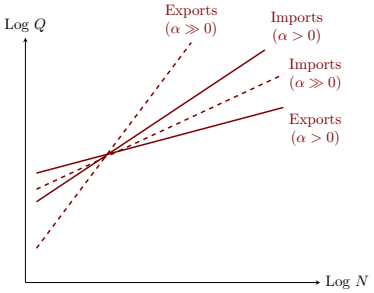
(b) Autarky: Market size and demand elasticity



(c) Quality and quantity depend on scale elasticity



(d) Exports and imports

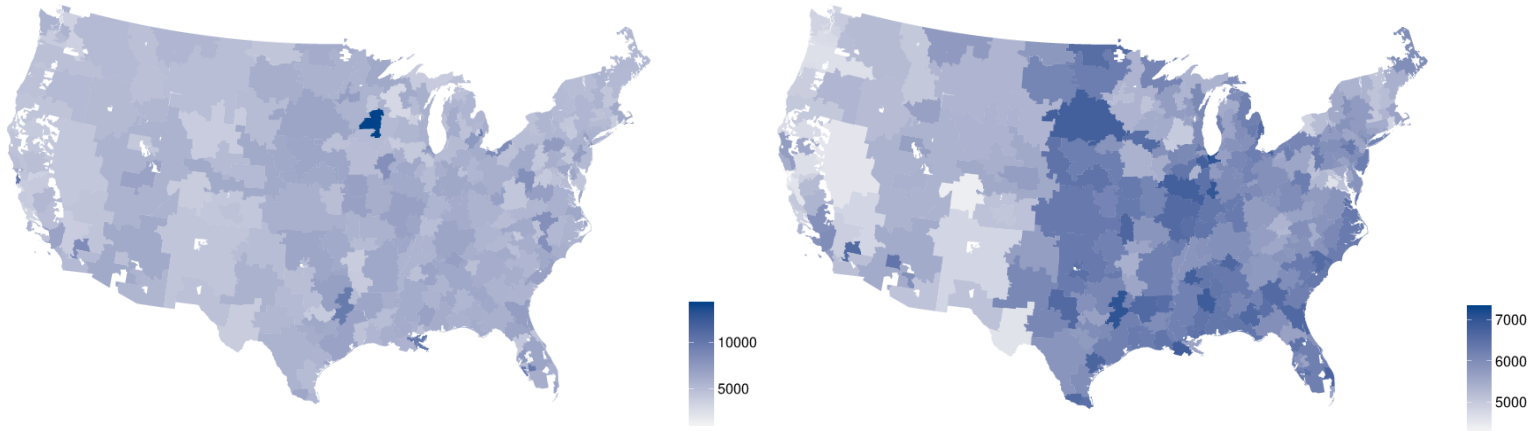


Notes: This figure depicts how increasing demand in one region affects its equilibrium outcomes. In panels a–c, quantity produced Q is on the horizontal axis and service quality δ is on the vertical axis. The black lines depict the free-entry isocost curve, $C = \overline{R}$, given by equation (2.3). The blue and cyan lines depict demand for the region's service, which we depict as log-linear for visual clarity. (The logit demand function is actually log-convex, which is consistent with all the depicted comparative statistics.) Equilibrium is the intersection of the demand and isocost curves. An increase in demand is the rightward shift from the solid to the dashed demand curve. This shift increases equilibrium quality from δ to δ'. Panel a shows that higher demand elicits higher quality if there are increasing returns to scale. Panel b shows that this quality improvement is larger when demand is more elastic. Panels c and d introduce trade and compare the extent of quality improvement under two different magnitudes of increasing returns ($\alpha > 0$ and $\alpha \gg 0$). These magnitudes govern the patterns of interregional trade, shown in panel d as a function of the number of potential patients N. Imports from other regions rise with N. With increasing returns to scale ($\alpha > 0$), exports to other regions also rise with N (a weak home-market effect). When the scale elasticity α is larger ($\alpha \gg 0$), the import curve is flatter and the export curve is steeper. With sufficiently strong increasing returns, an increase in local demand causes a greater increase in exports than imports (a strong home-market effect).

Figure 2.2: Production, consumption, and trade across regions

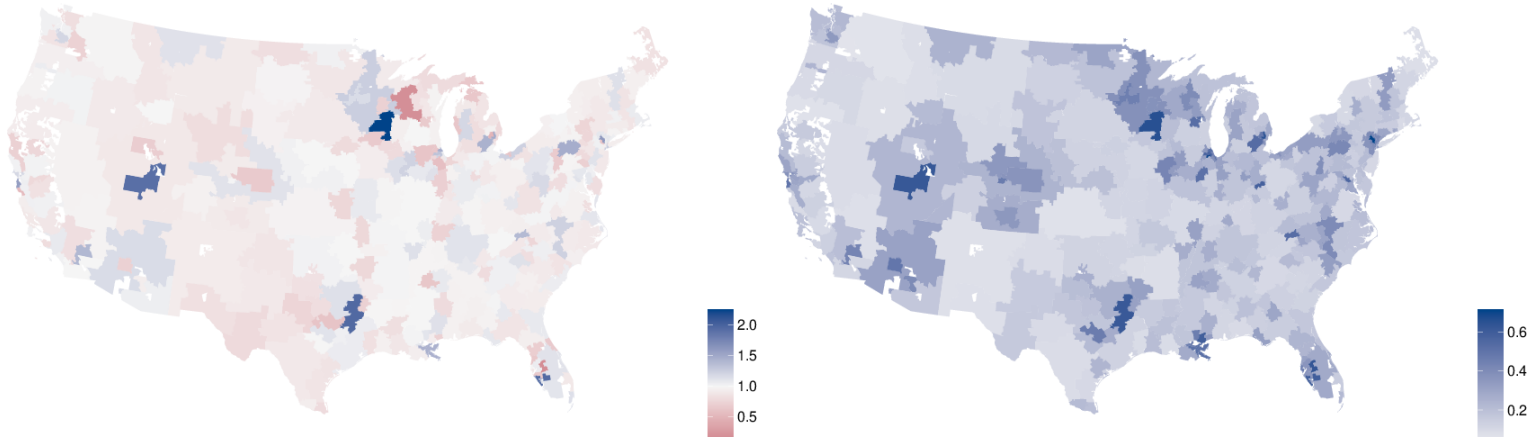
(a) Production per capita

(b) Consumption per capita



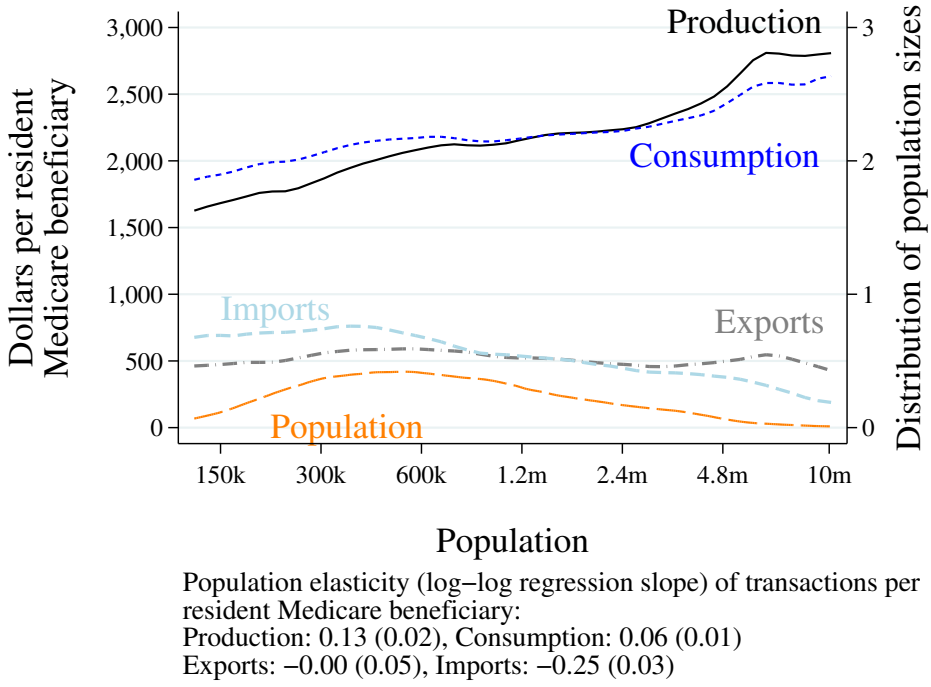
(c) Production divided by consumption

(d) Gross exports relative to production



Notes: Panel a shows production per capita, including professional and facility fees. The hospital referral region (HRR) of production is the location where the service is provided. Panel b shows consumption per capita, including professional and facility fees. The HRR of consumption is based on the patient's residential address. Colors depict deciles of production per capita in both Panels a and b. Panel c shows the ratio of production per capita to consumption per capita for professional services. Panel d shows gross exports as a share of total production by HRR for professional services. Data come from the Medicare 20% carrier Research Identifiable Files. All calculations exclude emergency-room care and skilled nursing facilities. Expenditures are computed by assigning each procedure its national average price. HRR definitions are from the Dartmouth Atlas Project.

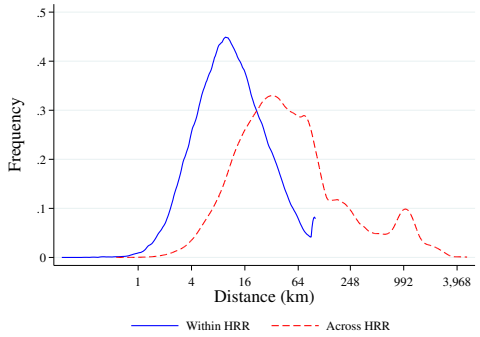
Figure 2.3: Production and consumption of medical care across regions



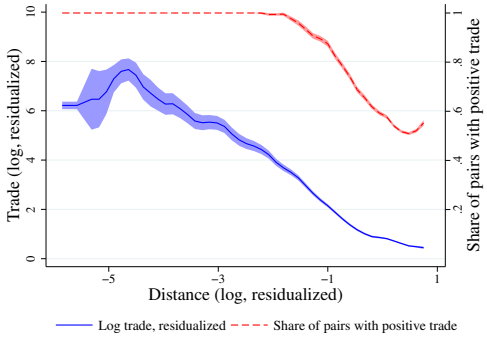
Notes: This figure shows production, consumption, and trade per capita of Medicare services across hospital referral regions (HRRs) of different sizes, all smoothed via local averages. We use the Medicare 20% carrier Research Identifiable Files to compute the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. The black series shows production of medical care per Medicare beneficiary residing in the HRR of production. The blue series shows consumption of medical care per Medicare beneficiary residing in the HRR of consumption. The dashed dark-gray series shows interregional "exports" of medical care and the dashed light-blue series shows interregional "imports" of medical care, again per Medicare beneficiary. The orange series depicts the distribution of HRR population sizes. HRR definitions are from the Dartmouth Atlas Project.

Figure 2.4: Patients travel between regions and trade declines with distance, moreso for lower-income patients

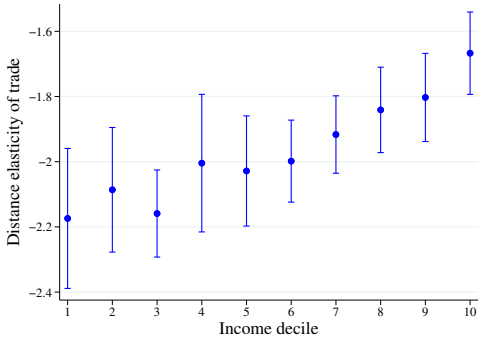
(a) Distribution of travel distances within and across HRRs



(b) Trade volume and extensive margin by distance



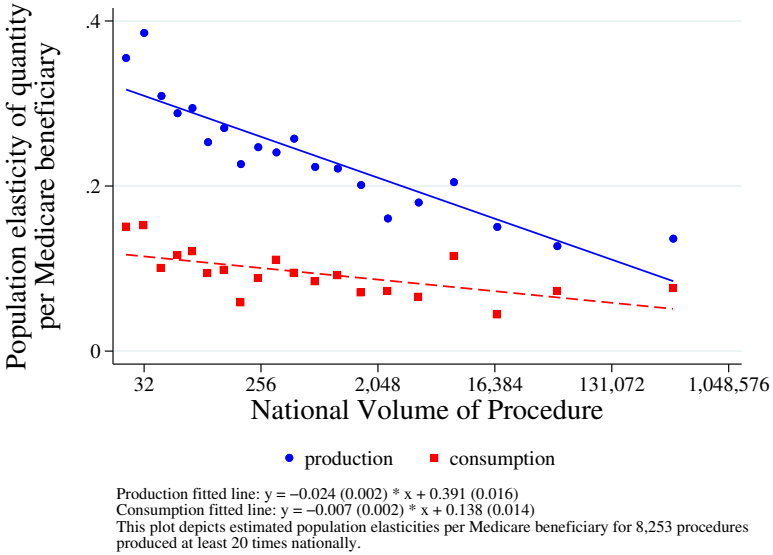
(c) Higher-income patients are less sensitive to distance



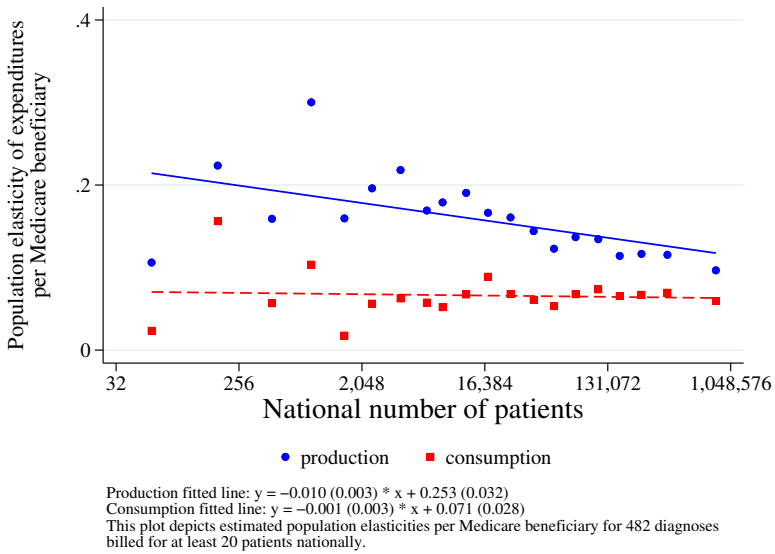
Notes: Panel a shows the distribution of patients' travel distances when patients obtain care within their home HRR (blue distribution) and when they travel across HRRs (red distribution). Travel distances measure the distance between home and treatment locations. For travel within a hospital referral region, the distance measure reflects the distance between the centroid of the patient's residential ZIP code and the ZIP code of the service location. We use ZCTA-to-ZCTA distances downloaded from the National Bureau of Economic Research; those exceeding 160 kilometers are winsorized at 160 kilometers. For travel across HRRs, we use ZCTA-to-ZCTA distances when they are within 160 kilometers and (for computational ease) use HRR-to-HRR distances beyond 160 kilometers. In Panel b, the blue series depicts the volume of trade against distance, after conditioning out the fixed effects in equation (2.12), for positive-trade pairs of locations. The red series shows the share of HRR pairs with positive trade as a function of the distance between them, after conditioning out the importer fixed effects and exporter fixed effects, as in equation (2.12). Panel c depicts the coefficient on log distance obtained by estimating equation (2.12) separately for each decile of the national ZIP-level median-household-income distribution. The 95% confidence intervals are computed using standard errors two-way clustered by both patient HRR and provider HRR. Patients from higher-income ZIP codes are less sensitive to distance.

Figure 2.5: Population elasticities of production and consumption

(a) Population elasticities by procedure frequency

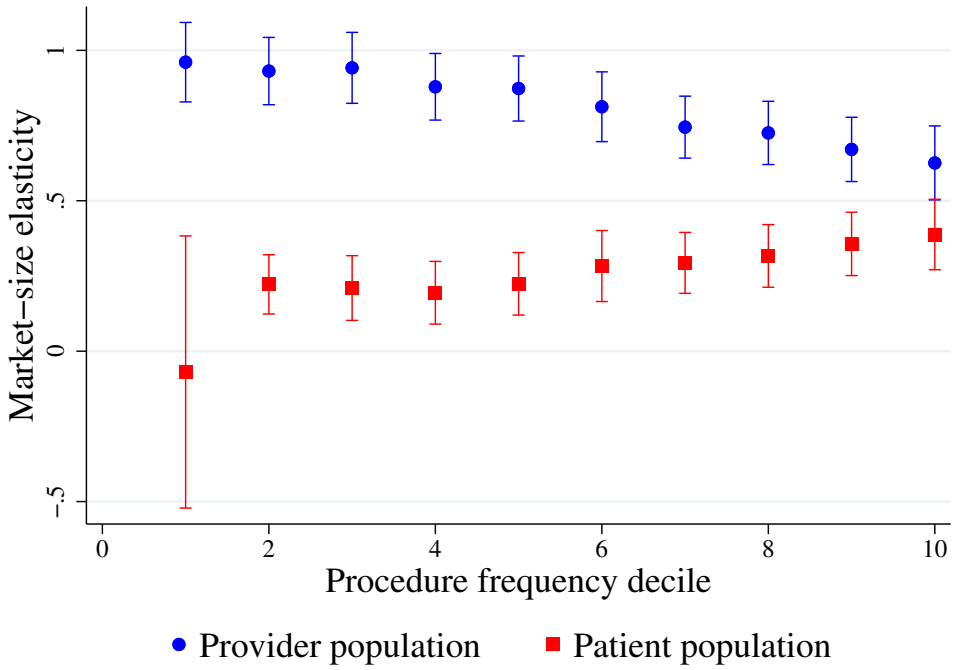


(b) Population elasticities by diagnosis frequency



Notes: The vertical axis of both panels plots the population elasticities of quantity of medical care produced and consumed per local Medicare beneficiary. The elasticities are computed using the Poisson models in equations (2.9) and (2.10) based on production location and patients' residential location, respectively. Panel a estimates these elasticities for each of the procedures provided at least 20 times nationally in the Medicare data. The horizontal axis shows the total national volume of physician services for the procedure. Panel b estimates the elasticities for care provided to treat each of the Clinical Classifications Software Refined (CCSR) diagnoses billed for at least 20 patients nationally in the Medicare data. Expenditures are computed from the Medicare 20% carrier Research Identifiable Files using the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. The horizontal axis shows the total number of patients nationally with the diagnosis. In both panels, the blue dots are a binned scatterplot of the estimated population elasticity of production per beneficiary as a function of the national volume. The red dots are the same for consumption (residential location)-based estimates. There is a significant negative relationship for production, indicating that production elasticities are highest for rare services and rare diseases. The relationship for consumption is much more modest. The difference between these estimates must be driven by trade between locations.

Figure 2.6: The home-market effect is stronger for rarer procedures

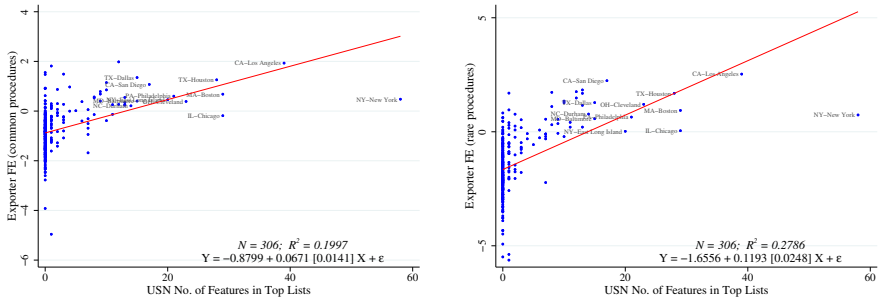


Notes: This figure groups non-emergency physician-provided services in the Medicare claims data into deciles based on the national frequency of each procedure. For each decile, we estimate equation (2.8), testing for a home market effect, and plot the estimated coefficients on provider and patient market log population with their 95% confidence intervals. The coefficients on provider-market size always exceed the respective coefficients on patient-market size, indicating a strong home-market effect. The coefficients on provider-market size monotonically decrease across the deciles. The coefficients on patient-market size monotonically increase across the deciles. Together, these two patterns show that the home-market effect is stronger the less common the procedure is, in line with the theoretical difference-in-difference prediction.

Figure 2.7: Estimated quality is positively correlated with total output and external quality metrics

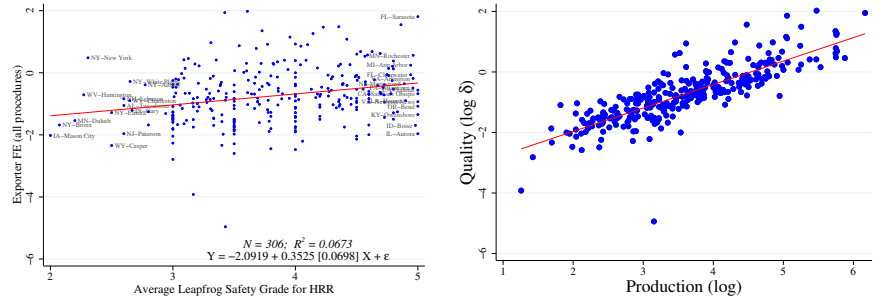
(a) *U.S. News* vs. quality estimated for common services

(b) *U.S. News* vs. quality estimated for rare services



(c) Leapfrog Safety Grade vs. estimated quality

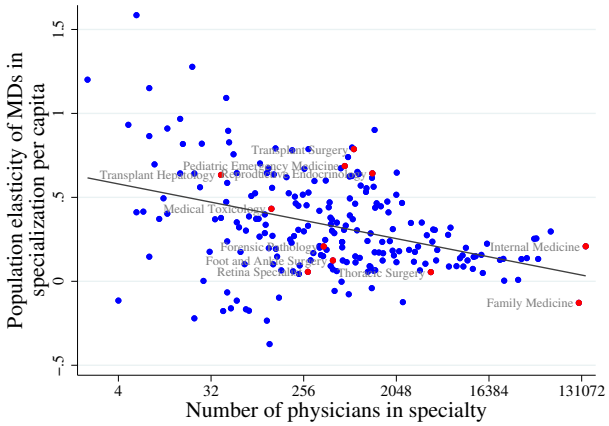
(d) Quality is higher in regions producing more output



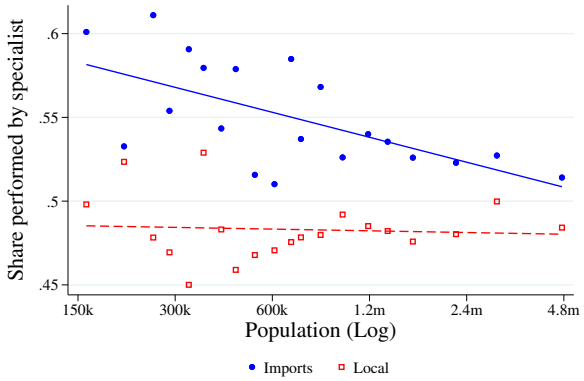
Notes: The first three panels show the relationship between the exporter fixed effects (our revealed-preference measure of quality) and external quality measures. The vertical axis shows the exporter fixed effects for each HRR estimated using trade in common services in Panel a, using trade in rare services in Panel b, and for all services in Panels c and d. The horizontal axis in Panels a and b is a count of the number of times each region's hospitals appear on the *U.S. News* list of best hospitals. *U.S. News* produces an overall ranking as well as rankings for 12 particular specialties. We count the number of times each HRR's hospitals appear on any of these 13 lists. Both panels show a positive relationship, indicating that patients travel farther to obtain care from regions highly ranked by *U.S. News*. The relationship is stronger for rare services, as the slope is nearly double that for common services. The horizontal axis in Panel c is the average safety grade for hospitals in an HRR (mapping A=5, B=4, etc.), for grades determined by the Leapfrog Group. These are positively correlated with exporter fixed effects. Panel d shows the relationship between production and the exporter fixed effects from equation (2.12), across HRRs. HRR production is measured as Medicare output produced (in millions US dollars) for non-emergency physician services in the 20% carrier file.

Figure 2.8: Imports are specialist-intensive, especially in smaller regions

(a) Population elasticities of physician specializations

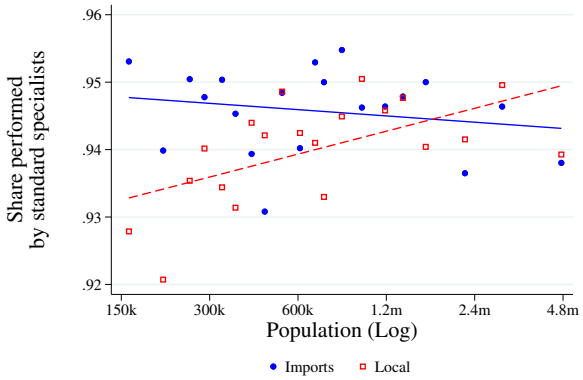


(b) Specialty care imports



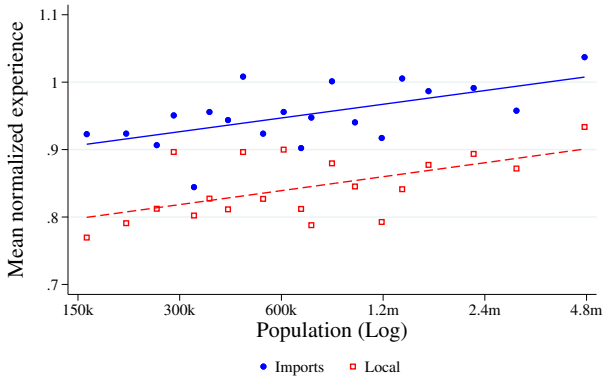
Imports: y = –0.021 (0.005) * x + 0.838 (0.071)
Local: y = –0.002 (0.004) * x + 0.509 (0.060)

(c) "Standard" specialty care



Imports: y = –0.001 (0.001) * x + 0.965 (0.018)
Local: y = 0.005 (0.002) * x + 0.876 (0.021)

(d) Provider experience

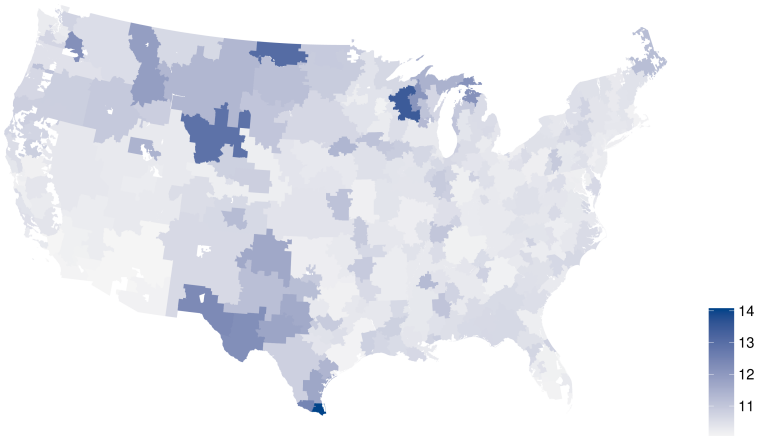


Imports: y = 0.030 (0.010) * x + 0.554 (0.137)
Local: y = 0.031 (0.014) * x + 0.427 (0.194)

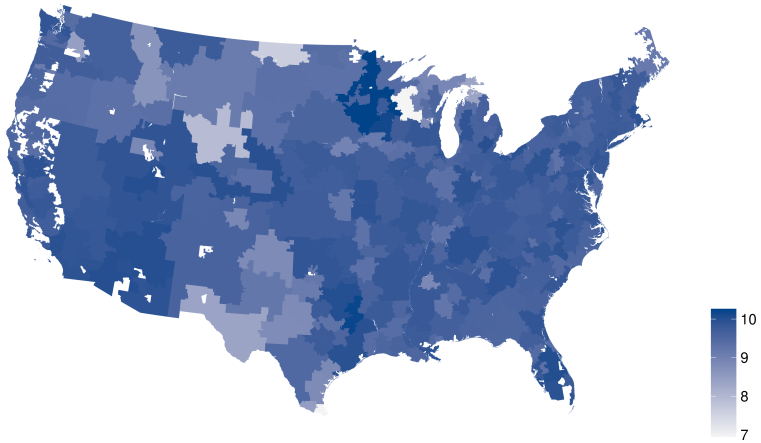
Notes: The vertical axis of Panel a depicts the population elasticities of quantity of physicians in an HRR. The population elasticities are computed for each specialty using the Poisson model in equation (2.14). The horizontal axis shows the nationwide number of physicians in each specialty. The negative relationship indicates that rare specialties are disproportionately concentrated in high-population regions. Panel b shows the share of procedures that are performed by a specialist, for imports and locally produced procedures, by market size. We define generalists as internal-medicine, general-practice, and family-practice physicians and define specialists as all other physicians. Imports are more likely to be performed by a specialist, and smaller markets' imports especially so. Panel c examines procedures that are typically performed by specialists, and classifies the "standard" specialists as the top two specialties performing the procedure nationally. It shows the shares of procedures performed by the "standard" specialties in imported specialty care and locally produced specialty care as a function of local population size. Imports are more likely to be performed by "standard" specialties, especially for smaller regions. Panel d shows the mean relative experience of providers for care produced locally and imported by population size of the patient's region. This panel describes only procedures that are performed in all hospital referral regions (143 procedures). In public-use Medicare data, we define a provider's experience for a given procedure as the number of times they performed the procedure for Traditional Medicare patients in the prior calendar year. Before aggregating to the regional level, we rescale experience in each procedure so that its mean is one. On average, patients in larger markets obtain treatment from more experienced providers. At all population levels, imported care is produced by more experienced providers than local care.

Figure 2.9: Counterfactual outcomes when reimbursements increase 10% everywhere

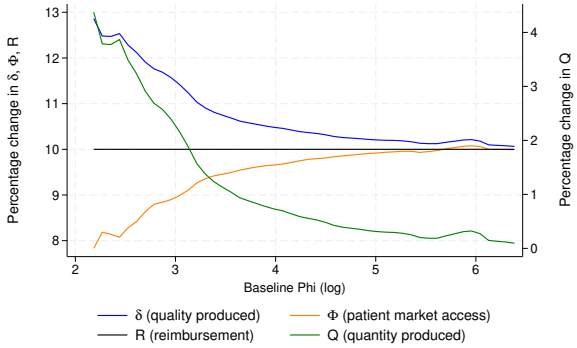
(a) Change (%) in output quality δ_i



(b) Change (%) in patient market access Φ_i



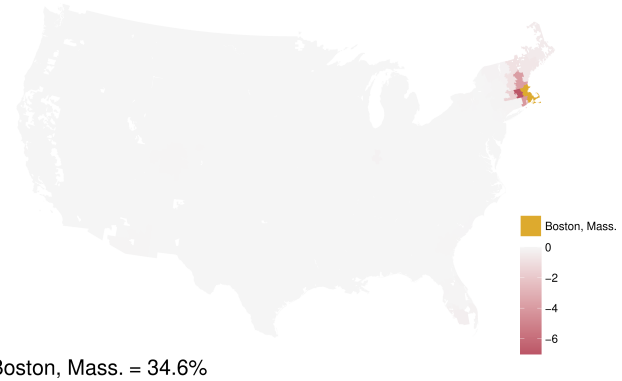
(c) Outcomes as a function of baseline patient market access Φ_i



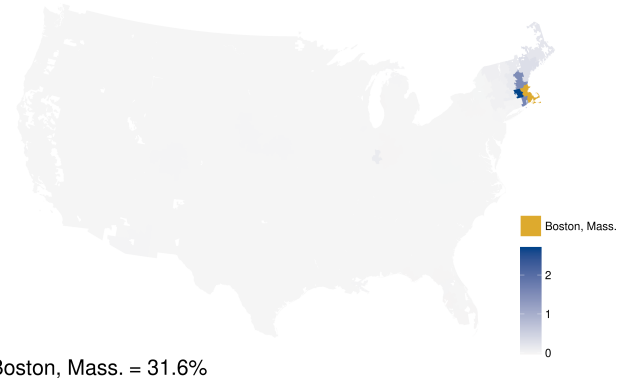
Notes: Panels a and b show the impacts of increasing reimbursements by 10% everywhere ($\hat{R}_i = 1.1\ \forall i$) based on our estimated model. Panel a depicts the percentage change in quality of care δ_i provided in each region. Panel b depicts the percentage change in the value of market access Φ_i for patients who live in a region. Panel c shows local linear regressions of the percentage changes in δ_i, Φ_i, and Q_i against the region's initial patient market access, Φ_i. There is a negative relationship between the percentage changes in δ and Φ across regions. Patients who live in the regions with the largest quality increases in δ tend to have the lowest gains in patients' market access, Φ. The exercise is described in detail in Section 2.7.

Figure 2.10: Counterfactual outcomes for higher reimbursements in one region

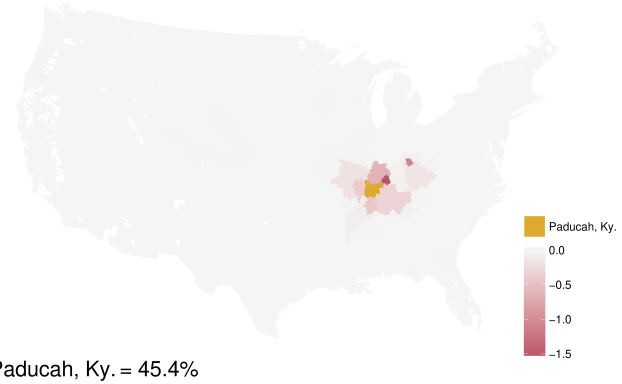
(a) Change (%) in output quality δ_i: higher reimbursement in Boston, Mass.



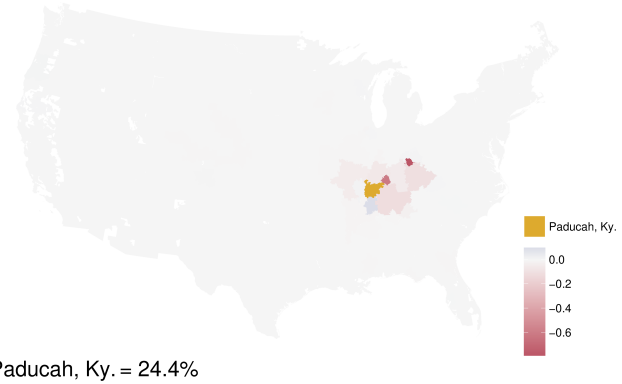
(b) Change (%) in market access Φ_i: higher reimbursement in Boston, Mass.



(c) Change (%) in output quality δ_i: higher reimbursement in Paducah, Ky.

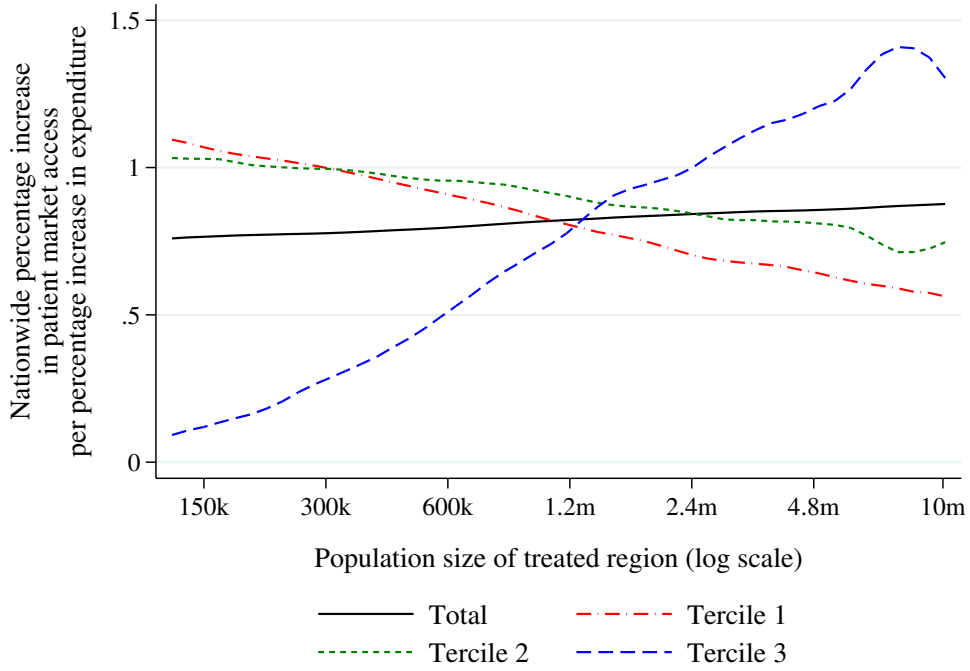


(d) Change (%) in market access Φ_i: higher reimbursement in Paducah, Ky.



Notes: Panels a and b show the impacts of increasing reimbursements by 30% in the Boston, Mass. HRR ($\hat{R}_i = 1.3$) based on our estimated model. Panel a illustrates the percentage change in quality of care δ_i provided in each region. Panel b illustrates the percentage change in the value of market access Φ_i for patients who live in an region. Panels c and d are analogous, but for a 30% increase in reimbursements in Paducah, Ky., a net importer. In all panels, the predicted change for the region whose reimbursement changes ("treated region") is listed on the map itself. In both cases, the quality produced in neighboring regions declines (Panels a and c). Patients in regions near Boston benefit from increased access to the treated region (Panel b), so there is a negative relationship between the percentage changes in δ and Φ across regions. In contrast, patients in regions near Paducah suffer a decrease in access (Panel d). The contrasting outcomes stem from Boston being a net exporter and Paducah being a net importer in the baseline equilibrium. The exercise is described in detail in Section 2.7.

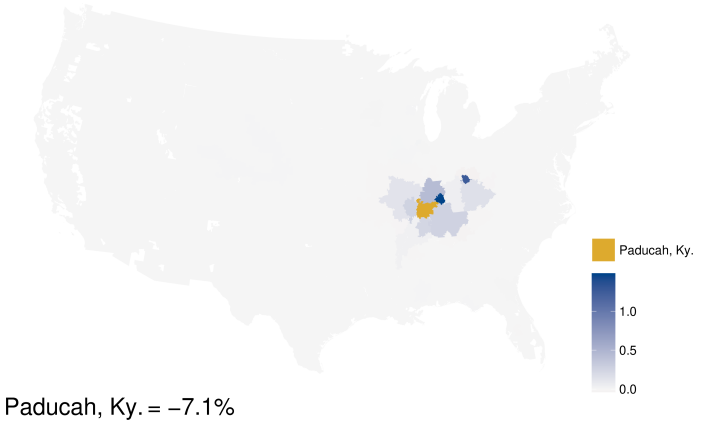
Figure 2.11: Changes in access $\hat{\Phi}_{j\kappa}$ by income when increasing reimbursements



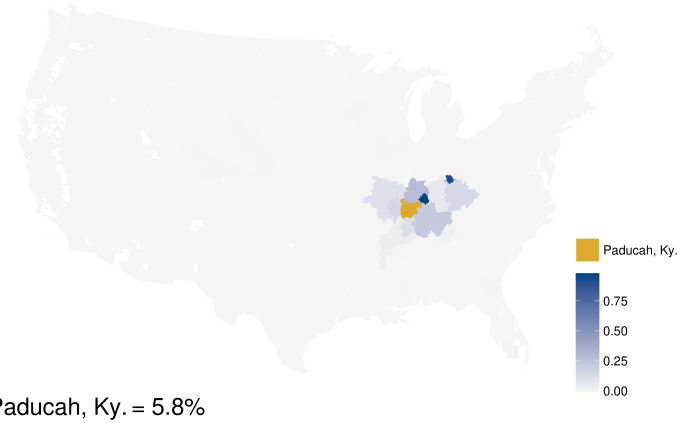
Notes: This figure summarizes the counterfactual outcomes of 30% higher reimbursements in one HRR as a function that HRR's population size. The nationwide return is the percentage increase in patient market access $\sum_\kappa \sum_j N_{j\kappa}\Phi_{j\kappa}$ per percentage increase in nationwide expenditures $\sum_i Q_i R_i$. The tercile-specific return is the increase in tercile-specific patient market access $\sum_j N_{j\kappa}\Phi_{j\kappa}$. Increasing reimbursements in more populous HRRs has the highest return when measured as impact on aggregate market access. Subsidies in less populous regions favor lower-income patients, primarily because there are more low-income patients living in and close to smaller regions.

Figure 2.12: Counterfactual outcomes when changing travel costs for Paducah, Ky. residents

(a) Change (%) in quality δ_i: reducing Paducah residents' travel costs by 30%



(b) Change (%) in access Φ_i: reducing Paducah residents' travel costs by 30%



Notes: Both panels show the impact of a 30% fall in travel costs for Paducah residents ($\hat{\rho}_{ij} = 1.3 \; \forall i \neq$ Paducah). Panel a illustrates the percentage change in quality of care δ_i provided in each region. Panel b illustrates the percentage change in the value of market access Φ_i for patients who live in a region. The note shows the change for Paducah itself. Reduced travel costs for Paducah residents improves their market access but reduces the quality of care produced in Paducah itself. The increase in imports by Paducah residents causes service quality in neighboring regions to increase because of scale effects. This higher quality in turn attracts additional patients from the ring surrounding them, reducing quality slightly in that distant ring.

