Market Size and Trade in Medical Services*

Jonathan I. Dingel   Joshua D. Gottlieb
Chicago Booth,       University of Chicago
CEPR, and NBER      and NBER

Maya Lozinski   Pauline Mourot
University of Chicago   Chicago Booth

July 2022

Abstract

We quantify the roles of increasing returns and trade costs in medical services. Using data on millions of Medicare claims, we document 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. Larger markets specialize in the production of less common procedures, and these procedures are more traded between regions. These patterns reflect economies of scale: larger regions produce higher-quality care because they serve more patients. Revealed-preference estimates of quality, which are positively related to external measures of quality of care, have a scale elasticity around 0.7. We use these estimates to evaluate the proximity-concentration tradeoff associated with various policy options for improving access to medical care.

Keywords: healthcare access, market-size effects, Medicare claims data, trade in services

JEL codes: F14, I11, R12

*We thank Cedric Elkouh, Feng Lin, Noah Sobel-Lewin, and especially Luke Motley and Vaidehi Parameswaran for excellent research assistance and seminar participants for helpful feedback. We thank the Becker Friedman Institute for support and enabling us to access the Medicare claims data. Dingel thanks the Cohen and Keenoy Faculty Research Fund at the University of Chicago Booth School of Business for support. Disclosure: Gottlieb has first-hand exposure to examples discussed in this paper. His family members have worked in healthcare in Baltimore. Healthcare exported from Baltimore saved a different family member’s eye. jdingel@chicagobooth.edu, jgottlieb@uchicago.edu, mayalozinski@uchicago.edu, pmourot@chicagobooth.edu
Rural Americans have worse health outcomes (Deryugina and Molitor, 2021; Finkelstein, Gentzkow, and Williams, 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” indeed go back decades (Newhouse et al., 1982a; Skinner et al., 2019). Properly evaluating this concern requires us to consider two economic mechanisms: economies of scale and trade costs. We estimate them and find that both interregional trade and scale economies play key roles in understanding spatial patterns of healthcare within the United States.

When medical services exhibit increasing returns to scale, there are benefits to geographically concentrating doctors. Indeed, physician specialization has long been suggested as a case in which division of labor is limited by the extent of the market (Baumgardner, 1988). But this concentration is less feasible if healthcare markets are geographically isolated. If medical care must be provided in patients’ home markets, the only way to serve patients in smaller markets is to ensure there are enough local doctors—even if this requires foregoing the benefits of scale.¹ For a patient suffering from a time-sensitive emergency, such as a stroke, severe trauma, or heart attack, this assumption is plausible.² But the overwhelming majority of medical spending is not for such emergencies.

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 geography of medical care may resemble other, tradable industries. This would imply a proximity-concentration trade-off: patients who import medical

---

¹ Many economists assume trade costs for these 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.”

² But Fischer, Royer, and White (2022) and Battaglia (2022) find that, even for care as time-sensitive as childbirth, the benefits of a higher-quality hospital outweigh the costs of travel.
services produced elsewhere incur trade costs but benefit from economies of scale.

We examine and quantify the roles of increasing returns and trade costs in medical services, two economic mechanisms essential to any analysis of geographic maldistribution. Using data on millions of Medicare claims, we show 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. Data on trade flows show that the exports come disproportionately from large markets, pointing towards scale economies in production.

The Mayo Clinic in Rochester, Minn. is a well-known example of the phenomenon we study. The Mayo Clinic is ranked first in the U.S. News Best Hospitals list, and three-quarters of services produced in the Rochester metropolitan area are exported to other regions. Patients from other metropolitan areas travel an average of 540 km to get care in Rochester. But, as a major healthcare exporter with a population of merely 220,000, Rochester is an outlier: people tend to travel farther for care from large regions. The average patient travels 515 km to Chicago and 620 km to New York City, compared with 140 km to Urbana-Champaign, Ill. or Charlottesville, Va. Other than Rochester, large metropolitan areas dominate the U.S. News rankings, and we find that they attract a disproportionate share of patients seeking care. Our theoretical framework links these patterns, as economies of scale in production interact with patients’ costs of traveling for care.