Table 2.1: Aggregate medical services exhibit a strong home-market effect

Estimation method:	(1) PPML	(2) PPML	(3) PPML	(4) IV
Provider-market population (log)	0.635	0.642	0.644	0.597
	(0.0622)	(0.0605)	(0.0453)	(0.0730)
Patient-market population (log)	0.380	0.376	0.405	0.360
	(0.0605)	(0.0581)	(0.0421)	(0.0517)
Distance (log)	-1.654	0.124		0.106
	(0.0497)	(0.289)		(0.255)
Distance (log, squared)		-0.181		-0.179
		(0.0283)		(0.0250)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.017	0.011	0.002	0.017
Observations	93,636	93,636	93,636	93,636
Distance elasticity at mean		-2.46		-2.46
Distance deciles			Yes	

Notes: This table reports estimates of equation (2.8), which estimates the presence of weak or strong home-market effects. The sample is all HRR pairs ($N = 306^2$), and the dependent variable in all regressions is the value of trade. The independent variables are patient- and provider-market log population, log distance between HRRs, and an indicator for same-HRR observations ($i = j$). The positive coefficient on provider-market log population implies a weak home-market effect, and the fact that this coefficient exceeds that on patient-market population implies a strong home-market effect. Column 2 makes the distance coefficient more flexible by adding a control for the square of log distance. Column 3 replaces parametric distance specifications with fixed effects for each decile of the distance distribution. Column 4 uses the provider-market and patient-market log populations in 1940 as instruments for the contemporaneous log populations when estimating by generalized method of moments (GMM). Trade flows are computed from the Medicare 20% carrier Research Identifiable Files, using the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.2: The home-market effect is stronger for rare procedures

	(1)	(2)	(3)	(4)	(5)	(6)
λ_X Provider-market population (log)	0.635	0.622	0.621		0.629	
	(0.0622)	(0.0601)	(0.0602)		(0.0592)	
λ_M Patient-market population (log)	0.380	0.381	0.382		0.379	
	(0.0605)	(0.0580)	(0.0581)		(0.0566)	
μ_X Provider-market population (log) × rare			0.302	0.291	0.316	0.288
			(0.0468)	(0.0453)	(0.0477)	(0.0455)
μ_M Patient-market population (log) × rare			-0.225	-0.220	-0.232	-0.212
			(0.0686)	(0.0669)	(0.0703)	(0.0657)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.017	0.019	0.020		0.014	
p-value for H_0: $\mu_X \leq \mu_M$			<0.001	<0.001	<0.001	<0.001
Observations	187,272	113,468	113,468	113,468	113,468	113,468
Distance controls	Yes	Yes	Yes	Yes		
Distance [quadratic] controls					Yes	Yes
Patient-provider-market-pair FEs				Yes		Yes

Notes: This table reports estimates of equation (2.11), which introduces interactions with an indicator for whether a procedure is "rare" (provided less often than the median procedure, when adding up all procedures provided nationally). The interactions with patient- and provider-market population reveal whether the home-market effect is larger for rare procedures. The unit of observation is {rare indicator, exporting HRR, importing HRR} so the number of observations is 2×306^2 in column 1, and the dependent variable in all regressions is the value of trade. Columns 2 onwards drop HRR pairs with zero trade in both procedure groups, and column 2 shows that this restriction has a negligible impact on the estimated log population coefficients. Columns 3 onwards include the rare indicator interacted with patient- and provider-market populations and distance covariates. Columns 1–4 control for distance using the log of distance between HRRs. Columns 5 and 6 add a control for the square of log distance. Columns 4 and 6 introduce a fixed effect for each ij pair of patient market and provider market, so these omit all covariates that are not interacted with the rare indicator. The positive coefficient on provider-market population × rare across all columns indicates that the home-market effect is stronger for rare than for common services. The negative coefficient on patient-market population × rare across all columns indicates that the *strong* home-market effect has a larger magnitude for rare services. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files, using the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.3: The stronger home-market effect for rare procedures is robust to instrumenting for population

	(1)	(2)	(3)	(4)	(5)	(6)
Geography:	HRR	HRR	CBSA	CBSA	CBSA	CBSA
Instrument:	1940 pop	1940 pop	1940 pop	1940 pop	Bedrock	Bedrock
Procedure Sample:	Common	Rare	Common	Rare	Common	Rare
Provider-market population (log)	0.595	1.081	0.716	0.895	1.157	1.753
	(0.0731)	(0.0914)	(0.0249)	(0.0388)	(0.307)	(0.524)
Patient-market population (log)	0.362	0.0477	0.396	0.328	0.182	-0.582
	(0.0518)	(0.115)	(0.0261)	(0.0344)	(0.373)	(0.580)
Distance (log)	0.102	0.992	-3.412	-1.378	-4.678	-4.631
	(0.255)	(0.442)	(0.294)	(0.989)	(1.049)	(2.520)
Distance (log, squared)	-0.179	-0.263	0.105	-0.0742	0.210	0.181
	(0.0251)	(0.0497)	(0.0287)	(0.0935)	(0.0845)	(0.199)
Observations	93,636	93,636	857,476	857,476	781,456	781,456
Distance elasticity at mean	-2.46	-2.77	-1.91	-2.43	-1.68	-2.05

Notes: This table reports estimates of equation (2.8), when separating procedures into those above- and below-median frequency and instrumenting for log population. The dependent variable in all regressions is the value of trade. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files, using the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. We report coefficients on provider market population, patient market population, log distance, and log distance squared. Every specification also includes a same-market ($i = j$) indicator variable. The odd-numbered columns are trade in above-median-frequency procedures; the even-numbered columns are trade in below-median-frequency procedures. In columns 1 and 2, the sample is all HRR pairs ($N = 306^2$). In columns 3 and 4, the sample is all CBSA pairs ($N = 926^2$). In columns 5 and 6, the sample is all CBSA pairs for which the bedrock-depth instrumental variable is available ($N = 844^2$). We use 1940 population counts to produce two instrumental variables: 1940 population in the patient market and 1940 population in the provider market are instruments for log population in the patient market and log population in the provider market, respectively. Similarly, we use bedrock depth to produce two instrumental variables for CBSAs. Both the strong home-market effect and its larger magnitude for rare procedures are robust to instrumenting for population, estimating by GMM. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.4: The home-market effect is stronger for rarer diagnoses

	(1)	(2)	(3)	(4)	(5)	(6)
λ_X Provider-market population (log)	0.635	0.622	0.616		0.624	
	(0.0625)	(0.0604)	(0.0588)		(0.0578)	
λ_M Patient-market population (log)	0.382	0.383	0.386		0.383	
	(0.0606)	(0.0580)	(0.0569)		(0.0555)	
μ_X Provider-market population (log) $\times$ rare			0.0719	0.0687	0.0763	0.0683
			(0.0547)	(0.0519)	(0.0561)	(0.0506)
μ_M Patient-market population (log) $\times$ rare			-0.0422	-0.0409	-0.0429	-0.0380
			(0.0419)	(0.0403)	(0.0440)	(0.0395)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.018	0.020	0.021		0.015	
p-value for H_0: $\mu_X \leq \mu_M$			0.114	0.113	0.113	0.115
Observations	187,272	112,626	112,626	112,626	112,626	112,626
Distance controls	Yes	Yes	Yes	Yes		
Distance [quadratic] controls					Yes	Yes
Patient-provider-market-pair FEs				Yes		Yes

Notes: This table augments equation (2.8) by adding interactions with an indicator for whether a diagnosis is "rare" (provided less often than the median diagnosis, when adding up all patients receiving the diagnosis nationally) or "common" (more often than median). The interactions with patient- and provider-market population reveal whether the home-market effect is larger for rare diagnoses. The unit of observation is {rare indicator, exporting HRR, importing HRR} so the number of observations is 2×306^2 in column 1, and the dependent variable in all regressions is the value of trade. Valid primary diagnoses observed in 1,000 distinct claims or more nationally in the professional fees 20% sample are included. Columns 2 onwards drop HRR pairs with zero trade, and column 2 shows that this restriction has a negligible impact on the estimated log population coefficients. Columns 1–4 control for distance using the log of distance between HRRs. Columns 5 and 6 add a control for the square of log distance. Columns 4 and 6 introduce a fixed effect for each ij pair of patient market and provider market, so these omit the patient- and provider-market population covariates. The positive coefficient on provider-market population $\times$ rare across all columns indicates that the home-market effect is stronger for rare than for common diagnoses. The negative coefficient on patient-market population $\times$ rare across all columns indicates that the *strong* home-market effect is especially true for rare diagnoses. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files, using the dollar value of physician services, excluding emergency-room care and assigning each procedure its national average price. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.5: Scale elasticity estimates

Panel A: All services	Baseline	No Diagonal	Controls
OLS	0.776	0.803	0.810
	(0.031)	(0.045)	(0.041)
2SLS: population (log)	0.712	0.798	0.724
	(0.031)	(0.050)	(0.039)
2SLS: population (1940, log)	0.533	0.663	0.545
	(0.072)	(0.098)	(0.067)
Panel B: Rare services			
OLS	0.947	1.083	0.927
	(0.030)	(0.045)	(0.035)
2SLS: population (log)	0.912	1.026	0.881
	(0.037)	(0.049)	(0.049)
2SLS: population (1940, log)	0.835	0.951	0.789
	(0.061)	(0.084)	(0.070)

Notes: This table reports estimates of $\hat{\alpha}$ from ordinary least squares (OLS) or two-stage least squares (2SLS) regressions of the form $\widehat{\ln \delta}_i = \alpha \ln Q_i + \ln R_i + u_i$, where $\widehat{\ln \delta}_i$ is estimated in equation (2.12), Q_i is region i's total production of non-emergency-room physician services for Medicare beneficiaries, R_i is Medicare's Geographic Adjustment Factor, and u_i is an error term. In the rows labeled "2SLS" we instrument for $\ln Q_i$ using the specified instruments. The $\ln R_i$ control is omitted in the columns labeled "no controls". Appendix Table 2.17 reports analogous estimates at the CBSA level, which allows us to also control for input costs (as input cost data are more reliable for CBSAs than for HRRs). In the columns labeled "no diag", Q_{ii} observations were omitted when estimating $\widehat{\ln \delta}_i$ in equation (2.12). Standard errors are robust to heteroskedasticity. Across all of the permutations of our method, we estimate substantial scale economies.

Table 2.6: Regression of $\hat{\Phi}_{j\mathfrak{k}}$ on tercile dummies and trade shares

	Nationwide Reimbursement Increase			Boston Reimbursement Increase		
	(1)	(2)	(3)	(4)	(5)	(6)
Income tercile = 2	1.080	1.085	-0.143	2.00e-05	0.00298	-0.00494
	(0.0554)	(0.0555)	(0.0318)	(0.00938)	(0.00966)	(0.00214)
Income tercile = 3	1.568	1.553	-0.240	0.0697	0.0649	-0.00732
	(0.0712)	(0.0698)	(0.0549)	(0.0349)	(0.0347)	(0.00266)
Imported share $(1 - m_{0j\kappa} - m_{jj\kappa})$			-0.519			
			(0.129)			
$m_{0j\kappa}$			-12.65			
			(0.422)			
$m_{\text{Boston},j\kappa}$						36.95
						(0.435)
Constant	8.763	8.767	11.10	0.0984	0.0989	-0.0691
	(0.0594)	(0.0403)	(0.0769)	(0.105)	(0.0127)	(0.00265)
Observations	885	885	885	885	885	885
R-squared	0.498	0.675	0.988	0.000	0.980	1.000
HRR fixed effects	No	Yes	Yes	No	Yes	Yes

Notes: This table uses linear regressions to summarize how market access changes across HRRs j and income terciles κ in response to two different counterfactual policies. The dependent variable in all columns is the percentage change in market access, $100 \times (\hat{\Phi}_{j\mathfrak{k}} - 1)$. Standard errors (in parentheses) are clustered by market. Columns 1, 2, and 3 consider a 10% reimbursement increase nationwide. Columns 4, 5, and 6 consider a 30% reimbursement increase in Boston only. The constant in the first regression reports the percentage change for the lowest income terciles, and the coefficients on the other terciles are the additional percentage point gain for those terciles relative to the lowest. Other controls include the outside option market share $m_{0j\kappa}$, imported share $1 - m_{0j\kappa} - m_{jj\kappa}$ (where $m_{jj\kappa}$ is local production), and Boston's market share $m_{\text{Boston},j\kappa}$. The coefficients are much smaller in columns 4, 5, and 6 because only Boston is treated, so most of the country is hardly affected. When we add market share controls, the coefficients indicating tercile differences become much smaller, indicating that baseline trade patterns drive the distributional impacts.

2.10 Theory appendix

2.10.1 Monopolistic competition with one firm per region

Suppose that there is a single firm in each region that offers fixed-price services to patients under monopolistic competition with the firms in other regions. Assume $K(\delta_i) = \delta_i$ and $H(Q_i) = Q_i^\alpha$. The profit-maximizing choice of quality δ_i by the firm in region i is

$$\max_{\delta_i} \pi_i = \left(\bar{R} - \frac{w_i \delta_i}{A_i Q_i^\alpha}\right) Q_i \quad \text{where } Q_i = \sum_j Q_{ij} = \delta_i \sum_j \frac{N_j}{\Phi_j} \rho_{ij}$$
$$\frac{\partial \pi_i}{\partial \delta_i} = 0 \implies \frac{\bar{R}}{2-\alpha} = \frac{w_i \delta_i}{A_i Q_i^\alpha} = C(Q_i, \delta_i; w_i, A_i)$$

This expression replaces the free-entry condition (2.4) in the definition of equilibrium. Changing the value of the constant on the left side of this equality does not change any of the subsequent theoretical predictions. In this respect, the monopolistic-competition model with one firm per region is isomorphic to the perfect-competition model with external economies of scale.

2.10.2 Model with multiple types of patients

This section extends the model to feature multiple types of patients who face different trade costs. There is a finite set of patient types, which are indexed by κ. A patient type is defined by the trade costs $\rho_{ij(k)} = \rho_{ij}^\kappa \forall k \in \kappa$. Qualities δ_i, including the outside option δ_0, are the same for all patient types. The demand by patients of type κ residing in location j for procedures performed by providers in location i is now given by

$$Q_{ij}^\kappa = \frac{\delta_i N_j^\kappa}{\Phi_j^\kappa} \rho_{ij}^\kappa.$$

The aggregate gravity equation is the sum of type-specific gravity equations:

$$Q_{ij} = \sum_{\kappa} Q_{ij}^{\kappa} = \delta_i \sum_{\kappa} \frac{N_j^{\kappa}}{\Phi_j^{\kappa}} \rho_{ij}^{\kappa}. \tag{2.15}$$

The free-entry condition (2.4) remains unchanged with the introduction of multiple patient types:

$$R_i = \frac{w_i \delta_i}{A_i Q_i^{\alpha}}.$$

In equilibrium, market clearing requires that

$$Q_i = \left(\frac{w_i \delta_i}{A_i R_i}\right)^{1/\alpha} = \delta_i \sum_j \sum_{\kappa} \frac{N_j^{\kappa}}{\Phi_j^{\kappa}} \rho_{ij}^{\kappa} \implies \delta_i = \left(\frac{A_i R_i}{w_i}\right)^{1/(1-\alpha)} \left(\sum_j \sum_{\kappa} \frac{N_j^{\kappa}}{\Phi_j^{\kappa}} \rho_{ij}^{\kappa}\right)^{\alpha/(1-\alpha)}.$$

2.10.3 Derivations of results in Section 2.2.5

Abusing notation so that $\mathcal{I}$ is both the set and number of regions, equations (2.2) and (2.3) together constitute $2\mathcal{I}$ equations with $2\mathcal{I}$ unknowns. For the special case of $H(Q_i) = Q_i^{\alpha}$ and $K(\delta_i) = \delta_i$, this reduces to the following $\mathcal{I}$ equations with the unknowns $\{\delta_i\}_{i=1}^{\mathcal{I}}$:

$$\delta_i = \left(\frac{\overline{R} A_i}{w_i}\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in \mathcal{I}} \frac{\rho_{ij}}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'} \rho_{i'j}} N_j\right)^{\frac{\alpha}{1-\alpha}}$$

Following Costinot et al. [2019], we examine the home-market effect in the neighborhood of a symmetric equilibrium. For brevity, assume $\frac{\overline{R} A_i}{w_i} = 1 \ \forall i$. Note that at the symmetric equilibrium:

$$\bar{\delta}^{\frac{1-\alpha}{\alpha}} = \frac{1}{1 + \bar{\delta} + \sum_{i' \neq i} \bar{\delta} \rho} \bar{N} + \sum_{j \neq i} \frac{\rho}{1 + \bar{\delta} + \sum_{i' \neq j} \bar{\delta} \rho} \bar{N} = \frac{1 + (\mathcal{I} - 1)\rho}{\bar{\Phi}} \bar{N} = \frac{\bar{\Phi} - 1}{\bar{\Phi}} \frac{\bar{N}}{\bar{\delta}}. \tag{2.16}$$

Given $\alpha > 0$, totally differentiating the above system of equations in terms of $\{\mathrm{d}\delta_i, \mathrm{d}N_i\}_{i=1}^{\mathcal{I}}$

and evaluating it at the symmetric equilibrium yields the following expression:

$$\frac{\bar{\Phi}^2}{\bar{N}}\frac{1-\alpha}{\alpha}\bar{\delta}^{\frac{1-2\alpha}{\alpha}}\mathrm{d}\delta_i = -\left[\mathrm{d}\delta_i + \rho\sum_{i'\neq i}\mathrm{d}\delta_{i'}\right] + \bar{\Phi}\frac{\mathrm{d}N_i}{\bar{N}} + \sum_{j\neq i} -\rho\left[\mathrm{d}\delta_j + \rho\sum_{i'\neq j}\mathrm{d}\delta_{i'}\right] + \sum_{j\neq i}\rho\bar{\Phi}\frac{\mathrm{d}N_j}{\bar{N}}.$$

Given $\mathrm{d}N_1 > 0$ and $\mathrm{d}N_j = 0\ \forall j \neq 1$, we obtain the following expression for $\mathrm{d}\ln\delta_1$:

$$\mathrm{d}\ln\delta_1 = \frac{\frac{\bar{\Phi}}{\bar{\delta}}\mathrm{d}\ln N_1 - (\mathcal{I}-1)(2\rho + ((\mathcal{I}-2)\rho^2))\mathrm{d}\ln\delta_{j\neq 1}}{\frac{\bar{\Phi}^2}{\bar{N}}\frac{(1-\alpha)}{\alpha}\bar{\delta}^{\frac{1-2\alpha}{\alpha}} + 1 + (\mathcal{I}-1)\rho^2}. \tag{2.17}$$

Further tedious algebra delivers the following expression for quality changes:

$$\mathrm{d}\ln\delta_1 - \mathrm{d}\ln\delta_{j\neq 1} = \frac{(1-\rho)}{\frac{\bar{\Phi}^2}{\bar{N}}\frac{(1-\alpha)}{\alpha}\bar{\delta}^{\frac{1-2\alpha}{\alpha}} + (1-\rho)^2}\frac{\bar{\Phi}}{\bar{\delta}}\mathrm{d}\ln N_1. \tag{2.18}$$

Equation (2.16) implies that $\frac{\bar{\Phi}^2}{\bar{N}}\frac{(1-\alpha)}{\alpha}\bar{\delta}^{\frac{1-2\alpha}{\alpha}} = \left(\frac{1-\alpha}{\alpha}\right)\frac{\bar{\Phi}(\bar{\Phi}-1)}{\bar{\delta}}$ and therefore

$$\begin{aligned}\mathrm{d}\ln\delta_1 - \mathrm{d}\ln\delta_{j\neq 1} &= \frac{(1-\rho)}{\left(\frac{1-\alpha}{\alpha}\right)\frac{\bar{\Phi}(\bar{\Phi}-1)}{\bar{\delta}} + (1-\rho)^2}\frac{\bar{\Phi}}{\bar{\delta}}\mathrm{d}\ln N_1 \\ &= \left[\frac{1-\alpha}{\alpha}\frac{(\bar{\Phi}-1)}{(1-\rho)\bar{\delta}} + \frac{(1-\rho)\bar{\delta}}{\bar{\Phi}}\right]^{-1}\mathrm{d}\ln N_1 > 0.\end{aligned}$$

The last expression above is reported in Section 2.2.5.

Prior to deriving the weak and strong home-market effects, we obtain an expression for $\frac{\mathrm{d}\ln\delta_j}{\mathrm{d}\ln N_1}$ for $j \neq 1$ around the symmetric equilibrium. Define $\bar{\mathcal{Q}} \equiv \frac{\bar{\Phi}^2}{\bar{N}}\frac{(1-\alpha)}{\alpha}\bar{\delta}^{\frac{1-2\alpha}{\alpha}} > 0$. Combining the expressions for $\mathrm{d}\ln\delta_1$ from equation (2.17) and for $\mathrm{d}\ln\delta_1 - \mathrm{d}\ln\delta_{j\neq 1}$ from equation (2.18) yields the following:

$$\frac{\mathrm{d}\ln\delta_{j\neq 1}}{\mathrm{d}\ln N_1} = \frac{\bar{\Phi}}{\bar{\delta}}\frac{\bar{\mathcal{Q}}\rho + \rho^3(\mathcal{I}-1) - \rho^2(\mathcal{I}-2) - \rho}{(\bar{\mathcal{Q}} + (1-\rho)^2)(\bar{\mathcal{Q}} + 1 + \rho^2 + 2\rho(\mathcal{I}-1) + \mathcal{I}\rho^2(\mathcal{I}-2))}$$

The weak home-market effect is derived as follows:

$$
\begin{aligned}
\ln Q_{1,j\neq 1} &= \alpha \ln Q_1 + \ln \rho - \ln \Phi_j + \ln N_j \\
\frac{\mathrm{d}\ln Q_{1,j\neq 1}}{\mathrm{d}\ln N_1} &= \alpha \frac{\mathrm{d}\ln Q_1}{\mathrm{d}\ln N_1} - \frac{\alpha}{\Phi_j}\left(\rho Q_1^{\alpha-1}\frac{\mathrm{d}Q_1}{\mathrm{d}\ln N_1} + Q_j^{\alpha-1}\frac{\mathrm{d}Q_j}{\mathrm{d}\ln N_1} + \rho\sum_{i'\neq 1,j} Q_{i'}^{\alpha-1}\frac{\mathrm{d}Q_{i'}}{\mathrm{d}\ln N_1}\right) \\
&= \frac{\mathrm{d}\ln \delta_1}{\mathrm{d}\ln N_1} - \frac{1}{\Phi_j}\left(\rho\delta_1\frac{\mathrm{d}\ln \delta_1}{\mathrm{d}\ln N_1} + \delta_j\frac{\mathrm{d}\ln \delta_j}{\mathrm{d}\ln N_1} + \rho\sum_{i'\neq 1,j}\delta_{i'}\frac{\mathrm{d}\ln \delta_{i'}}{\mathrm{d}\ln N_1}\right) \\
&= \left(\frac{\bar{N}-Q_{1j}}{\bar{N}}\right)\frac{\mathrm{d}\ln \delta_1}{\mathrm{d}\ln N_1} - \left(\frac{\bar{N}-Q_{0j}-Q_{1j}}{\bar{N}}\right)\frac{\mathrm{d}\ln \delta_j}{\mathrm{d}\ln N_1} \\
&= \left(\frac{\bar{N}-Q_{1j}}{\bar{N}}\right)\left[\frac{\mathrm{d}\ln \delta_1}{\mathrm{d}\ln N_1} - \frac{\mathrm{d}\ln \delta_j}{\mathrm{d}\ln N_1}\right] + \frac{Q_{0j}}{\bar{N}}\frac{\mathrm{d}\ln \delta_j}{\mathrm{d}\ln N_1} \\
&= \frac{\Phi}{\bar{\delta}\bar{N}}\frac{1}{\bar{\mathcal{Q}}+(1-\rho)^2}\Big[(Q_{jj}+(\mathcal{I}-2)Q_{1j})(1-\rho) \\
&+ \frac{Q_{0j}}{\bar{\mathcal{Q}}+1+\rho^2+2\rho(\mathcal{I}-1)+\mathcal{I}\rho^2(\mathcal{I}-2)} \\
&\quad\times\left\{\bar{\mathcal{Q}}+(\rho-1)^2+2(\mathcal{I}-1)(\rho-\rho^2)+(\mathcal{I}-1)(\mathcal{I}-2)[\rho^2-\rho^3]\right\}\Big] \\
&> 0.
\end{aligned}
$$

The condition for the strong home-market effect is derived as follows:

$$
\begin{aligned}
Q_{1,j\neq 1} - Q_{j\neq 1,1} &= \frac{Q_1^\alpha \rho}{1+Q_1^\alpha\rho+Q_j^\alpha+\sum_{i\neq 1,j}Q_i^\alpha\rho}N_j \\
&\quad - \frac{Q_j^\alpha\rho}{1+Q_1^\alpha+Q_j^\alpha\rho+\sum_{i\neq 1,j}Q_i^\alpha\rho}N_1 \\
\mathrm{d}\ln Q_{1,j\neq 1} - \mathrm{d}\ln Q_{j\neq 1,1} &= \mathrm{d}\ln N_j - \mathrm{d}\ln N_1 + \alpha\left[1+(1-\rho)\frac{\bar{Q}^\alpha}{\bar{\Phi}}\right]\left(\mathrm{d}\ln Q_1 - \mathrm{d}\ln Q_j\right) \\
&= -\mathrm{d}\ln N_1 + \left[1+(1-\rho)\frac{\bar{\delta}}{\bar{\Phi}}\right]\left(\mathrm{d}\ln \delta_1 - \mathrm{d}\ln \delta_j\right) \\
&= \left[\frac{1-\frac{1-\alpha}{\alpha}\frac{1+(\mathcal{I}-1)\rho}{1-\rho}}{\frac{1-\alpha}{\alpha}\frac{(1+(\mathcal{I}-1)\rho)}{(1-\rho)}+\frac{(1-\rho)\bar{\delta}}{1+(1+(\mathcal{I}-1)\rho)\bar{\delta}}}\right]\mathrm{d}\ln N_1.
\end{aligned}
$$

There is a strong home-market effect in the neighborhood of the symmetric equilibrium

if and only if $\mathrm{d}\ln Q_{1,j\neq 1} - \mathrm{d}\ln Q_{j\neq 1,1} > 0$.

$$\left[\frac{1-\frac{1-\alpha}{\alpha}\frac{1+(\mathcal{I}-1)\rho}{1-\rho}}{\frac{1-\alpha}{\alpha}\frac{(1+(\mathcal{I}-1)\rho)}{(1-\rho)}+\frac{(1-\rho)\bar{\delta}}{1+(1+(\mathcal{I}-1)\rho)\bar{\delta}}}\right]\mathrm{d}\ln N_1 > 0 \iff \frac{\alpha}{1-\alpha} > \frac{1+(\mathcal{I}-1)\rho}{1-\rho}$$

This is true if α is large enough and ρ is small enough.

Our difference-in-differences prediction concerns how the effect of market size on net exports varies with the number of potential patients $\bar{N}$. Given the scale elasticity α and (inverse) trade costs ρ, the denominator of the right side of equation (2.6) is increasing in the symmetric-equilibrium quality $\bar{\delta}$. For two procedures that both exhibit a strong home-market effect because they have the same scale elasticity and trade costs, the effect of population size on net exports will be larger for the procedure with lower service quality. The symmetric-equilibrium service quality is increasing in the number of potential patients $\bar{N}$ because there are increasing returns (see equation (2.16)). Thus, in the neighborhood of the symmetric equilibrium, the strength of a strong home-market effect is decreasing in the number of potential patients.

2.11 Data appendix

2.11.1 Procedure frequency in main sample compared with aggregate and private data

Medicare provides two public-use files based on 100 percent claims. The first one contains the complete count of procedures billed by HCPCS code but does not have information about providers. We use it to confirm that procedure counts based on the confidential data do not suffer substantial sampling bias. In Figure 2.21, we split procedure codes into deciles based on their national frequencies, separately in the confidential and public datasets. This generates a 100-cell matrix by decile pair. We plot the share of procedures in each cell in

this matrix to determine how well the two datasets align. The vast majority of the codes are on the diagonal, with almost all of the remainder adjacent to the diagonal. This suggests that sampling error is not causing us to mischaracterize procedure frequency.

Medicare provides a second public file at the level of physician-by-procedure (HCPCS code). This summary does not contain any patient-level information so cannot be used to study trade flows, but we can use it to replicate analyses based on the location of production and physician experience. This file is censored such that physician-by-procedure pairs with 10 or fewer observations per year are suppressed, which makes for a more complicated bias than simple 20 percent random sampling. Nevertheless, all of the results that can be tested on this sample confirm those found in the 20 percent sample.

Since our procedure frequency measures rely on Medicare data, we would mismeasure frequency if the Medicare population uses a substantially different composition of care from the broader population. For example, childbirth is less common among Medicare beneficiaries. So our frequency measures may not capture the true national frequency of a procedure.

We address this by comparing procedure frequencies between the Medicare public data and private data from the Health Care Cost Institute (HCCI). The HCCI data contain claims for about 55 million privately insured patients (about 35% of individuals with employer-based insurance). We only consider HCPCS codes performed on at least eleven patients in the HCCI data. Note that frequencies are computed for all providers here, not only MDs and DOs. The authors acknowledge the assistance of the Health Care Cost Institute (HCCI) and its data contributors, Aetna, Humana, and Blue Health Intelligence, in providing the claims data analyzed in this section.

We examine whether procedures classified as above median frequency in one dataset are above median frequency in the other dataset. Table 2.18 shows that 88% of the services above median frequency in Medicare are also as above median frequency in the HCCI data. Similarly, 82% of the services below median frequency in Medicare are also below median

frequency in the HCCI data.

We next compare classifications of procedures' frequency deciles in Figure 2.22. Analogous to Figure 2.21, this plot visualizes the share of procedures which fall into each of pair of frequency decile bins in HCCI and Medicare data. The two classifications appear to coincide relatively well, with slightly stronger agreement for very frequent procedures compared to rarer procedures in the Medicare public-use data. Overall, the frequency classifications of procedures coincide well between Medicare public-use data and HCCI data.

2.11.2 Additional details on data sources

Physician earnings. The Gottlieb et al. [2020] earnings data depicted in Appendix Figure 2.14 are only available for 111 commuting zones. The American Community Survey (ACS) covers far more CBSAs, but this source top-codes income for a substantial share of doctors.

U.S. News and World Report. The publication produces an overall ranking and rankings for 12 particular specialties. We count the number of times each HRR's hospitals appear on any of these 13 lists.[54] Thus, higher ranking on the horizontal axis indicates a region has some combination of more top-ranked hospitals, or each of its hospitals performs well in many specialty areas.

2.11.3 Geographic price adjustments

Professional fees. To adjust for geographic price variation in the professional fees, we compute a national average price per Healthcare Common Procedure Coding System (HCPCS) code as the sum of the line allowed amount, which includes the line item's Medicare-paid and beneficiary-paid amounts (i.e., deductible, copayment, and coinsurance), divided by the

54. Results are similar when we use other methods to aggregate the rankings information, including when we account for the ordered nature of the lists.

sum of the line service count per HCPCS code nationally. We then apply this average price to all billing for the HCPCS code when computing total spending across services.

Hospital inpatient fees. We use the field "final standard payment amount" in the MedPAR file, which is computed as described in Finkelstein et al. [2016] and Gottlieb et al. [2010b]. This represents "a standard Medicare payment amount, without the geographical payment adjustments and some of the other add-on payments that go to the hospitals" according to the data documentation.

Hospital outpatient fees. To adjust for geographic price variation in hospital outpatient fees, we compute a national average price per Healthcare Common Procedure Coding System (HCPCS) code, Ambulatory Payment Classifications (APC) code, and revenue center code. HCPCS codes reflect the procedure performed and APC codes reflect a prospective payment system applicable to outpatient analogous to Diagnosis Related Groups (DRGs) for inpatient claims. Revenue center contains information on the place of service, e.g. rehabilitation or acute care, so we consider two procedures performed in different revenue centers as different procedures for price adjustment purposes.

The total amount per claim line is calculated as the sum of the claim (Medicare) payment amount, the primary payer amount, the Part B beneficiary co-insurance amount, the beneficiary Part B deductible amount, and the beneficiary blood deductible amount. These amounts are summed nationally for each {HCPCS code, APC code, revenue center code} triplet, and divided by the frequency of that triplet to obtain a national average price. We then apply this average price to all instances of that {HCPCS code, APC code, revenue center code} combination when computing total spending across services.

2.11.4 Residential measurement error

This appendix uses two methods to investigate potential measurement error in patients' residential location. The first source of potential error is "snowbird" patients, who have multiple residences and therefore may appear to travel farther than they actually do. They may need medical care while spending months in a warmer HRR that is not the one listed as their main residence (or vice versa). Our results are robust to two methods of removing potential snowbirds: excluding Arizona, California, and Florida, following Finkelstein et al. [2016], and excluding the 10% of HRRs with the highest share of second homes in American Community Survey data. These results are in Tables 2.10 and 2.11. The results are little changed by these sample restrictions.

We test for more general location measurement error by examining how far patients appear to travel for dialysis. Since Medicare patients requiring dialysis must generally visit a dialysis center thrice weekly, they are unlikely to go substantial distances for this service. Table 2.12 compares travel distances for dialysis with other care. Dialysis patients appear to travel less than one-quarter as often as other patients—and even less when excluding snowbird states—suggesting that our residential location assignment is largely accurate.

2.11.5 Scale elasticity estimation with unobserved market segments

Our data only contain procedure-level production and consumption in Traditional Medicare (TM), not for Medicare Advantage (MA) or non-Medicare (NM) patients. We quantify how this biases our estimate of the scale elasticity, α, based on geographic variation. Suppose the production function is

$$\ln \delta_i = \alpha \ln Q_i + u_i,$$

where $Q_i = Q_i^{\text{TM}} + Q_i^{\text{MA}} + Q_i^{\text{NM}}$ is the total quantity produced in region i, of which we only observe Q_i^{TM}. When we estimate the scale elasticity α using Q_i^{TM} as a proxy for Q_i, our

regression coefficient may be biased:

$$\frac{\text{Cov}(\ln \delta_i, \ln Q_i^{\text{TM}})}{\text{Var}(\ln Q_i^{\text{TM}})} = \frac{\text{Cov}(\alpha \ln Q_i, \ln Q_i^{\text{TM}})}{\text{Var}(\ln Q_i^{\text{TM}})} + \frac{\text{Cov}(u_i, \ln Q_i^{\text{TM}})}{\text{Var}(\ln Q_i^{\text{TM}})} = \alpha\zeta,$$

where ζ, which governs the bias, is the regression coefficient from $\ln Q_i = \zeta \ln Q_i^{\text{TM}} + \mathfrak{u}_i$.

To compute ζ we differentiate the identity $Q_i = Q_i^{\text{TM}} \left(1 + \frac{Q_i^{\text{MA}}}{Q_i^{\text{TM}}} + \frac{Q_i^{\text{NM}}}{Q_i^{\text{TM}}}\right)$ with respect to Q_i^{TM}, which we observe:

$$\frac{\text{d} \ln Q_i}{\text{d} \ln Q_i^{\text{TM}}} = 1 + s_i^{\text{MA}} \varrho_i^{\text{MA}} + s_i^{\text{NM}} \varrho_i^{\text{NM}},$$

where $s_i^{\text{MA}} \equiv \frac{Q_i^{\text{MA}}}{Q_i^{\text{TM}} + Q_i^{\text{MA}} + Q_i^{\text{NM}}}$ is the MA share of production in region i, $\varrho_i^{\text{MA}} \equiv \frac{\text{d} \ln \frac{Q_i^{\text{MA}}}{Q_i^{\text{TM}}}}{\text{d} \ln Q_i^{\text{TM}}}$ is the TM production elasticity of relative production, and s_i^{NM} and ϱ_i^{NM} are similarly defined for non-Medicare (NM) insurance. To make it feasible to estimate these elasticities, we assume that they are constant across regions. If relative quantities produced are uncorrelated with the Traditional Medicare quantity produced ($\varrho^{\text{MA}} = \varrho^{\text{NM}} = 0$), then $\zeta = 1$ and $\alpha\zeta$ is an unbiased estimate of the scale elasticity α.[55] Otherwise, we need estimates of the average production shares $\bar{s}^{\text{MA}}$ and $\bar{s}^{\text{NM}}$ and the regression coefficients ϱ^{MA} and ϱ^{NM} to compute ζ.

We compute the production shares using data on aggregate expenditures and price deflators from prior research. Medicare, including both TM and MA, paid for $153 billion of the $525 billion spent nationally on physician services in 2017 [Centers for Medicare and Medicaid Services, 2022]. Per capita spending and prices are similar between the two parts of Medicare [Berenson et al., 2015, Gupta et al., 2022]. Given this similarity, we apportion Medicare's production between TM and MA based on relative enrollment and obtain $\bar{s}^{\text{MA}} = 0.111$. Next we consider Non-Medicare (NM) production. Private insurance spent

55. A special case would be if the quantity of care produced outside of TM is perfectly correlated with volume inside TM, so the shares s_i^{MA} and s_i^{NM} are constant.

$226 billion, which we deflate by a factor of 1.43 to account for the higher prices private insurance pays to make quantities comparable to Medicare [Lopez and Jacobson, 2020]. Medicaid spent roughly $41 billion, which we deflate by its relative price of 0.72 [Zuckerman et al., 2021]. We incorporate other residual categories of production without price adjustments.[56] Combining these, we obtain an average $\bar{s}^{\text{NM}} = 0.676$.

To estimate ϱ^{MA} and ϱ^{NM}, we assume that relative production is proportionate to relative resident beneficiaries. We obtain the number of TM beneficiaries and number of MA beneficiaries by HRR from Medicare enrollment data and compute the number of NM patients as total population minus Medicare enrollees.[57] Regressing the respective beneficiary ratios on log TM production yields $\hat{\varrho}^{\text{MA}} = 0.073$ and $\hat{\varrho}^{\text{NM}} = 0.069$. Putting these together means $\hat{\zeta} = 1.055$, so our estimated $\alpha\zeta = 0.66$ from Table 2.5 implies a scale elasticity of $\alpha = \frac{0.66}{1.055} = 0.63$.

2.12 Details of counterfactual calculations

Section 2.12.1 describes how we compute counterfactual equilibrium outcomes relative to baseline equilibrium outcomes in the model. Section 2.12.2 describes the assumptions we make to infer the number of potential patients N_j and hence import shares m_{ij}, which are inputs into these calculations. Section 2.12.3 describes how to compute counterfactual outcomes in the model when there are multiple (observed) types of patients who differ in their trade costs. Section 2.12.4 describes how we infer the number of potential patients of

56. These other categories in the National Health Expenditure data are labeled Other Health Insurance Programs and Other Third Party Payers, along with out-of-pocket spending. Our simplifying approach here amounts to assuming Medicare prices for these residual categories.

57. Ideally we would like to use the quantity of production in NM and MA markets, but we do not have this available at the HRR level. Beneficiaries might seem like a problematic proxy because the composition of NM beneficiaries varies widely across space, with some regions having a high Medicaid share and others a high private share. In aggregate, these two markets turn out to have similar per capita quantities of physician service spending: while private spending is $1,118 per capita and Medicaid spending is $550 per capita, the price adjustments mentioned above the quantities are relatively similar at $782 and $764, respectively, when valued at Medicare prices.

each type.

2.12.1 Computing equilibrium outcomes in counterfactual scenarios

We compute counterfactual equilibrium outcomes relative to baseline equilibrium outcomes by rewriting the equilibrium system of equations in terms of the initial allocation, constant elasticities, relative exogenous parameters, and relative endogenous equilibrium outcomes, a technique known as "exact hat algebra" in the trade literature.

If $K(\delta) = \delta$ and $H(Q) = Q^\alpha$, an equilibrium is a set of quantities and qualities $\{Q_i, \delta_i\}_{i \in \mathcal{I}}$ that simultaneously satisfy equations (2.4) and (2.1) and $Q_i = \sum_j Q_{ij}$. Consider two equilibria: the baseline equilibrium and the counterfactual equilibrium. Define export shares $x_{ij} \equiv \frac{Q_{ij}}{\sum_{j'} Q_{ij'}}$ and import shares $m_{ij} \equiv \frac{Q_{ij}}{N_j}$ in the baseline equilibrium. Denote the counterfactual parameters and equilibrium outcomes by primes. Plugging $Q_i = \sum_j Q_{ij}$ into equation (2.4), we can write the system of equations for each equilibrium as

$$\begin{aligned} \delta_i' &= \left(\frac{R_i' A_i'}{w_i'}\right)\left(\textstyle\sum_j Q_{ij}'\right)^\alpha \qquad & Q_{ij}' &= \delta_i' \frac{\rho_{ij}'}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'}' \rho_{i'j}'} N_j' \\ \delta_i &= \left(\frac{R_i A_i}{w_i}\right)\left(\textstyle\sum_j Q_{ij}\right)^\alpha & Q_{ij} &= \delta_i \frac{\rho_{ij}}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'} \rho_{i'j}} N_j \end{aligned}.$$

Define $\hat{y} \equiv \frac{y'}{y}$ for every variable y. For example, $\hat{\delta}_i \equiv \frac{\delta_i'}{\delta_i}$.