The geographic scope of the market for a medical procedure depends on its national scale: doctors performing rare procedures export their services across a broader geographic scope, sometimes serving patients who reside thousands of kilometers away. Rarer procedures are disproportionately produced and exported by large markets. For example, half of the patients in whom cardiothoracic surgeons implant left ventricular assist devices (LVADs) to restore heart function in patients with congestive heart failure come from outside the surgeon’s region. By contrast, only 18 percent of colonoscopies are performed on patients outside their home region. We show that this pattern of specialization and trade reflects
economies of scale: serving a larger volume of patients reduces the quality-adjusted cost of producing a service, which causes larger markets to export medical services.

Local increasing returns imply an important tradeoff for healthcare policies concerning the geographic distribution of care. If total physician supply is fixed—as it may be in the United States due to entry restrictions limiting physician residencies—relocating production to rural or “under-served” places reduces the agglomeration benefits attained by co-locating physicians in larger markets. Policies that aim to equalize “access to care” must grapple with the trade-off between fragmenting the production of medical services and reducing distances between producers and patients, as in Glaeser and Gottlieb (2008). Travel subsidies may be more effective than subsidizing rural providers in many contexts. But if patients differ in their ability or willingness to travel for higher-quality services, the most efficient solution to this tradeoff may not be the most equitable. More broadly, the large share of healthcare that is imported from other regions has implications for how researchers measure disparities in access, define the relevant market when measuring concentration, and estimate the contribution of place-based factors to health outcomes.

Section 1 develops a model of trade in medical services to guide our empirical analysis of US healthcare. We adapt standard models of agglomeration and trade to a setting in which the government sets prices, so quality and travel patterns clear markets. Our model demonstrates that local increasing returns can generate a home-market effect even under fixed prices. When local increasing returns are sufficiently strong relative to market size, the model predicts that larger markets will have an endogenous comparative advantage and be net exporters of medical services. With sufficiently strong scale economies, healthcare can serve as an export base for these urban economies. If average cost declines faster at lower quantities, the market-size effect on exports is larger for less common procedures, a prediction we investigate using a difference-in-differences design.

Section 2 describes the Medicare claims data we use to study the spatial pattern of production and trade within the United States. Medicare, the federal government’s insurance
program for the elderly and disabled and the largest insurer in the United States, pays bills submitted by medical service providers. These claims distinguish between more than 10,000 distinct medical procedures and report the place of service and where the patient lives. We use a random 20-percent sample of all physician claims paid by traditional (fee-for-service) Medicare in 2017 to construct procedure-level trade matrices for medical services.

We begin our empirical investigation in Section 3 by examining how production and consumption vary with market size. Production is geographically concentrated in larger markets, while consumption is much less correlated with population size. 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 in medical services, we estimate a gravity model of bilateral gross trade flows, similar to 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 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 4 investigates the model’s predictions for common and rare services. 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. The estimated home-market effect is substantially stronger for less common procedures: a larger residential population drives a greater increase in exports for rarer services. In our theoretical framework, this occurs because less common procedures are on the steeper part of the local-increasing-returns production function.

We investigate the quality of care traded across regions in Section 5. By estimating patients’ willingness to travel to each exporting region for medical services, we recover revealed-preference estimates of service quality. We show that these revealed-preference estimates are positively related to external quality measures, such as U.S. News hospital rankings. Esti-
mated quality is increasing in market size, with an elasticity around 0.7. These results suggest that the benefits of market size manifest in higher-quality services, and trade expands access to this higher-quality care.

Local increasing returns could reflect a variety of mechanisms (Marshall, 1890), and we consider one of the most classic: physicians' increased ability to specialize in larger markets. We use specialization to illustrate how patient travel opens up markets for better care: imports are more likely to be provided by a specialist, and by the appropriate specialist, than locally produced services. With specialists disproportionately located in larger markets, the opportunity to specialize is a likely source of local increasing returns in medicine and higher-quality care in larger markets.

Section 6 uses our estimates to explore quantitatively the proximity-concentration trade-off that our results suggest. We consider two counterfactual policies to improve access to care in smaller markets. First, we quantify the marginal benefit of reallocating care across hospital referral regions, accounting for the geographic distribution of patients served. This reallocation worsens access for most patients, as weakening agglomeration reduces the quality of care available to those in more populous regions. Second, we consider the alternative policy of subsidizing patient travel to reduce spatial inequality in consumption of medical services. It is essential to account for trade and increasing returns when predicting the impacts of these policy alternatives.

The higher-quality care available in larger markets may not benefit all patients equally. We next show that patients trade off between the benefits of geographically concentrated production and incurring travel costs differently depending on socioeconomic status. Splitting patients into deciles based on the median household income in their residential ZIP code, we estimate a separate gravity regression for each decile. We find that patients residing in lower-income neighborhoods are less likely to travel farther for medical care. This result 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
efficiency and quality gains due to local increasing returns do not benefit all patients equally.

This paper builds on research in urban, trade, and health economics. Urban economists have highlighted skill-biased agglomeration effects in production as knowledge workers have become increasingly important and concentrated in skilled cities (Berry and Glaeser, 2005; Moretti, 2011; Diamond, 2016; Davis and Dingel, 2020; Eckert, Ganapati, and Walsh, 2020). But connecting this to production of, and trade in, personal services has been more difficult. We show that—even in a service-based economy—the sizes of both local and potential export markets influence local production and quality.

The trade literature has examined market-size effects in manufacturing, but investigated services much less. For internationally traded goods, Davis and Weinstein (2003), Hanson and Xiang (2004), and Bartelme et al. (2019) link market size to export patterns, in line with the home-market effect laid out in Krugman (1980) and Helpman and Krugman (1985). For intranational trade in goods, Dingel (2017) shows that market-size effects account for a substantial share of observed quality specialization. For pharmaceuticals, market-size effects have been demonstrated using demographic variation over time (Acemoglu and Linn, 2004) and across countries (Costinot et al., 2019). Market-size effects in services have been far less investigated, likely in part because of the paucity of reliable data on trade in services (Lipsey, 2009; Muñoz, 2022). Medical services are distinctive among services because they could be traded, we observe the locations of both providers and patients, and they are reported in terms of a standard classification.

The importance of medical care for health, life expectancy, and welfare generate substantial public-policy interest. In rural locations, health outcomes are worse but there are fewer doctors per capita. An important series of papers by Newhouse et al. (1982a,b,c), Newhouse (1990) and Rosenthal, Zaslavsky, and Newhouse (2005) considered this issue and, for similar reasons to ours, argued that a uniform geographic distribution of physicians is the wrong benchmark. Building on these studies, we demonstrate that medical services are traded between regions and connect this trade to economies of scale. We observe the lo-
ocations to which patients import care, allowing us to estimate the impact of geography on patient access. Critically, we leverage modern trade theory, which guides our formal model, estimation strategy, and informs the analysis of counterfactual policies.

More generally, the healthcare literature has documented enormous variation in care provision across space (Fisher et al., 2003a,b), and shown that much of it emerges from the supply side rather than patient demand (Finkelstein, Gentzkow, and Williams, 2016). But determining the underlying reasons for this variation has proven challenging. Chandra and Staiger (2007, 2020) have shown that different levels and styles of production are efficient in different places. But what determines regional production styles or “medical culture” (Gawande, 2009; Cutler et al., 2019; O’Malley, Bubolz, and Skinner, 2021)? Our results show that market size and increasing returns play an important role.

1 Theoretical framework

This section develops a model of trade in medical services tailored to guide our empirical analysis of US healthcare. Our framework features a discrete-choice model of patients selecting quality-differentiated services in the presence of trade costs. Local 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.

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, which continues to deliver a gravity equation for trade flows
and to predict home-market effects. The model shows that price controls in an increasing-returns world can have spatially heterogeneous impacts on the quality of output and the value of consumers’ choice sets. This framework delivers testable predictions about spatial variation in services quality and trade patterns when prices are fixed.

1.1 Demand

We use a logit model of an individual choosing a provider for a given procedure. Patients and providers are in regions indexed by \( i \) or \( j \), with \( I \) denoting the set of regions. Let \( N_j \) denote the number of patients residing in \( j \) who make a choice. All providers in a region are identical.