We now rewrite the counterfactual equilibrium equations in terms of baseline equilibrium shares x_{ij}, m_{ij}, the scale elasticity α, (relative) counterfactual exogenous parameters $\hat{A}, \hat{R}, \hat{w}, \hat{\rho}, \hat{N}$, and (relative) counterfactual endogenous qualities $\hat{\delta}$.

First, divide the counterfactual free-entry condition by the baseline free-entry condition to obtain an expression for relative quality:

$$\frac{\delta_i'}{\delta_i} = \frac{\hat{R}_i \hat{A}_i}{\hat{w}_i}\left(\frac{\sum_{j \in \mathcal{I}} Q_{ij}'}{\sum_{j \in \mathcal{I}} Q_{ij}}\right)^\alpha = \frac{\hat{R}_i \hat{A}_i}{\hat{w}_i}\left(\sum_{j \in \mathcal{I}} \frac{Q_{ij}}{\sum_{j \in \mathcal{I}} Q_{ij}} \frac{Q_{ij}'}{Q_{ij}}\right)^\alpha = \frac{\hat{R}_i \hat{A}_i}{\hat{w}_i}\left(\sum_{j \in \mathcal{I}} x_{ij} \frac{Q_{ij}'}{Q_{ij}}\right)^\alpha \tag{2.19}$$

Second, divide the counterfactual gravity equation by the baseline gravity equation to obtain an expression for relative bilateral flows:

$$\frac{Q'_{ij}}{Q_{ij}} = \frac{\delta'_i}{\delta_i}\left(\frac{\frac{\rho'_{ij}}{\sum_{i'\in 0\cup\mathcal{I}}\delta'_{i'}\rho'_{i'j}}N'_j}{\frac{\rho_{ij}}{\sum_{i'\in 0\cup\mathcal{I}}\delta_{i'}\rho_{i'j}}N_j}\right) = \frac{\frac{\delta'_i}{\delta_i}\frac{\rho'_{ij}}{\rho_{ij}}\frac{N'_j}{N_j}}{\sum_{i'\in 0\cup\mathcal{I}}\frac{\delta_{i'}\rho_{i'j}}{\sum_{i'\in 0\cup\mathcal{I}}\delta_{i'}\rho_{i'j}}\frac{\delta'_{i'}\rho'_{i'j}}{\delta_{i'}\rho_{i'j}}}$$
$$= \frac{\hat{\delta}_i\hat{\rho}_{ij}\hat{N}_j}{\sum_{i'\in 0\cup\mathcal{I}}\frac{Q_{i'j}}{N_j}\hat{\delta}_{i'}\hat{\rho}_{i'j}} = \frac{\hat{\delta}_i\hat{\rho}_{ij}\hat{N}_j}{m_{0j}+\sum_{i'\in\mathcal{I}}m_{i'j}\hat{\delta}_{i'}\hat{\rho}_{i'j}}$$

Plug this expression for relative bilateral flows into equation (2.19) and rearrange terms to obtain the following system of $\mathcal{I}$ equations with unknowns $\{\hat{\delta}_i\}_{i=1}^{\mathcal{I}}$:

$$\hat{\delta}_i = \left(\hat{R}_i\hat{A}_i/\hat{w}_i\right)^{\frac{1}{1-\alpha}}\left(\sum_{j\in\mathcal{I}}\frac{x_{ij}\hat{\rho}_{ij}\hat{N}_j}{m_{0j}+\sum_{i'\in\mathcal{I}}m_{i'j}\hat{\delta}_{i'}\hat{\rho}_{i'j}}\right)^{\frac{\alpha}{1-\alpha}}. \tag{2.20}$$

2.12.2 Inferring the number of potential patients

A baseline calibration of our model requires α, x_{ij}, and m_{ij} in order to use equation (2.20) to compute relative counterfactual outcomes. We have estimated α. The export shares $x_{ij} \equiv \frac{Q_{ij}}{\sum_j Q_{ij}}$ are easily computed using the observed trade matrix.[58] The challenge is computing import shares $m_{ij} \equiv \frac{Q_{ij}}{N_j}$ because we do not observe N_j; while we observe the number of Medicare beneficiaries in region j, not all beneficiaries are in the market for all services. This section describes the assumptions we make in order to infer the values of the relevant market size $N_j\ \forall j \in \mathcal{I}$. Specifically, we assume per capita demand is uniform, outside-option quality is constant across regions, and the average outside-option share is 10%, as described below.

We have estimated $\theta_j = N_j/\Phi_j$ in equation (2.12). We observe the number of beneficiaries

58. Dingel and Tintelnot [2021] document overfitting problems when calibrating gravity models using noisy observed shares. We obtain similar counterfactual outcomes when calibrating our model using gravity-predicted shares.

enrolled in Traditional Medicare in region j, which we denote S_j^{TM}. By definition, $m_{0j} = \frac{\delta_{0j}}{\Phi_j}$. We assume $\delta_{0j} = \delta_0 \ \forall j$ and $N_j \propto S_j^{\text{TM}}$. This implies

$$m_{0j} = \frac{\delta_{0j}}{\Phi_j} = \frac{\delta_0 \theta_j}{N_j} = \frac{\delta_0 \theta_j}{\mathfrak{s} S_j^{\text{TM}}},$$

where $\mathfrak{s}$ is a constant of proportionality. We set $\frac{\delta_0}{\mathfrak{s}}$ such that the average outside-option share is 10%, $\frac{1}{\mathcal{I}} \sum_j m_{0j} = 0.1$. This requires $\frac{\delta_0}{\mathfrak{s}} = \frac{0.1 \times \mathcal{I}}{\sum_j \theta_j / S_j^{\text{TM}}}$. With m_{0j} in hand, we can infer N_j:

$$m_{0j} = 1 - \sum_{i \in \mathcal{I}} m_{ij} = 1 - \frac{1}{N_j} \sum_{i \in \mathcal{I}} Q_{ij} \implies N_j = \frac{1}{1 - m_{0j}} \sum_{i \in \mathcal{I}} Q_{ij}.$$

With N_j in hand, we can compute all import shares, $m_{ij} = \frac{Q_{ij}}{N_j} \ \forall i \in 0 \cup \mathcal{I}, \forall j \in \mathcal{I}$.

We exclude the Anchorage, Alaska HRR from our counterfactual computations. The entire state of Alaska is one (geographically isolated and very large) HRR. The average within-Alaska-HRR procedure incurs more than 60 kilometers of travel. In the gravity regression, Alaska has the smallest exporter fixed effect: very few patients travel to Alaska for care. Alaska's importer fixed effect is quite large because Alaskans import about 15% of their services and the average import traverses 3,616 kilometers. As a result, the implied outside-option share would exceed one when we set the nationwide average to 10%. We therefore exclude the Alaska HRR from the economy when computing counterfactual outcomes. Given its considerable geographic isolation, Alaska would have little influence on outcomes in other regions.

The qualitative and spatial patterns of counterfactual outcomes are the same if we assume the average outside-option share is 20% rather than 10%.

2.12.3 Counterfactual outcomes with multiple patient types

This section describes how to compute counterfactual equilibrium outcomes relative to baseline equilibrium outcomes when there are multiple patient types who face heterogeneous trade costs. The derivation is very similar to that of Section 2.12.1. Define import shares $m_{ij\kappa} \equiv \frac{Q_{ij\kappa}}{N_{j\kappa}}$ in the baseline equilibrium. Define patient-type shares $n_{j\kappa} \equiv \frac{N_{j\kappa}}{N_j}$. We rewrite the system of baseline and counterfactual gravity equations (2.15) and free-entry condition (2.4) as follows:

$$\begin{aligned} \delta_i' &= \left(\frac{R_i' A_i'}{w_i'}\right)\left(\textstyle\sum_j Q_{ij}'\right)^{\alpha} \qquad & Q_{ij}' &= \delta_i' \textstyle\sum_\kappa \frac{\rho_{ij\kappa}'}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'}' \rho_{i'j\kappa}'} N_{j\kappa}' \\ \delta_i &= \left(\frac{R_i A_i}{w_i}\right)\left(\textstyle\sum_j Q_{ij}\right)^{\alpha} & Q_{ij} &= \delta_i \textstyle\sum_\kappa \frac{\rho_{ij\kappa}}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'} \rho_{i'j\kappa}} N_{j\kappa} \end{aligned}.$$

As above, dividing the counterfactual free-entry condition by the baseline free-entry condition yields the expression for relative quality in equation (2.19). Second, divide the counterfactual gravity equation by the baseline gravity equation to obtain an expression for relative bilateral flows:

$$\frac{Q_{ij}'}{Q_{ij}} = \frac{\delta_i'}{\delta_i}\left(\frac{\sum_\kappa \frac{\rho_{ij\kappa}'}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'}' \rho_{i'j\kappa}'} N_{j\kappa}'}{\sum_\kappa \frac{\rho_{ij\kappa}}{\sum_{i' \in 0 \cup \mathcal{I}} \delta_{i'} \rho_{i'j\kappa}} N_{j\kappa}}\right) = \frac{\delta_i'}{\delta_i}\left(\frac{\sum_\kappa \frac{\rho_{ij\kappa}'}{\Phi_{j\kappa}'} N_{j\kappa}'}{\sum_\kappa \frac{\rho_{ij\kappa}}{\Phi_{j\kappa}} N_{j\kappa}}\right) = \hat{\delta}_i \sum_\kappa \frac{n_{j\kappa}}{\frac{m_{ij}}{m_{ij\kappa}}} \frac{\hat{\rho}_{ij\kappa}}{\hat{\Phi}_{j\kappa}} \hat{N}_{j\kappa}$$

Plugging this expression for relative bilateral flows into equation (2.19) and then rearranging terms yields the following system of $\mathcal{I}$ equations with unknowns $\{\hat{\delta}_i\}_{i=1}^{\mathcal{I}}$:

$$\hat{\delta}_i = \left(\hat{R}_i \hat{A}_i / \hat{w}_i\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in \mathcal{I}} x_{ij} \left(\sum_\kappa \frac{n_{j\kappa}}{m_{ij}} \frac{m_{ij\kappa} \hat{\rho}_{ij\kappa}}{m_{0j\kappa} + \sum_{i' \in \mathcal{I}} m_{i'j\kappa} \hat{\delta}_{i'} \hat{\rho}_{i'j\kappa}} \hat{N}_{j\kappa}\right)\right)^{\frac{\alpha}{1-\alpha}}. \qquad (2.21)$$

2.12.4 Inferring the number of potential patients of each type

Because we do not observe patients who select the outside option, we make assumptions that allow us to infer $N_{j\kappa}$ and thus $n_{j\kappa}$ and $m_{ij\kappa}$, which are needed to compute counterfactual outcomes using equation (2.21). We start from a type-specific variant of the gravity equation (2.7) with fixed effects, as in the single-type equation (2.12). The estimating equation is

$$\ln \mathbb{E}\left(\overline{R}Q_{ij\kappa}\right) = \ln \delta_i + \ln\left(\frac{N_{j\kappa}}{\Phi_{j\kappa}}\right) + \gamma^{\kappa} X_{ij} = \ln \delta_i + \ln \theta_{j\kappa} + \gamma^{\kappa} X_{ij}.$$

This yields an estimate of $\theta_{j\kappa} = N_{j\kappa}/\Phi_{j\kappa}$.

As in the single-type case above, we assume per capita demand is uniform and outside-option quality is constant across regions. We observe the number of beneficiaries of type κ enrolled in Traditional Medicare in region j, which we denote $S_{j\kappa}^{\text{TM}}$. We assume $\delta_{0j} = \delta_0 \ \forall j$ and $N_{j\kappa} = \mathfrak{s} S_{j\kappa}^{\text{TM}}$, where $\mathfrak{s}$ is a constant of proportionality that is common across types. This implies

$$m_{0j\kappa} = \frac{\delta_0}{\Phi_{j\kappa}} = \frac{\delta_0 \theta_{j\kappa}}{N_{j\kappa}} = \frac{\delta_0 \theta_{j\kappa}}{\mathfrak{s} S_{j\kappa}^{\text{TM}}}.$$

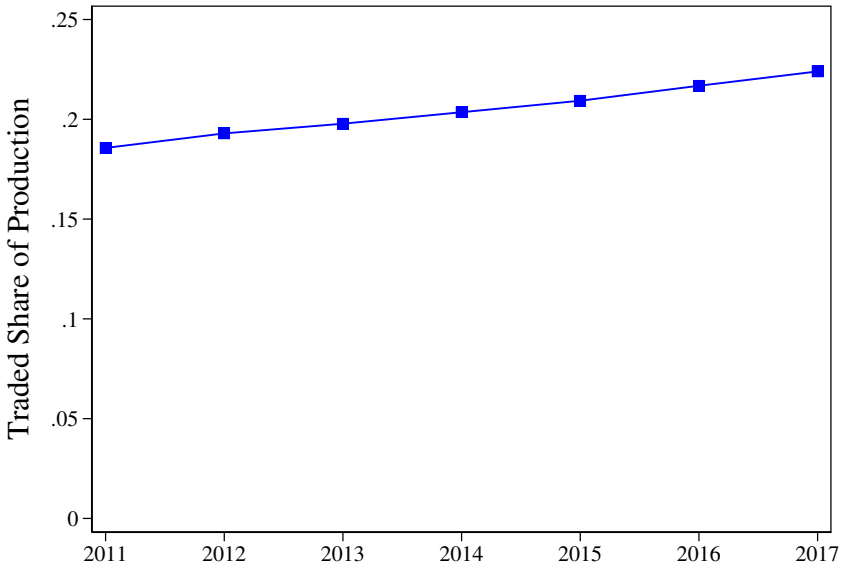
Let $\mathcal{K} = \sum_{\kappa} 1$ denote the number of patient types. We set $\frac{\delta_0}{\mathfrak{s}}$ such that the average outside-option share, across all types, is 10%, $\frac{1}{\mathcal{I}\mathcal{K}} \sum_{j\kappa} m_{0j\kappa} = 0.1$. This implies

$$m_{0j\kappa} = 0.1 \times \frac{\theta_{j\kappa}/S_{j\kappa}^{\text{TM}}}{\frac{1}{\mathcal{I}\mathcal{K}} \sum_{j'\kappa'} \theta_{j'\kappa'}/S_{j'\kappa'}^{\text{TM}}}.$$

Using the resulting $N_{j\kappa} = \frac{1}{1-m_{0j\kappa}} \sum_{i \in \mathcal{I}} Q_{ij\kappa}$ allows us to compute all import shares.

2.13 Additional exhibits

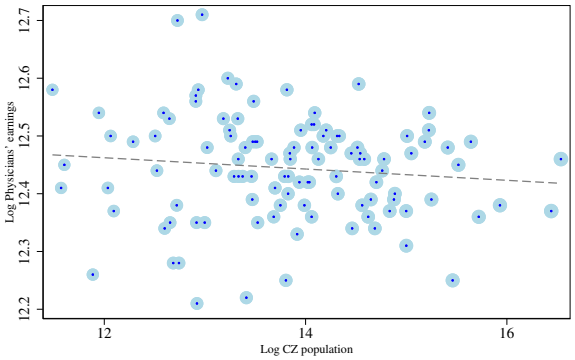
Figure 2.13: Trade in medical services has increased over time



Note: This figure shows the annual exported share of production from 2011 to 2017. Production and trade are computed using the Medicare 20% carrier Research Identifiable Files for the relevant years. Production is exported when the patient's address and the service location are in different HRRs. HRR definitions are from the Dartmouth Atlas Project.

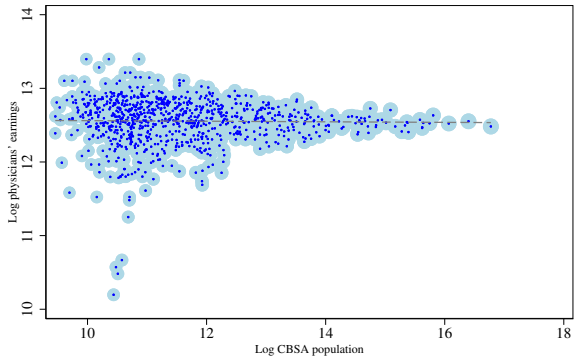
Figure 2.14: Population elasticities of input costs

(a) Physicians' earnings (commuting zones)



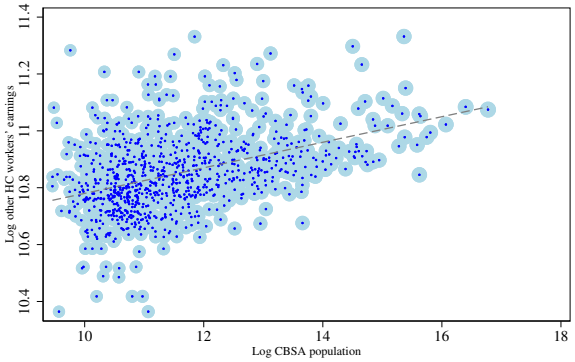
Y = 12.5794 – 0.0098 [0.0082] Log CZ population + ε
N = 111; R^2 = 0.0127

(b) Physicians' earnings (CBSAs)



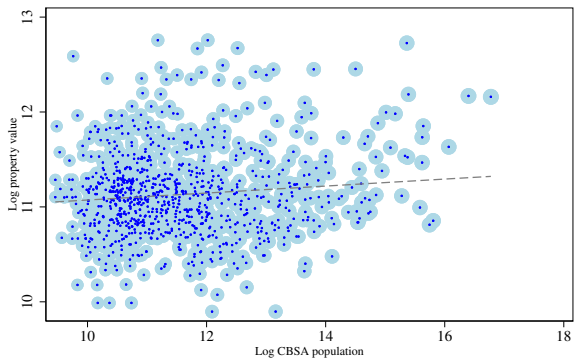
Y = 12.5994 – 0.0038 [0.0082] Log CBSA population + ε
N = 893; R^2 = 0.0002

(c) Other healthcare workers' earnings



Y = 10.3307 + 0.0449 [0.0033] Log CBSA population + ε
N = 896; R^2 = 0.1748

(d) Median house value

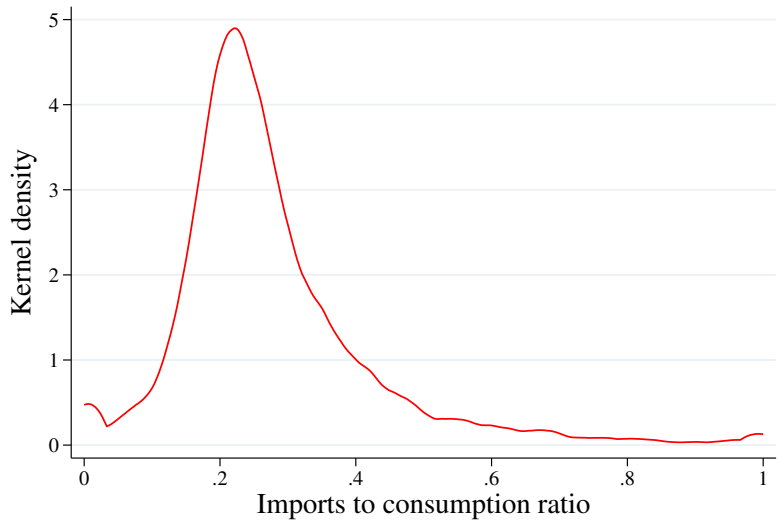


Y = 10.7048 + 0.0368 [0.0116] Log CBSA population + ε
N = 896; R^2 = 0.0110

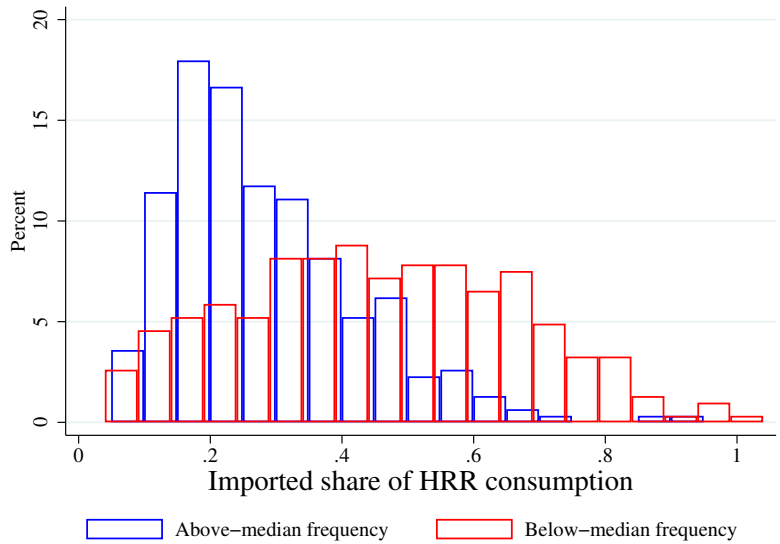
Notes: This figure depicts relationships between input costs and population sizes. Panel a shows physicians' earnings across 111 commuting zones using data from Gottlieb et al. [2020]. Panels b, c, and d show variation across CBSAs in physicians' earnings, other healthcare workers' earnings, and median house values (a proxy for real estate and other locally priced inputs) using data from the 2015–2019 American Community Survey.

Figure 2.15: Variation in trade shares across procedures and regions

(a) Distribution of import share by procedure

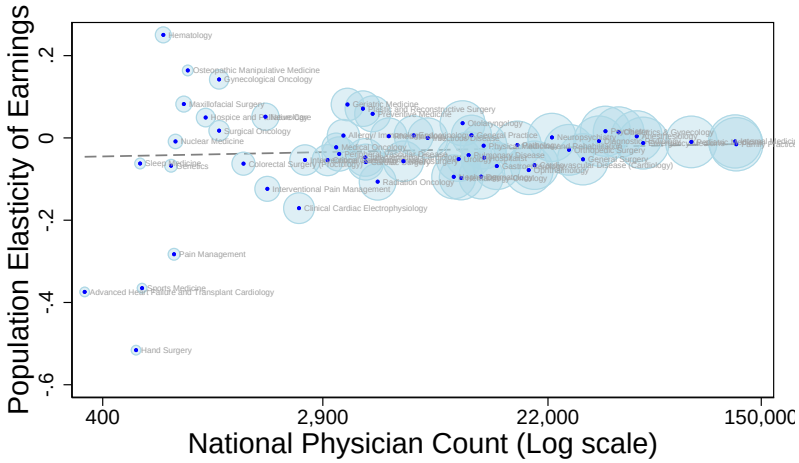


(b) Distributions of import shares for common and rare services



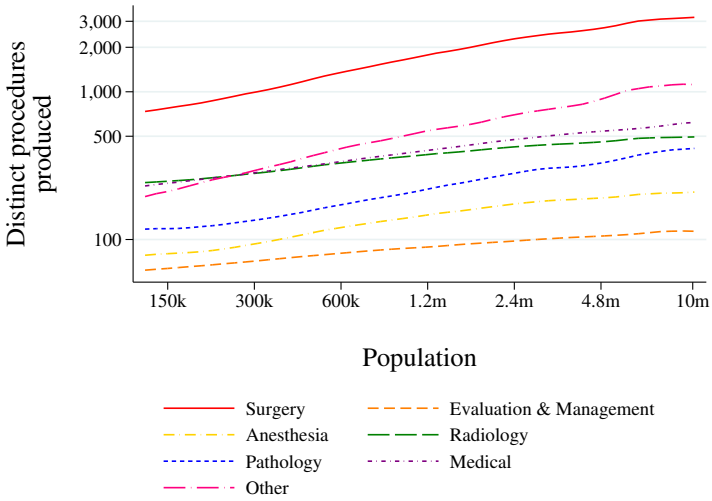
Notes: Panel a shows the distribution of the imported consumption share across procedures for procedures performed at least 20 times (in our 20% sample of Medicare claims). Imports are defined as care provided to a patient who lives in one HRR at a service location in a different HRR. Panel b splits all services into two groups based on how often they are performed nationally. Those performed less often than the median are shown in red, and those performed more often than the median service are shown in blue. Import shares are substantially higher for the rarer services.

Figure 2.16: Specialists' income patterns do not explain the output-population gradient



Notes: This figure shows the population elasticity of income for different medical specialties against the total number of physicians in those specialties. For each specialty, we estimate the elasticity of income with respect to population across commuting zones, using data from Gottlieb et al. [2020]. The graph shows that these elasticities are unrelated to the total national count of physicians in those specialties.

Figure 2.17: Larger markets produce a greater variety of procedures

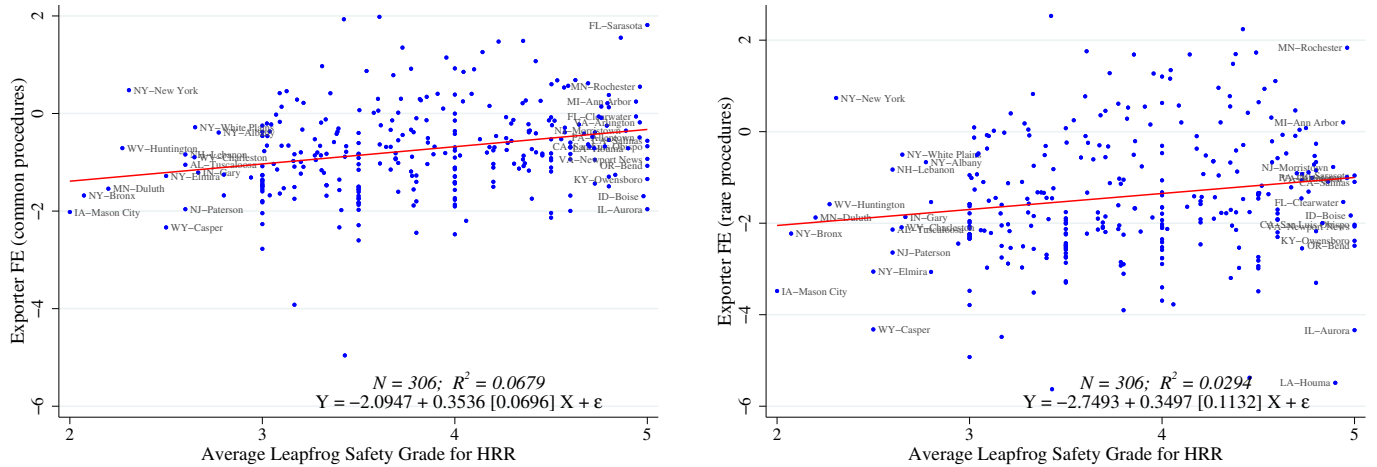


Notes: This figure shows the local relationship between the number of distinct services performed in the Medicare data in a given HRR and that HRR's population. More populous HRRs perform more unique services; Table 2.20 reports the population elasticities. We use procedure classifications from the American Academy of Professional Coders, which groups codes into surgeries, anesthesia, radiology, pathology, medical, and evaluation & management services [AAPC, 2021]. We combine Category II codes, Category III codes and Multianalyte Assays into "other."

Figure 2.18: Leapfrog Safety Grade vs. estimated quality: common and rare

(a) Leapfrog Safety Grade vs. quality for common services

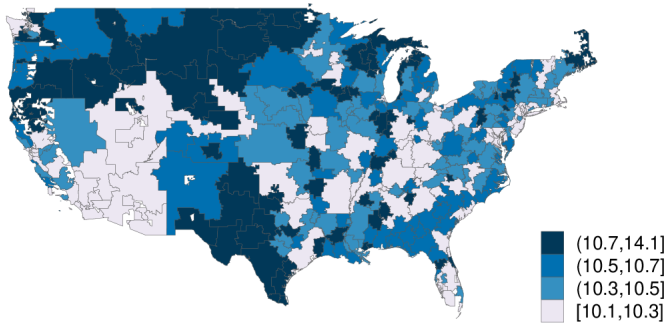
(b) Leapfrog Safety Grade vs. quality for rare services



Notes: This figure shows the relationship between exporter fixed effects, estimated separately for common and rare services, and the Leapfrog Safety Grade. The vertical axis shows the exporter fixed effects for each HRR estimated from equation (2.12), in Panel a using trade in common services, and in Panel b using trade in rare services. The horizontal axis in both panels is the average safety grade for hospitals in an HRR, determined by the Leapfrog Group. The Leapfrog Safety Grades range from A to F, which we scale as integers from 1 (for F) to 5 (for A). We then compute the mean score for all hospitals in the HRR. The Safety Grades are positively associated with the exporter fixed effects for both rare and common procedures.

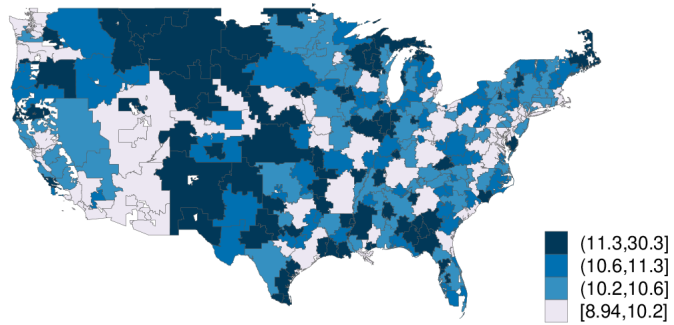
Figure 2.19: Counterfactual change in quality δ for rare vs. common services when increasing reimbursement by 10% everywhere

(a) Change (%) in quality δ for common services



Cutoffs: Percentiles 25, 50, 75.

(b) Change (%) in quality δ for rare services



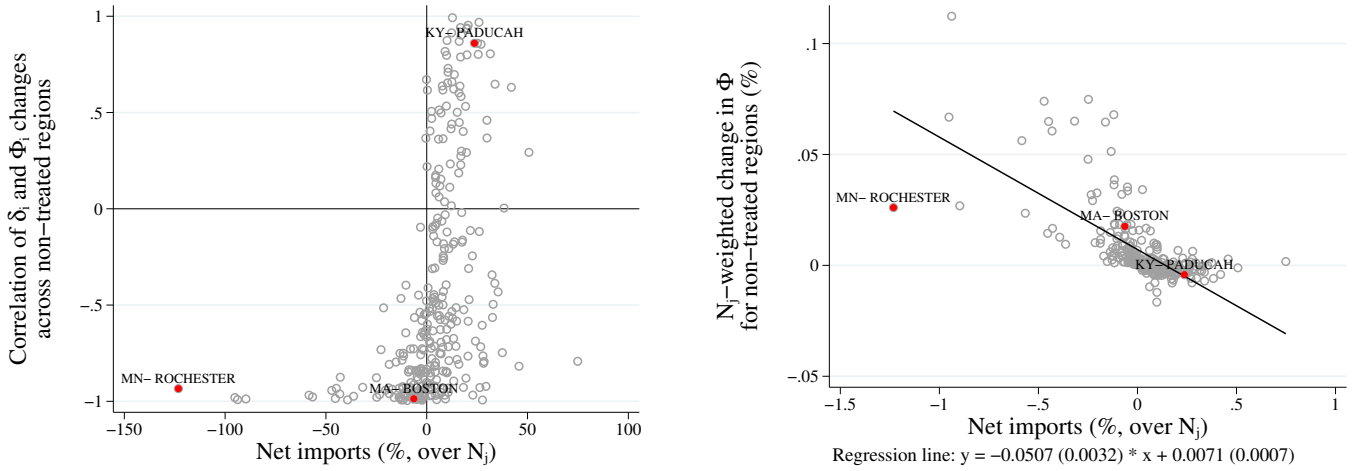
Cutoffs: Percentiles 25, 50, 75.

Notes: Both panels show the impacts of increasing reimbursements by 10% everywhere ($\hat{R}_i = 1.1$ for all i) on the quality of production in each region, δ_i. Panel a illustrates the change for common services, and Panel b for rare services. Each panel is based on the baseline trade matrix for the respective set of services. Panel a uses an agglomeration elasticity of $\alpha = 0.6$ and Panel b uses $\alpha = 0.9$. The common-services scenario excludes the Alaska HRR and the rare-services scenario excludes four HRRs (Alaska, Hawaii, Houma, La., and Minot, N.D.). The pattern of outcomes is qualitatively similar but the magnitudes vary more for rare services.

Figure 2.20: Spillovers from higher reimbursements in one region depend on that region's net imports

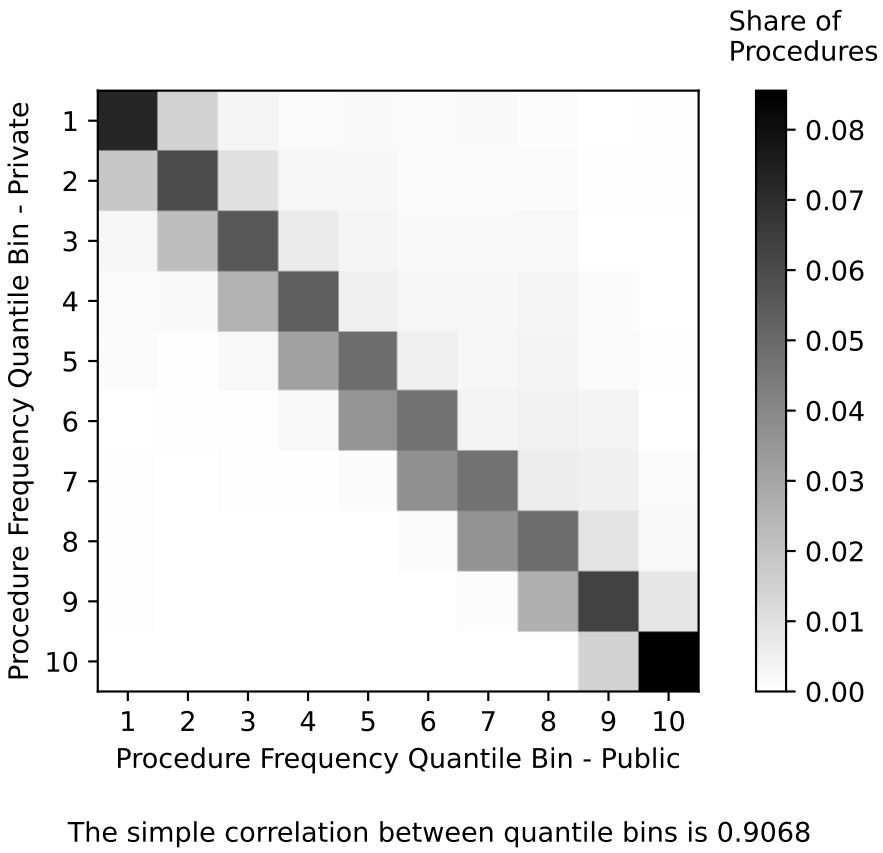
(a) Correlation of $\hat{\delta}_i$ and $\hat{\Phi}_i$ across non-treated regions

(b) Change in non-treated regions' aggregate market access



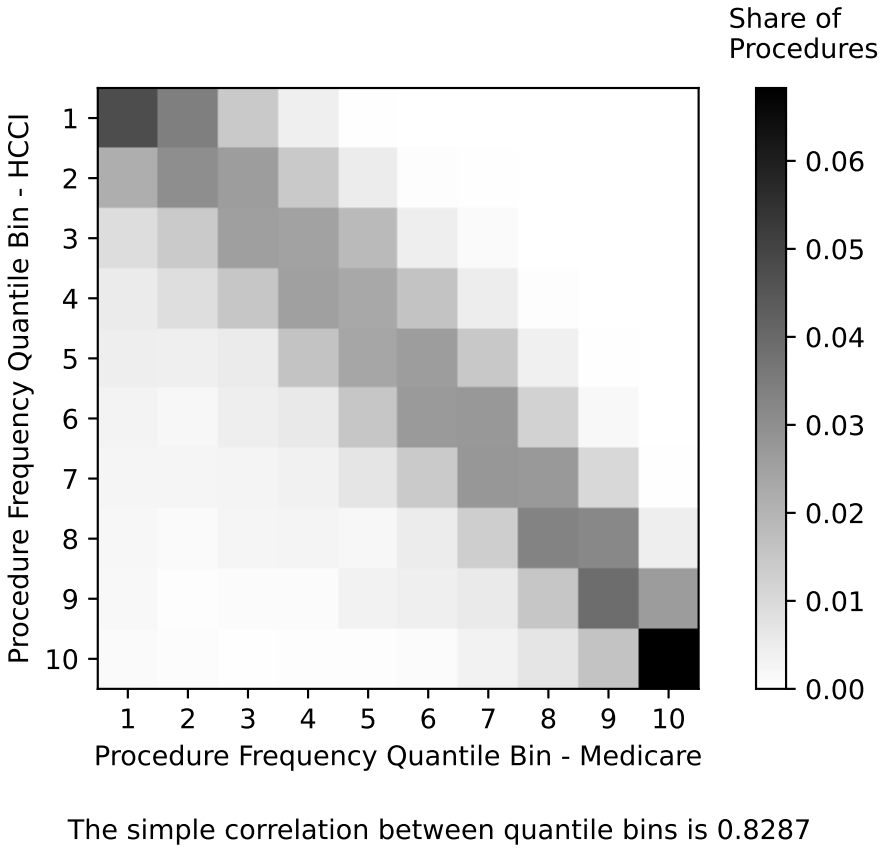
Notes: This figure characterizes counterfactual outcomes when raising reimbursements by 30 percent in one HRR. We conduct this exercise for each region, one at a time, and each observation in each panel represents one such counterfactual scenario. Panel a illustrates the contrast in spillovers as a function of net imports of the treated region. The vertical-axis value for each observation reports the correlation—across *all regions other than the treated one* for the exercise in question—between the counterfactual changes $\hat{\delta}_i$ and $\hat{\Phi}_i$. The scatterplot relates these correlations to the *treated* region's net import share, which is plotted on the horizontal axis. When the treated region is a net exporter, changes in quality δ_i and in market access Φ_i for non-treated regions move in opposite directions: a region whose output quality declines experiences an increase in market access through imports from the treated region. However, increasing reimbursements in a net-importing region often has the opposite effect: neighboring regions with quality reductions also experience lower market access, (changes in δ_i and Φ_i are positively correlated). For each counterfactual, the vertical-axis value in Panel b shows the aggregate impact on patient market access *excluding the treated region.* The panel relates this impact to the *treated* region's net imports, shown on the horizontal axis. When the treated region is a net importer, the aggregate impact on market access for non-treated regions tends to be smaller or even negative.

Figure 2.21: Deciles of Procedure Frequency in Confidential and Public Medicare Data



Notes: This figure shows the share of procedures in each frequency decile in the Medicare public data compared to the Medicare confidential data. The classification of procedures by frequency deciles appears largely consistent between the two data sources for Medicare patients.

Figure 2.22: Deciles of Procedure Frequency in Medicare and Private Insurance Data



Notes: This figure shows the share of procedures in each frequency decile in the Medicare versus privately insured data. The classification of procedures by frequency deciles appears largely consistent when comparing public Medicare data with data on privately insured patients from the Health Care Cost Institute (HCCI).

Table 2.7: Higher-income patients are less sensitive to distance: Procedure-level estimates

	(1) 25min visit	(2) cataract removal	(3) knee joint repair	(4) heart artery bypass	(5) gallblader removal
Distance (log)	-2.075	-2.281	-2.255	-2.246	-2.135
	(0.0790)	(0.0829)	(0.0947)	(0.0876)	(0.0855)
Distance (log) × income tercile 2	0.0946	0.143	0.171	0.0987	0.205
	(0.0610)	(0.0819)	(0.0754)	(0.0823)	(0.0685)
Distance (log) × income tercile 3	0.206	0.287	0.227	0.402	0.314
	(0.0777)	(0.0914)	(0.0937)	(0.0927)	(0.0907)
Observations	271,728	268,400	262,352	240,352	250,800
Patient market-income FE & Provider market FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the coefficient on log distance for each income tercile from gravity regressions estimated separately for five procedures varying in frequency: 25 min office visit (HCPCS 99214), cataract removal (66984), knee joint repair (27447), heart artery bypass (33533), and gallblader removal (47562). The dependent variable in all regressions is the number of procedures traded. Each regression includes log distance interacted with an income tercile indicator, an indicator for same-HRR observations ($i = j$), an exporting HRR fixed effect, and an income-tercile-importing-HRR fixed effect. The coefficients for higher income terciles are positive, indicating that patients residing in higher-income ZIP codes are less sensitive to distance. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.8: Estimates of a strong home-market effect by CBSA

	(1)	(2)	(3)	(4)	(5)	(6)
Estimation method:	PPML	PPML	PPML	IV	PPML	IV
Instrument:				1940 pop		Bedrock
Provider-market population (log)	0.734	0.739	0.703	0.716	0.739	1.161
	(0.0232)	(0.0234)	(0.0205)	(0.0249)	(0.0259)	(0.307)
Patient-market population (log)	0.395	0.393	0.417	0.396	0.394	0.178
	(0.0290)	(0.0292)	(0.0264)	(0.0261)	(0.0311)	(0.373)
Distance (log)	-2.311	-3.464		-3.403	-3.400	-4.677
	(0.0493)	(0.324)		(0.295)	(0.347)	(1.056)
Distance (log, squared)		0.110		0.104	0.105	0.210
		(0.0323)		(0.0288)	(0.0346)	(0.0850)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.000	0.000	0.000	0.000	0.000	0.063
Observations	857,476	857,476	857,476	857,476	781,456	781,456
Sample:	All CBSAs	All CBSAs	All CBSAs	All CBSAs	Bedrock data	Bedrock data
Distance elasticity at mean		-1.90		-1.92	-1.90	-1.68
Distance deciles			Yes			

Notes: This table reports estimates of equation (2.8), which estimates the presence of weak or strong home-market effects. The dependent variable in all regressions is the value of trade computed by assigning each procedure its national average price. The independent variables are patient- and provider-market log population, log distance between CBSAs, and an indicator for same-CBSA observations ($i = j$). The positive coefficient on provider-market log population implies a weak home-market effect, and the fact that this coefficient exceeds that on patient-market population implies a strong home-market effect. Column 2 makes the distance coefficient more flexible by adding a control for the square of log distance. Column 3 replaces parametric distance specifications with fixed effects for each decile of the distance distribution. Column 4 uses the provider-market and patient-market log populations in 1940 as instruments for the contemporaneous log populations when estimating by generalized method of moments. Column 5 reports the PPML estimate on the subsample of regions for which we have data on depth to bedrock available ($N = 884^2$). Column 6 uses depth to bedrock in the importing and exporting regions as instruments for current log population in those regions, respectively. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.9: Estimates of a strong home-market effect including facility spending

Estimation method:	(1) PPML	(2) PPML	(3) PPML	(4) IV
Provider-market population (log)	0.687	0.700	0.689	0.829
	(0.0576)	(0.0525)	(0.0382)	(0.0586)
Patient-market population (log)	0.226	0.217	0.255	0.266
	(0.0571)	(0.0507)	(0.0345)	(0.0470)
Distance (log)	-1.635	0.877		0.932
	(0.0495)	(0.316)		(0.253)
Distance (log, squared)		-0.254		-0.258
		(0.0319)		(0.0250)
Same hrr	0.434	1.791	4.685	
	(0.174)	(0.238)	(0.0637)	
Observations	93,636	93,636	93,636	93,636
Distance elasticity at mean		-2.76		-2.76
Distance deciles			Yes	

Notes: This table reports estimates of equation (2.8), which estimates the presence of weak or strong home-market effects, when including professional and facility fees. The sample is all HRR pairs ($N = 306^2$). The dependent variable in all regressions is the value of trade when including professional and facility (inpatient and outpatient) fees at national average prices. The independent variables are patient- and provider-market log population, log distance between HRRs, and an indicator for same-HRR observations ($i = j$). The positive coefficient on provider-market log population implies a weak home-market effect, and the fact that this coefficient exceeds that on patient-market population implies a strong home-market effect. Column 2 makes the distance coefficient more flexible by adding a control for the square of log distance. Column 3 replaces parametric distance specifications with fixed effects for each decile of the distance distribution. Column 4 uses the provider-market and patient-market log populations in 1940 as instruments for the contemporaneous log populations when estimating by generalized method of moments. Trade flows are computed from the Medicare 20% carrier, MedPAR, and outpatient claims Research Identifiable Files, excluding emergency-room care and skilled nursing facilities. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.10: Estimates of a strong home-market effect excluding AZ, FL, CA

Estimation method:	(1) PPML	(2) PPML	(3) PPML	(4) IV
Provider-market population (log)	0.647	0.647	0.649	0.663
	(0.0811)	(0.0701)	(0.0425)	(0.0626)
Patient-market population (log)	0.375	0.383	0.414	0.400
	(0.0809)	(0.0693)	(0.0421)	(0.0570)
Distance (log)	-1.748	1.690		1.707
	(0.0608)	(0.428)		(0.397)
Distance (log, squared)		-0.360		-0.361
		(0.0429)		(0.0396)
Observations	67,600	67,600	67,600	67,600
Distance elasticity at mean		-3.35		-3.35
Distance deciles			Yes	

Notes: This table reports estimates of equation (2.8), which estimates the presence of weak or strong home-market effects, excluding snowbird states. The sample is all HRR pairs, excluding those in Arizona, Florida, or California. The dependent variable in all regressions is the value of trade computed by assigning each procedure its national average price. The independent variables are patient- and provider-market log population, log distance between HRRs, and an indicator for same-HRR observations ($i = j$). The positive coefficient on provider-market log population implies a weak home-market effect, and the fact that this coefficient exceeds that on patient-market population implies a strong home-market effect. Column 2 makes the distance coefficient more flexible by adding a control for the square of log distance. Column 3 replaces parametric distance specifications with fixed effects for each decile of the distance distribution. Column 4 uses the provider-market and patient-market log populations in 1940 as instruments for the contemporaneous log populations when estimating by generalized method of moments. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.11: Estimates of a strong home-market effect excluding HRRs with high second-home share

Estimation method:	(1) PPML	(2) PPML	(3) PPML	(4) IV
Provider-market population (log)	0.654	0.663	0.661	0.679
	(0.0652)	(0.0641)	(0.0453)	(0.0571)
Patient-market population (log)	0.369	0.362	0.392	0.382
	(0.0639)	(0.0619)	(0.0424)	(0.0564)
Distance (log)	-1.675	0.364		0.372
	(0.0509)	(0.307)		(0.279)
Distance (log, squared)		-0.210		-0.211
		(0.0300)		(0.0273)
Observations	76,176	76,176	76,176	76,176
Distance elasticity at mean		-2.64		-2.64
Distance deciles			Yes	

Notes: This table reports estimates of equation (2.8), which estimates the presence of weak or strong home-market effects, excluding HRRs with a high second-home share. The sample is all HRR pairs excluding those in the top 10% based on the share of housing units that are vacant for seasonal/recreational purposes in the 2013–2017 American Community Survey. See Table 2.10 notes on the variables, instruments, geographic units, and standard errors.

Table 2.12: Travel for dialysis

Distance (km)	Share of output				
	All (Professional)	All (Facility)	All (Dialysis)	No snowbird states (Dialysis)	Snowbird states (Dialysis)
$[0, 50)$	0.77	0.77	0.94	0.94	0.93
$[50, 100)$	0.12	0.12	0.03	0.03	0.03
$[100, .)$	0.11	0.11	0.03	0.02	0.04

Notes: For the care described in each column and the distance intervals in each row, the entries in this table report the share of patients traveling that distance from their residential ZIP code to the service location's ZIP code. The first column shows professional claims (from Medicare's "carrier" file), the second column shows facility (hospital) claims, and the third column shows dialysis claims. The remaining columns split dialysis claims between "snowbird" states (AZ, CA, and FL, following Finkelstein et al. 2016) and other states. In non-snowbird states, the table shows that 94% of patients travel less than 50 km from their home for dialysis, and only 2% more than 100 km. This is less than one-fifth as much as for other facility or professional care, suggesting that residential location is recorded correctly for almost all patients.

Table 2.13: Contrasting geographies of colonoscopies and LVAD insertions

	Colonoscopy	LVAD Insertion
Code	G0121	33979
N	58,798	333
Physicians	13,475	177
$\hat{\beta}_p^{\text{production}}$	0.00	0.71
$\hat{\beta}_p^{\text{consumption}}$	-0.01	0.03
Share traded (HRR)	0.15	0.50
Share traded (CBSA)	0.15	0.48
Median distance traveled (km)	18.44	65.50
Share $>$ 100km	0.06	0.37

Notes: This table reports statistics for two HCPCS codes: screening colonoscopy (G0121) and LVAD insertion (33979). We report the number of times the procedure is performed in 2017 in our 20% sample of Medicare patients and the number of distinct physicians performing it. The population elasticities of production and consumption are estimated using the Poisson models in equations (2.9) and (2.10) based on production HRR and patients' residential HRR, respectively. We also report the shares of procedures in which the patient and service location are in different HRRs or CBSAs, the median distance traveled for all care, and the share in which the patient and service location are more than 100 kilometers apart.

Table 2.14: Estimates of a stronger home-market effect for rare diagnoses including facility spending

	(1)	(2)	(3)	(4)	(5)	(6)
λ_X Provider-market population (log)	0.665	0.659	0.649		0.662	
	(0.0573)	(0.0560)	(0.0557)		(0.0516)	
λ_M Patient-market population (log)	0.240	0.241	0.246		0.239	
	(0.0567)	(0.0551)	(0.0548)		(0.0498)	
μ_X Provider-market population (log) $\times$ rare			0.223	0.209	0.231	0.207
			(0.0243)	(0.0206)	(0.0236)	(0.0204)
μ_M Patient-market population (log) $\times$ rare			-0.0793	-0.0753	-0.0841	-0.0700
			(0.0216)	(0.0158)	(0.0210)	(0.0161)
Observations	187,272	147,814	147,814	147,814	147,814	147,814
Distance controls	Yes	Yes	Yes	Yes		
Distance [quadratic] controls					Yes	Yes
Patient-provider-market-pair FEs				Yes		Yes

Notes: This table reports estimates of equation (2.11), which introduces interactions with an indicator for whether a diagnosis is "rare" (provided to less patients than the median diagnosis, when adding up all diagnoses nationally). The dependent variable in all regressions is the value of trade when including professional and facility (inpatient and outpatient) fees at national average prices. The interactions with patient- and provider-market population reveal whether the home-market effect is larger for rare diagnoses. The unit of observation is {rare indicator, exporting HRR, importing HRR} so the number of observations is 2×306^2 in column 1. All diagnoses are included. Columns 2 onwards drop HRR pairs with zero trade in both diagnosis groups, which leads to a larger sample than in Table 2.4 because trade in facility fees is included in addition to professional fees for all diagnoses. Column 2 shows that this restriction has a negligible impact on the estimated log population coefficients. Columns 1–4 control for distance using the log of distance between HRRs. Columns 5 and 6 add a control for the square of log distance. Columns 4 and 6 introduce a fixed effect for each ij pair of patient market and provider market, so these omit all covariates that are not interacted with the rare indicator. The positive coefficient on provider-market population $\times$ rare across all columns indicates that the home-market effect is stronger for rare than for common services. The negative coefficient on patient-market population $\times$ rare across all columns indicates that the *strong* home-market effect has a larger magnitude for rare services. Trade flows are computed from the Medicare 20% carrier, MedPAR, and outpatient claims Research Identifiable Files, excluding emergency-room care and skilled nursing facilities. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.15: Home-market effect is stronger for rare services controlling for patient engagement

	(1)	(2)
Provider-market population (log) × common × high engagement	-0.0355	-0.0354
	(0.0349)	(0.0347)
Provider-market population (log) × rare × low engagement	0.231	0.244
	(0.0482)	(0.0370)
Provider-market population (log) × rare × high engagement	0.481	0.360
	(0.0808)	(0.141)
Patient-market population (log) × common × high engagement	0.0440	0.0450
	(0.0257)	(0.0255)
Patient-market population (log) × rare × low engagement	-0.146	-0.125
	(0.0374)	(0.0243)
Patient-market population (log) × rare × high engagement	-0.477	-0.575
	(0.0923)	(0.271)
Distance (log) × common × high engagement	-0.0548	0.146
	(0.0209)	(0.118)
Distance (log) × rare × low engagement	0.0488	0.716
	(0.0375)	(0.171)
Distance (log) × rare × high engagement	-0.127	2.458
	(0.0814)	(2.764)
Distance (log, squared) × common × high engagement		-0.0193
		(0.0100)
Distance (log, squared) × rare × low engagement		-0.0615
		(0.0152)
Distance (log, squared) × rare × high engagement		-0.277
		(0.324)
Observations	226,936	226,936
Distance controls	Linear	Quadratic
Patient-provider-market-pair FEs	Yes	Yes
Additional distance elasticity at mean for high engagement: common procedures	-0.05	-0.13
Additional distance elasticity at mean for high engagement: rare procedures	-0.18	-1.34

Notes: This table reports estimates of a variant of equation (2.11), which adds interactions with indicators for whether a procedure is "rare" (provided less often than the median procedure) and for whether a procedure is "high engagement" (median number of distinct claims per patient for the procedure in a given year is above one) or low engagement. The unit of observation is {rare indicator, high-engagement indicator, exporting HRR, importing HRR}, and the dependent variable is the value of trade. Each column includes fixed effects for each ij pair of patient market and provider market, rare versus common procedures, and high- versus low-engagement procedures, plus indicators for three categories (common × high-engagement, rare × low-engagement, and rare × high-engagement) interacted with patient- and provider-market populations and distance covariates. Covariates for common × low-engagement procedures are omitted, since they would lead to collinearity with the ij fixed effects. Column 2 adds a control for the square of log distance and its interactions. The negative coefficient on provider-market population and the positive coefficient on patient-market population for common and high-engagement procedures indicate that the home-market effect is slightly less *strong* compared to common and low-engagement procedures, even though these effects are not all statistically different from zero. The positive coefficient on provider-market population × rare and the negative coefficient on patient-market population × rare for both high- and low-engagement procedures indicates that the *strong* home-market effect is stronger for rare services, whether they are high- or low-engagement. The distance elasticity is more negative for high-engagement procedures (both rare and common). Trade flows are computed from the Medicare 20% carrier Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 2.16: Gravity regression by procedure: individual procedures exhibit a strong home-market effect

	(1)	(2)	(3)	(4)	(5)	(6)
Procedure:	Colonoscopy	Cataract surgery	Brain tumor	Brain radiosurgery	LVAD	Colon removal
HCPCS code:	G0121	66982	61510	61798	33979	44155
Provider-market population (log)	0.515	0.466	0.928	1.148	1.251	0.992
	(0.0690)	(0.0729)	(0.0884)	(0.119)	(0.168)	(0.165)
Patient-market population (log)	0.351	0.436	0.191	0.165	0.181	-0.143
	(0.0692)	(0.0690)	(0.0726)	(0.0817)	(0.141)	(0.147)
Distance (log)	0.446	0.965	1.018	1.484	2.176	3.097
	(0.395)	(0.495)	(0.534)	(0.686)	(0.910)	(1.630)
Distance (log, squared)	-0.217	-0.270	-0.268	-0.304	-0.366	-0.500
	(0.0394)	(0.0491)	(0.0564)	(0.0697)	(0.0922)	(0.171)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.100	0.407	0.000	0.000	0.000	0.000
Observations	93,636	93,636	93,636	93,636	93,636	93,636
Distance elasticity at mean	-2.66	-2.90	-2.82	-2.87	-3.07	-4.07
Total count	58,798	43,604	1,922	752	333	112

Notes: This table reports estimates of equation (2.8) for procedure-level trade for six selected HCPCS codes, which vary in how common they are. For all procedures, the sample is all HRR pairs ($N = 306^2$). The dependent variable in all regressions is the value of trade in the procedure (computed using each procedure's national average price). The independent variables are patient- and provider-market log population, log distance and square of log distance between HRRs, and an indicator for same-HRR observations ($i = j$). The positive coefficient on provider-market log population implies a weak home-market effect, and the fact that this coefficient exceeds that on patient-market population implies a strong home-market effect. Trade flows are computed from the Medicare 20% carrier Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market. The bottom row reports the total national count of the procedure in our sample. Common procedures include screening colonoscopy (column 1) and cataract surgery (column 2). In a screening colonoscopy, the physician visualizes the large bowel with a camera to look for cancer. In a cataract surgery, the surgeon removes a cloudy lens from the eye to improve vision. Relatively rare procedures include brain radiosurgery (column 3), brain tumor removal (column 4), left ventricular assist device (LVAD) implantation (column 5) and colon removal (column 6). In brain radiosurgery, an area of the brain is irradiated, often to kill a tumor. In an LVAD implantation, a pump is implanted in the chest to assist a failing heart in pumping blood. Brain tumor and colon removals involve surgical removal of the respective structures.

Table 2.17: Scale elasticity estimates for CBSAs

Panel A: All services	Baseline	No Diagonal	Controls
OLS	0.888	1.052	0.907
	(0.009)	(0.017)	(0.010)
2SLS: population (log)	0.845	1.023	0.852
	(0.010)	(0.016)	(0.013)
2SLS: population (1940, log)	0.848	0.928	0.851
	(0.014)	(0.025)	(0.017)
2SLS: bedrock depth	0.810	0.762	0.812
	(0.038)	(0.099)	(0.044)
Panel B: Rare services			
OLS	0.941	1.108	0.945
	(0.010)	(0.028)	(0.011)
2SLS: population (log)	0.914	1.106	0.909
	(0.013)	(0.026)	(0.016)
2SLS: population (1940, log)	0.942	1.019	0.941
	(0.017)	(0.044)	(0.022)
2SLS: bedrock depth	0.814	0.095	0.807
	(0.063)	(0.393)	(0.078)

Notes: This table reports estimates of α from ordinary least squares (OLS) or two-stage least squares (2SLS) regressions of the form $\widehat{\ln \delta}_i = \alpha \ln Q_i + \ln R_i + \ln w_i + u_i$ using core-based statistical areas (CBSAs) as the geographic units. The dependent variable $\widehat{\ln \delta}_i$ is estimated in equation (2.12), Q_i is region i's total production for Medicare beneficiaries, R_i is Medicare's Geographic Adjustment Factor, the w_i covariate includes mean two-bedroom property value and mean annual earnings for non-healthcare workers, and u_i is an error term. In the rows labeled "2SLS" we instrument for $\ln Q_i$ using the specified instruments. The $\ln R_i$ and $\ln w_i$ controls are omitted in the columns labeled "no controls". In the columns labeled "no diag", Q_{ii} observations were omitted when estimating $\widehat{\ln \delta}_i$ in equation (2.12). Standard errors (in parentheses) are robust to heteroskedasticity.

Table 2.18: Classification of rare and common procedures in Medicare vs. private insurance data

Above median HCCI Above median CMS	0	1	total
0	82	18	100
1	12	88	100

Notes: This table compares the percentage of procedures classified as rare (above median frequency equals one) or common (above median frequency equals zero) in the public Medicare data versus the private insurance data from the Health Care Cost Institute (HCCI). Classifying procedures as rare versus common is consistent when using Medicare or privately insured data.

Table 2.19: Specialization earnings and frequency

	(1)	(2)	(3)
Dependent variable: Per capita population elasticity			
Number of physicians in specialization (log, national)	-0.0716		-0.0677
	(0.0139)		(0.0137)
Mean earnings (log)		-0.245	-0.174
		(0.0697)	(0.0543)
Observations	209	209	209
R-squared	0.199	0.050	0.223

Notes: This table reports estimates of a regression of per capita population elasticity of physician count on the national count of physicians and mean earnings. Each observation is an NPPES taxonomy code. Earnings (wage and business income) data from Gottlieb et al. [2020] are reported by Medicare specialty groups. We use a crosswalk to map Medicare specialty groups to NPPES taxonomy codes. The estimation sample excludes 11 taxonomy codes that are not mapped to any Medicare specialty. Standard errors (in parentheses) are robust to heteroskedasticity.

Table 2.20: Larger markets produce a greater variety of procedures

	(1) All	(2) Anesthesia	(3) E&M	(4) Medical	(5) Other	(6) Pathology	(7) Radiology	(8) Surgery
Population (log)	0.357	0.292	0.169	0.294	0.428	0.358	0.201	0.400
	(0.00736)	(0.0132)	(0.00528)	(0.00663)	(0.0115)	(0.0204)	(0.00610)	(0.00959)
Observations	306	306	306	306	306	306	306	306

Notes: This table reports the population elasticity of the number of distinct service codes produced in a region, estimated using Poisson pseudo-maximum likelihood (PPML). Column 1 shows the coefficient including all service types. The remaining columns show the coefficients for specific categories of service types. We use procedure classifications from the American Academy of Professional Coders, which groups codes into surgeries, anesthesia, radiology, pathology, medical, and evaluation & management ("E&M") services [AAPC, 2021]. We combine Category II codes, Category III codes and Multianalyte Assays into "other."

CHAPTER 3

KNOWLEDGE GROWTH AND SPECIALIZATION

3.1 Introduction

Many fields, such as computer science, molecular biology, and medicine, have a rapidly growing knowledge base. Do expert workers respond to growth in knowledge by becoming more specialized? We study this empirically in the context of oncology, which has experienced explosive growth in knowledge. Using a panel of Medicare claims data and historical cancer treatment guidelines, we test if New advancements occur at a breakneck pace in computer science, basic sciences, and medicine[1]. Experts must stay at the frontier of this rapidly growing knowledge base. However, staying up to date may be costly. Specialization limits the information one must keep up with and the required working knowledge to be at the frontier to a manageable scope. Several economic theories predict that workers should become more specialized as general knowledge increases [Becker and Murphy, 1992, Jones, 2009]. Do expert workers respond to growth in the depth of knowledge by increasing worker specialization?

We empirically evaluate how specialization responds to growth in knowledge in the context of medical oncology. Medical oncologists are physicians who treat cancer patients medically (i.e., non-surgically). Cancer is a condition with a vast demand for treatment; Cancer is the second leading cause of death in the United States [Centers for Disease Control and Prevention, 2022] and the lifetime risk of cancer is 40% [American Cancer Society, 2020].

As a result, there has been substantial research and innovation in cancer drugs over the last few decades. For example, we find that the total length of National Comprehensive Cancer Network Guidelines, a "complete library" of cancer guidelines, has grown from 1,075 pages in 2002 to 5,760 pages in 2020. Similarly, the total number of FDA-approved anti-cancer drugs increased from 90 to 243 (data from Pantziarka et al. [2021]).

1. Research in collaboration with Pauline Mourot.

We conceptualize the relationship between general knowledge growth and specialization in this context using the model from Becker and Murphy [1992]. Oncologists can treat every type of cancer patient or collaborate with other oncologists to divide patients by cancer type. If they divide patients by cancer type, they will produce a higher quantity or quality of patient care. For example, an oncologist specializing in breast cancer may spend less time keeping up to date with various clinical guidelines, allowing higher patient throughput. They may also have more experience with breast cancer therapies, leading to higher quality of care. However, oncologists may also face coordination costs that limit their ability to specialize. These include the types of challenges that many organizations face, including contracting, incentive, and agency issues. The optimal team size and degree of specialization balance the benefits of greater specialization with the increased coordination costs of larger teams. Becker and Murphy [1992] postulate that increasing general knowledge increases the returns to specialization and pushes the equilibrium worker to a more specialized role in a larger team. We aim to estimate this relationship empirically.

To do so, we combine a novel collection of historical cancer guidelines data with a 21-year panel of Medicare claims data. We proxy the growth in clinically relevant knowledge with the increase in the length of comprehensive cancer guidelines. We construct a simulated instrument of exposure to knowledge growth. In this instrument, we sum the increase in guidelines across all fields of cancer that a physician works in. As such, this instrument combines two sources of variation: differences in the sets of cancer subfields that physicians work in and the differential rates of knowledge growth observed across these subfields.

This simulated instrument suffers from omitted variables bias, as noted generally of formula instruments in Borusyak and Hull [2020a]. In particular, oncologists who work in more fields are mechanically more exposed to higher levels of knowledge growth. To purge this bias, we recenter this simulated instrument using the general approach developed by Borusyak and Hull [2020a]. Our recentering procedure eliminates the correlation between

the simulated instrument and key physician observables, including physicians' region size and the number of other oncologists in the same organization.

We then test how exposure to knowledge growth influences oncologist specialization. Our primary measure of specialization is a physician-level Herfindahl–Hirschman index (HHI) calculated over the share of clinical management work in each cancer subfield. Oncologists who start specialized become significantly more specialized in response to knowledge growth. In contrast, oncologist who start out as general oncologists do not become more specialized in response to knowledge growth. Such findings indicate the possibility of high coordination costs in oncology when interpreted through the framework of Becker and Murphy [1992].

First, we contribute to the literature on specialization. Explanations for specialization are as old as economics itself [Smith, 1776]. In a seminal paper, Becker and Murphy [1992] hypothesized that specialization is limited by coordination costs and the extent of general knowledge. They model an increase in general knowledge as an increase in the returns to time spent learning skills to do specific subtasks of production. They then show that their model predicts that specialization and team size will increase if the amount of general knowledge increases. This seminal work establishes a model but does not include an empirical analysis. To our current knowledge, the proposed project would be the first to empirically test the predictions of this foundational model.

This work also contributes to the health economics literature on the causes of medical specialization. Specialization has been increasing in medicine for the last century [Dalen et al., 2017]. Baumgardner [1988b] shows that market size influences the scope of practice, showing that general practitioners perform a narrow range of activities in larger markets. Meltzer and Chung [2010] show that specialization between two types of clinical settings is driven by how frequently the same patients use both settings. To our knowledge, we contribute the first robust empirical examination of the extent to which knowledge growth contributes to growing specialization.

We also contribute to the literature about the consequences of knowledge growth. [Jones, 2009] documents that a growing "burden on knowledge" on inventors, who are knowledge producers, leads to greater specialization in fields with deeper knowledge. We extend the investigation of the "burden of knowledge" to physicians, who are primarily users of knowledge [Jain et al., 2019]. Our results suggest that the burden of knowledge weighs less heavily on users than on producers of knowledge.

Finally, this work contributes a novel examination of the cumulative impact of technological change to an extensive empirical literature on the consequences of technological change in medicine. Most of this literature focuses on single technologies, each with varied economic impacts on providers and patients Skinner [2011], Chandra et al. [2014], Skinner and Staiger [2015], Arrow et al. [2020]. Our work is novel in that it considers the cumulative effect of technology on the organization of production in healthcare.

3.2 Data

3.2.1 Guidelines Data

To measure growth in knowledge, we create a novel dataset of historical cancer guidelines data from the National Comprehensive Cancer Network. The National Comprehensive Cancer Network (NCCN) is a prominent "not-for-profit alliance of 33 leading cancer centers devoted to patient care, research, and education" [NCCN, 2023a]. In 1996, they started releasing treatment guidelines for common types of cancer. In 2001, they released the first "complete library" of cancer treatment guidelines [NCCN, 2023b]. The NCCN guidelines are the leading set of cancer treatment guidelines. They represent an expert consensus on approaches to cancer treatment and are updated multiple times a year to reflect the most up-to-date evidence. They are widely used by oncologists; Per one survey, "96% of oncologists think that the NCCN guidelines are important to use when making decisions about patient care"

[McGivney, 2008].

The guidelines contain recommendations for all stages of the cancer treatment process, including "diagnosis, imaging, drug therapies, radiation, [and] surgery" [McGivney, 2008]. All of this information is relevant to medical oncology, as oncologists have two key roles in cancer care. The first role is managing "systemic therapies", primarily drug treatments like chemotherapy. The second role is to serve as the central node in a network of other specialists who may be involved in a patient's cancer care, e.g., cancer surgeons or pathologists. As such, oncologists are trained in "comprehensive" cancer care management and must stay up to date with the high-level recommendations for all types of cancer treatments [Popescu et al., 2014].

The guidelines serve as a proxy for the volume of relevant clinical knowledge. The guidelines are organized anatomically, with separate documents for different cancer types, such as breast and colorectal cancer. We proxy the volume of knowledge in a given year for a type of cancer as the page length of the first set of guidelines per year for that cancer type.

This proxy relies on several key assumptions to be a valid measure of the volume of knowledge. First, this proxy assumes that each page represents an equal volume of knowledge. Reassuringly, over the main time period for our research design (2008-2014), the formatting of the guidelines was unchanged, suggesting that increases in page length result from more content, not changes in formatting. Second, the proxy assumes that any relevant knowledge not included in the guidelines is either exactly proportional to the volume of information the guideline pages contain or remains constant over time. For example, the guidelines may not include all relevant information about a chemotherapy drug, but we assume it contains a consistent share of the information about each drug. Similarly, the cancer treatment guidelines are built on a foundation of pre-existing knowledge in fields like internal medicine. We assume that the relevant volume of knowledge in internal medicine does not substantially increase over this time. Therefore, the change in the length of guidelines reflects the full

change in the volume of relevant knowledge.

3.2.2 Medicare Claims Data

We also use a 21-year panel of Medicare claims data. Medicare provides federally sponsored health insurance coverage for disabled adults and most U.S. adults over sixty-five. This age group bears the highest cancer incidence in the U.S., making it the ideal population to study cancer care for Disease Control and Prevention [2019]. We use a 20% sample of Medicare Part B Claims from 1999-2019.

Using this data, we construct an annual panel of oncologists who meet minimum patient counts. First, we restrict our sample to physicians whose most commonly listed specialty in a year is Hematology, Medical Oncology, or Hematology-Oncology. For our main analysis, we restrict to a balanced panel of oncologists from 2008 to 2014 who meet minimum patient volumes as described below.

Next, we measure which types of cancer an oncologist manages. For this, we restrict our sample to the evaluation and management claim lines billed by each oncologist. As such, this restricts our sample to instances where the oncologist managed care and decision-making for a patient, which requires high-level medical knowledge and judgment.

We then assign each evaluation and management visit to a cancer type based on the patient's primary cancer type. We determine a patient's primary cancer as the specific type of cancer for which they have the most diagnosis codes in that year from any type of procedure. For example, if a patient has diagnosis codes for both "Unspecified" cancer and breast cancer, they are considered to have breast cancer. If an oncologist sees that patient, they are denoted as having managed a breast cancer patient. Among this sample, we further restrict to physicians who evaluated and managed at least 25 unique cancer patients in that year. This minimum patient count ensures that our measurements are based on a sufficient patient volume to capture a significant share of the types of cancer an oncologist treats.

We also measure the share of time they spend managing each type of cancer. Each claim line in our sample contains an evaluation and management HCPCS code, which we map to work Relative Value Units (wRVUs). Work Relative Value Units are measures Medicare uses to estimate the amount of physician work various HCPCS codes require. In the context of evaluation and management codes, work Relative Value Units reflect the expected time required to perform various types of evaluation and management. We use this to proxy the share of time that each oncologist spends managing each type of cancer, which we use later to measure physician specialization.

3.2.3 Medicare Data on Provider Practice and Specialty

We also use the Medicare Data on Provider Practice and Specialty (MD-PPAS) to measure physician characteristics. This variable allows us to observe the physician's birth year and back out their age at the start of our panel period. It also contains the top two Tax ID numbers from which each oncologist bills. Multiple physicians can share the same tax ID number, which implies they are in the same organization or group practice. As such, this variable provides information about if a physician is a solo provider or in a bigger organization.

3.2.4 Measuring relevant knowledge growth for each oncologist

Next, we combine the Medicare claims and guidelines data to measure our key treatment variable, the growth in relevant knowledge for each oncologist. First, for each oncologist, we determine the set of cancer types they manage in a baseline year, in this case, 2008. The set of cancers for oncologist i in 2008 can denoted with C_i. Each cancer, c, has an increase in relevant guideline page count Δk_c over the following five years (2009 to 2014). We assume that the full guideline and change in guidelines for cancer c is relevant to a physician if they treat cancer c in the baseline year. With this assumption, we can calculate oncologist i's exposure to knowledge growth as $\Delta k_i = \sum_{C_i} \Delta k_c$.

For example, suppose oncologist i treats only lung and breast cancer in 2008. Also, suppose that over the following five years, the lung cancer guidelines increased by 50 pages ($\Delta k_{lung} = 50$) and the breast cancer guidelines increased by 100 pages ($\Delta k_{breast} = 100$). Then, oncologist i would be exposed to 150 pages of knowledge growth over the following five years (2009 to 2014).

We custom group ICD-9 diagnosis codes to map from cancer types treated in claims to guidelines. ICD-9 diagnosis codes are too granular to be used directly as types of cancer treated. For example, they contain separate codes for breast cancer in the right breast versus the left breast. We wish to group these into the category of "breast cancer".

We create a set of custom diagnosis code groupings that align with how cancers are grouped in guidelines. We start with a list of cancer diagnosis codes from the Clinical Classification Software (CCS). The CCS also has groupings to the level of cancer type, but these groupings have several challenges that we rectify with our custom grouping. First, we drop the CCS diagnosis codes for cancer treatment and consider only those for specific types of cancer. Second, we make our groupings more specific. For example, the CCS groupings classify colon cancer of "unknown behavior" as an "unknown" cancer type. We classify this as colon cancer. Third, we make our custom grouping more granular to align with the guidelines as much as possible. We divide the CCS category of "Leukemia" into "Myeloid Leukemia" and "Lymphocytic Leukemia" (guidelines are even further split by acute and chronic within each subtype of leukemia). Lastly, we rearrange groupings that are misaligned with clinical groupings. For example, the CCS groupings group rectal cancer with anal cancer. However, guidelines group rectal cancer with colon cancer into a single set of guidelines for colorectal cancer. Our custom grouping mirrors the guidelines, with a single category for colorectal cancer. These mappings are based on ICD-9 codes. In mid-2015, the claims data switches from ICD-9 to ICD-10 codes. In these years, we map from ICD-10 to ICD-9 codes, then to our custom cancer grouping.

Notably, these groupings are largely anatomical, which mirrors the primary dimension of cancer subtyping. The anatomical origin of the source of the cancer is the first dimension along which cancers are grouped in guidelines and diagnosis code groupings. For example, breast cancers are cancers where the origin tumor site is the breast. Leukemias are cancers that originate from the white blood cells of the immune system. Within each anatomical origin, there are further subclassifications along numerous other dimensions, e.g., stage and biomarkers. However, we are reassured that anatomic origin is the first dimension along which cancers are initially subtyped by the fact that NCCN guidelines, which leading cancer care organizations create, are also organized anatomically. This observation suggests that our measures of cancer types align with how the field of oncology distinguishes among subtypes.

3.2.5 Measuring Physician specialization

Next, we discuss how we measure our outcome of interest: oncologist specialization. Our main measure is a Herfindahl–Hirschman Index (HHI) over the share of time that an oncologist spends managing each type of cancer. We calculate HHI as follows, where T_{ic} is time (proxied by work RVUs) that oncologist i spends managing cancer c and T_i is the total time they spend managing cancer: $HHI_i = 10,000 \sum_C \left(\frac{T_{ic}}{T_i}\right)^2$. Observe that HHI_i has a maximum value of 10,000, achieved if an oncologist spends all their time managing one type of cancer. If an oncologist divides their time equally between two types of cancer, then HHI_i is 5,000. With four cancers equally split, HHI is 2,500, and with five, it is 2,000. The lower limit of HHI_i is 0, which is approached as an oncologist spends an infinitesimally small amount of time on an infinite number of cancer types.

This measure has several key advantages. First, HHI is continuous and based on the work that an oncologist performs in a year. This measure contrasts with more common measures of specialization, which are often based on job titles. Second, it is sensitive to subtle intensive margin changes in specialization. The measure will capture if an oncologist

shifts towards focusing on one particular type of cancer, even if they do not entirely stop seeing other types of cancers. As such, it can capture differences in specialization for the same worker over time and across workers with identical job titles.

3.3 Identification

The following section discusses our identification strategy. We regress changes in specialization on exposure to knowledge growth. Our proxy of exposure to knowledge growth is a formula instrument and suffers from omitted variables bias. We discuss our approach to remove this bias by recentering the instrument for expected exposure to knowledge growth.

3.3.1 Estimating Equation

First, we discuss the main estimating equation. Denote physician i's change in specialization between 2009 and 2014 as ΔHHI_i. Denote their exposure to knowledge growth as Δk_i. Recall that exposure to knowledge growth is calculated as $\Delta k_i = \sum_{C_i} \Delta k_c$, where C_i is the set of cancers that physician i treats in 2008 and Δk_c is the change in page length for each type of cancer between 2009 and 2014. Also, denote the vector of other physician characteristics as X_i. The main estimating equation is as follows:

$$\Delta HHI_i = \beta \Delta k_i + \theta X_i + \epsilon_i$$

We estimate this equation using ordinary least squares with robust standard errors.

This estimating equation has several key features. First, the outcome uses within physician variation in specialization over time. This within-physician design requires a balanced panel. It also allows for specialization to respond immediately to knowledge growth. If the specialization response is highly lagged, that may attenuate our estimates.

The main coefficient of interest is β, the effect of knowledge growth on specialization. β

is estimated using variation in exposure to knowledge growth (Δk_i), which is a simulated or "Bartik-style" instrument.

Variation in exposure to knowledge growth (Δk_i) comes from two sources. The first source of variation is that physicians work in different sets of cancer subfields. For example, some treat breast cancer, some treat leukemia, and some treat both. This variation is analogous to variation in industry shares in a Bartik instrument. The second source of variation is that different cancer subfields have different amounts of knowledge growth. For example, leukemia may have more knowledge growth than breast cancer, so physicians who treat only leukemia are more exposed to knowledge growth than those who treat only breast cancer. This variation is analogous to variation in industry-level shocks in a Bartik instrument.

3.3.2 Recentering instrument to correct for omitted variables bias

Importantly, simulated instruments generally suffer from omitted variables bias, which must be corrected for to estimate accurate causal effects [Borusyak and Hull, 2020b]. In this context, the omitted variables bias arises from the fact that physicians who treat more types of cancer are mechanically exposed to more knowledge growth. For example, someone who treats only breast cancer will be less exposed to knowledge growth than someone who treats both breast cancer and leukemia. The breast cancer only oncologist may differ in many important ways from the oncologist who treats both cancers. For example, breast cancer only oncologists may be more likely than general oncologists to work at a large medical center that enables such specialization. Oncologists in large medical centers may have different specialization trends than those in solo practice for reasons unrelated to knowledge growth. This variation in organization size will be correlated with both exposure to knowledge growth (Δk_i) and with change in specialization (ΔHHI_i), leading to omitted variables bias in the estimate of β.

We address this omitted variables bias using the recentering approach proposed by

Borusyak and Hull [2020b]. This approach removes the non-random exposure to knowledge growth that drives the omitted variables bias. In our case, the non-random exposure results from the breast cancer specialist being predictably less exposed to knowledge growth than the generalist.

To remove the non-random variation, we estimate expected exposure to knowledge growth ($\mathbb{E}[\Delta k_i]$) and remove it from realized exposure to knowledge growth Δk_i. Our measure of realized exposure to knowledge growth, Δk_i, can be decomposed into two components.

$$\Delta k_i = \underbrace{\mathbb{E}[\Delta k_i]}_{\text{OVB}} + \underbrace{\Delta \tilde{k}_i}_{\text{Remaining Variation}}$$

The first component, $\mathbb{E}[\Delta k_i]$, is the expected exposure to knowledge growth and contains the omitted variables bias. The second component is the remaining variation and is assumed to be exogenous and uncorrelated with other factors that impact specialization except for knowledge growth.

This expression can be rearranged as follows:

$$\begin{aligned}\Delta \tilde{k}_i &= \Delta k_i - E[\Delta k_i] \\ &= \sum_{C_i} (\Delta k_c - \mathbb{E}[\Delta k_c])\end{aligned}$$

We can estimate the residual exposure to knowledge growth, $\Delta \tilde{k}_i$, by subtracting the expected exposure ($\mathbb{E}[\Delta k_c]$) from the realized exposure (Δk_i). Realized exposure (Δk_i) is calculated exactly from our data. Calculating expected exposure to knowledge growth ($\mathbb{E}[\Delta k_c]$) requires us to make assumptions about the expected knowledge growth per cancer subtype $\mathbb{E}[\Delta k_c]$. In the next section, we will discuss how we construct this expectation and the implications for the substance of the identification assumptions.

Once we have this expectation, we can then substitute our realized exposure, Δk_i, with

the residual exposure, $\Delta\tilde{k}_i$, for an unbiased and efficient estimate of β, as below:

$$\Delta HHI_i = \beta\Delta\tilde{k}_i + \theta X_i + \epsilon_i$$

The formal identification assumption here is that this residual exposure to knowledge growth is uncorrelated with the contents of the error term, including omitted variables ($\Delta\tilde{k}_i \perp \epsilon_i$). The real-world content of this mathematical assumption depends on exactly how we form our expectation of knowledge growth by cancer type ($\mathbb{E}[\Delta k_c]$).

3.3.3 Estimating Expected Knowledge Growth by Cancer Type

This section discusses different assumptions around expected knowledge growth by cancer type and the corresponding identification assumptions.

Our first approach assumes all cancer fields have the same expected knowledge growth. Specifically, we assume that $\mathbb{E}[\Delta k_c] = \hat{\mu}$ where $\hat{\mu}$ is the mean page increase across cancer types, weighted by the number of oncologists who manage that cancer type. Let n_i denote the number of cancer types that an oncologist treats. $\Delta\tilde{k}_i$ is estimated as follows:

$$\begin{aligned}\Delta\tilde{k}_i &= \sum_{C_i}(\Delta k_c - \hat{\mu}) \\ &= \Delta k_i - n_i\hat{\mu}\end{aligned}$$

This approach is equivalent to controlling for the number of cancers an oncologist treats [Borusyak and Hull, 2020b]. Essentially, it compares doctors who treat the same number of cancers. The variation in $\Delta\tilde{k}_i$ arises because oncologists vary in the composition of the set of cancers they treat, holding the set size constant, and some cancers have more innovation than others.

Critically, this approach assumes that cancer-level knowledge shocks are unpredictable

and uncorrelated with physician sorting. We are particularly concerned that cancer-level knowledge shocks are not entirely unpredictable. A key concern is that more common cancers likely draw more investment in innovation [Acemoglu and Linn, 2004b]. In addition, more common cancers may also support more growth in specialization for the oncologists who work in that field since it is easier to find sufficient patients. Addressing this potential bias motivates our next approach.