A patient (indexed by \( k \)) choosing a provider in region \( i \) would obtain utility \( U_{i,k} \), which has a provider-region-specific component, a region-pair component, and an idiosyncratic component:

\[
U_{i,k} = \ln \delta_i + \ln \phi_{ij(k)} + \epsilon_{ik}.
\]

The provider-region-specific component \( \delta_i \) would usually include a product’s characteristics and price. In the Medicare context, as we discuss below, the government pays reimbursement rates that it sets administratively. So the \( \delta_i \) relevant for the patient is the quality of the providers in region \( i \). The region-pair component \( \phi_{i,j} \) represents (inverse) bilateral trade costs. The idiosyncratic component \( \epsilon_{ik} \) is independently and identically drawn from a Type 1 extreme value distribution, so the probability that patient \( k \) selects a provider in region \( i \) is

\[
\text{Pr}(U_{i,k} > U_{i',k} \forall i' \neq i) = \frac{\exp\left(\ln \delta_i + \ln \phi_{ij(k)}\right)}{\sum_{i' \in 0,i \notin I} \exp\left(\ln \delta_{i'} + \ln \phi_{i'j(k)}\right)}.
\]

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

\[^3\text{This formulation of demand is familiar from the hospital competition literature, which has studied competition among hospitals on price and quality. The literature tends to assume competition occurs within a specified geographic radius (e.g., Kessler and McClellan, 2000; Cooper et al., 2018), within a market defined}\]
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 \( Q_{0j} \) to denote the number of patients in region \( j \) selecting the outside option. Because each patient selects at most one provider, \( N_j = \sum_{i \in I \cup \{0\}} Q_{ij} \). The demand by patients residing in \( j \) for procedures performed by providers in \( i \) is

\[
Q_{ij} = \frac{\delta_i N_j}{\Phi_j} \phi_{i,j},
\]

where \( \Phi_j \equiv \sum_{i' \in 0 \cup I} \delta_{i'} \phi_{i'j} \) is the logit price index for patients in region \( j \). Equation (1) is a gravity equation with an origin \( i \) component, a destination \( j \) component, and an \( ij \) pair component. We refer to \( \Phi_j - 1 \), the value of the choice set excluding the outside option, as “patient market access.”

1.2 Production

Local increasing returns are crucial to our account of market size and trade in medical services. We assume competitive production of services with local external economics of scale and free entry. Each price-taking provider in region \( i \) chooses its output quality and quantity given total regional production, denoted \( Q_i \), exogenous factor price \( w_i \), and an exogenous productivity shifter \( A_i \).

In the peculiar institutional setting of US healthcare, prices are not an equilibrium object determined by the intersection of supply and demand. Instead, Medicare determines the “reimbursement rates” it pays based on a regulatory formula. It sets these prices regardless by metropolitan area or similar definition (e.g., Ho, 2009; Gowrisankaran, Nevo, and Town, 2015; Clemens and Gottlieb, 2017; Lewis and Pflum, 2017; Ho and Lee, 2019; Dafny, Ho, and Lee, 2019; Garthwaite, Ody, and Starc, 2022), or have data from certain states (e.g., Town and Vistnes, 2001; Gaynor and Vogt, 2003; Capps, Dranove, and Satterthwaite, 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, Kleiner, and Vogt, 2013; Dranove and Ody, 2016) and are natural if one assumes healthcare demand is local—as has been standard (see footnote 1). 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.
of quality or quantity, so we treat Medicare’s reimbursement rate as a fixed, exogenous value $\overline{R}$. While there is some geographic variation in reimbursements, it is not very large and we disregard it for the moment.

The profit-maximization problem for a provider in region $i$ involves choosing both $L$, the quantity of the composite input employed, and $\delta$, the output quality:

$$\max_{L, \delta} \overline{R} A_i \frac{H(Q_i)}{K(\delta)} L - w_i L.$$  

We assume producing a higher-quality service is more costly, so $K(\delta)$ is increasing. The local-increasing-returns function $H(Q_i)$ depends on total regional production, $Q_i$, which competitive firms take as given when choosing their own output levels (Chipman, 1970). We assume that $H(Q_i)$ is an increasing, concave function. The special case of local constant returns to scale is $\frac{dH}{dQ_i} = 0$. The idiosyncratic productivity shifter $A_i$ captures any other influences on the region’s productivity, such as historical investments.