Our second approach allows expected knowledge growth to vary with cancer characteristics. Acemoglu and Linn [2004b] finds that conditions with more potential patients had more innovation. Budish et al. [2015] finds that cancers with surrogate endpoints have more innovation. We incorporate both of these factors into our estimate of expected knowledge growth, $\mathbb{E}[\Delta k_c]$

Specifically, using a Poisson regression, we predict the expected amount of knowledge growth for a cancer type based on these cancer characteristics. Let p_c denote the share of cancer patients in Medicare with each type of cancer in 2008. Also, define e_c as an indicator for if a cancer type has surrogate endpoints per Budish et al. [2015] ($e_c = \mathbb{1}(\textit{surrogate endpoints})$). We include both of these in a Poisson regression to predict the expected page count increase for each type of cancer, as follows:

$$E[ln\Delta k] = \beta_0 + \beta_1 e_c + \beta_2 p_c + \beta_3 e_c p_c$$

Note that this equation also controls for knowledge shocks shared across all cancer types (loaded on β_0)

Denote the predicted knowledge growth from this regression as $\Delta\hat{k}_c$. Then $\Delta\tilde{k}_i$ is estimated as follows:

$$\Delta\tilde{k}_i = \sum_{C_i} (\Delta k_c - \Delta\hat{k}_c) \tag{3.1}$$

This approach controls for the number of cancers an oncologist treats, weighting each cancer

by the expected innovation.

This approach makes several key assumptions. First, it assumes that the volume of patients, presence of secondary endpoints, and shared knowledge shocks are the key factors correlated to highly predictable variation in innovation across cancer types. Any residual variation in knowledge growth by cancer type is unrelated to oncologists' specialization decisions except through its direct effect on specialization. We attribute the residual variation to randomness and uncertainty in the research and development process, e.g., some clinical trials fail and others succeed. In the next section, we empirically test if this recentering approach purges measurable omitted variables bias.

3.4 Descriptive Patterns

In this section, we describe key characteristics of our sample and patterns of specialization over time.

3.4.1 Knowledge growth

Analysis of historical guidelines reveals that cancer treatment guidelines have become many times longer since 2002. Figure 3.1a displays the length of guidelines between 2000 and 2020 on average and for selected cancer types. Appendix Table 3.8 lists the page count for each type of guideline in 2002 and 2020. The total length of guidelines has increased over five-fold in this period, from 1,075 pages in 2002 to 5,760 pages in 2020.

This growth in guidelines may be driven by many factors, but at least one is an explosion in the number of drugs available to treat cancer. The number of FDA-approved anti-cancer drugs increased from 90 in 2002 to 243 in 2020. Figure 3.1b displays the number of U.S. Food and Drug Administration approved anti-cancer drugs over time. Not only has the total number of drugs been increasing steadily over time, the rate at which new drugs are introduced has been accelerating. Data on anti-cancer drugs is from Pantziarka et al. [2021]

and uses Anatomical Therapeutic Chemical Level 5 classifications to define individual drugs.

3.4.2 Physician Summary Statistics

Next, we describe summary statistics for our balanced panel of physicians from 2008 to 2014. Table 3.1 shows summary statistics for key variables of interest for a balanced panel of 4,854 oncologists. On average, they see 69 unique cancer patients in the baseline year, 2008. On average, they see 15 distinct types of cancer, but some see as few as one type of cancer, and others see as many as 27 types. Their average age at the beginning of the panel is 50 years old. The mean oncologist HHI is 2,737. Over half of oncologists are in the largest one-fifth of hospital referral regions by population. On average, 20 other oncologists work at their Tax ID numbers, but this can be as low as zero and as high as 245.

In the initial year, the total length of NCCN guidelines is 1,595 pages, but only 1,014 pages of these are relevant to each oncologist, on average. Over the following five years, the total and relevant number of guidelines more than double, increasing by 1,745 pages in total and 1,088 relevant pages on average.

3.4.3 Trends in Oncologist Specialization Over Time

Next, we characterize the distribution of oncologist specialization over time. Figure 3.2 shows the 10th, 25th, 50th, 75th, and 90th percentile of oncologist HHI between 1999 and 2019. It also plots the HHI overall, using all evaluation and management services from oncologists. Several striking trends stand out.

First, there is a sharp, discontinuous decline in HHI at all points in the distribution between 2014 and 2016, corresponding to the switch from ICD-9 to ICD-10 midway through 2015. This clear trend break motivates the decision to end our main balanced panel in 2014, which prevents coding changes from contaminating our estimates of change in HHI.

Second, the typical oncologist has not become substantially more specialized over time.

Median specialization and below is flat or declining slightly over this period. The distribution's 10th, 25th, and 50th percentiles are all close to each other and the HHI of the field. This result suggests that most oncologists are "general" oncologists who see a similar distribution of cancer patients as the field overall. The 75th percentile becomes slightly more specialized, but this increase is modest. This percentile of physician HHI increases by a modest 158 points between 1999 and 2014 and another 182 points between 2016 and 2019. This lack of growth in specialization is surprising, given the explosive growth in knowledge and treatment options for cancer over these 21 years.

In contrast, subspecialized oncologists have become dramatically more specialized over time. The 90th percentile of oncologist HHI increases by over 1000 HHI points from 1999 to 2014 and another 371 points from 2016 to 2019.

These contrasting time trends suggest that "generalist" and "specialist" oncologists may respond to knowledge growth differently. Specialists may respond; generalists likely do not.

3.4.4 Differences Between More and Less Specialized Oncologists

We split our sample into "generalist" and "specialist" oncologists to understand how they differ and to later test if they respond differently to knowledge growth. Table 3.2 shows the characteristics of each group in our balanced panel. We define generalists as those below the 75th percentile of physician HHI in the initial year of the panel. We define specialists as those at or above the 75th percentile of the distribution.

Substantial differences are present between these two groups in the baseline year. In 2008, specialists have an average HHI of 4,637, over double that of the generalists, who have an HHI of 2,104 on average. Specialists are also slightly lower volume; they see 59 unique patients per year in the dataset, compared to 73 for the generalists. Specialists are also slightly more concentrated in big cities and work in slightly larger organizations. These differences are consistent with large firms and large markets facilitating specialization.

3.5 Results

In this next section, we discuss the results. First, we evaluate our measures of exposure to knowledge growth for their correlation with omitted variables and strength. Next, we test for the effect of knowledge growth on specialization using the recentered instrument for specialist and generalist oncologists.

3.5.1 Instrument Correlation with Omitted Variables

In this section, we evaluate our measures of exposure to knowledge growth. We find that the unadjusted measure strongly correlates with omitted variables, but the Poisson recentered measure removes this correlation. The Poisson recentered measure is also a strong instrument for the unadjusted measure. Both of these findings make the Poisson recentered measure our preferred measure.

First, we test if our measures of exposure to knowledge growth correlate with observable omitted variables that likely contribute to omitted variables bias. We follow the strategy in [Borusyak and Hull, 2020b] to test for such correlations. We regress our unadjusted measure (Δk_i) and recentered measures ($\Delta \tilde{k}_i$) of physician-level exposure to knowledge growth on key physician characteristics. These characteristics are market size (measured as the population of the physician's hospital referral region or HRR), firm size (proxied as the number of other oncologists at their tax ID numbers), and physician age, all in the baseline year.

These variables are all sources of concern about omitted variables bias. Table 3.3 shows the results for regressions of the raw measure (Δk_i), mean recentered measure ($\Delta \tilde{k}_{i,mean}$), and Poisson recentered measure ($\Delta \tilde{k}_{i,Poisson}$) on these omitted variables for the population overall. Appendix tables 3.9 and 3.10 show the results separately for specialist and generalist oncologists.

As hypothesized, the unadjusted measure of knowledge growth is significantly correlated with all three regressors. For example, an increase in the HRR population of 1 million

is associated with a decrease in exposure to knowledge growth of -23.1 pages (σ = 1.9) Physicians in larger markets likely treat fewer types of cancer and are mechanically exposed to less innovation. A regression using this unadjusted measure of exposure to knowledge growth will also inappropriately compare physicians in larger and smaller markets. These physicians may have different specialization trends for reasons unrelated to knowledge growth. Overall, the three regressions explain about 5% of the variation in exposure to knowledge growth. These results substantiate concerns about omitted variables bias in formula instruments and suggest that the recentering procedure is necessary for unbiased estimation.

The mean recentered measure of knowledge growth is substantially less correlated with all three regressors (market size, firm size, and physician age). For example, the correlation with market size is still significant but much weaker, with a coefficient of 5.7 ($\sigma = 0.8$). The regressors now explain only 1.6% of the variation in exposure to knowledge growth.

Finally, the Poisson recentered measure of knowledge growth largely purges all correlation between the measure and the regressors. The correlation with market size is much smaller and statistically indistinguishable from zero at -.05 (σ = .51). The same is true for our proxy of firm size. In the regression on the unadjusted measure, the coefficient is -.56 ($sigma = .08$). The coefficient shrinks for the Poisson adjustment measure and is statistically indistinguishable from zero at -.01 ($\sigma = .02$).

The magnitude of the correlation with physician age has also decreased substantially. Unfortunately, there is still a modest correlation, with a coefficient of -.561 ($\sigma = .01$). In other words, a 60-year-old oncologist is predicted to be exposed to 17 fewer pages of knowledge growth than a 30-year-old oncologist (on a mean exposure to knowledge growth of 1,088 pages). The combined regressors explain about .5% of the total variation in exposure to knowledge growth. As such, this correlation's size and explanatory power is small. We control for age in later specifications.

Overall, the Poisson recentering procedure nearly eliminates the correlation between the

exposure to knowledge growth and observable omitted variables. The fact that the correlation with observable predictors of exposure has been largely eliminated reassures us that the correlation with unobservable predictors has also largely been eliminated. The Poisson recentered measure is our favored measure of exposure to knowledge growth.

3.5.2 Instrument Strength

Next, we test if our Poisson recentered measure of knowledge growth is a "strong" instrument for the unadjusted measure. We have removed variation in this measure by recentering; how much remains?

To assess this, we regress our Poisson recentered measure ($\Delta\tilde{k}_{i,Poisson}$) on our unadjusted measure. (Δk_i) in a "quasi-first stage" regression. This regression is not a true first-stage regression because we do not use the predicted values in a second-stage regression to estimate the coefficient of interest. Rather, we use the recentered measure $\Delta\tilde{k}_{i,Poisson}$ directly to measure the coefficient of interest, per [Borusyak and Hull, 2020b]. However, the "quasi-first stage" regression can inform us about the extent of the variation that remains after recentering.

Figure 3.3 shows a binned scatter plot of the regression of the Poisson recentered measure ($\Delta\tilde{k}_{i,Poisson}$) on the unadjusted measure (Δk_i). Appendix figures 3.7a and 3.7b show the same results separately for specialist and generalist oncologists. The recentered measure is a highly significant predictor of the unadjusted measure ($\beta = 0.77$; $\sigma = .04$). The overall R^2 of the regression is 5.6%, which suggests that the recentered measure removes 94.4% of the variation in the unadjusted measure. However, the F statistic is 306, suggesting that the recentered measure is a "strong instrument" for the unadjusted measure by traditional metrics.

3.5.3 Estimates of Main Effects

Next, we estimate the effect of exposure to knowledge growth on physician specialization. All our estimates presented here use the Poisson recentered measure of exposure to knowledge growth in a balanced panel of oncologists from 2008 to 2014.

Figure 3.4 shows binned scatter plots and key regression estimates for the univariate relationship between the exposure to knowledge growth and change in specialization separately for specialist and generalist oncologists.

Figure 3.4a shows this relationship for oncologists who start in the top 25% of the initial specialization distribution, or "specialist" oncologists. The relationship between exposure to knowledge growth and increased specialization is large and significant. For every 100-page increase in relevant guidelines, these oncologists increase their specialization by 219 HHI points ($\sigma = 75$). Relative to baseline means, this represents a 4% increase in HHI for a 12% increase in relevant guidelines. Table 3.4 shows the regression table of results, adding fixed effects for age bin, hospital referral region, and tax ID size. The coefficient estimates and standard errors are unchanged after age and region-fixed effects are added. The coefficient estimate is slightly attenuated after adding fixed effects for Tax ID size, falling just below the threshold for significance at the 10% level (137.5, $\sigma = 84.0$). This pattern suggests that organization size mediates some of this relationship between knowledge growth and specialization.

Notably, the increases in specialization are driven by intensive margin changes in specialization. We test the effect of knowledge growth on extensive margin specialization - the count of unique types of cancer treated. Appendix table 3.6 displays the results for a regression of the change in the unique types of cancer treated by an oncologist on that oncologists' exposure to knowledge growth for specialized oncologists. The estimate is statistically indistinguishable from zero. The estimates are precise enough to rule out that exposure to 100 more pages of guidelines reduces the count of cancer types managed by one

or more. Specialized oncologists do not become more specialized on the extensive margin; the observed changes in HHI must be driven by changes on the intensive margin. In other words, specialized oncologists focus their practice more on some types of cancer but do not entirely stop seeing other types. This pattern could reflect oncologists continuing to see established patients but limiting their new patients to those with cancers in a narrower area of specialization.

Next, we consider how knowledge growth impacts general oncologists. Figure 3.4b shows the relationship between exposure to knowledge growth and HHI for oncologists starting in the bottom 75% of the initial specialization distribution, or "general" oncologists. In contrast to the results for specialists, general oncologists do not become specialized in response to knowledge growth. The estimated effect on HHI is more precise than the effects for specialists and statistically indistinguishable from zero at $\beta = -17.4$ ($\sigma = 23.3$). Table 3.5 shows the regression table of results, adding fixed effects for age bin, hospital referral region, and tax ID size. After adding these fixed effects, the coefficient estimates and standard errors are again unchanged. The coefficient estimate is still statistically indistinguishable from zero. In addition, there is no effect on the count of cancer types treated (results in figure 3.7)

Appendix figure 3.8 and appendix table 3.12 replicates the above analysis for oncologists overall. In aggregate, the effect of knowledge growth on both HHI and the count of cancer types treated is not significantly different from zero.

3.6 Discussion

This section discusses possible reasons for the heterogeneous response to knowledge growth in the context of classic theoretical constraints to specialization. We focus on two potential constraints to specialization: market size and coordination costs.

Smith [1776] was the first in a long line of many economists to observe that in many contexts, specialization is limited by the extent of the market. If the extent of the market

limits specialization, then workers should not become more specialized in response to knowledge growth. Instead, oncologist specialization should increase with market size. Figure 3.5a shows this theoretical relationship.

Becker and Murphy [1992] highlight that coordination costs between specialized workers may also limit specialization. On the one hand, workers become more productive as they become more specialized. On the other hand, production becomes split between greater numbers of workers, leading to higher coordination costs and ultimately constraining specialization. If coordination costs are a binding constraint in our setting, the oncologist specialization will remain flat across market size. Figure 3.5a shows this theoretical relationship.

In this framework, workers should become more specialized in response to knowledge growth, as knowledge growth raises the returns to specialization and thus increases the optimal degree of specialization. Of course, market size constraints may also bind for sufficiently small markets. In that case, specialization will increase only in large markets with sufficient market size to support the new, higher optimal level of specialization. Figure 3.5a also shows this theoretical prediction. Lastly, even within the Becker and Murphy [1992] framework, workers may not respond to knowledge growth appreciably if the returns to specialization are very low or coordination costs are very high. In that case, specialization will remain constant across space, even with substantial knowledge growth.

Each of these constraints to specialization produces differing predictions about the degree of specialization over market size and time. Next, we attempt to ascertain which theories are most consistent with the data and, thus, which constraints might be driving the observed heterogeneous responses to knowledge growth. We graph the median and 90th percentile of specialization by market size quintile in 1999 and 2014 in Figure 3.6a. Figure 3.6b shows the same graph for 2016 and 2019.

Based on these graphs, the median oncologist seems highly constrained by coordination

costs. Median specialization remains entirely flat across the distribution of market size in 1999. The median oncologist is not more specialized in the largest markets than the median in the smallest markets. This pattern is highly consistent with coordination costs limiting the specialization for the typical oncologist, as in Becker and Murphy [1992]. This pattern is the same in 2014, 2016, and 2019. The specialization of the median oncologist stays the same over time, too, even though there has been explosive growth in the volume of knowledge throughout this 21-year period. That suggests that either the returns to specialization are quite low or coordination costs are quite high in this context.

We find it somewhat implausible that low returns to specialization drive oncologists' lack of response to knowledge growth. An extensive body of literature documents that physician human capital is highly domain-specific and decays rapidly, consistent with high returns to specialization. Anecdotally, physicians at top academic hospitals are typically sub-specialists Gesme and Wiseman [2011], Graham et al. [2021] Perusing the physician directories of top cancer hospitals confirms this observation. The observation that top-ranked cancer hospitals largely employ cancer specialists suggests (but, of course, does not prove) that specialization has meaningful returns to quality. There may be low returns to specialization, but it seems unlikely to us, given that oncology is a rapidly changing and highly knowledge-intensive medical field.

Notably, the patterns for highly specialized oncologists suggest a different set of constraints to specialization. In our results, specialist oncologists respond to knowledge growth by becoming more specialized. This pattern implies that they are also constrained by coordination costs, as in Becker and Murphy [1992]. However, the finding that they respond to knowledge growth suggests that these specialists face either higher returns to specialization or lower coordination costs than generalists.

Figure 3.6a also displays the 90th percentile of specialization by market size quintile in 1999 and 2014. Figure 3.6a shows it for 2016 and 2019.

Both Figure 3.6a and 3.6b show that the 90th percentile of physician specialization is rising over time, but only in the largest markets. In 1999, the 90th percentile of specialization is nearly the same in the largest markets as in the smallest ones, hovering just over 4,000 HHI points. By 2014, the 90th percentile of specialization has risen in the largest markets to over 5,500 HHI points and fallen in the smallest markets to roughly 3,500. The specialization gap between the top and bottom 20% of markets is over 2,000 HHI points. Between 2016 and 2019, 90th percentile specialization rises even further in the largest markets, while remaining unchanged in the smallest markets.

Together, these results suggest that specialized oncologists are constrained by both coordination costs and market size. They become more specialized when knowledge grows, but only when they work in sufficiently large markets. The cross-sectional patterns in Figure 3.6 also suggest that economics of scale in subspecialization are growing over time.

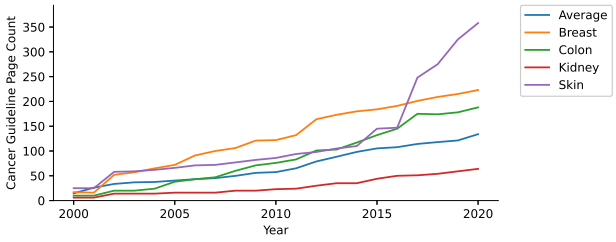
3.7 Conclusion

Oncology is a field with rapid, overwhelming knowledge growth. In this paper, we examine how this growth in knowledge impacts worker specialization. Studying a balanced panel of oncologists, we observe that the typical oncologist is a general oncologist and does not become more specialized in response to knowledge growth. The median oncologist in large and small markets remains a generalist. We interpret these findings to suggest that coordination costs for most oncologists are very high, limiting their scope to achieve the gains from specialization. However, we also find that ex-ante specialized oncologists become significantly more specialized in response to knowledge growth. However, only specialized oncologists in large markets become more specialized. This pattern suggests that knowledge growth increases market-level economies of scale in knowledge work.

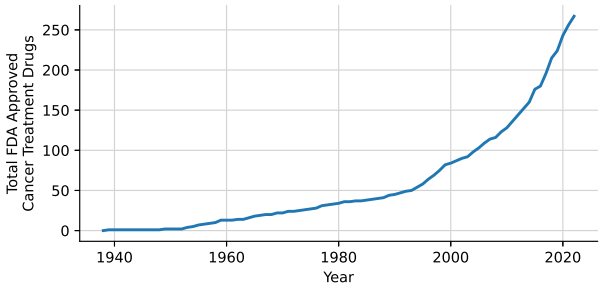
3.8 Exhibits

Figure 3.1: Oncology has experience massive increases in available knowledge and technologies

(a) Oncology treatment guidelines have become considerably longer

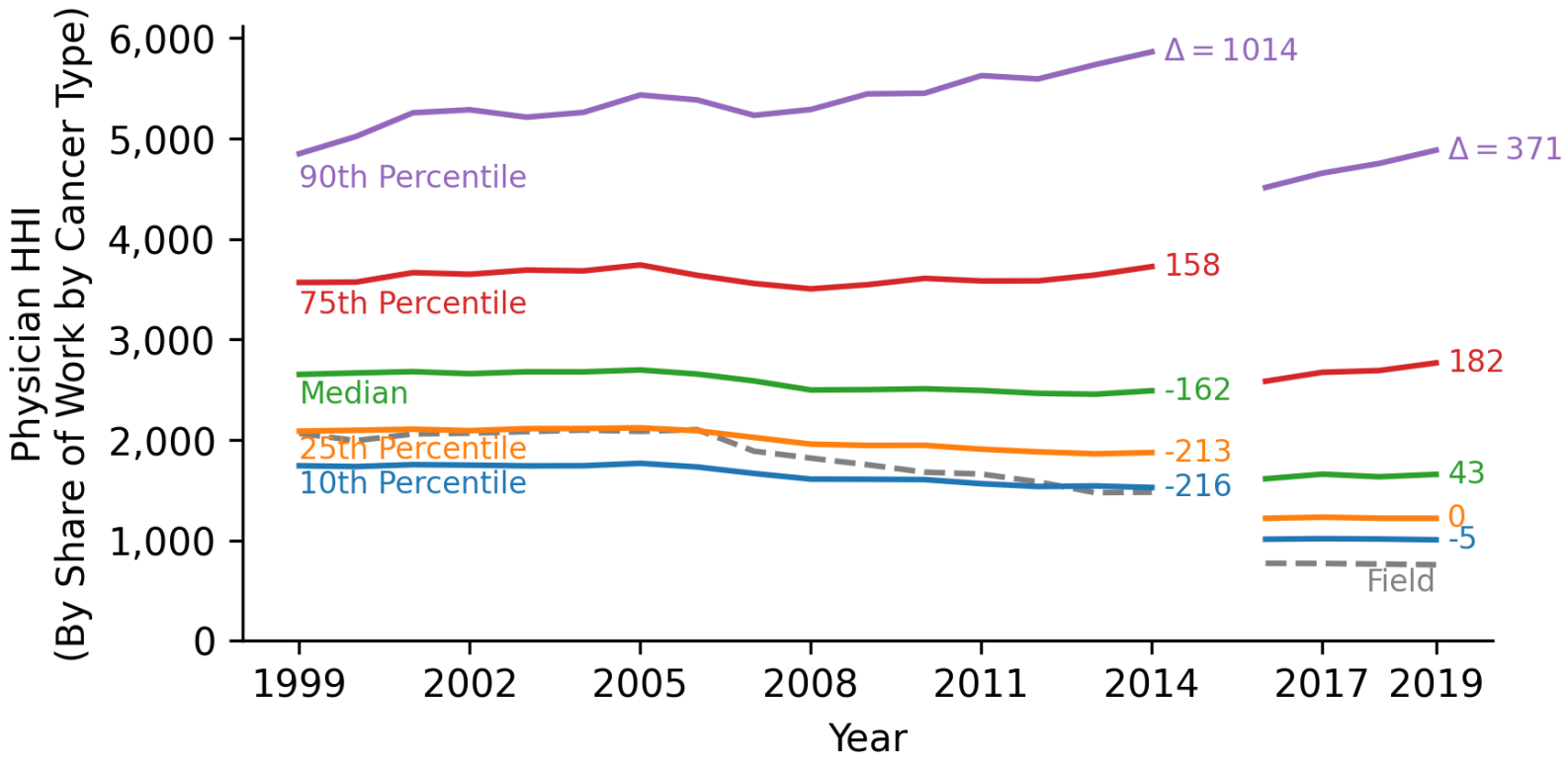


(b) The number of FDA-approved anti-cancer drugs has greatly increased



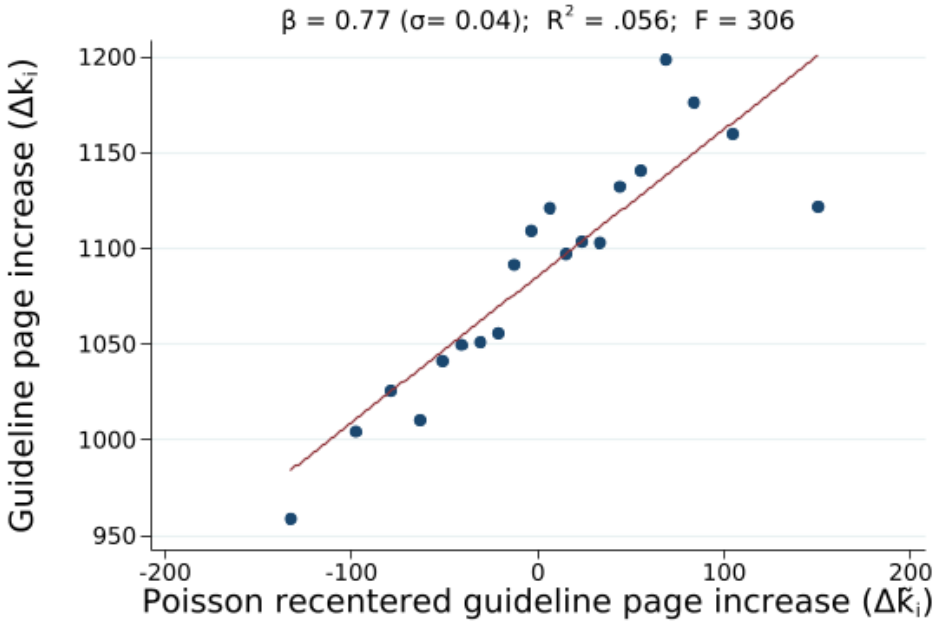
Notes: Panels a and b show the growth in cancer guideline length and pharmaceutical treatments over time. Panel a displays the page count of the first set of treatment guidelines in each year from the National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology. Panel b displays the number of unique anti-cancer drugs, where unique drugs are distinguished by unique Anatomical Therapeutic Chemical Level 5 Classification Codes. Note that the sample is limited to drugs with direct anti-cancer uses which are approved by the U.S. Food and Drug Administration. It does not include combinations of drugs, drugs used for cancer symptom relief or diagnosis, or unapproved drugs that are being investigated. The data in this panel is from Pantziarka et al. [2021].

Figure 3.2: Specialized Oncologists are Growing More Specialized Over Time



Notes: This figure shows the distribution of physician-level Herfindahl—Hirschman index (HHI), a measure of specialization, from 1999-2019. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who saw at least 25 unique cancer patients in the year in a 20% sample of Medicare Part B claims. Physician HHI is calculated within physician using shares of time spent evaluating and managing different types of cancer. Time is proxied using the work Relative Value Units for each evaluation and management HCPCS code. The 10th, 25th, 50th, 75th, and 90th percentile are displayed and labeled. The dashed grey line represents the HHI of the field overall. The trend break in 2015 is attributed to the switch in diagnosis coding schemes, from ICD-9 to ICD-10, and 2015 is dropped from the graph for visual clarity. Overall, the 10th, 25th, and 50th percentile of oncologists see a roughly representative distribution of patients, suggesting they are "general" oncologists. As such, most oncologists remain generalists over this period. However, the most specialized oncologists, those near the 90th percentile of the distribution, become substantially more specialized over time.

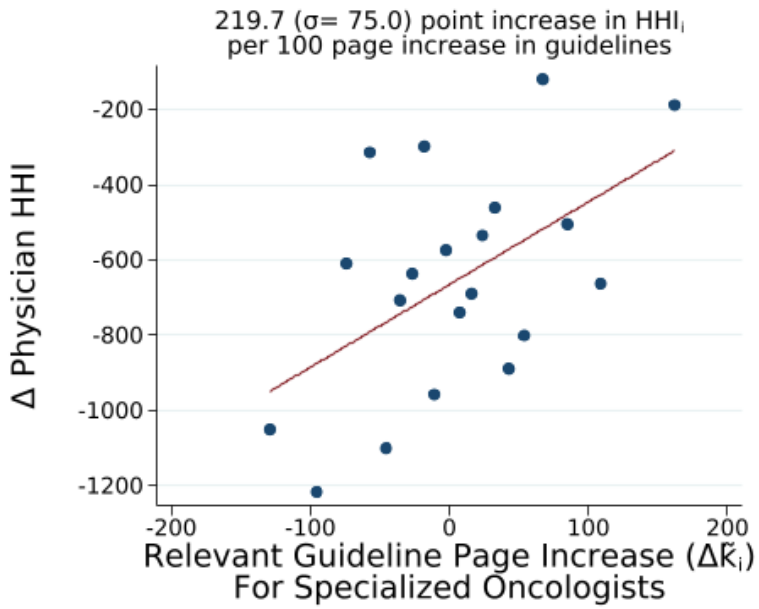
Figure 3.3: Quasi first stage suggests recentered instrument ($\Delta\tilde{k}_i$) is a strong "instrument" for unadjusted instrument (Δk_i)



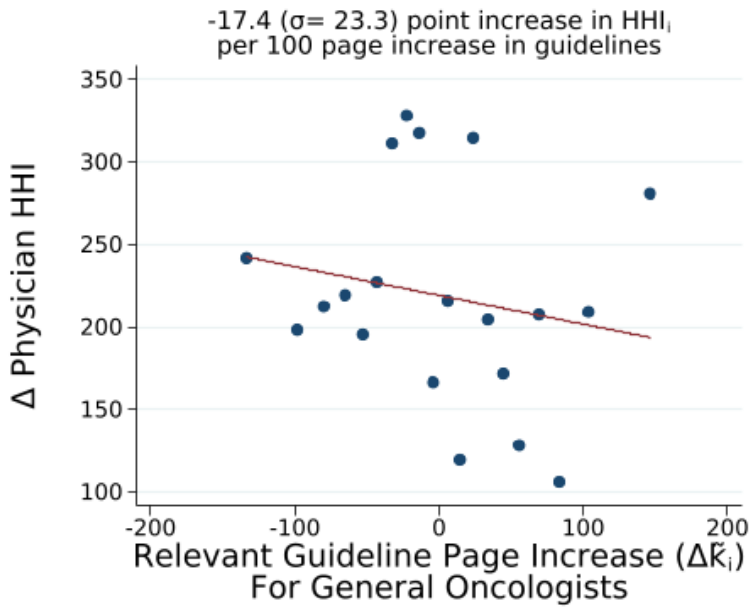
Notes: This figure shows the results for a quasi first stage regression of the recentered measure of exposure to knowledge growth on the unadjusted measure. It displays a binned scatter plot and key regression outputs for the bivariate relationship between the raw measure and the Poisson recentered measure of physician-level exposure to knowledge growth. The raw measure of physician level exposure to knowledge growth is calculated using the cancer types a physician treated in 2008 and the subsequent growth in knowledge in those fields over the next five years (2009-2014), based on realized changes in guideline lengths. In the Poisson recentered measure, the predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. Regression standard errors are robust. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Figure 3.4: Knowledge growth increases specialization only for those who start specialized

(a) Knowledge growth increases specialization for oncologists who start specialized



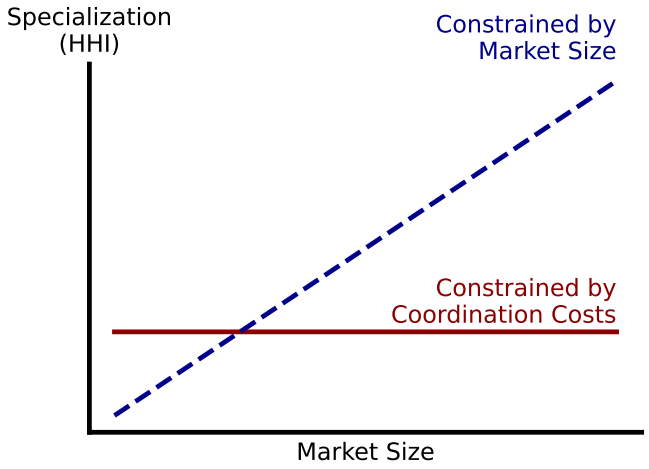
(b) But not for general oncologists



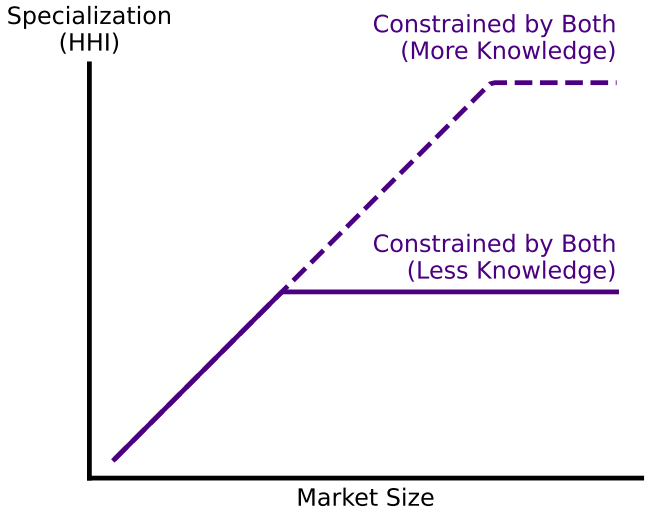
Notes: Panels a and b shows binned scatterplots and key regression estimates for the univariate relationship between the Poisson recentered exposure to knowledge growth and change in specialization. Panel a shows this relationship for oncologists who start in the top 25% of the initial specialization distribution based on a Herfindahl—Hirschman index (HHI) over types of cancer managed in 2009. Panel b shows this relationship for oncologists who start in the bottom 75% of the initial specialization distribution based on a Herfindahl—Hirschman index (HHI) over types of cancer managed in 2009. Exposure to knowledge growth. In the Poisson recentered measure, the predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. Regression standard errors are robust. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Figure 3.5: Growth in knowledge (Δk) can increase market-level economies of scale

(a) Market Size and Coordination Cost Constraints to Specialization Lead to Different Patterns of Specialization By Market Size



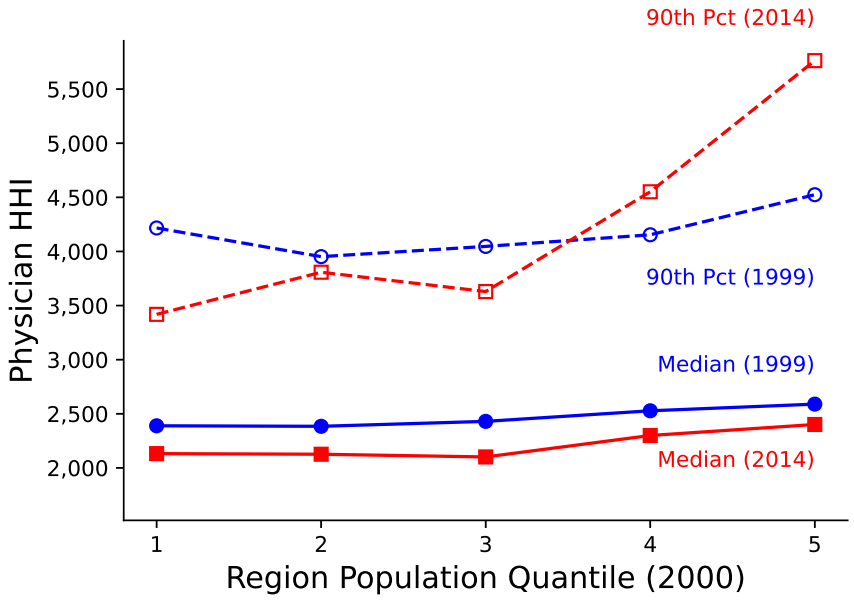
(b) When Both Constraints Are Exist, More Knowledge Can Lead to Larger Spatial Differences in Specialization



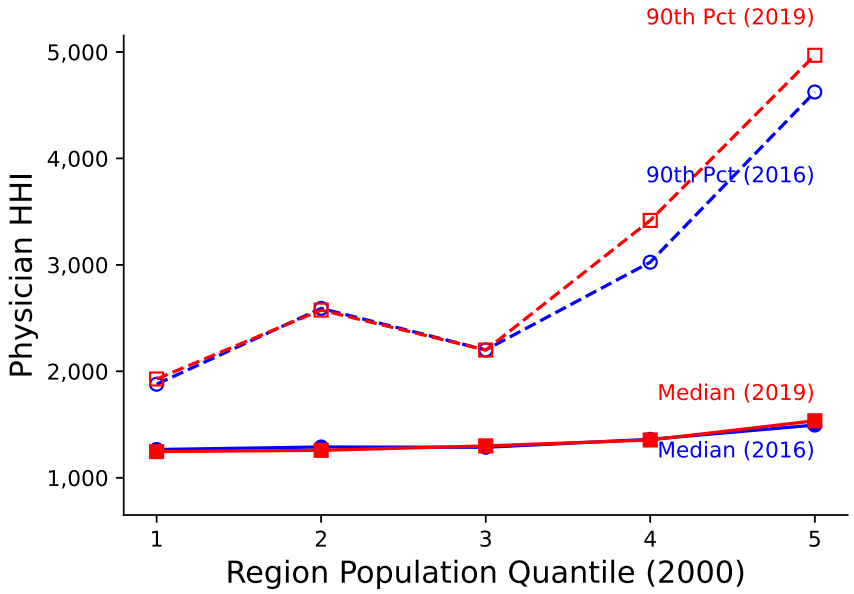
Notes: Panels a and b show theoretical predictions. Panel a predicted relationship between specialization and market size if specialization is constrained by coordination costs and if it is constrained by market size. Panel b shows the change in the relationship between specialization and market size when knowledge increases in a world where both coordination costs and market size constrain specialization.

Figure 3.6: Median Specialization is Stable Over Time and Across Market Size, but Extreme Specialization is Increasing in Largest Markets

(a) 1999—2014



(b) 2016—2019



Notes: Panels a and b show the median and 90th percentile of oncologist specialization by 2000 market size quintile. Panel a shows these values for 1999 and 2014, while Panel b shows these values for 2016 and 2019. We separate these two time periods because diagnosis coding changed in 2015, the year between the two periods. Physician specialization is measured using HHI over the types of cancer that the physician evaluated and managed in that year. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in the year.

Table 3.1: Summary Statistics for Balanced Panel of Oncologists (2008-2014)

	Mean	SD	Min	Max	N
N	4,854				
Unique Patients (2008)	69	33	25	297	4,854
Count of Cancer Types (2008)	15	4	1	27	4,854
HHI (2009)	2,737	1,452	954	10,000	4,854
Guideline Length (2009)	1,789				
Relevant Guideline Length (2009)	1,159	242	42	1,702	4,854
Change in Length of Guidelines (2009-2014)	1,745				
Change in Length of Relevant Guidelines (2009-2014)	1,088	225	24	1,612	4,854
Share in Largest 20% of Regions	1	0	0	1	4,854
Count Oncologists at Main Tax ID (2008)	19	40	1	219	4,854
Physician age (2008)	49	9	30	77	4,854

Notes: This table shows summary statistics for a balanced panel of oncologists from 2008 to 2014. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014. More information about each variable is calculated is detailed in section 3.2.

Table 3.2: Specialized Oncologists Work in Larger Regions and Organizations than General Oncologists (2008-2014)

	Top 25% Most Specialized Docs		Other 75%	
	Mean	Std	Mean	Std
N	1,214		3,640	
Unique Patients (2008)	59	(29)	73	(34)
Count of Cancer Types	13	(5)	16	(3)
HHI (2009)	4,637	(1,729)	2,104	(460)
Share in Largest 20% of Regions	0.63	(0.48)	0.50	(0.50)
Count Oncologists at Main Tax ID (2008)	25	(41)	17	(40)
Physician age (2008)	50	(9)	49	(9)
NCI Cancer Center Zip	0.16	(0.37)	0.04	(0.18)
Relevant Guideline Length (2008)	860	(279)	1,065	(165)
Change in Length of Relevant Guidelines (2009-2014)	932	(283)	1,139	(174)

Notes: This table compares key summary statistics for "specialist" and "general" oncologists. Specialist oncologists are defined as those in the top 25% of the specialization distribution as measured by a Herfindahl—-Hirschman index (HHI) over types of cancer managed in 2009. Specialist oncologists are defined as those in the bottom 75% of the specialization distribution as measured by a Herfindahl—-Hirschman index (HHI) over types of cancer managed in 2009. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014. More information about how each variable is detailed in section 3.2.

Table 3.3: Physician characteristics become substantially less predictive of exposure to knowledge growth after recentering

	(1) Unadjusted (Δk_i)	(2) Remove physician variation ($\Delta \tilde{k}_i$)	(3) Remove expected cancer variation ($\Delta \tilde{k}_i$)
Region population (millions)	-23.10	5.649	-0.0669
	(1.909)	(0.805)	(0.511)
Count oncologists at main org	-0.537	0.0792	0.0165
	(0.0881)	(0.0411)	(0.0240)
Physician age in 2008	0.976	-0.772	-0.561
	(0.362)	(0.170)	(0.112)
Observations	4,854	4,854	4,854
R-squared	0.055	0.016	0.005
Recentered		Yes	Yes

Notes: This table shows the results of a regression of exposure to knowledge growth on observable physician characteristics. The physician characteristics include the population of the physicians' Hospital Referral Region in millions, the number of other oncologists at the Tax IDs that they work at, and the physicians age. All physician characteristics are from 2008, the baseline year for the panel. In Column 1, the outcome variable is the raw measure of physician level exposure to knowledge growth based on the cancer types they treat in 2008 and the subsequent growth in knowledge in those fields over the next five years (2009-2014), based on changes in guideline lengths. In Column 2, the outcome variable is the mean recentered version of the measure of exposure to knowledge growth. In this measure, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated as the number of cancer types treated times the mean knowledge growth per cancer type, weighted by the number of oncologists who treat the cancer type. In Column 3, the outcome variable is a Poisson recentered version of the measure of exposure to knowledge growth. As in Column 2, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. However, the predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014. All regression standard errors are robust.