Provider optimization and free entry imply a regional average cost function that we denote $C(Q_i, \delta_i; w_i, A_i)$. Provider size $L$ is indeterminate (and unimportant) given the assumed linear production function and price-taking behavior. The first-order and free-entry conditions mean every provider in region $i$ chooses the unique quality $\delta_i$ that satisfies

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

(2)

Given the factor price and productivity shifter, $C(Q_i, \delta_i; w_i, A_i) = \overline{R}$ 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 $i$. Along an isocost curve, quality varies with quantity according to $\frac{d\delta}{dQ} = -\frac{\partial C}{\partial Q} / \frac{\partial C}{\partial \delta}$. Local increasing returns make the isocost curve upward-sloping in $(Q, \delta)$ space. With free entry and fixed prices, the benefits of scale are realized as higher-quality services in regions that
produce greater quantities. If $H(Q_i)$ is concave and $K(\delta)$ is convex, the upward-sloping isocost curve is concave.

While our assumptions thus far suffice for some qualitative results, we later assume functional forms when estimating quantitative magnitudes. In this case, we specify that $K(\delta_i) = \delta_i$ and $H(Q_i) = Q_i^\alpha$, where the scale elasticity $\alpha$ is between zero and one. This particular case of the free-entry condition (2) is

$$\overline{R} = \frac{w_i \delta_i}{A_i Q_i^\alpha}. \tag{3}$$

### 1.3 Equilibrium

Equilibrium equates supply and demand in each region, $Q_i = \sum_j Q_{ij}$. Using the demand system given by equation (1), this can be written as

$$Q_i = \delta_i \sum_j \frac{N_j}{\Phi_j} \phi_{ij} \tag{4}$$

Given exogenous parameters $\overline{R}$, $\{w_i, A_i, N_i\}_{i \in I}$, and $\{\phi_{ij}\}_{(i,j) \in (I,I)}$, an equilibrium is a set of quantities and qualities $\{Q_i, \delta_i\}_{i \in I}$ that simultaneously satisfy equations (2) and (4).

### 1.4 Scale effects in autarky

We first consider equilibrium in autarky: patients can choose whether to receive care, but they cannot travel between regions ($\phi_{ij} = 0$ for $i \not\in \{0, j\}$). In an autarkic equilibrium for region $j$, the equilibrium quantity $Q_{jj}$ and $\delta_j$ produced must be such that $N_j = Q_{0j} + Q_{jj}$. Per equation (4), the equilibrium quantity and quality must satisfy

$$Q_{jj} = \frac{\delta_j \phi_{jj}}{1 + \delta_j \phi_{jj}} N_j.$$

They must also satisfy the free-entry condition (2).
Panel (a) of Figure 1 illustrates the autarkic equilibrium when production exhibits increasing returns for two regions, one small \((i = S)\) and one big \((i = B)\). We assume they have the same factor prices, \(w_S = w_B\), and productivity shifters, \(A_S = A_B\), so their isocost curves are the same. As noted above, the isocost curve is upward-sloping when there are increasing returns. Demand is also upward-sloping, since \(Q_{jj}\) is increasing in \(\delta_j\). An equilibrium is given by the intersection of these two curves.\(^4\) In the presence of local increasing returns, the more populous region produces a higher-quality service in the autarkic equilibrium. As a result, consumption (and production) per capita is higher in the larger region. If there were constant returns to scale, \(\frac{\partial C}{\partial Q} = 0\), then the isocost curve would be a horizontal line and equilibrium quality would be independent of quantity and solely determined by the reimbursement rate \(\bar{R}\). Absent increasing returns, output per capita would be the same in the two regions.

1.5 Home-market effect when two regions trade

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

Panel (b) of Figure 1 illustrates the equilibrium pattern of trade between two regions that differ in size. The autarkic differences in quality associated with regional size are amplified by trade. The larger region’s higher-quality service attracts patients from the smaller region, shifting its residual demand curve to the right. This elicits an increase in its \(^5\)The logit demand curve in equation (1) runs through the origin. For visual clarity, Figure 1 uses upward-sloping demand curves with non-zero intercepts.

\(^5\)For the equilibrium to be stable, the supply curve cannot exhibit too strong increasing returns at the equilibrium quantity-quality combination. At this point, the demand curve must be steeper than the isocost curve. Otherwise, there exists a nearby quantity-quality combination where demand intersects an isocost curve with unit cost strictly less than \(\bar{R}\).
quantity produced, improving its quality further. Symmetrically, the smaller region suffers lower demand when patients travel to the larger region for care, and the smaller quantity of production deteriorates the quality of its output. This is the home-market effect: differences in the scale of demand interact with scale economies to make the larger region a net exporter.

Panel (b) of Figure 1 depicts demand conditions in which the equilibrium involves positive production in both regions, but the smaller region could cease production entirely. With lower trade costs, or smaller population size, there could be no values of \( \delta_S \) in which demand is large enough to cover the costs of production. This would show up in the graph as a residual demand curve that does not intersect the isocost curve, as in Panel (c).

Next, we contrast equilibrium outcomes for common and rare medical procedures. Panel (c) of Figure 1 shows the outcomes in autarky. We assume that both procedures have identical isocost curves, but the common procedure is in high demand in both regions, while the rare procedure has much lower demand. Because \( H(Q_i) \) is concave and \( K(\delta) \) is convex, the isocost curve is flatter at higher quantities. The small and large regions thus produce comparable qualities of the common procedure in autarky. If we allowed trade between the two regions, this similarity implies mean that there would be little motivation for trade based on regional variation in quality. By contrast, the rare procedure is not produced at all in the smaller region in autarky, yielding substantial quality variation.

Panel (d) of Figure 1 introduces trade for both of these procedures. Residual demand for the rare procedure increases dramatically in the larger region, since the smaller-region residents can now import their services. Residual demand for the common procedure is less affected, since the very modest differences in quality mean that trade is motivated by patients’ idiosyncratic preferences. When returns to scale are decreasing in quantity, we predict a larger effect of market size on trade flows for less common procedures.

While we have illustrated the economic logic using two regions, our empirical investigation will examine Medicare-covered services throughout the United States. We briefly discuss the many-region case before turning to the data.
1.6 Home-market effect with many regions

We now examine the many-region equilibrium in the special case of $H(Q_i) = Q_i^\alpha$ and $K(\delta_i) = \delta_i$. Abusing notation so that $I$ is both the set and number of regions, equations (2) and (4) together constitute $2I$ equations with $2I$ unknowns. For the special case, this reduces to the following $I$ equations with the unknowns $\{\delta_i\}_{i=1}^I$:

$$\delta_i = \left(\frac{RA_i}{w_i}\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in I} \frac{\phi_{ij}}{\sum_{i' \in \emptyset \cup I} \delta_{i'}\phi_{i'j}} N_j\right)^{\frac{\alpha}{1-\alpha}}$$

Following Costinot et al. (2019), we examine the home-market effect in the neighborhood of a symmetric equilibrium. Suppose all regions are the same size, $N_i = \bar{N} \forall i$, and trade costs are symmetric: $\phi_{ii} = 1$ and $\phi_{ij} = \phi \in (0, 1) \forall i \neq 0, j$. For brevity, we assume that $\frac{RA_i}{w_i} = 1 \forall i$. To examine the effect of market size, we totally differentiate the above system of equations in terms of $\{d\delta_i, dN_i\}_{i=1}^I$, assume $dN_1 > 0$ and $dN_j = 0 \forall j \neq 1$, assume there are increasing returns ($\alpha > 0$), and evaluate this system at the symmetric equilibrium. After tedious algebra (see Appendix A), one obtains the following expression for quality changes, where $\bar{\delta}$ and $\bar{\Phi}$ denote quality and the logit price index in the symmetric equilibrium:

$$d \ln \delta_1 - d \ln \delta_{j\neq1} = \left[\frac{1 - \alpha}{\alpha} \frac{(\bar{\Phi} - 1)}{(1 - \phi)\bar{\delta}} + \frac{(1 - \phi)\bar{\delta}}{\bar{\Phi}}\right]^{-1} d \ln N_1 > 0.$$  

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. This quality increase causes an increase in its exports to every other region:

$$\frac{d \ln Q_{1j}}{d \ln N_1} = \left(\frac{N-Q_{1j}}{N}\right) \left[\frac{d \ln \delta_1}{d \ln N_1} - \frac{d \ln \delta_j}{d \ln N_1}\right] + \frac{Q_{1j}}{N} \frac{d \ln \delta_j}{d \ln N_1} > 0.$$  

This consequence of greater demand in the presence of local increasing returns is the weak home-market effect in a fixed-price environment. Of course, the increase in region 1’s population size also increases its demand for imports. The effect on the region’s net exports is

$$d \ln Q_{1,j\neq1} - d \ln Q_{j\neq1,1} = \frac{\alpha}{N} - (1 - \alpha) \frac{1+(\bar{I}-1)\phi}{1-\phi} \frac{d \ln \delta_j}{d \ln N_1} + \frac{Q_{1j}}{N} \frac{d \ln \delta_j}{d \ln N_1}.$$  