Table 3.4: Knowledge growth increases specialization for oncologists in the top 25% of the initial specialization distribution

	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	196.2	195.5	187.7	159.7
	(73.58)	(73.87)	(82.04)	(83.33)
Observations	1,163	1,163	1,163	1,163
R-squared	0.006	0.008	0.135	0.148
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and change in specialization for oncologists who start in the top 25% of the initial specialization distribution. Specialization is measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed in 2009. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Table 3.5: Knowledge growth does not increase specialization for oncologists in the bottom 75% of the initial specialization distribution

	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	-16.71	-16.40	-8.400	-9.892
	(23.33)	(23.40)	(25.03)	(24.89)
Observations	3,623	3,623	3,623	3,623
R-squared	0.000	0.000	0.077	0.078
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and change in specialization for oncologists who start in the bottom 75% of the initial specialization distribution. Specialization is measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed in 2009. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Table 3.6: For specialized oncologists, knowledge growth does not reduce the number of types of cancers managed

	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	-0.0790	-0.117	-0.0318	-0.0469
	(0.143)	(0.143)	(0.158)	(0.159)
Observations	1,163	1,163	1,163	1,163
R-squared	0.000	0.008	0.158	0.161
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and the number of cancers managed for specialized oncologists. Specialized oncologists are defined as oncologists in the top 25% of the specialization distribution in 2009, measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Table 3.7: For general oncologists, knowledge growth does not reduce the number of types of cancers managed

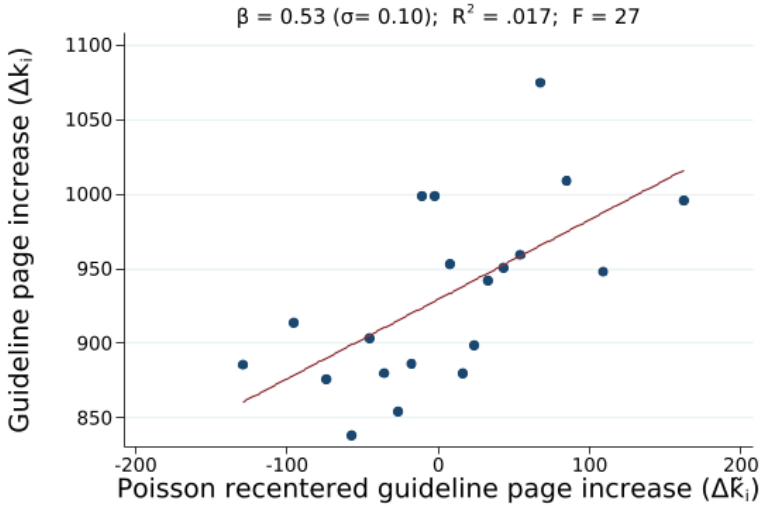
	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	-0.105	-0.139	-0.135	-0.129
	(0.0776)	(0.0773)	(0.0813)	(0.0814)
Observations	3,623	3,623	3,623	3,623
R-squared	0.000	0.018	0.127	0.129
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and the number of cancers managed for general oncologists. General oncologists are defined as oncologists in the bottom 75% of the specialization distribution in 2009, measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

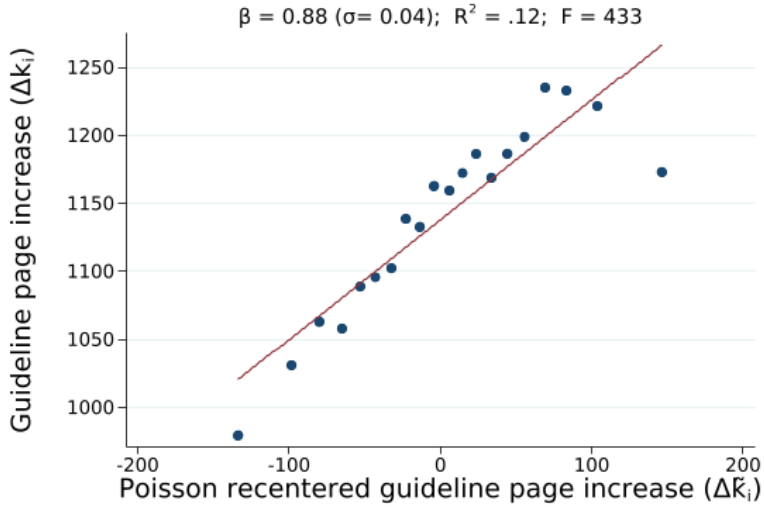
3.9 Additional exhibits

Figure 3.7: Quasi first stage suggests recentered instrument ($\Delta\tilde{k}_i$) is a strong "instrument" for unadjusted instrument (Δk_i)

(a) Specialized Oncologists

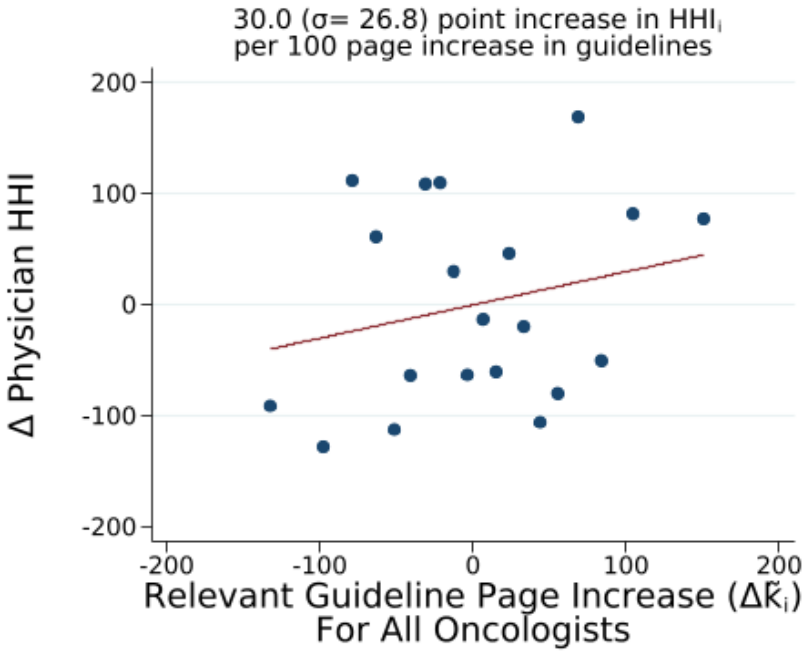


(b) General Oncologists



Notes: Panels a and b show the results for a quasi first stage regression of the recentered measure of exposure to knowledge growth on the unadjusted measure. Panel a displays the results for specialized oncologists, defined as those in the top 25% of the specialization distribution in 2009, as measured by a Herfindahl—Hirschman index over types of cancers managed. Panel b displays the results for general oncologists, defined as those in the bottom 75% of the specialization distribution. Both figures display a binned scatter plot and key regression outputs for the bivariate relationship between the raw measure and the Poisson recentered measure of physician-level exposure to knowledge growth. The raw measure of physician level exposure to knowledge growth is calculated using the cancer types a physician treated in 2008 and the subsequent growth in guideline length in those fields over the next five years (2009-2014). In the Poisson recentered measure, the predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated based on the cancer types a physician treated in 2008 and the predicted growth in knowledge for each cancer type over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. Regression standard errors are robust. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Figure 3.8: Knowledge growth does not increase specialization on average



Notes: Figure 3.8 shows a binned scatter plot and key regression estimates for the univariate relationship between the Poisson recentered exposure to knowledge growth and change in specialization across all oncologists. In the Poisson recentered measure, the predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. Regression standard errors are robust. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Table 3.8: The length of cancer treatment guidelines has increased substantially over time for many types of cancer

Year Site	2002	2020
ALL	0	117
AML	23	138
Anal	9	55
Bladder	40	113
Bone	25	96
Breast	52	223
Breast Risk	22	64
Breast Screening	33	77
CML	24	80
CNS	51	165
Cervical	24	86
Colon	20	188
Esophageal	19	154
GTN	0	40
Gastric	18	125
Hairy Cell	0	24
Head And Neck	86	233
Hepatobiliary	32	151
Hodgkins	30	85
Kaposi	0	39
Kidney	14	64
Lung	68	294
MDS	22	91
MPN	0	101
Mastocytosis	0	67
Mesothelioma	0	46
Myeloma	27	95
NHL	49	658
Neuroendocrine	46	141
Ovarian	26	371
Pancreatic	23	160
Penile	0	49
Prostate	18	167
Prostate Detection	22	63
Rectal	18	167
Sarcoma	31	146
Skin	58	358
Small Bowel	0	48
Testicular	34	78
Thymic	0	41
Thyroid	69	136
Uterine	42	108
Vulvar	0	58
All	**1,075**	**5,760**

Notes: Figure 3.8 displays the page count of the first set of treatment guidelines in 2002 and 2020 from the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology. The NCCN considered the 2002 guidelines to be a "complete library" of oncology guidelines [NCCN, 2023b]. However, in the years since, they have added guidelines for additional types of cancer.

Table 3.9: For specialized oncologists, physician characteristics become substantially less predictive of exposure to knowledge growth after recentering

	(1) Unadjusted (Δk_i)	(2) Remove physician variation ($\Delta \tilde{k}_i$)	(3) Remove expected cancer variation ($\Delta \tilde{k}_i$)
Region population (millions)	-37.78	5.036	0.0634
	(4.496)	(1.679)	(0.972)
Count oncologists at main org	-1.141	0.171	0.0133
	(0.253)	(0.0833)	(0.0441)
Physician age in 2008	1.261	-0.466	-0.797
	(0.875)	(0.351)	(0.227)
Observations	1,213	1,213	1,213
R-squared	0.128	0.017	0.010
Recentered		Yes	Yes

Notes: This table shows the results of a regression of exposure to knowledge growth on observable physician characteristics for specialized oncologists. Specialized oncologists are defined as oncologists in the top 25% of the specialization distribution in 2009, measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed. The physician characteristics include the population of the physicians' Hospital Referral Region in millions, the number of other oncologists at the Tax IDs that they work at, and the physicians age. All physician characteristics are from 2008, the baseline year for the panel. In Column 1, the outcome variable is the raw measure of physician level exposure to knowledge growth based on the cancer types they treat in 2008 and the subsequent growth in knowledge in those fields over the next five years (2009-2014), based on changes in guideline lengths. In Column 2, the outcome variable is the mean recentered version of the measure of exposure to knowledge growth. In this measure, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated as the number of cancer types treated times the mean knowledge growth per cancer type, weighted by the number of oncologists who treat the cancer type. In Column 3, the outcome variable is a Poisson recentered version of the measure of exposure to knowledge growth. As in Column 2, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. However, the predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014. All regression standard errors are robust.

Table 3.10: For general oncologists, physician characteristics become substantially less predictive of exposure to knowledge growth after recentering

	(1) Unadjusted (Δk_i)	(2) Remove physician variation ($\Delta \tilde{k}_i$)	(3) Remove expected cancer variation ($\Delta \tilde{k}_i$)
Region population (millions)	-9.590	3.966	-0.250
	(1.481)	(0.861)	(0.609)
Count oncologists at main org	-0.0555	-0.00741	0.0115
	(0.0673)	(0.0449)	(0.0288)
Physician age in 2008	1.343	-1.003	-0.492
	(0.322)	(0.189)	(0.129)
Observations	3,641	3,641	3,641
R-squared	0.016	0.013	0.004
Recentered		Yes	Yes

Notes: This table shows the results of a regression of exposure to knowledge growth on observable physician characteristics for general oncologists. General oncologists are defined as oncologists in the bottom 75% of the specialization distribution in 2009, measured using a Herfindahl—-Hirschman index (HHI) over types of cancer managed. The physician characteristics include the population of the physicians' Hospital Referral Region in millions, the number of other oncologists at the Tax IDs that they work at, and the physicians age. All physician characteristics are from 2008, the baseline year for the panel. In Column 1, the outcome variable is the raw measure of physician level exposure to knowledge growth based on the cancer types they treat in 2008 and the subsequent growth in knowledge in those fields over the next five years (2009-2014), based on changes in guideline lengths. In Column 2, the outcome variable is the mean recentered version of the measure of exposure to knowledge growth. In this measure, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Predicted exposure is estimated as the number of cancer types treated times the mean knowledge growth per cancer type, weighted by the number of oncologists who treat the cancer type. In Column 3, the outcome variable is a Poisson recentered version of the measure of exposure to knowledge growth. As in Column 2, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. However, the predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014. All regression standard errors are robust.

Table 3.11: Knowledge growth does not detectably increase specialization for oncologists overall

	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	29.30	27.96	44.02	45.12
	(26.88)	(26.95)	(28.36)	(28.35)
Observations	4,842	4,842	4,842	4,842
R-squared	0.000	0.001	0.061	0.062
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and change in specialization for all oncologists. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

Table 3.12: For all oncologists, knowledge growth does not reduce the number of types of cancers managed

	(1) Recentered ($\Delta\tilde{k}_i$)	(2) Add Age FE	(3) Add HRR FE	(4) Add Tax ID Size FE
Increase in Guideline Length (100 Pages)	-0.0813	-0.115	-0.106	-0.113
	(0.0680)	(0.0677)	(0.0706)	(0.0707)
Observations	4,842	4,842	4,842	4,842
R-squared	0.000	0.014	0.106	0.106
Age FE		Yes	Yes	Yes
HRR FE			Yes	Yes
Tax ID Size FE				Yes

Notes: This table shows regression estimates for the relationship between exposure to knowledge growth and the number of cancers managed for all oncologists. The measure of exposure to knowledge growth is the Poisson recentered version of this measure. Column 1 shows the coefficient estimate for the univariate regression. The regression standard errors are robust. Columns 2, 3 and 4 add fixed effects for the physicians age group, hospital referral region and the number of other oncologists at their Tax IDs, respectively. In the Poisson recentered measure of exposure to knowledge growth, predicted exposure to knowledge growth is subtracted from raw exposure to knowledge growth. Raw exposure to knowledge growth is calculated as the total increase in the page count of guidelines over 2009 to 2014 for the types of cancers a physician treats in 2008. Predicted exposure is estimated based on the cancer types a physician treats in 2008 and the predicted, rather than realized, growth in knowledge over the next five years. The predicted knowledge growth for each type of cancer is estimated based on a Poisson regression of the number of Medicare patients with the cancer types in 2008, an indicator for if the cancer type has well-established surrogate endpoints, and the interaction of the two. All physicians in the sample are hematologist-oncologists, oncologists, or hematologists who evaluated and managed at least 25 unique cancer patients in a 20% sample of Medicare Part B claims in 2008, 2009 and 2014.

CHAPTER 4

ACCURACY AND INTERPRETABILITY IN GOVERNMENT PAYMENT ALGORITHMS

4.1 Introduction

Algorithmic predictions are used in government decision-making to determine the flow of billions of public dollars. These algorithms commonly rely on regression models to predict values such as property values for tax purposes or a patient's annual total healthcare expenditures in public health insurance. Policymakers often face a trade-off between using simple, transparent statistical models and achieving higher predictive accuracy. While machine learning techniques offer potentially increased prediction accuracy, they often require policymakers to use complex non-parametric models.

Medicare Advantage exemplifies this tension. In Medicare Advantage, risk adjustment modifies payment amounts to health insurers based on expected patient costs, in order to address the "market for lemons", or adverse selection. In 2021, its risk adjustment formulas determined the allocation of over $300 billion, or 1% of US GDP [Statista Research Department, 2022, Cubanski and Neuman, 2023]. Given the program's scale, even small changes to these formulas can change the allocation of billions in public funds.

Given the vast public funds involved, simple and interpretable risk adjustment is a key policymaker objective. For example, during the first 15 years of Medicare Advantage, Medicare used a model with only a few demographic variables, known as the "Demographic Model". This model was used even though it explained minimal variation in spending (roughly 1%). However, evidence mounted that this too-simple model contributed to selection and substantially increased Medicare's costs. Ultimately, Medicare added diagnosis information to risk adjustment and created the current Hierarchical Condition Category (HCC) model [McGuire et al., 2011]. Although this model added complexity, the developers of the HCC model still

emphasized that preserving interpretability and parsimony were central design of the model [Pope et al., 2004]. More recently, in reports to Congress, Medicare has emphasized that risk adjustment models should be transparent, interpretable, and clinically meaningful with "face-validity" [MedPAC, 2014, 2021], likely because this makes them more defensible in the face of public scrutiny.

However, the program's current risk adjustment techniques have been criticized for their inaccuracy, leading to over- and under-payments [Rose et al., 2017, Zink and Rose, 2020, MedPAC, 2021]. Such payment discrepancies distort market dynamics and compromise healthcare access for vulnerable populations [Geruso and Layton, 2017]. The HCC model and associated policy reforms appear to have mitigated selection [Newhouse et al., 2015] but the remaining selection continues to be substantial, costly, and harmful to patients [Ryan et al., 2023, Zhu et al., 2023]. A 2017 review concluded that despite theoretical and practical challenges, improvements to risk adjustment offer "the best tool we have to address selection across plans in competitive health insurance markets" [Geruso and Layton, 2017].

Machine learning provides an avenue toward more accurate risk adjustment to address selection. However, machine learning models are frequently complex and hard to understand, causing them to fall short of stated yet mathematically informal interpretability requirements.

In light of these challenges, this paper aims to empirically examine the trade-off between the accuracy and interpretability of risk adjustment models in Medicare Advantage. Specifically, I evaluate whether machine learning models, despite their complexity, offer improvements in prediction accuracy that justify their use.

To measure this trade-off, I introduce a formal measure of model complexity tailored for government payment policy. I argue that model complexity (or lack of interpretability) is best proxied in this context as the number of coefficients in a model. This proxy assumes that the number of coefficients reflects the number of objects, or "cognitive chunks," that

stakeholders consider when interpreting a risk adjustment model.

Using Medicare claims data, I fit and evaluate multiple risk adjustment models. These models include both conventional and machine learning models, and they are fitted on Medicare claims data using standard risk adjustment variables. The models use the same underlying variables and differ solely in their functional forms. For each model, I estimate its complexity and out-of-sample accuracy.

The results show the clear trade-off between model accuracy and complexity. I find that the non-linear models provide the largest improvements in accuracy. A gradient-boosted tree changes mean absolute error (MAE) by $-1,352 (CI: $-1,392, $-1,316) relative to the current Medicare model, the HCC model. This is roughly three-fourths the size as past model changes. I also find that predictions from the gradient-boosted tree are more stable in the presence of simulated upcoding than predictions from the HCC model. However, this model also increases complexity dramatically, to 187,389, from 113 in the current HCC model.

To assess whether this increased complexity is justifiable, I use past changes in Medicare risk adjustment to estimate a range of plausible preferences of accuracy relative to complexity. I find that policymakers have accepted model changes that reduce MAE by $17.97 per additional coefficient. New models provide a reduction of $0.00722 per additional coefficient. As such, for new models to be acceptable, policymakers would need to be willing to accept model changes that are considerably less efficient at reducing error than they have in the past. These conclusions are robust to using mean squared error and relaxing preference assumptions about the disutility of complexity. They are limited to the extent that greater accuracy reduces selection incentives. As a whole, these findings suggest that standard machine learning models alone are unlikely to provide acceptable solutions to current issues with risk adjustment accuracy.

This work has several contributions. First, it provides a direct comparison between cur-

rent and proposed risk adjustment models for Medicare Advantage. While prior research has suggested the potential for machine learning models to improve the accuracy of risk adjustment in Medicare Advantage [Rose, 2016, Park and Basu, 2018, Kan et al., 2019, McGuire et al., 2021, Zink and Rose, 2020, Irvin et al., 2020], these studies rely on different datasets, typically commercial claims data that cover a younger, healthier population. They also often focus on risk adjustment in other settings, such as ACA exchanges. Given that individuals on Medicare exhibit more complex patterns of health conditions and spending, the value-add of machine learning is expected to be higher in this setting. Therefore, the conclusion—that machine learning is not worth the additional complexity—is more compelling, as it comes from a direct comparison with standard Medicare Advantage models where the value-add is expected to be larger.

This paper also contributes to the literature identifying trade-offs in the design of health insurance risk adjustment formulas. Ellis and McGuire [2007] observe a trade-off between "fit" (accuracy), "power," and "balance" generated by different risk adjustment formulas. Zink and Rose [2020] observe a trade-off between global accuracy and accuracy for certain patient subgroups in risk adjustment. Layton et al. [2018] argue that risk adjustment should maximize a welfare-grounded objective function, rather than R^2, highlighting a trade-off between accuracy and welfare. This paper contributes by identifying a key trade-off between accuracy and interpretability, which arises due to governance constraints on risk adjustment.

This study also has broader implications for policy settings in the US and internationally. Risk adjustment is used widely in US healthcare policy, with roughly two-thirds of Medicare and Medicaid dollars allocated via risk-adjusted payments, totaling over $1 trillion annually [Medicare Trustees, 2022, KFF, 2023]. Other countries, including Germany, Netherlands, Switzerland, and Chile [Kautter et al., 2014, Henriquez et al., 2023], also employ risk adjustment in their publicly regulated health insurance markets. Payment formulas are also used to assess property taxes in the US and worldwide [Norregaard, 2013, Berry, 2021]. Hence,

the trade-off between accuracy and complexity is broadly relevant, and these findings can inform efforts to incorporate machine learning into payment policy in those settings as well.

This work also introduces a new domain application to the machine learning interpretability literature. Existing research has considered model interpretability across numerous domains [Rudin, 2019] but payment policy remains relatively unexplored despite its growing importance in the US and other countries. Rose [2016] and McGuire et al. [2021] consider how to simplify risk adjustment, but their motivation is to reduce opportunities for gaming, not to improve interpretability. They arrive at a different measure, which leads to substantially different conclusions about the complexity and policy feasibility of machine learning models. More broadly, this paper highlights that interpretability is a key barrier to using machine learning in payment policy and identifies it as an important domain for future research.

4.2 Complexity in risk adjustment

4.2.1 Risk adjustment accuracy

The goal of risk adjustment is not perfect accuracy but rather to maximize accuracy conditional on using only "appropriate" variation. Broadly, appropriate variation is variation that contributes to selection incentives (i.e., higher expected costs) but does not lead to manipulation or moral hazard by insurance companies. This distinction is operationalized, albeit imperfectly, by including only variables with "appropriate" variation in risk adjustment, such as demographics and health conditions, and excluding ones with inappropriate variation, like past spending [Geruso and Layton, 2017].

4.2.2 The need for interpretability

Maintaining model interpretability and face validity is also a policy priority. Risk adjustment rate setting controls the flows of vast amounts of public funds to private companies, so these models are subject to intense public scrutiny. Medicare publicly releases all model parameters, a contrast to other policy algorithms, such as bail decisions, where parameters are often proprietary [Rudin, 2019]. Tweaks in models are closely followed by trade press and private companies, down to small changes in coefficient values.[1] As such, this need for transparent, face-valid models appears to stem from political and governance constraints.

Medicare risk adjustment also has additional requirements, or what the machine learning literature refers to as "auxiliary criteria." For example, per reports to Congress, risk adjustment must have face validity, a criterion I interpret as having two parts. First, a model should be interpretable, and second, upon interpretation, it must pass unspecified "sniff tests." These auxiliary criteria are not fully formalized and therefore are not included in the objective function. Interpretability allows the Centers for Medicare and Medicaid Services (CMS) to ex-post assess—and, if needed, to enforce—these auxiliary criteria that have not been included in the objective function [Doshi-Velez and Kim, 2017].

The presence of auxiliary criteria also explains why a simple rule like "minimize mean squared error" is an insufficient criterion by which to judge alternative risk adjustment models. Narrowly focusing on MSE makes one liable to generate models that do not meet necessary auxiliary criteria and are therefore acceptable to policymakers.

Of note, I must clarify what impact, if any, interpretability has on gaming incentives. Interpretability will likely worsen selection incentives if it necessitates simpler, less accurate models, as discussed above. However, it is theoretically ambiguous how complex, non-linear models change the incentives to upcode.

1. To quote one recent trade press article: "How will ... the changes in the proposed coefficients financially impact your organization? ... The time to act is now! CMS will be accepting commentary through Friday, March 3, 2023" [James et al., 2023].

4.2.3 Measure of model complexity

This study introduces a precise measure of model complexity, defined here as non-interpretability, for payment policy contexts: the number of coefficients in a model. Simply put, the model asks how many parameters are necessary to generate the full range of predictions.

For linear models, this measure is the L0 norm, or the number of non-zero coefficients. For example, a linear model with an intercept and a coefficient for female would have a complexity of two.

For tree-based models, this measure of complexity is determined by the number of unique, feasible decision paths, or equivalently, the number of combinations of variable values that lead to distinct predictions. For example, consider a regression tree that splits only sex and then, for men only, the presence of heart failure. This model can be represented with three coefficients: one for women, one for men with heart failure, and one for men without heart failure. Each coefficient represents the predicted spending for each group, giving the model a complexity of three.

The approach to calculating complexity differs subtly for tree-based models like gradient-boosted trees and random forests, which are both collections of trees where the predictions of each tree are combined in sums or averages. Consider a second tree that splits only on sex and then, for women only, on the presence of diabetes. This second tree also has an individual complexity of three. However, only four coefficients are needed to express the predictions made by combining trees, one for each of the following groups: men with and without heart failure and women with and without diabetes. As such, the combined trees have a complexity of four.

4.2.4 Strengths and limitations of the complexity measure

Why is the number of coefficients an appropriate measure of model complexity (or non-interpretability) in this context? The machine learning interpretability literature argues

that interpretability should be considered through an explanation's basic units or "cognitive chunks" [Doshi-Velez and Kim, 2017], which are domain specific [Rudin, 2019, Doshi-Velez and Kim, 2017]. These cognitive chunks reflect the complexity of explaining a model, not fitting it. In risk adjustment, coefficients vary payment rates, which makes them likely to be highly cognitively salient to stakeholders. In addition, coefficients have interpretable labels, e.g., heart failure [CMS, 2023], suggesting that stakeholders are inspecting and interpreting the model at this level. Lastly, the HCC model developers used the number of coefficients as an informal measure of model complexity, observing that the HCC model was relatively parsimonious at "fewer than 200 parameters" [Pope et al., 2004].

There are two key limitations. First, the evidence for this measure is based on secondary interpretation of policy documents and papers. Definitive assessment requires human subjects research [Doshi-Velez and Kim, 2017], which is outside this paper's scope.

Second, this definition is a measure of *global* complexity, the model's overall complexity, to use parlance from Doshi-Velez and Kim [2017]. It does not directly measure local complexity, the complexity of explaining individual payment decisions. A measure of local complexity would be desirable given that policymakers often need to justify individual decisions. Unfortunately, it is difficult to pin down reasonable and comparable measures of local complexity across models because the interpretation of any set of coefficients depends on the structure of the rest of the model.

4.2.5 Alternative measures of model complexity

An alternative measure of global model complexity is the number of substantive input variables used in a model, as in Rose [2016], McGuire et al. [2021]. Under this definition, a linear model with 100 input variables would be considered equally interpretable as a random forest with the same 100 input variables and hundreds of thousands of interaction terms. This is because this definition does not count interaction terms in the salient cognitive chunks. I ar-

gue that this is an incomplete assessment of what matters to stakeholders in risk adjustment. Interaction terms impact payments, so stakeholders are likely to demand explanations that account for them. This makes such terms highly relevant cognitive chunks when assessing interpretability.

Definitive arbitration of the saliency of interaction terms would require human subject experiments [Doshi-Velez and Kim, 2017]. But in the absence of such experiments, circumstantial evidence will have to suffice. The original developers of the HCC model specifically considered the clinical face-validity of interaction terms, implying they believe interactions are subject to interpretation and scrutiny [Pope et al., 2004].

One other measure of global model complexity from a related literature is the number of unique potential predictions of a model [Kleinberg and Mullainathan, 2019]. This definition argues the salient object is the prediction itself and that the process used to arrive at the prediction is irrelevant. According to this definition, a linear model with 100 indicator variables has a complexity of 1.3×10^{30} (2^{100}, or one nonillion possible unique predictions). A fully saturated tree-based model will have the same number of possible predictions, though most tree-based models will have fewer. As such, the linear model is weakly *more* complex than the random forest. This definition of complexity, when applied to the models used by Medicare policymakers, suggests that Medicare policymakers are currently using a maximally complex model. Therefore, they have no distaste for complexity whatsoever, rendering this exercise unnecessary.

Why not just try to explain non-interpretable models rather than require interpretable models? Simplified explanations of complex models are necessarily inaccurate; if they were perfectly accurate, they would be complex. As such, the explanation must be wrong sometimes and is therefore not entirely trustworthy or transparent [Rudin, 2019].

4.3 Medicare data

Next, I discuss the data used. The analysis uses Medicare fee-for-service claims data, which include diagnoses and the amount paid for each service. Using these data, I closely, though not identically, follow the sample selection and variable creation procedures used in Medicare Advantage risk adjustment models. The predictor variables include demographics and health conditions in 2018. The outcome variable is annualized healthcare spending in 2019.

I restrict all models to standard Medicare risk adjustment variables so that they use variation already deemed acceptable by Medicare. As such, the new models differ from standard models primarily in their functional form, not in the variation they can use.

Appendix 4.9 provides more details on the data, sample selection, and variable construction.

The sample is split into a training set (80%), a validation set (10%), and a test set (10%). Models are fit in the training set, and model performance is currently evaluated on the validation set. The test set remains untouched and is available for future use.

The sample includes 4,002,909 individuals, of which 3,202,327 are included in the training sample. Appendix Tables 4.2 and 4.2 show summary statistics for the training and validation samples. The average patient has 2.60 (SD = 3.56) payment-relevant health conditions. The average annualized spending is \$13,449.85 (SD = \$35,441.55).

4.4 Model fitting and evaluation

I first fit risk adjustment models using standard and machine learning models. Then, for each model, I measure the model complexity using the number of coefficients and evaluate model accuracy using out-of-sample MAE. Last, I evaluate the trade-off between accuracy and complexity by estimating the marginal reduction in error per additional model coefficient.

4.4.1 Model specifications

Broadly, I fit three types of models: standard Medicare models, alternative linear models, and tree-based models.

Standard models: First, I refit standard Medicare models in my sample. This approach allows me to make a direct comparison between standard and new models that is uncontaminated by any potential differences in the underlying data used in model fitting. The fitted models include the Demographic model and the HCC model.

Alternative linear models: I include a selection of parametric models using OLS and lasso, designed to incorporate health condition interactions and reduce overfitting. To account for comorbidities, I fit an OLS model which interacts health conditions with counts of other comorbidities. I refer to this as the "HCCxCount" model. I include both the HCC and HCCxCount variables in Lasso regression as well. Lasso regression is a standard model for reducing overfitting. It sets some coefficients to zero if they provide insufficient explanatory power while biasing the remaining coefficients toward zero [Hastie et al., 2009].

Non-parametric (tree-based) machine learning models: I also fit non-parametric tree-based models—specifically, regression trees, random forests, and gradient-boosted trees. The single regression tree model is the simplest type of tree model. It splits the data into groups based on column values (e.g., males with heart failure and without diabetes) and generates a prediction for each group. Random forests fit multiple trees, each on random samples of the data and columns, and then average the predictions across trees. Random forests have performed well in risk adjustment models in commercial claims data [Rose, 2016]. Gradient-boosted trees fit regression trees sequentially, with each tree fit on the residuals of the previous tree [Hastie et al., 2009]. Gradient-boosted trees have been used successfully to predict heart attacks, bail violations, and missed diagnoses [Mullainathan and Obermeyer, 2021, Kleinberg et al., 2018, Chan et al., 2022].

These models allow for flexible functional forms and variable interactions, allowing them

to capture complex, non-linear interactions. Often, they yield higher-quality predictions than linear models [Hastie et al., 2009, Rose, 2016]. I train two versions of each model, one that minimizes MSE and another that minimizes MAE. Appendix 4.10 provides more details about model fitting and tuning.

4.4.2 Model accuracy

The primary accuracy metric I use is MAE, defined as the prediction's average distance from the realized value ($\sum_i^n \frac{|y_i - \hat{y}_i|}{n}$). A key advantage of this measure is its meaningful units: a 10-unit decrease in MAE indicates that predictions are, on average, $10 more accurate per person. Additionally, MAE treats all errors equally, unlike metrics based on squared error. However, one disadvantage is that MAE differs from MSE, the metric that most linear models are trained on. To address this, in Appendix 4.12 I replicate the main results with MSE as the accuracy metric. In addition, for the main analysis, I assume that improved accuracy serves as a sufficient proxy for reduced selection incentives. The robustness of this assumption is evaluated in the results section.

To generate confidence intervals for model accuracy, I use a bootstrapping approach. Specifically, I generate 100 samples of the validation dataset, drawn with replacement. I then calculate the metrics of interest for each sample and use the 2.5% and 97.5% percentile values as 95% confidence interval bounds. This "quasi-Monte Carlo" approach, adapted from Park and Basu [2018], provides confidence intervals while preserving computational feasibility.

4.4.3 Model complexity

I measure model complexity as described previously. For linear models, model complexity is measured by the L0 norm, or the number of non-zero coefficients. For non-parametric, tree-based models, model complexity is measured as the number of unique and feasible decision

paths. Equivalently, this measure is the number of combinations of variable values that lead to distinct predictions. For single trees, I measure this exactly as the number of "leaves," or terminal nodes, on the tree. For random forests and gradient-boosted trees, I estimate this as the number of unique predictions in the training data due to computational limitations. This estimate provides a lower bound to the model's complexity.

4.4.4 Marginal value of additional complexity

How should policymakers decide how to trade off complexity versus accuracy? For this analysis, I assume model complexity and accuracy are two of the many factors that policymakers value and that they have a constant marginal utility of both. I then consider relative preferences between the two, holding all else equal.

The returns to complexity can be characterized as the marginal increase in accuracy per marginal increase in complexity. Policymakers will accept (or at least seriously consider) new models if the returns to complexity, in terms of error reduction, are sufficiently large.

Medicare's risk adjustment approach has evolved from the simpler Demographic model to the more accurate but complex HCC model. I use this transition to estimate bounds on preferences for complexity relative to accuracy. I estimate the marginal reduction in error per marginal coefficient from this change. I interpret this value as a revealed preference measure of error reduction per additional coefficient that policymakers are willing to accept.

Next, I estimate the error reduction per additional coefficient for new models. Focusing on models on the Pareto frontier, I compute the marginal reduction in error per marginal increase in coefficients for each model. I then compare this value with the value from past model changes. For robustness, I also consider alternative functional forms of preferences over complexity.

4.5 Results

In this section, I first estimate the accuracy and complexity of each model. One model significantly reduces error relative to the current Medicare model but it also substantially increases the number of coefficients. The reduction in error per additional coefficient is much less than that of past changes to Medicare models, suggesting that policymakers would likely not find this model preferable to the status quo. The results are robust to alternative accuracy measures, alternative preference assumptions, and upcoding. They are limited to the extent that accuracy is a sufficient proxy for selection incentives.

4.5.1 Model accuracy and complexity

Figure 4.1 displays the prediction accuracy for different models, and Figure 4.2 shows their varying levels of complexity. Figure 4.3 outlines the Pareto frontier of accuracy and complexity.

Among models on the Pareto frontier, complexity increases with accuracy. A prediction of the mean has a single parameter and therefore a complexity of 1. The average MAE is \$15,792 (CI:15,683, 15,907). The Demographic model increases complexity to 13 coefficients and reduces MAE by \$-540.20 (CI:\$-553.70, \$-525.90), or -3.4%. The HCC model adds more complexity, raising it to 113 coefficients. The additional complexity earns a larger decrease in MAE, reducing it by \$-2,337 (CI:\$-2,370, \$-2,307), or -15%, relative to the mean.

The HCCxCount model achieves modest improvements in performance with modest increases in complexity. It increases complexity to 184. It also reduces MAE, but this reduction is not significantly different of that from HCC model (\$-0.87; CI:\$-7.24, \$4.63).

The gradient-boosted tree trained on MAE reduces error the most, by \$-3,690 (CI:\$-3,714, \$-3,667), or -23%, relative to the mean. This is a difference of \$-1,352 (CI:\$-1,392, \$-1,316) relative to the HCC model. The change from the HCC model to the gradient-boosted tree is almost as large (roughly three-fourths the size) as the change from the Demographic to the

HCC model, suggesting it is economically significant. However, this error reduction comes with an enormous degree of complexity: 187,389, or 165,831% that of the HCC model.

4.5.2 Marginal value of complexity

Is this increase in accuracy worth the large increase in complexity? To assess, I assume constant marginal utility for both accuracy and complexity. I use past changes in risk adjustment to infer bounds on relative preferences between the two. I find that policymakers would need to be willing to accept model changes that are considerably less efficient at reducing error than they have in the past for new models to be acceptable.

I first estimate a bound on acceptable changes. The HCC model reduces MAE by $17.97 per additional coefficient relative to the Demographic model. The transition from HCC to Demographic model was adopted, which suggests that policymakers are willing to accept at least this rate of error reduction.

Figure 4.4 displays the marginal reduction in error per marginal increase in coefficients for Pareto models. The reduction in error falls with each additional increase in complexity.

The gradient-boosted tree (MAE) improves accuracy substantially, but inefficiently. Due to its complexity, it reduces error by a comparatively tiny amount per coefficient: $0.00722 , relative to the next best model (and similarly, $0.0072 relative to the HCC model). This is only 0.04% of reduction in error per additional coefficient from the transition to the HCC model. For this model to be acceptable, policymakers would have to be willing to accept a tiny fraction of the error reduction per coefficient than they have for past model changes.

4.5.3 Robustness

Next, I evaluate the robustness of these results with respect to alternative preference assumptions, measures of accuracy, and selection incentives. I find that they are largely robust to alternative assumptions that increase policymakers' tolerance of marginal complexity. I

also find that they are robust to alternative measures of accuracy, but that accuracy is an imperfect proxy for selection incentives. Finally, I find that the improved accuracy is likely robust to upcoding.

Alternative preference assumptions

I consider alternative preference assumptions which progressively relax policymaker distaste for complexity. First, I assume policymakers care about the relative percentage increase in complexity. Then the return to complexity for the gradient-boosted tree is only 0.57% of the return from the switch to the HCC model. Second, I assume policymakers care about log complexity. Then the return to complexity is 23.5% of the return from the switch to the HCC model. Even with mild distaste for complexity, policymakers would need to accept much lower returns to complexity than before.

Alternative measures of accuracy

I next assess how the results change when using MSE to measure accuracy since MAE and MSE weight error differently. Appendix Figures 4.6, 4.7, and 4.8 present the main analyses using MSE instead of MAE. The results are extremely similar. Again, gradient-boosted trees (this time trained on MSE) provide the largest reduction in MSE. The returns to complexity are remain low; policymakers would need to accept and error reduction per coefficient of 14.03, which is considerably less than past accepted changes of 1.403e+06.

Accuracy and selection incentives

Next, I assess the extent to which greater accuracy is a sufficient proxy for reduced selection incentives. Appendix Figure 4.9 shows the performance of different models on a range of selection incentive measures. The measures are calculated overall and for patient subgroups

thought to be subject to strong selection incentives, such as those with multiple chronic conditions [Zink and Rose, 2020, MedPAC, 2021][2].

The extent to which accuracy proxies for selection incentives depends on the type of selection. For example, one selection incentive measure is the scope for selection conditional on risk score, as measured by the sum of positive residuals, which is highlighted by Brown et al. [2014]. The gradient-boosted tree (MAE) substantially reduces the scope for this type of selection overall. It also greatly reduces differences across subgroups, reducing the incentive to select certain subgroups of patients. However, other metrics provide a different result. One such metric is tail risk, the probability that a patient's costs substantially exceed predictions, which is highlighted by Park and Basu [2018]. I find that the gradient-boosted tree (MAE) increases tail risk overall and increases differences across subgroups, creating larger incentives to avoid certain subgroups. As such, the main results are limited to the extent that accuracy proxies for the dominant types of selection, which the literature is inconclusive on.

Robustness to Upcoding

One key limitation of the main analysis is that it considers model accuracy without considering strategic diagnosis coding. However, strategic diagnosis coding is prevalent in Medicare Advantage [Geruso and Layton, 2020]. This limitation stems from using the Medicare claims data as they do not contain strategic diagnosis coding.

To understand how upcoding might affect model accuracy, I estimate the increase in predicted spending when adding a specific diagnosis to a patient's record, which also captures the returns from such upcoding. I focus on a subset of diagnoses listed in industry promotional material as the "biggest HCC coding opportunities" [RCX Rules, 2023]. Figure 4.5 shows the distributions of predicted spending increases for both the HCC model

2. Appendix 4.11 provides more details on the methods.

and gradient-boosted tree trained on MAE. Appendix Figure 4.10 replicates this with the gradient-boosted tree trained on MSE.

I find that the HCC model consistently shows higher increases in predicted spending. For the six diagnoses, upcoding in the HCC model raises spending by several thousand dollars more than in the gradient-boosted tree. This observation suggests that the gradient-boosted trees generate more stable, and therefore more accurate, predictions in the presence of upcoding and lower incentives to upcode.