14
The larger population size of region 1 makes it a net exporter of the medical procedure if increasing returns are sufficiently strong ($\alpha$ is large enough) and initial quality is low enough ($\bar{N}$ is small enough). That is, the effect on the region’s net exports is positive if (and only if)

$$\frac{\alpha}{1 - \alpha} > \frac{1 + (I - 1)\phi}{1 - \phi} \bar{N}.$$

When this inequality holds, the procedure has a strong home-market effect. When it fails, there is a weak home-market effect but not a strong one. This, while the existence of some agglomeration economies 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, there is a natural case that healthcare can support their economies; rather than exporting manufactured goods, as in decades past, larger cities can reinvent themselves (Glaeser, 2005) and export medical services. In the absence of a strong effect, the larger populations would demand more imports than they produce, and healthcare would have a harder time acting as an economic base.

Given a scale elasticity $\alpha$, this effect is diminishing in $\bar{N}$. Thus, even if all procedures have the same scale elasticity $\alpha$, rare treatments could exhibit a strong home-market effect while common procedures may not. If rare procedures also have stronger economies of scale—for example, because they are still experiencing significant learning-by-doing—that would amplify their stronger market-size effect. This result motivates a difference-in-differences research design: we compare the market-size effects of common and rare procedures.

2 Data description

Our primary dataset is 2017 claims data from the Medicare insurance program in the United States. Medicare is the federal government’s insurance program for the elderly and disabled and is the largest insurer in the U.S. It does not directly employ physicians or run its own
hospitals, but instead pays bills submitted by independent physicians, physician groups, hospitals, and other medical service providers. These bills—called “claims” in industry terminology—report the specific service provided using a 5-digit code from the Healthcare Common Procedure Coding System (HCPCS), along with other detailed information about the visit. There are over 10,000 distinct HCPCS codes, which identify individual procedures at a granular level. The payment amount for each claim is determined by federal regulation, so it does not reflect any pricing decision by the physician or hospital. Alternative analyses use groupings of patient diagnoses to account for potential substitution between treatments.

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 use place-of-service information to exclude care provided in an Emergency Department. Because Medicare rarely reimbursed telehealth in 2017, this trade involves traveling to receive a service delivered in-person. 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 medical care based on 1992–93 data. We construct a bilateral trade matrix for HRRs by interpreting the patient’s residential HRR as the importing region and the place of service’s HRR as the exporting region. Since HRRs group together patient ZIP codes based on where major cardiovascular procedures are performed and are defined such that each HRR contains at least one hospital performing major cardiovascular procedures and neurosurgery, the construction of these geographic units should tend to minimize trade between different HRRs.

---

6For instance, an office visit of medium complexity (generally 20–29 minutes) is distinct from one of higher or lower complexity. Office visits with new patients are distinct from those with established patients, and nursing facility or emergency department visits are distinct from those in office. 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 it is intramuscular or intranasal.

7We 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 488 clinical categories present in our data, which we split at the median frequency to separate common from rare diagnoses.

8In 2012, Medicare spent only $5 million — less than 0.001% of its expenditures — on telehealth services (Neufeld and Doarn, 2015). Medicare is laggard in this regard (Dorsey and Topol, 2016).

9We investigate our results’ robustness to the use of alternative geographies, including core-based sta-
Physicians, hospitals, pharmacies, and other healthcare providers submit different types of claims. We use a random 20 percent sample of all physician claims paid by Traditional (fee-for-service) Medicare in 2017.\textsuperscript{10} Medicare only provides a 20 percent sample because of the size of the file; even our one-year sample contains 229 million services, representing $19 billion in spending. The sampling is randomized by patient, making it a representative sample of Medicare activities in all relevant dimensions. 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 same features of healthcare production and consumption as the rest of the economy.

The fact that we only see a sample of Medicare data (and hence, an even smaller share of overall medical care) raises the possibility of not observing the existence of physicians or procedures that are so rare that a 20 percent 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 (based on a 100 percent sample) to replicate some of our analyses and obtain extremely similar findings.\textsuperscript{11} Second, we use physician registry data to study the geographic patterns of production by specialty. These data describe 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 even a one-to-many mapping of specialties to services, since many services can be provided by statistical areas (CBSAs) and metropolitan statistical areas, a subset of CBSAs that eliminates the smaller micropolitan areas. (Not all results reported.)