4.6 Discussion

Gradient-boosted trees improve accuracy relative to standard risk adjustment models, consistent with results from other settings and features of gradient-boosted trees. However, they increase complexity to a degree that they are unlikely to be acceptable to policymakers. Key limitations of this work are that it does not test all potential model functional forms and uses accuracy as a proxy for varied selection incentives.

Why do the gradient-boosted trees offer such large increases in accuracy and complexity, and to what extent is this surprising? Regarding accuracy, gradient-boosted trees have consistently proven to be the best at structured machine learning problems. For example, at Kaggle, an online platform where people compete to solve machine learning problems, gradient-boosted trees were found to win the most competitions for supervised learning problems across a range of domains [Harasymiv, 2015, Nielsen, 2016]. Regarding complexity, gradient-boosted trees impose minimal functional form assumptions, which allows them to capture non-linear relationships but also requires many more parameters. In addition, they are ensembles of multiple trees, which greatly increases parameters relative to using a single tree or linear model.

Another reason why the gradient-boosted trees perform well is that they are fit to optimize MAE, while the linear models are optimizing MSE. While this could seem an unfair

comparison, it highlights a key strength of this type of model. Standard linear models are largely tied to the objective of minimizing MSE. In contrast, gradient-boosted trees can target any custom objective that is a function of y and $\hat{y}$, leading them to perform better at that objective. This is a useful feature given recent literature suggesting alternative objective functions for risk adjustment [Layton et al., 2018, Zink and Rose, 2020].

The results show that for gradient-boosted trees to be acceptable, policymakers would need to accept substantially lower returns to complexity than they have in the past. This finding is robust to alternative preference assumptions that reduce the marginal distaste for complexity and to different accuracy metrics. Thus, more accurate machine learning models can be reasonably rejected as improvements over current models, assuming policymakers' dislike of complexity has not fallen dramatically over time. Policymakers would need to be over a hundred times more tolerant of complexity than they have currently revealed themselves to be for these models to be acceptable.

One limitation of this study is that I have not tested every possible functional form that could use these variables. While other models (or versions of these models) may exist that provide similar accuracy for much less complexity, they would need to provide the observed increase in accuracy with less than 1/100th of the additional complexity of the models I examine. This degree of improvement seems unlikely to result from functional form changes alone. Therefore, standard machine learning methods are likely insufficient to substantially improve risk adjustment accuracy without increasing complexity to unacceptable levels.

Another limitation is that accuracy is an imperfect proxy for varied types of selection incentives. In line with past literature, I find that the optimal model depends on the relative importance of different types of selection and the data moment used to proxy it [Park and Basu, 2018, Zink and Rose, 2020]. While this study is not the first to observe the limitations of minimizing prediction error as an approach to risk adjustment [Layton et al., 2018, Zink and Rose, 2020, Geruso and Layton, 2017], this method remains the conventional approach

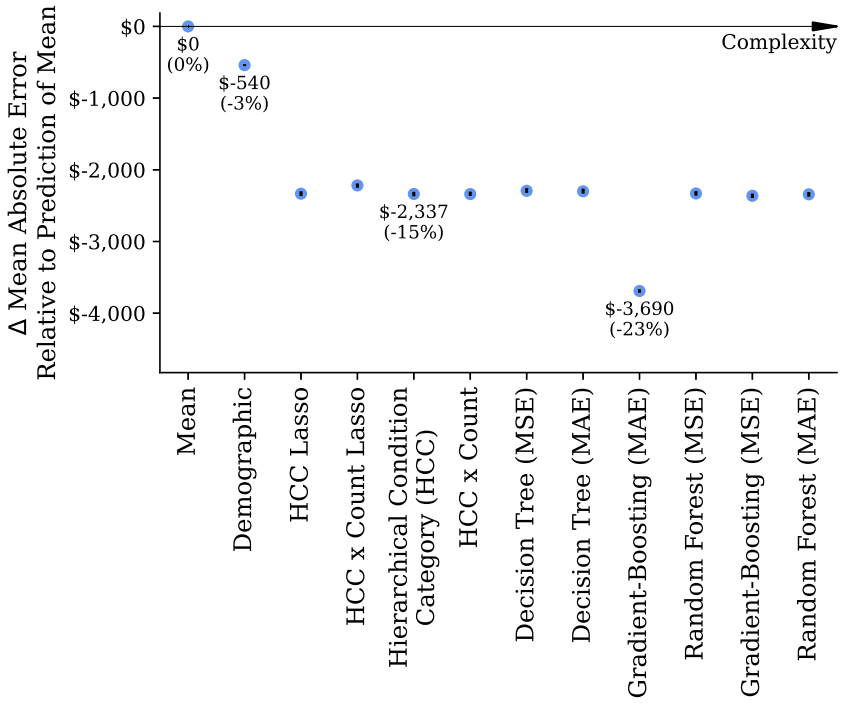
[MedPAC, 2021]. Geruso and Layton [2017] review a number of these challenges and conclude that despite its limitations, conventional risk adjustment remains among the best tools available to address selection, as evidenced by its wide adoption.

4.7 Conclusion

This paper examines the trade-off between accuracy and interpretability in risk adjustment models for Medicare Advantage. I find that although interpretability is a documented criterion, it has been previously informal and unquantified in the context of payment policy. To address this, I introduce a concrete measure of model complexity (or non-interpretability). By quantifying interpretability in this context, I formalize an important auxiliary criterion, enabling the formal objective to be more fully specified. Using the same data and variables employed in Medicare Advantage risk adjustment, I assess both traditional and machine learning models. My analysis reveals that machine learning models can significantly enhance prediction accuracy and improve robustness to upcoding but introduce unprecedented levels of complexity. These models provide very small increases in accuracy relative to their increase in complexity. As such, they would require policymakers to accept substantially smaller reductions in error per additional coefficient than they have in the past. Most likely, policymakers would not consider these models improvements over the status quo. As such, when accounting for auxiliary criteria in policy contexts, like interpretability, the optimal choice of models is likely to change meaningfully. Consequently, future research should explore how to use the advances offered by machine learning in ways that align with policy constraints in this critical domain.

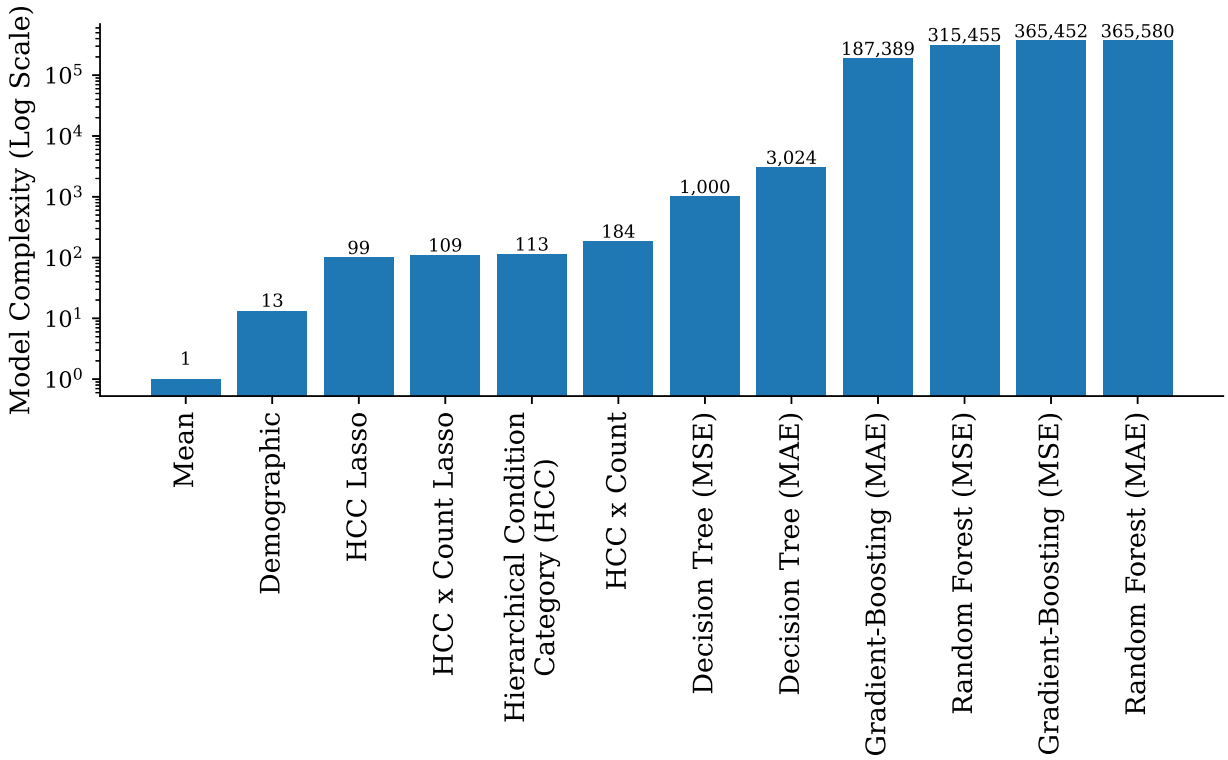
4.8 Exhibits

Figure 4.1: Difference in MAE of Model Predictions Relative to Predicting the Mean



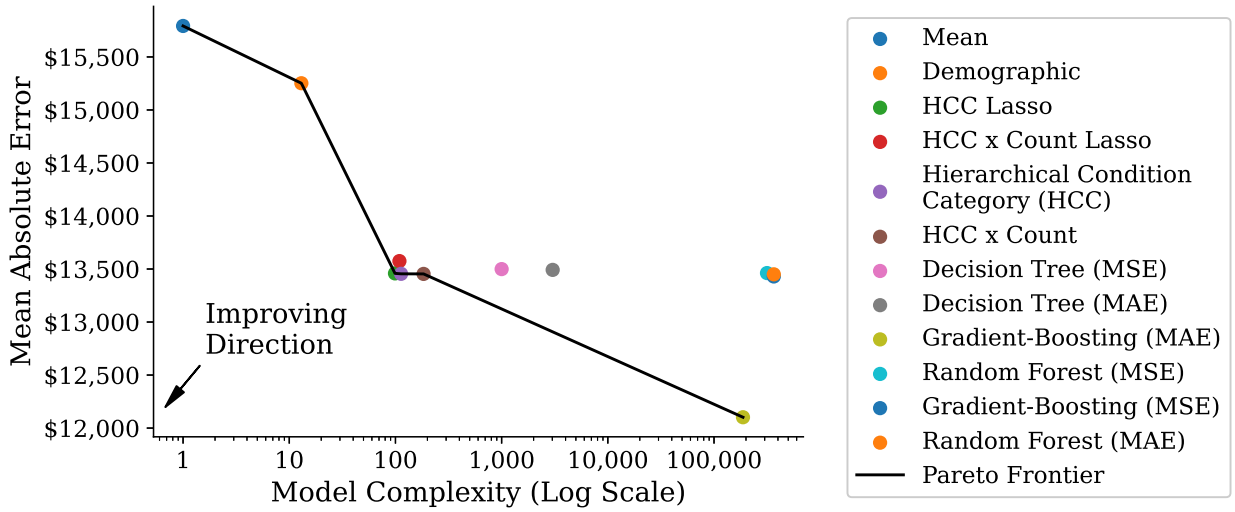
Notes: This graph shows the change in MAE for model predictions relative to predicting the mean. The x-axis shows the names of various models, ordered by increasing complexity. The "Hierarchical Condition Category" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function that the model was trained on, mean absolute error or mean squared error. The y-axis shows the reduction in the out-of-sample MAE, relative to always predicting the mean. MAE is calculated out of sample in 10% of the available data. The point estimate of MAE for each model is represented by the round marker. Standard errors are calculated with bootstrapped samples in the out-of-sample data. Ninety-five percent confidence intervals are shown as black bars; note that they are narrower than the height of the round markers. The values in parentheses are the percentage reductions in MAE relative to always predicting the mean.

Figure 4.2: Model Complexity



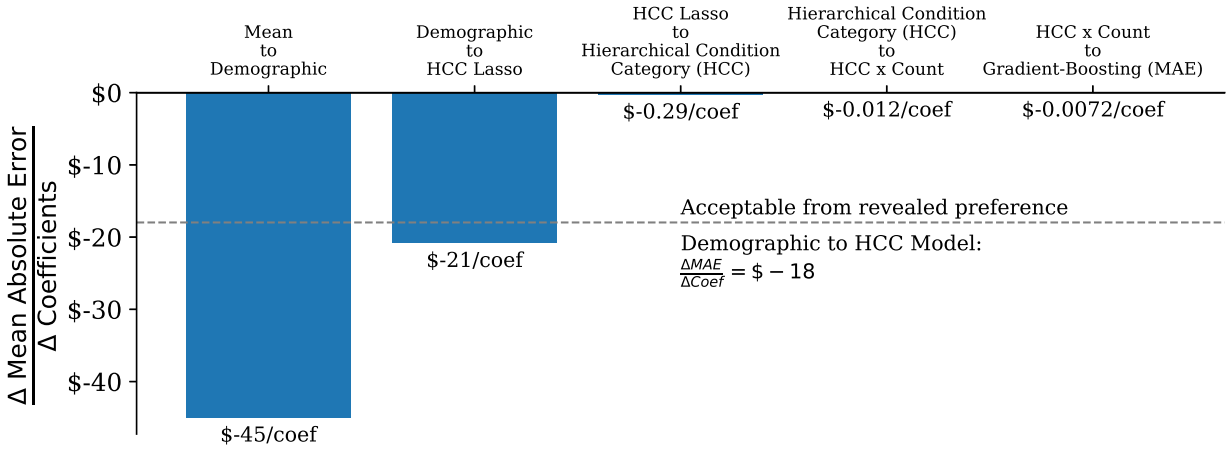
Notes: This graph shows the complexity of selected risk adjustment models fit in the data. The x-axis shows the names of various models. The "Hierarchical Condition Category (HCC)" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function that the model was trained on, mean absolute error or mean squared error. The y-axis shows the "complexity," or non-interpretability, of a model. Model complexity is calculated as the number of coefficients required to characterize the range of outputs. For linear models, complexity is the number of non-zero coefficients, or the L0 norm. For tree-based models, complexity is the number of unique and feasible decision paths in the model. Note that the y-axis is in log scale, not linear scale, to display the full range of complexity values.

Figure 4.3: Pareto Frontier of Accuracy (MAE) and Complexity (Number of Coefficients)



Notes: This graph shows a scatterplot of the accuracy and complexity of different risk adjustment models. The x-axis, in log scale, shows model complexity, measured as the number of coefficients. The y-axis shows model accuracy as measured by MAE out of sample. The black line shows the Pareto frontier of accuracy and complexity. The improving direction is toward the origin, or left and down, toward zero MAE and zero complexity.

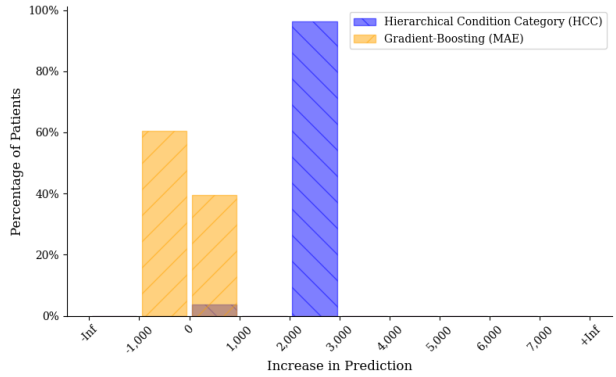
Figure 4.4: Marginal Change in MAE per Coefficient by Model for Subset of Pareto Models in Terms of Complexity and MAE



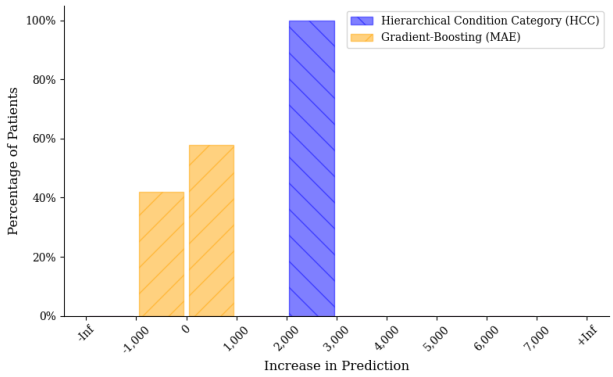
Notes: This figure restricts attention to models that are on the Pareto frontier of MAE and complexity. The x-axis lists model pairs in order of increasing complexity, and the y-axis shows the marginal decrease in MAE per additional model coefficient for each pair of models. The dotted horizontal gray line shows the marginal decrease in MAE per additional model coefficient for past risk adjustment model changes, specifically the change from the Demographic to the HCC model. These past changes were acceptable to policymakers. The more accurate, more complex models offer a much smaller decrease in MAE per additional coefficient. For these new models to be acceptable, policymakers would have to be willing to accept much smaller error reduction per coefficient than they have in the past.

Figure 4.5: Increase in Predicted Patient Cost from Upcoding

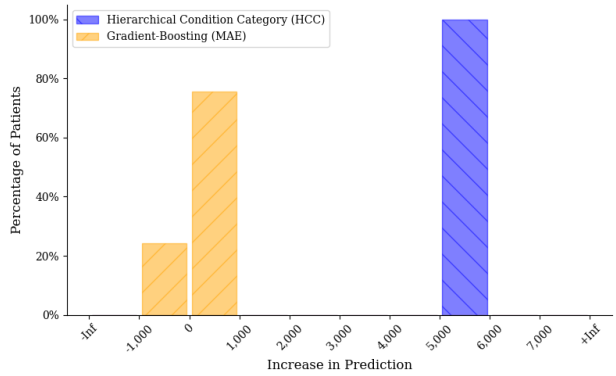
(a) Diabetes with Chronic Conditions (HCC 19)



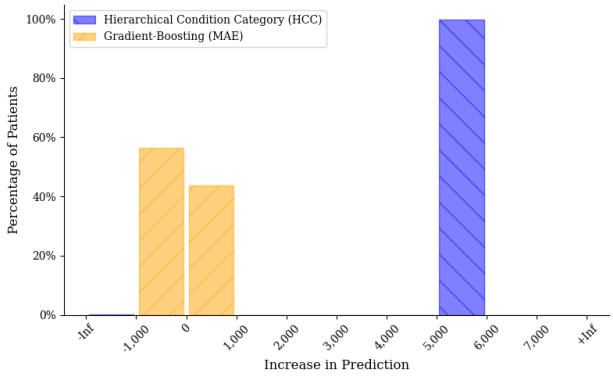
(b) Morbid Obesity (HCC 22)



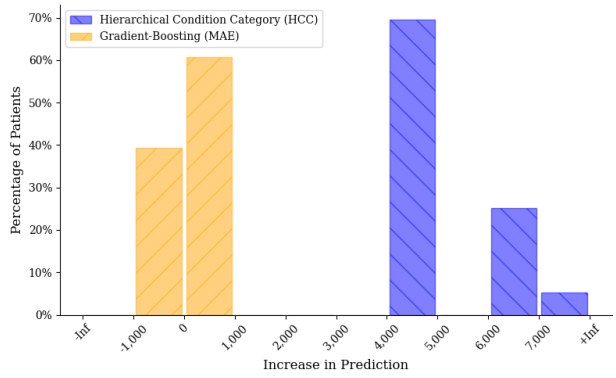
(c) Rheumatoid Arthritis (HCC 40)



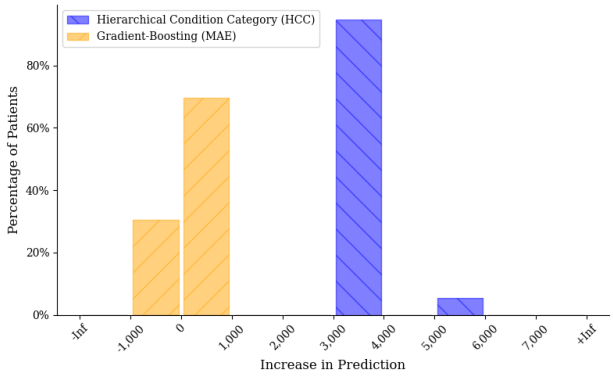
(d) Coagulation Defects (HCC 48)



(e) Congestive Heart Failure (HCC 85)



(f) Specified Heart Arrhythmias (HCC 96)



Notes: Panel a shows the distribution of the change in predicted patient costs from adding diabetes with chronic complications (HCC 19) to patient records without HCC 19. If the patients have HCC 17 or HCC 18, milder types of diabetes, on their record, then those are set to zero. If they do not have HCC 17 or 18, their count of HCCs is increased by one, and any relevant diabetes interaction terms are set to one. The x-axis contains bins for the change in predicted spending. The y-axis shows the fraction of patients in the validation sample who fall into the bin. Bins with 10 or fewer individuals are suppressed per CMS requirements. Panels b, c, d, e, and f show the same analysis for adding morbid obesity (HCC 22), rheumatoid arthritis (HCC 40), coagulation defects (HCC 48), congestive heart failure (HCC 85), and heart arrhythmias (HCC 96), respectively.

4.9 Sample and variable construction

Data overview. The Medicare data contain records of healthcare usage and diagnoses. I use diagnoses from the Carrier (physician services), Home Health, Outpatient (outpatient facility fees), and MedPAR (inpatient facility fees) claims files. I calculate spending from these same files.

Sample selection. The sample selection is as follows. Individuals must have been enrolled in both Medicare Part A and B for all of 2018 and at least one month of 2019. They are included if they are 65 years or older and qualified for Medicare by age, and they are excluded if they have end-stage renal disease or enrolled in Medicaid or Medicare Advantage at any time between 2018 and 2019. Individuals are also excluded if their gender is unknown or their state of residence is not a US state (e.g., a territory). Medicare also restricts its sample to those for whom Medicare is a primary payer and subsets the sample based on whether a patient is institutionalized long term. I cannot observe these variables, but I assume the vast majority of patients have Medicare as a primary payer and are not institutionalized.

Spending calculations. To calculate total patient spending, I add up all spending by Medicare, the patient, and other sources across all files in 2019. Spending is annualized by the number of months the patient was enrolled in Medicare Part A and B. I do not perform any price adjustments. Prices in Medicare are administratively set and vary slightly across geography due to geographic adjustments and other rate-setting tools. However, this price variation is small compared to variation in private insurance prices, and it does not drive meaningful variation in spending [CBO, 2022, Gottlieb et al., 2010a].

Note that Medicare risk adjustment takes the additional step of dividing total spending by mean spending such that spending outcomes are a percentage of mean spending (e.g., 200% of the mean). I do not implement this step, which keeps outcome units in 2019 dollars and aids in interpretability. The difference does not otherwise affect the results.

Predictor variable structure for non-parametric models. The machine learning models use the same variables as the standard HCC model. However, certain variables are formatted differently for the non-parametric models (tree, random forest, gradient-boosted tree) than in the HCC linear model. Broadly, variables are formatted as single variables (e.g., age, sex) rather than as a series of saturated indicators with pre-specified interactions.

- **Age.** Age information is binned into the same groupings as used in the HCC model. However, age group is provided to the model as a single ordinal variable, containing an ordered group number, rather than as a series of indicator variables interacted with sex. This preserves the ordinal information in the age group variable.

- **Sex.** Sex is provided to the machine learning models as a single binary variable rather than as a series of indicator variables interacted with the age group. This allows sex to be interacted with any variable in the process of model fitting rather than restricting it to interactions with the age group.

- **HCC groupings.** Each HCC grouping (e.g., diabetes of any severity) is provided to the machine learning models as a single indicator variable rather than as a series of interaction variables between specified HCC groupings. This allows HCC groupings to be interacted with any other HCC grouping or variable rather than restricting the interaction to a specified subset of other HCC groupings.

4.10 Model fitting and tuning

This section discusses the implementation of risk adjustment model fitting in greater detail to aid in replicability. The code is available upon request.

4.10.1 Standard Medicare Models

Demographic and HCC Models. The Demographic model uses five-year age bins interacted with sex indicators. This HCC model adds 86 hierarchical health condition indicators, 7 health condition interactions, and health condition count indicator variables for counts 4 through 10+. I use the 2023 V24 model version and fit it for community-dwelling, aged, non-disabled, non-Medicaid beneficiaries [CMS, 2023].

Validation of recalibrated Medicare models To verify the reliability of the Medicare models used in this analysis, I confirm that the recalibrated Medicare models have a very similar in-sample R^2 to the R^2 reported by Medicare. Medicare reports indicate an R^2 of 0.77% for the Demographic model, slightly lower than the R^2 of 1.61% in this sample. For the V24 HCC model, Medicare reports indicate an R^2 of 12.57, slightly higher than the R^2 of 11.03% in this sample [MedPAC, 2021]. I attribute the small differences in R^2 to differences in sample year and construction and to the inherent variance in the R^2 estimates.

4.10.2 Alternative linear models:

I include a selection of parametric models using OLS and lasso, designed to incorporate health condition interactions and reduce overfitting.

Lasso. The first model employs lasso regression with standard HCC model coefficients. Lasso regression is a standard model for reducing overfitting and therefore improving out-of-sample performance. Like linear regression, it minimizes MSE but adds a constraint on the max value of the sum of the absolute values of coefficients. As a result, it sets some coefficients to zero if they provide insufficient explanatory power while biasing the remaining coefficients toward zero [Hastie et al., 2009].

Of note, the lasso implements an L1 penalized sparse regression (limits the sum of the absolute value of coefficients). The ideal regression here would implement an L0 constraint, which limits the number of non-zero coefficients. However, L0 regression is notoriously

computationally inefficient and likely infeasible on a dataset of this size [Bertsimas et al., 2016, Hazimeh and Mazumder, 2020].

Lasso regressions are fit using Scikit-Learn [Pedregosa et al., 2011]. The optimal sparsity parameter for lasso is determined via cross-validation using the approach implemented in Pedregosa et al. [2011], "LassoCV."

"HCCxCount" model Another variant I introduce is "HCCxCount," with linear and lasso regression versions. This set of models is motivated by the observation that Medicare risk adjustment still consistently underpredicts costs for patients with multiple comorbidities and is known to inadequately account for comorbidity interactions [MedPAC, 2021]. These models add a set of regression variables that multiply each HCC indicator with the total count of HCCs, allowing the effect of health conditions on costs to increase with a patient's overall comorbidity burden. They also include standard HCC, HCC interaction, HCC count, and demographic variables. I fit both a standard linear regression and a lasso regression with these variables.

4.10.3 Tree-Based Machine Learning Models

All machine learning models are fit using Scikit-Learn [Pedregosa et al., 2011] except for gradient-boosted trees, which are fit using XGBoost [Chen and Guestrin, 2016].

Regression tree and random forest. The hyperparameters are determined via a random search of hyperparameter values, and selected hyperparameters are those that generate models that perform best in threefold cross-validation. The models are trained to minimize either MSE or MAE, and they are evaluated in cross-validation accordingly. Once the best set of hyperparameters are chosen, the model is refit with these parameters on the full training dataset.

Gradient-boosted regression tree. The hyperparameter tuning process is the same as for regression trees and random forest, with one additional step. Once the best set of hyper-

parameters are chosen, the model is refit with these parameters on the full training dataset with early stopping criteria. Early stopping prevents the model from fitting additional trees once additional trees stop improving out-of-sample fit.

4.11 Selection incentives analysis

The main text's analysis focuses on the trade-off between complexity and accuracy. In that context, I assume that a lower MAE serves as a sufficient statistic for both improved accuracy and reduced selection incentives. Figure 4.9 evaluates the validity of this assumption.

In addition to estimating MAE, I also calculate several common measures of selection incentives pulled from the literature, both overall and for select patient subgroups. This section describes the methods for this analysis and key results.

I first calculate MAE for different subgroups of patients where there is plausible means and motivation for selection. Evaluated subgroups include individuals with chronic mental health disorders, chronic substance use disorders, multiple chronic conditions, and no chronic conditions. The first three consistently cost more than predicted, while the last one consistently costs less [Zink and Rose, 2020, MedPAC, 2021].

To construct these groups, I use "chronic condition" variables taken from the Master Beneficiary Summary File Chronic Conditions and Other Chronic Conditions files. These variables are constructed using diagnoses from multiple years of claims data and prescription information [ResDAC, 2023]. They differ from the standard HCC health condition variables, which use only diagnoses from one year of claims data. As such, these chronic condition variables contain additional information that likely predicts spending and may be available to insurance plans. However, because these variables are not included in risk adjustment models, they reflect dimensions along which insurance companies have the incentives and means to influence patient selection. These considerations make them important factors for assessing selection incentives.

Figure 4.9a shows the MAE for these patient subgroups for different risk adjustment models. Overall, the MAE within each subgroup largely declines as model complexity increases, reaching its lowest with the gradient-boosted tree (MAE).

The first type of selection incentive I consider is selection conditional on predicted risk. Some patients may incur a lot of costs, yet their predicted costs could be even higher. These patients represent opportunities for positive selection. The sum of positive residuals captures the potential opportunity for positive selection [Brown et al., 2014]. Figure 4.9b shows the results by subgroup. Interestingly, the sum of positive residuals holds roughly constant or declines as model complexity increases. It drops dramatically with the gradient-boosted tree (MAE), both overall and by subgroup, and leads to much smaller differences across subgroups.

The second type of selection is selection to avoid tail risk. If insurance companies have some risk aversion, they will avoid patients with a high risk of costing substantially more than predicted and be less concerned about small deviations from predicted costs. Figure 4.9c shows the probability that a patient costs over twice as much as predicted, following Park and Basu [2018]. Broadly, differences in tail risk across groups are narrow with more complex models. The one exception to this is the gradient-boosted tree (MAE), which raises tail risk particularly for high-cost groups and leads to larger group-level differences.

The last type of selection incentive I examine is selection by risk, e.g., by expected compensation. Some patient groups consistently incur costs that exceed their predicted values. This discrepancy is typically measured as the net compensation for a group, representing the expected loss or gain per patient ($\frac{\sum_i \hat{y}_i - \sum_i y_i}{N}$). Accordingly, I calculate net compensation for various subgroups under different risk adjustment models. Figure 4.9d shows the results. The HCC model successfully narrows differences across groups in net compensation relative to predicting the mean. The gradient-boosted tree (MAE) leads to negative net compensation overall and larger differences across subgroups.

4.12 Additional exhibits

Table 4.1: Training Sample Summary Statistics

Variable	Mean	StdDev
Age	75.62	7.50
Female	0.56	0.50
Count of HCCs	2.60	3.56
Annualized Spending	13,449.85	35,441.55
Mortality Rate	0.04	0.19

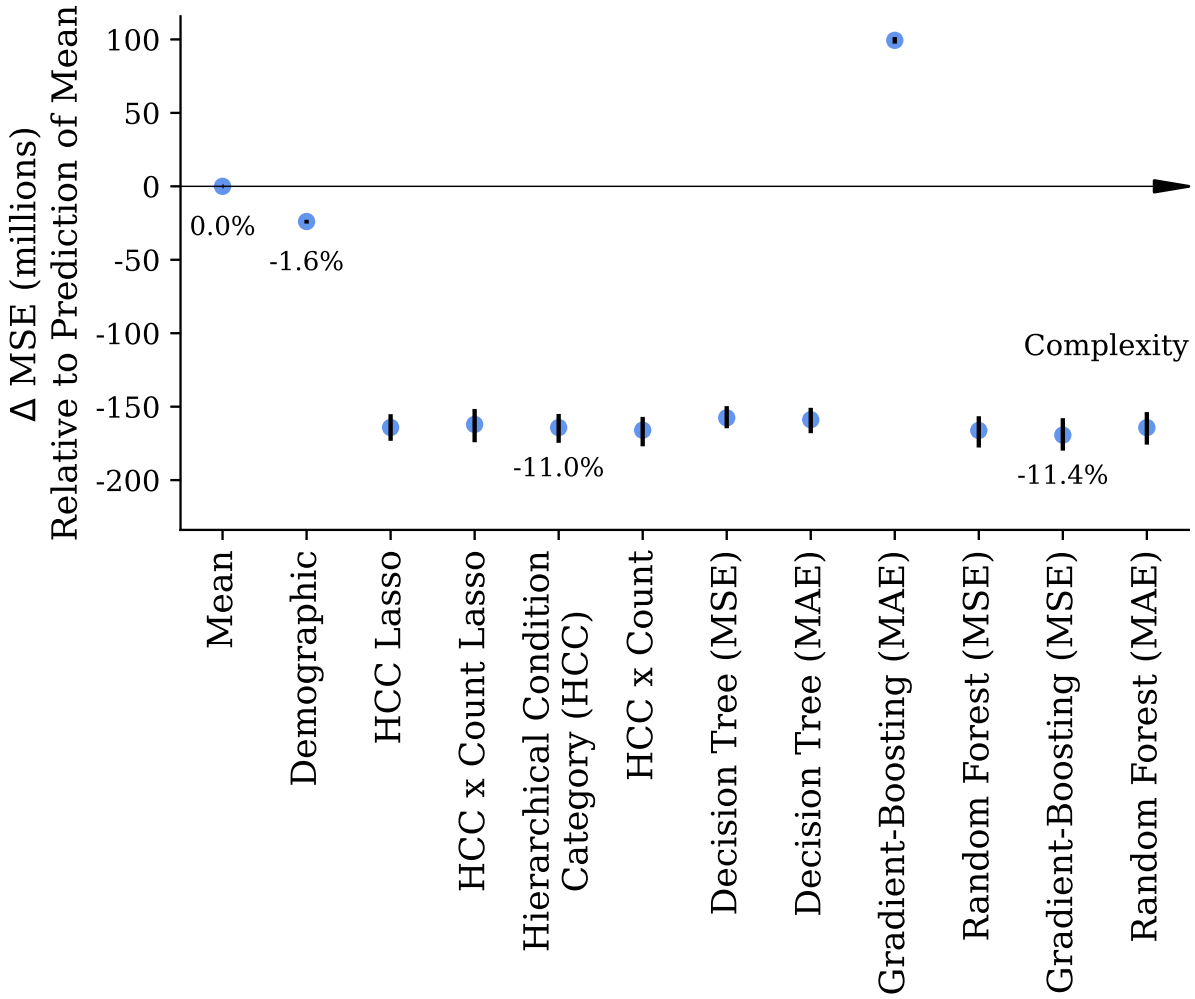
Notes: This table shows summary statistics for the data the models are trained on. The count of HCCs is the count of HCCs per person. These are determined based on diagnosis codes in claims in 2018, the "base year." Annualized spending and death rates are calculated in 2019, the "outcome year" for prospective payments.

Table 4.2: Validation Sample Summary Statistics

Variable	Mean	StdDev
Age	75.61	7.50
Female	0.56	0.50
Count of HCCs	2.61	3.57
Annualized Spending	13,512.97	38,587.86
Mortality Rate	0.04	0.19

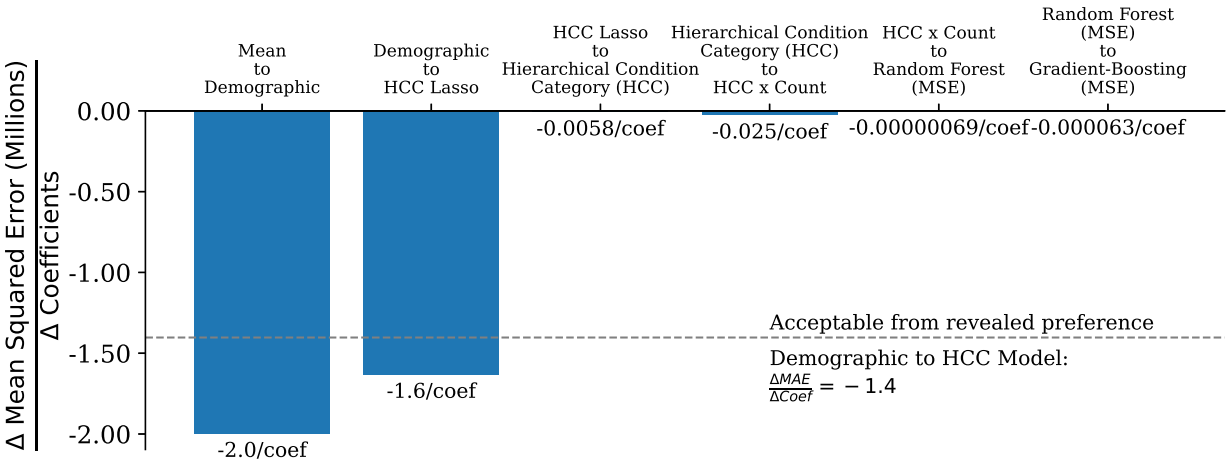
Notes: This table shows summary statistics for the validation data, i.e., the data where model accuracy is assessed out of sample. It is analogous to Table , which shows summary statistics for the training data. The count of HCCs is the count of HCCs per person and is determined based on diagnosis codes in claims in 2018, the "base year." Annualized spending and death rates are calculated in 2019, the "outcome year" for prospective payments.

Figure 4.6: Difference in MSE of Model Predictions Relative to Predicting the Mean



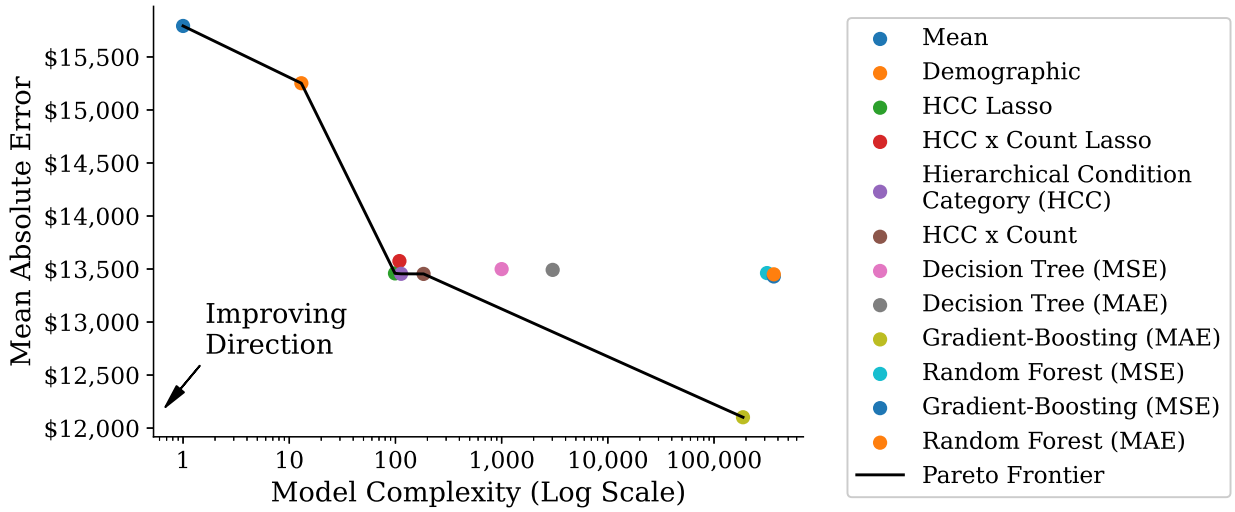
Notes: This graph shows the change in the MSE for model predictions relative to predicting the mean. It is analogous to Figure , but shows MSE rather than MAE. The x-axis shows the names of various models, ordered by increasing complexity. The "Hierarchical Condition Category" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function on which the model was trained, the mean absolute error, or the mean squared error. The MSE is calculated out of sample in 10% of the available data. The point estimate is the round marker, and standard errors are calculated with bootstrapped samples in the out-of-sample data. Ninety-five percent confidence intervals are shown as black bars; note that in some cases they are narrower than the height of the round markers. The values in parentheses are the percentage reductions in the MSE relative to always predicting the mean.

Figure 4.7: Marginal Change in MSE per Coefficient by Model for Subset of Pareto Models in Terms of Complexity and MAE



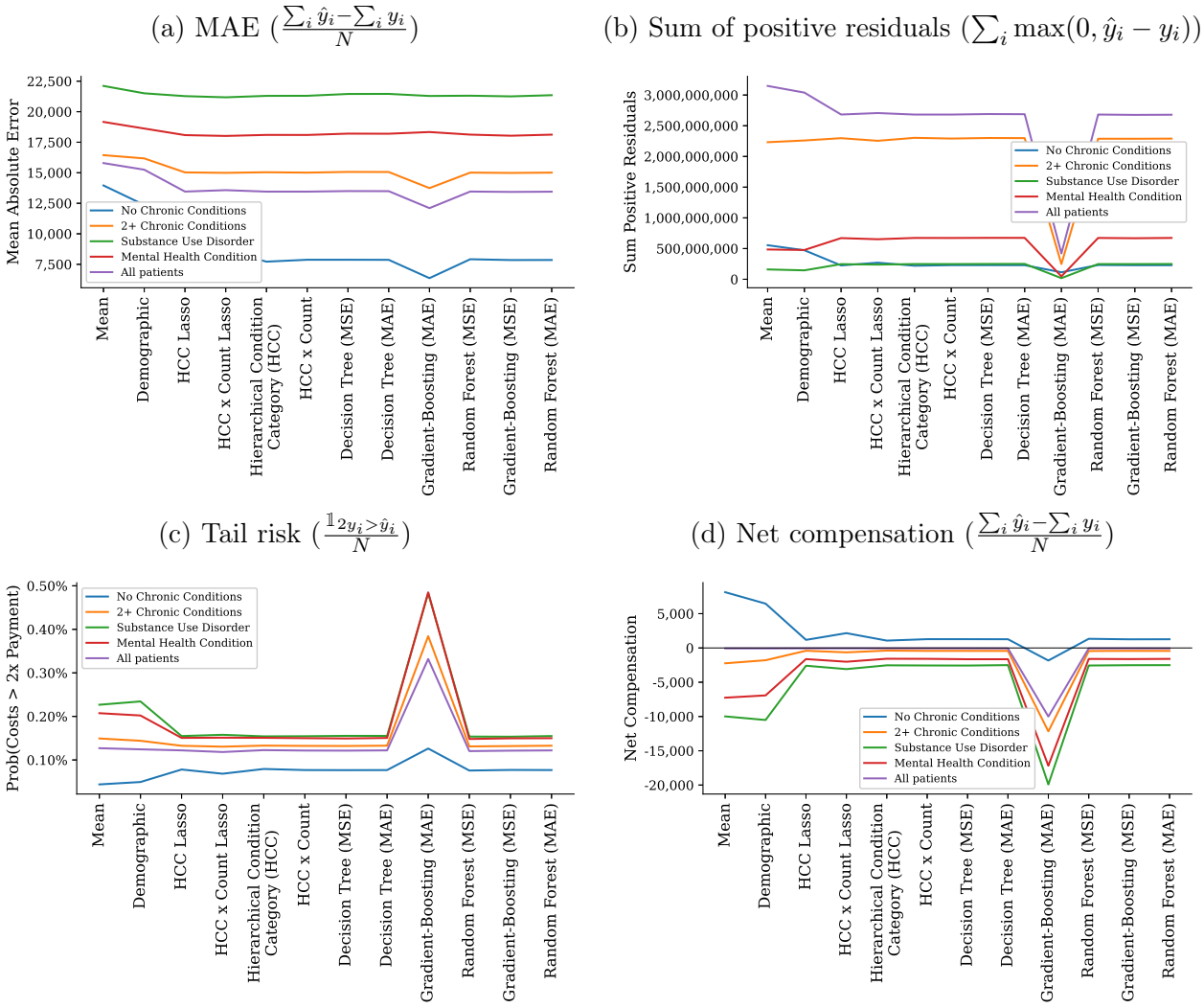
Notes: This figure restricts attention to models that are on the Pareto frontier of the mean squared error (MSE) and complexity. It is analogous to Figure but uses MSE rather than MAE. The x-axis lists model pairs in order of increasing complexity, and the y-axis shows the marginal decrease in the MSE per additional model coefficient for each pair of models. The dotted horizontal gray line shows the marginal decrease in the MSE per additional model coefficient for past risk adjustment model changes, specifically the change from the Demographic to the HCC model. I assume that model changes that offer less than 10% of this decrease are not acceptable to policymakers. As such, none of the models that are more complex than the HCC model are worth their additional complexity.

Figure 4.8: Pareto Frontier of Accuracy (MSE) and Complexity (Number of Coefficients)



Notes: This graph shows a scatterplot of the accuracy and complexity of different risk adjustment models. It is analogous to figure but shows MSE rather than MAE. The x-axis, in log scale, shows model complexity, measured as the number of coefficients. The y-axis shows model accuracy as measured by the mean squared error (MSE) out of sample. The black line shows the Pareto frontier of accuracy and complexity. The improving direction is toward the origin, or left and down, toward zero MSE and zero complexity.

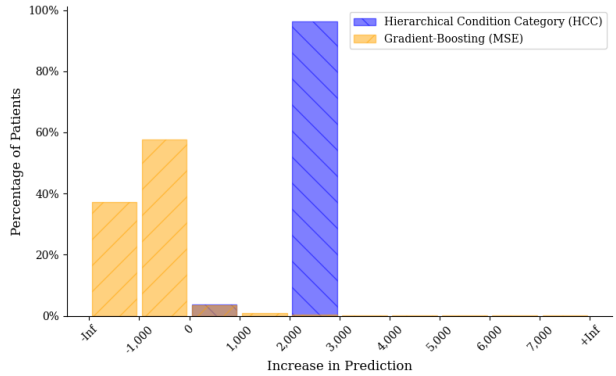
Figure 4.9: Alternative Measures of Model Performance and Selection Incentives by Model and Patient Subgroup



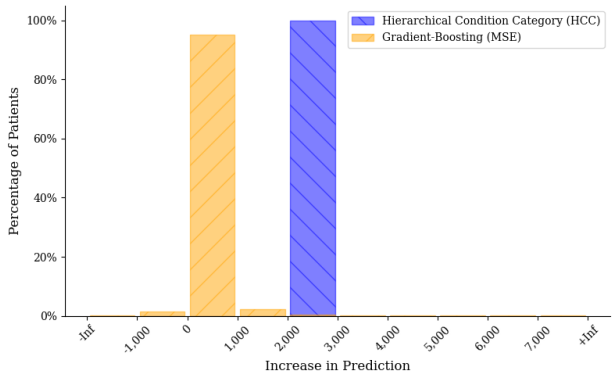
Notes: Panel a shows MAE on the y-axis for specified patient groups by risk adjustment model. Groups include patients with 0 chronic conditions, 2+ chronic conditions, mental health disorders, and substance use disorders, as identified by the Medicare Master Beneficiary Summary File Chronic Conditions and Other Conditions Files. These variables are based on multiple prior years of claims diagnosis data and prescription data, unlike the HCC variables, which are based only on one prior year of claims diagnosis data. Risk adjustment models are ordered on the x-axis by increasing complexity. Panel b is the same but with the sum of positive residuals on the y-axis. Panel c shows the tail risk, or the probability that realized expenditures substantially (2x) exceeds predicted spending and, therefore, payments. Panel d is similar but shows net compensation for the specified patient subgroups on the y-axis.

Figure 4.10: Increase in Predicted Patient Cost from Upcoding

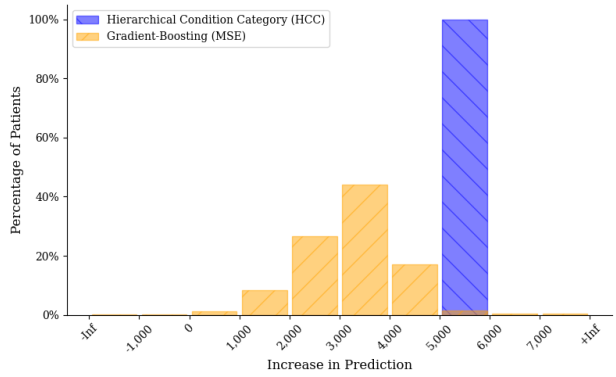
(a) Diabetes with Chronic Conditions (HCC 19)



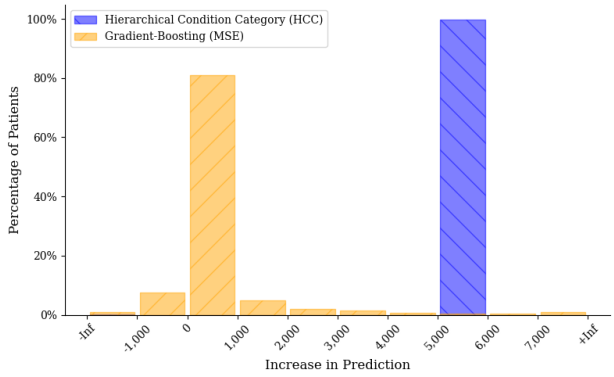
(b) Morbid Obesity (HCC 22)



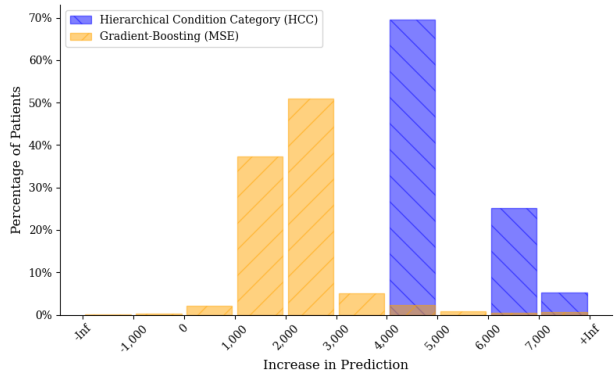
(c) Rheumatoid Arthritis (HCC 40)



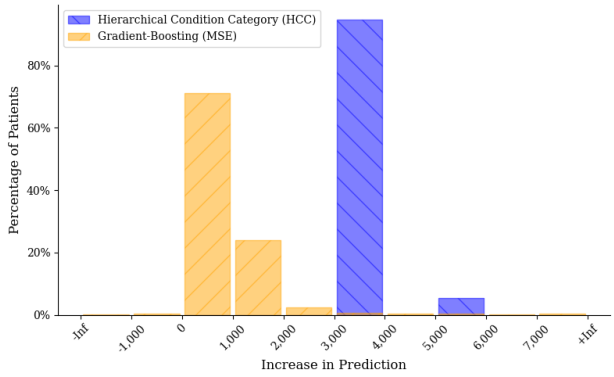
(d) Coagulation Defects (HCC 48)



(e) Congestive Heart Failure (HCC 85)



(f) Specified Heart Arrhythmias (HCC 96)



Notes:
Panel a shows the distribution of the change in predicted patient costs from adding diabetes with chronic complications (HCC 19) to patient records without HCC 19. It is analogous to Figure a but shows results for a gradient-boosted tree trained to minimize MSE, rather than MAE. If the patients have HCC 17 or HCC 18, milder types of diabetes, on their record, then those are set to zero. If they do not have HCC 17 or 18, their count of HCCs is increased by one, and any relevant diabetes interaction terms are set to one. The x-axis contains bins for the change in predicted spending. The y-axis shows the fraction of patients in the validation sample who fall into the bin. Bins with 10 or fewer individuals are suppressed per CMS requirements. Panels b, c, d, e, and f show the same analysis for adding morbid obesity (HCC 22), rheumatoid arthritis (HCC 40), coagulation defects (HCC 48), congestive heart failure (HCC 85), and heart arrhythmias (HCC 96), respectively.

REFERENCES

AAPC. What is cpt?, Dec 2021. URL https://www.aapc.com/resources/medical-coding/cpt.aspx. Available online at https://www.aapc.com/resources/medical-coding/cpt.aspx.

Daron Acemoglu and Joshua Linn. Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry. *The Quarterly Journal of Economics*, 119(3):1049–1090, 2004a. URL https://ideas.repec.org/a/oup/qjecon/v119y2004i3p1049-1090..html.

Daron Acemoglu and Joshua Linn. Market size in innovation: theory and evidence from the pharmaceutical industry. *The Quarterly journal of economics*, 119(3):1049–1090, 2004b.

Sumit Agarwal, J. Bradford Jensen, and Ferdinando Monte. Consumer Mobility and the Local Structure of Consumption Industries. NBER Working Paper 23616, 2020. URL https://ideas.repec.org/p/nbr/nberwo/23616.html.

Treb Allen, Simon Fuchs, Sharat Ganapati, Alberto Graziano, Rocio Madera, and Judit Montoriol-Garriga. Urban welfare: Tourism in barcelona. 2021.

American Cancer Society. Lifetime Risk of Developing or Dying From Cancer, January 2020.

James E. Anderson, Catherine A. Milot, and Yoto V. Yotov. How much does geography deflect services trade? canadian answers. *International Economic Review*, 55(3):791–818, 2014. ISSN 00206598, 14682354. URL http://www.jstor.org/stable/24517967.

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973, 1963. ISSN 00028282. URL http://www.jstor.org/stable/1812044.

Kenneth J Arrow, L Kamran Bilir, and Alan Sorensen. The impact of information technology on the diffusion of new pharmaceuticals. *American Economic Journal: Applied Economics*, 12(3):1–39, 2020.

Katherine Baicker and Amitabh Chandra. *Understanding Agglomerations in Health Care*, pages 211–236. University of Chicago Press, February 2010. doi:10.7208/chicago/9780226297927.001.0001. URL http://www.nber.org/chapters/c7986.

Dominick G. Bartelme, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare. The Textbook Case for Industrial Policy: Theory Meets Data. NBER Working Paper 26193, National Bureau of Economic Research, Aug 2019. URL https://ideas.repec.org/p/nbr/nberwo/26193.html.

Timothy Bartik and George Erickcek. Higher education, the health care industry, and metropolitan regional economic development: What can 'eds & meds' do for the economic fortunes of a metro area's residents? Upjohn Working Papers 08-140, W.E. Upjohn

Institute for Employment Research, 2007. URL https://EconPapers.repec.org/RePEc:upj:weupjo:08-140.

Emily Battaglia. The effect of hospital closures on maternal and infant health, March 2022. University of Delaware, mimeo. Available online at https://emilybattaglia.github.io/Battaglia/Battaglia_JMP.pdf (accessed June 29, 2022).

James R Baumgardner. Physicians' Services and the Division of Labor across Local Markets. *Journal of Political Economy*, 96(5):948–982, October 1988a. doi:10.1086/261571. URL https://ideas.repec.org/a/ucp/jpolec/v96y1988i5p948-82.html.

James R. Baumgardner. Physicians' Services and the Division of Labor Across Local Markets. *Journal of Political Economy*, 96(5):948–982, October 1988b. ISSN 0022-3808, 1537-534X. doi:10.1086/261571. URL https://www.journals.uchicago.edu/doi/10.1086/261571.

G. S. Becker and K. M. Murphy. The Division of Labor, Coordination Costs, and Knowledge. *The Quarterly Journal of Economics*, 107(4):1137–1160, November 1992. ISSN 0033-5533, 1531-4650. doi:10.2307/2118383. URL https://academic.oup.com/qje/article-lookup/doi/10.2307/2118383.

Robert A. Berenson, Jonathan H. Sunshine, David Helms, and Emily Lawton. Why medicare advantage plans pay hospitals traditional medicare prices. *Health Affairs*, 34(8):1289–1295, 2015.

Christopher R Berry. Reassessing the property tax. *Available at SSRN 3800536*, 2021.

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

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, pages 813–852, 2016.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research, 2020a.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research, 2020b.

Jason Brown, Mark Duggan, Ilyana Kuziemko, and William Woolston. How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program. *American Economic Review*, 104(10):3335–3364, October 2014. ISSN 0002-8282. doi:10.1257/aer.104.10.3335. URL http://pubs.aeaweb.org/doi/10.1257/aer.104.10.3335.

Eric Budish, Benjamin N Roin, and Heidi Williams. Do firms underinvest in long-term research? evidence from cancer clinical trials. *American Economic Review*, 105(7):2044–2085, 2015.

Ariel Burstein, Sarah Lein, and Jonathan Vogel. Cross-border shopping: evidence and welfare implications for switzerland. 2022.

Cory Capps, David Dranove, and Mark Satterthwaite. Competition and market power in option demand markets. *The RAND Journal of Economics*, pages 737–763, 2003.

CBO. The prices that commercial health insurers and medicare pay for hospitals' and physicians' services. Technical Report 57422, Congressional Budget Office, 2022. URL https://www.cbo.gov/system/files/2022-01/57422-medical-prices.pdf.

Centers for Disease Control and Prevention. FastStats - Leading Causes of Death, January 2022. URL https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

Centers for Medicare and Medicaid Services. National health expenditure data, Sep 2022. URL https://www.cms.gov/files/zip/nhe-tables.zip.

David C. Chan, Matthew Gentzkow, and Chuan Yu. Selection with Variation in Diagnostic Skill: Evidence from Radiologists. *The Quarterly Journal of Economics*, 137(2):729–783, 2022.

Amitabh Chandra, David Malenka, and Jonathan Skinner. The diffusion of new medical technology: The case of drug-eluting stents. In *Discoveries in the Economics of Aging*, pages 389–403. University of Chicago Press, 2014.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi:10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

John S. Chipman. External Economies of Scale and Competitive Equilibrium. *The Quarterly Journal of Economics*, 84(3):347–385, 1970. URL https://ideas.repec.org/a/oup/qjecon/v84y1970i3p347-385..html.

Helen Clapesattle. *The Doctors Mayo*. Mayo Foundation for Medical Education, 1969.

Jeffrey Clemens and Joshua D. Gottlieb. Do physicians' financial incentives affect treatment patterns and patient health? *American Economic Review*, 104(4):1320–1349, April 2014.

Jeffrey Clemens and Joshua D. Gottlieb. In the shadow of a giant: Medicare's influence on private payment systems. *Journal of Political Economy*, 125(1):1–39, February 2017.

CMS. Cms-hcc software v2422.86.p2. https://www.cms.gov/files/zip/2023-initial-icd-10-mappings.zip, 2023. accessed April 28, 2023.

Zack Cooper, Stuart V. Craig, Martin Gaynor, and John Van Reenen. The price ain't right? hospital prices and health spending on the privately insured. *The Quarterly Journal of Economics*, 134(1):51–107, 09 2018. ISSN 0033-5533. doi:10.1093/qje/qjy020. URL https://doi.org/10.1093/qje/qjy020.

Arnaud Costinot, Dave Donaldson, Margaret Kyle, and Heidi Williams. The more we die, the more we sell? a simple test of the home-market effect. *The Quarterly Journal of Economics*, 134(2):843–894, 2019. URL https://EconPapers.repec.org/RePEc:oup:qjecon:v:134:y:2019:i:2:p:843-894.

Juliette Cubanski and Tricia Neuman. The Facts on Medicare Spending and Financing. *Henry J. Kaiser Family Foundation, San Francisco*, 2023. URL https://www.kff.org/medicare/issue-brief/what-to-know-about-medicare-spending-and-financing/.

Leemore Dafny, Kate Ho, and Robin S Lee. The price effects of cross-market mergers: theory and evidence from the hospital industry. *The RAND Journal of Economics*, 50(2):286–325, 2019.

James E. Dalen, Kenneth J. Ryan, and Joseph S. Alpert. Where Have the Generalists Gone? They Became Specialists, Then Subspecialists. *The American Journal of Medicine*, 130(7):766–768, July 2017. ISSN 00029343. doi:10.1016/j.amjmed.2017.01.026. URL https://linkinghub.elsevier.com/retrieve/pii/S0002934317301341.

Donald R. Davis and Jonathan I. Dingel. The comparative advantage of cities. *Journal of International Economics*, 123(C), 2020. doi:10.1016/j.jinteco.2020.10. URL https://ideas.repec.org/a/eee/inecon/v123y2020ics0022199620300106.html.

Donald R. Davis and David E. Weinstein. Market access, economic geography and comparative advantage: an empirical test. *Journal of International Economics*, 59(1):1–23, 2003. URL http://ideas.repec.org/a/eee/inecon/v59y2003i1p1-23.html.

Donald R. Davis, Jonathan I. Dingel, Joan Monras, and Eduardo Morales. How Segregated Is Urban Consumption? *Journal of Political Economy*, 127(4):1684–1738, 2019. doi:10.1086/701680. URL https://ideas.repec.org/a/ucp/jpolec/doi10.1086-701680.html.

Tatyana Deryugina and David Molitor. The causal effects of place on health and longevity. *Journal of Economic Perspectives*, 35(4):147–70, November 2021. doi:10.1257/jep.35.4.147. URL https://www.aeaweb.org/articles?id=10.1257/jep.35.4.147.

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

Jonathan I. Dingel. The Determinants of Quality Specialization. *Review of Economic Studies*, 84(4):1551–1582, 2017. URL https://ideas.repec.org/a/oup/restud/v84y2017i4p1551-1582..html.

Jonathan I. Dingel and Felix Tintelnot. Spatial economics for granular settings. Working Paper 27287, National Bureau of Economic Research, January 2021. URL http://www.nber.org/papers/w27287.

Anne-Célia Disdier and Keith Head. The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48, February 2008. URL http://ideas.repec.org/a/tpr/restat/v90y2008i1p37-48.html.

E. Ray Dorsey and Eric J. Topol. State of telehealth. *New England Journal of Medicine*, 375(2):154–161, 2016. doi:10.1056/NEJMra1601705. URL https://doi.org/10.1056/NEJMra1601705. PMID: 27410924.

Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.

David Dranove and Christopher Ody. Evolving measures of provider market power. *American Journal of Health Economics*, 2(2):145–160, 2016.

David Dranove and Mark A. Satterthwaite. The industrial organization of health care markets. *Handbook of Health Economics*, 1:1093–1139, 2000.

David Dranove, Mark Shanley, and Carol Simon. Is hospital competition wasteful? *The RAND Journal of Economics*, 23(2):247–262, 1992.

Fabian Eckert, Sharat Ganapati, and Conor Walsh. Skilled Scalable Services: The New Urban Bias in Economic Growth. CESifo Working Paper Series 8705, 2020. URL https://ideas.repec.org/p/ces/ceswps/_8705.html.

Randall P. Ellis and Thomas G. McGuire. Predictability and Predictiveness in Health Care Spending. *Journal of Health Economics*, 26(1):25–48, 2007. ISSN 0167-6296. doi:https://doi.org/10.1016/j.jhealeco.2006.06.004. URL https://www.sciencedirect.com/science/article/pii/S0167629606000725.

Keith Marzilli Ericson and Amanda Starc. Measuring consumer valuation of limited provider networks. *American Economic Review*, 105(5):115–19, 2015.

Elena Falcettoni. The determinants of physicians' location choice: Understanding the rural shortage, January 2021. Mimeo, Federal Reserve. Available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3493178.

Amy Finkelstein, Matthew Gentzkow, and Heidi Williams. Sources of geographic variation in health care: Evidence from patient migration. *Quarterly Journal of Economics*, 131(4):1681–1726, 2016.

Amy Finkelstein, Matthew Gentzkow, and Heidi Williams. Place-based drivers of mortality: Evidence from migration. *American Economic Review*, 111(8):2697–2735, August 2021. doi:10.1257/aer.20190825. URL https://www.aeaweb.org/articles?id=10.1257/aer.20190825.

Stefanie J. Fischer, Heather Royer, and Corey D. White. Health care centralization: The health impacts of obstetric unit closures in the us. NBER Working Paper 30141, June 2022. URL http://www.nber.org/papers/w30141.

Elliott S. Fisher, David E. Wennberg, Thrse A. Stukel, Daniel J. Gottlieb, F. Lee Lucas, and Etoile L. Pinder. The implications of regional variations in medicare spending. part 1: the content, quality, and accessibility of care. *Annals of Internal Medicine*, 138(4):273–287, 2003a.

Elliott S. Fisher, David E. Wennberg, Thrse A. Stukel, Daniel J. Gottlieb, F. Lee Lucas, and Etoile L. Pinder. The implications of regional variations in medicare spending. part 2: health outcomes and satisfaction with care. *Annals of Internal Medicine*, 138(4):288–298, 2003b.

Abraham Flexner. Medical education in the united states and canada: A report to the carnegie foundation for the advancement of teaching. Bulletin Number Four, Carnegie Foundation, 1910.

Centers for Disease Control and Prevention. 10 Leading Causes of Death, United States 2020, Both Sexes, All Ages, All Races, 2019. URL `https://wisqars.cdc.gov/data/lcd/home`.

Masahisa Fujita and Paul Krugman. When is the economy monocentric?: von thünen and chamberlin unified. *Regional Science and Urban Economics*, 25(4):505–528, 1995.

Craig Garthwaite, Christopher Ody, and Amanda Starc. Endogenous quality investments in the us hospital market. *Journal of Health Economics*, page 102636, 2022.

Martin Gaynor and William B Vogt. Competition among hospitals. *The RAND Journal of Economics*, 34(3):764–785, Winter 2003.

Martin S. Gaynor, Samuel A. Kleiner, and William B. Vogt. A structural approach to market definition with an application to the hospital industry. *The Journal of Industrial Economics*, 61(2):243–289, 2013.

Michael Geruso and Timothy Layton. Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.

Michael Geruso and Timothy J. Layton. Selection in Health Insurance Markets and Its Policy Remedies. *Journal of Economic Perspectives*, 31(4):23–50, November 2017. ISSN 0895-3309. doi:10.1257/jep.31.4.23. URL `http://pubs.aeaweb.org/doi/10.1257/jep.31.4.23`.

Dean H Gesme and Marian Wiseman. Subspecialization in community oncology: option or necessity? *Journal of Oncology Practice*, 7(3):199, 2011.

Edward L. Glaeser. Reinventing boston: 1630–2003. *Journal of Economic Geography*, 5(2): 119–153, 2005.

Daniel J. Gottlieb, Weiping Zhou, Yunjie Song, Kathryn Gilman Andrews, Jonathan S. Skinner, and Jason M. Sutherland. Prices Don't Drive Regional Medicare Spending Variations. *Health Affairs*, 29(3):537–543, 2010a.

Daniel J Gottlieb, Weiping Zhou, Yunjie Song, Kathryn Gilman Andrews, Jonathan S Skinner, and Jason M Sutherland. Prices don't drive regional medicare spending variations. *Health Affairs*, 29(3):537–543, 2010b.

Joshua D. Gottlieb, Maria Polyakova, Kevin Rinz, Hugh Shiplett, and Victoria Udalova. Who values human capitalists' human capital? healthcare spending and physician earnings. University of Chicago, mimeo., July 2020.

Antonio M. Gotto and Jennifer Moon. *Weill Cornell Medicine: A History of Cornell's Medical School*. Ithaca: Cornell University Press, 2016.

Gautam Gowrisankaran, Aviv Nevo, and Robert Town. Mergers when prices are negotiated: Evidence from the hospital industry. *American Economic Review*, 105(1):172–203, 2015.

Laura S Graham, Alexandra O Sokolova, Ali Raza Khaki, Qian "Vicky" Wu, and Nancy E Davidson. Gender differences in faculty rank and subspecialty choice among academic medical oncologists. *Cancer investigation*, 39(1):21–24, 2021.

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

Atul Gupta. Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, 111(4):1241–83, April 2021. doi:10.1257/aer.20171825. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20171825`.

Atul Gupta, Amol Navathe, and Aaron Schwartz. How does medicare advantage affect health care use? evidence from beneficiary transitions, 2022. University of Pennsylvania, mimeo.

Gordon H. Hanson and Chong Xiang. The home-market effect and bilateral trade patterns. *American Economic Review*, 94(4):1108–1129, 2004. URL `http://ideas.repec.org/a/aea/aecrev/v94y2004i4p1108-1129.html`.

Vasyl Harasymiv. Lessons from 2 million machine learning models on kaggle, 2015. URL `https://www.kdnuggets.com/2015/12/harasymiv-lessons-kaggle-machine-learning.html`.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

Elhanan Helpman and Paul R. Krugman. *Market Structure and Foreign Trade*. MIT Press, 1985.

Josefa Henriquez, Richard C. van Kleef, Andrew Matthews, Thomas McGuire, and Francesco Paolucci. Combining risk adjustment with risk sharing in health plan payment systems: Private health insurance in australia. Technical report, National Bureau of Economic Research, 2023.

Kate Ho and Robin S. Lee. Insurer competition in health care markets. *Econometrica*, 85 (2):379–417, 2017.

Kate Ho and Robin S. Lee. Equilibrium provider networks: Bargaining and exclusion in health care markets. *American Economic Review*, 109(2):473–522, 2019.

Katherine Ho. Insurer-provider networks in the medical care market. *American Economic Review*, 99(1):393–430, March 2009. doi:10.1257/aer.99.1.393. URL `https://www.aeaweb.org/articles?id=10.1257/aer.99.1.393`.

Chang-Tai Hsieh and Esteban Rossi-Hansberg. The Industrial Revolution in Services. Working Papers 21-34, Center for Economic Studies, U.S. Census Bureau, October 2021. URL `https://ideas.repec.org/p/cen/wpaper/21-34.html`.

Jeremy A. Irvin, Andrew A. Kondrich, Michael Ko, Pranav Rajpurkar, Behzad Haghgoo, Bruce E. Landon, Robert L. Phillips, Stephen Petterson, Andrew Y. Ng, and Sanjay Basu. Incorporating Machine Learning and Social Determinants of Health Indicators into Prospective Risk Adjustment for Health Plan Payments. *BMC Public Health*, 20:1–10, 2020.

Mukesh K. Jain, Vivian G. Cheung, Paul J. Utz, Brian K. Kobilka, Tadataka Yamada, and Robert Lefkowitz. Saving the Endangered Physician-Scientist — A Plan for Accelerating Medical Breakthroughs. *New England Journal of Medicine*, 381(5):399–402, August 2019. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMp1904482. URL `http://www.nejm.org/doi/10.1056/NEJMp1904482`.

Melissa James, Michael Stearns, and Kimberly Rykaczewski. A First Look at the 2024 CMS Advance Notice. `https://www.wolterskluwer.com/en/expert-insights/a-first-look-at-the-2024-cms-advance-notice`, 2023. accessed June 7, 2023.

J. Bradford Jensen and Lori G. Kletzer. Tradable services: Understanding the scope and impact of services offshoring. *Brookings Trade Forum*, pages 75–116, 2005. URL `http://www.jstor.org/stable/25058763`.

Benjamin F Jones. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *Review of Economic Studies*, page 35, 2009.

Hong J. Kan, Hadi Kharrazi, Hsien-Yen Chang, Dave Bodycombe, Klaus Lemke, and Jonathan P. Weiner. Exploring the Use of Machine Learning for Risk Adjustment: A Comparison of Standard and Penalized Linear Regression Models in Predicting Health Care Costs in Older Adults. *PloS One*, 14(3):e0213258, 2019.

John Kautter, Gregory C Pope, and Patricia Keenan. Affordable care act risk adjustment: Overview, context, and challenges. *Medicare & Medicaid Research Review*, 4(3), 2014.

Daniel P. Kessler and Mark B. McClellan. Is hospital competition socially wasteful? *The Quarterly Journal of Economics*, 115(2):577–615, 2000.

KFF. Total medicaid mco spending, 2023. URL https://www.kff.org/other/state-indicator/total-medicaid-mco-spending/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D. Accessed on May 25, 2023.

Stephanie Khoury, Jonathan M. Leganza, and Alex Masucci. Health professional shortage areas and physician location decisions, February 2022. UCSD, mimeo. Available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3701160.

Jon Kleinberg and Sendhil Mullainathan. Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808, 2019.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

Patrick Kline and Enrico Moretti. Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority. *The Quarterly Journal of Economics*, 129(1):275–331, 11 2013. ISSN 0033-5533. doi:10.1093/qje/qjt034. URL https://doi.org/10.1093/qje/qjt034.

Paul Krugman. Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, 70(5):950–59, 1980.

Timothy J. Layton, Thomas G. McGuire, and Richard C. Van Kleef. Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets. *Journal of Health Economics*, 61:93–110, 2018.

Sanghoon Lee. Ability sorting and consumer city. *Journal of Urban Economics*, 68(1):20–33, 2010. ISSN 0094-1190. doi:https://doi.org/10.1016/j.jue.2010.03.002. URL https://www.sciencedirect.com/science/article/pii/S0094119010000136.

Antoine Levy and Jacob Moscona. Specializing in density: Spatial sorting and the pattern of trade, 2020. Mimeo, MIT. Available online at https://economics.mit.edu/files/16986.

Matthew S. Lewis and Kevin E. Pflum. Diagnosing hospital system bargaining power in managed care networks. *American Economic Journal: Economic Policy*, 7(1):243–74, 2015.

Matthew S. Lewis and Kevin E. Pflum. Hospital systems and bargaining power: Evidence from out-of-market acquisitions. *The RAND Journal of Economics*, 48(3):579–610, 2017.

Robert E. Lipsey. Measuring international trade in services. In Marshall B. Reinsdorf and Matthew J. Slaughter, editors, *International Trade in Services and Intangibles in the Era of Globalization*, pages 27–74. University of Chicago Press, 2009. doi:doi:10.7208/9780226709604-003. URL https://doi.org/10.7208/9780226709604-003.

Eric Lopez and Gretchen Jacobson. How much more than medicare do private insurers pay? a review of the literature, Apr 2020. URL https://www.kff.org/medicare/issue-brief/how-much-more-than-medicare-do-private-insurers-pay-a-review-of-the-literature/.

Alfred Marshall. *Principles of Economics*. London: Macmillan and Co., 1890.

Bill McGivney. The nccn compendium for cancer management. *American Health and Drug Benefits*, 1(5):40–44, 2008. Interview with Bill McGivney, PhD.

Thomas G. McGuire, Joseph P. Newhouse, and Anna D. Sinaiko. An Economic History of medicare part c. *The Milbank Quarterly*, 89(2):289–332, 2011.

Thomas G McGuire, Anna L Zink, and Sherri Rose. Improving the performance of risk adjustment systems: Constrained regressions, reinsurance, and variable selection. *American Journal of Health Economics*, 7(4):497–521, 2021.

Medicare Trustees. 2022 annual report of the boards of trustees of the federal hospital insurance and federal supplementary medical insurance trust funds, 2022. URL https://www.cms.gov/files/document/2022-medicare-trustees-report.pdf.

MedPAC. Report to the Congress: Medicare and the Health Care Delivery System. Technical report, Medicare Payment Advisory Commission, 2014.

MedPAC. Report to the Congress: Risk adjustment in medicare advantage. Technical report, Medicare Payment Advisory Commission, December 2021.

David O. Meltzer and Jeanette W. Chung. U.S. Trends in Hospitalization and Generalist Physician Workforce and the Emergence of Hospitalists. *Journal of General Internal Medicine*, 25(5):453–459, May 2010. ISSN 0884-8734, 1525-1497. doi:10.1007/s11606-010-1276-2. URL http://link.springer.com/10.1007/s11606-010-1276-2.

Yuhei Miyauchi, Kentaro Nakajima, and Stephen J. Redding. The economics of spatial mobility: Theory and evidence using smartphone data. Working Paper 28497, National Bureau of Economic Research, February 2021. URL http://www.nber.org/papers/w28497.

Enrico Moretti. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics*, 121(1-2):175–212, 2004.

Enrico Moretti. Local labor markets. In *Handbook of Labor Economics*, volume 4, pages 1237–1313. Elsevier, 2011.

Mathilde Muñoz. Trading non-tradables: The implications of europe's job posting policy. 2022.

Sendhil Mullainathan and Ziad Obermeyer. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *The Quarterly Journal of Economics*, 137(2):679–727, 12 2021. ISSN 0033-5533. doi:10.1093/qje/qjab046. URL `https://doi.org/10.1093/qje/qjab046`.

NCCN. About the national comprehensive cancer network. `https://www.nccn.org/home/about`, 2023a. URL `https://www.nccn.org/home/about`. Accessed on January 15, 2024.

NCCN. History of the national comprehensive cancer network. `https://www.nccn.org/home/about/nccn-history`, 2023b. URL `https://www.nccn.org/home/about/nccn-history`. Accessed on January 15, 2024.

Jonathan D. Neufeld and Charles R. Doarn. Telemedicine spending by medicare: A snapshot from 2012. *Telemed J E Health*, 21(8):686–693, Aug 2015. ISSN 1556-3669 (Electronic); 1530-5627 (Linking). doi:10.1089/tmj.2014.0185.

Joseph P. Newhouse. Geographic access to physician services. *Annual Review of Public Health*, 11(1):207–230, 1990.

Joseph P. Newhouse, Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. Does the geographical distribution of physicians reflect market failure? *The Bell Journal of Economics*, 13(2):493–505, 1982a.

Joseph P. Newhouse, Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. The geographic distribution of physicians: Is the conventional wisdom correct?, 1982b. Santa Monica: RAND Corp. Publ. No. R-2734.

Joseph P. Newhouse, Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. Where have all the doctors gone? *Journal of the American Medical Association*, 247(17): 2392–2396, 1982c.

Joseph P. Newhouse, Mary Price, John Hsu, J. Michael McWilliams, and Thomas G. McGuire. How much favorable selection is left in Medicare Advantage? *American Journal of Health Economics*, 2015.

Didrik Nielsen. Tree Boosting with XGBoost—Why Does XGBoost Win "Every" Machine Learning Competition? Master's thesis, NTNU, 2016.

Mr John Norregaard. *Taxing immovable property revenue potential and implementation challenges*. International Monetary Fund, 2013.

Pan Pantziarka, Rica Capistrano I, Arno De Potter, Liese Vandeborne, and Gauthier Bouche. An open access database of licensed cancer drugs. *Frontiers in pharmacology*, 12:627574, 2021.

Sungchul Park and Anirban Basu. Alternative Evaluation Metrics for Risk Adjustment Methods. *Health Economics*, March 2018. ISSN 10579230. doi:10.1002/hec.3657. URL http://doi.wiley.com/10.1002/hec.3657.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Nathan Petek. The marginal benefit of hospitals: Evidence from the effect of entry and exit on utilization and mortality rates. *Journal of Health Economics*, 86:102688, 2022.

Gregory C. Pope, John Kautter, Randall P. Ellis, Arlene S. Ash, John Z. Ayanian, Lisa I. Iezzoni, Melvin J. Ingber, Jesse M. Levy, and John Robst. Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model. *Health Care Financing Review*, 25(4): 119, 2004.

RA Popescu, Robert Schäfer, Raffaele Califano, Robert Eckert, Robert Coleman, J-Y Douillard, Andrés Cervantes, Paolo G Casali, Cristiana Sessa, Eric Van Cutsem, et al. The current and future role of the medical oncologist in the professional care for cancer patients: a position paper by the european society for medical oncology (esmo). *Annals of Oncology*, 25(1):9–15, 2014.

RCX Rules. An overview of medicare, September 2023. URL rcxrules.com. Accessed September 25, 2023.

ResDAC. 30 ccw chronic conditions algorithms: Mbsf_chronic_{YYYY}. Technical report, Centers for Medicaid and Medicare, 07 2023.

Sherri Rose. A Machine Learning Framework for Plan Payment Risk Adjustment. *Health Services Research*, 51(6):2358–2374, December 2016. ISSN 00179124. doi:10.1111/1475-6773.12464. URL http://doi.wiley.com/10.1111/1475-6773.12464.

Sherri Rose, Savannah L. Bergquist, and Timothy J. Layton. Computational Health Economics for Identification of Unprofitable Health Care Enrollees. *Biostatistics*, 18(4):682–694, October 2017. ISSN 1465-4644, 1468-4357. doi:10.1093/biostatistics/kxx012. URL http://academic.oup.com/biostatistics/article/18/4/682/3077114/Computational-health-economics-for-identification.

R. A. Rosenblatt and L. G. Hart. Physicians and rural america. *The Western Journal of Medicine*, 173(5):348–351, 11 2000. doi:10.1136/ewjm.173.5.348. URL https://pubmed.ncbi.nlm.nih.gov/11069878.

Meredith B. Rosenthal, Alan Zaslavsky, and Joseph P. Newhouse. The geographic distribution of physicians revisited. *Health Services Research*, 40(6p1):1931–1952, 2005. doi:https://doi.org/10.1111/j.1475-6773.2005.00440.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6773.2005.00440.x.

Stuart S. Rosenthal and William C. Strange. The attenuation of human capital spillovers. *Journal of Urban Economics*, 64(2):373–389, 2008.

Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler, and Matthew Sobek. IPUMS USA. Technical Report Version 12.0, Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D010.V12.0.

Andrew M. Ryan, Zoey Chopra, David J. Meyers, Erin C. Fuse Brown, Roslyn C. Murray, and Travis C. Williams. Favorable selection in medicare advantage is linked to inflated benchmarks and billions in overpayments to plans: Study examines medicare advantage favorable selection, benchmarks, and payments to plans. *Health Affairs*, 42(9):1190–1197, 2023.

J. M. C. Santos Silva and Silvana Tenreyro. The Log of Gravity. *The Review of Economics and Statistics*, 88(4):641–658, November 2006. URL https://ideas.repec.org/a/tpr/restat/v88y2006i4p641-658.html.

David Silver and Jonathan Zhang. Impacts of basic income on health and economic well-being: Evidence from the va's disability compensation program. Working Paper 29877, National Bureau of Economic Research, March 2022. URL http://www.nber.org/papers/w29877.

Jonathan Skinner. Causes and consequences of regional variations in health care. In *Handbook of health economics*, volume 2, pages 45–93. Elsevier, 2011.

Jonathan Skinner and Douglas Staiger. Technology diffusion and productivity growth in health care. *Review of Economics and Statistics*, 97(5):951–964, 2015.

Lucy Skinner, Douglas O. Staiger, David I. Auerbach, and Peter I. Buerhaus. Implications of an aging rural physician workforce. *New England Journal of Medicine*, 381(4):299–301, 2019.

Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations, volume 1*, volume 1. Oxford: Clarendon Press, 1776.

Statista Research Department. U.s. gross domestic product: Forecast 2021-2032, 2022. URL https://www.statista.com/statistics/216985/forecast-of-us-gross-domestic-product/. Accessed: 2023-05-25.

Robert Town and Gregory Vistnes. Hospital competition in hmo networks. *Journal of Health Economics*, 20(5):733–753, 2001.

Jane M. Zhu, Mark Katz Meiselbach, Coleman Drake, and Daniel Polsky. Psychiatrist networks in medicare advantage plans are substantially narrower than in medicaid and aca markets. *Health Affairs*, 42(7):909–918, 2023. doi:10.1377/hlthaff.2022.01547. URL `https://doi.org/10.1377/hlthaff.2022.01547`. PMID: 37406238.

Anna Zink and Sherri Rose. Fair Regression for Health Care Spending. *Biometrics*, 76(3): 973–982, 2020.

Stephen Zuckerman, Laura Skopec, and Joshua Aarons. Medicaid physician fees remained substantially below fees paid by medicare in 2019. *Health Affairs*, 40(2):343–348, 2021.