\textsuperscript{10}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.\textsuperscript{11}Medicare provides summary data from the complete (100 percent) records 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. 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.
physicians of different specialties—but we expect many of the same economic forces to apply at the level of physician specialties.

3 Is there a home market effect in medical services?

This section describes our empirical strategy for estimating the roles of economies of scale and trade costs in the production and consumption of medical services. Section 3.1 documents size-related spatial variation in production and consumption of procedures. Section 3.2 shows that bilateral trade declines with distance. Section 3.3 describes our empirical strategy, which identifies the consequences of market size using gravity equations to model bilateral trade flows of medical services. Section 3.4 reports the empirical estimates, which demonstrate a strong home-market effect.

3.1 Spatial variation in production and consumption

Figure 2 shows maps of healthcare production and consumption across space. Panel (a) shows production and Panel (b) shows consumption of medical care per resident Medicare beneficiary. The consumption map shows patterns that have been well-documented by the Dartmouth Atlas and related literature on geographic variation in healthcare (Fisher et al., 2003a,b; Finkelstein, Gentzkow, and Williams, 2016). The production map shows even more pronounced variation: more production in large urban agglomerations and less in rural places. There is substantial variation in production even between neighboring regions, while consumption exhibits smoother spatial variation.

The subsequent panels show patterns of trade, which constitute the difference between production and consumption. Panel (c) shows gross exports as a share of local production, and Panel (d) indicates whether each HRR is a net importer or net exporter. We see exports from major urban agglomerations, plus places like Rochester, Minn. and Hanover, N.H. that specialize in healthcare. Nationally, 22% of all production is exported to a patient in another
HRR. For manufactured goods, the export share across CBSAs is 68%.

Figure 3 plots the average aggregate production and consumption per capita across HRRs of different sizes. Both rise monotonically with population, but production rises about twice as steeply, with population elasticities of 0.06 for consumption and 0.13 for production. The difference between production and consumption is net trade: larger markets are net exporters and smaller markets are net importers. Gross trade flows are larger than net trade flows, with imports comprising about one-third of consumption in the smallest regions. Exports per capita are approximately flat with respect to population, which means total exports are increasing with local population. By contrast, imports per capita decline with an elasticity of $-0.24$ with respect to population.

### 3.2 Bilateral trade and bilateral distance

Despite the clear patterns in Figure 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), Hinsdale, Ill. (just west of Chicago), and Royal Oak, Mich. (just north of Detroit). Other than Rochester—home to the Mayo Clinic, which serves patients from across the globe—these exporting regions are all in major metropolitan areas. We thus turn to examining bilateral trade flows.

Figure 4 depicts how trade varies with the distance between the patient and place of service. Panel (a) shows the distribution of distances patients travel for care, distinguishing between places of service in the patient’s home region and other regions.$^{12}$ Within HRRs, we see a narrow distribution of distances that peaks around 10 km. When traveling to

$^{12}$For distances within a hospital referral region, these kernel densities are based on the distance between the centroid of the patient’s residential ZIP code and the ZIP code of the provider. We use the Census Bureau’s corresponding ZIP code tabulation area (ZCTA) to obtain geographic centroids. 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.
providers in a different HRR, patients travel a much greater 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 distances of thousands of kilometers, evidence of substantial long-distance travel for care. Patients’ willingness to travel these distance will inform our revealed-preference estimates of regional service quality.

Panel (b) of Figure 4 suggests that these patterns follow a gravity relationship. The blue series depicts the volume of trade against distance (for pairs of HRRs with positive trade flows) after removing fixed effects for each exporter and each importer. The relationship is nearly log-linear for short distances, but flattens out for longer distances. The red series captures the extensive margin by showing the share of positive-trade pairs as a function of distance. This share is 100 percent for short distances and about 60 percent for the most distant HRR pairs. Together, these patterns motivate the inclusion of distance covariates $X_{ij}$ in the subsequent gravity-based analysis.

### 3.3 Gravity-based empirical strategy

The heart of our empirical investigation examines trade flows. While we focus on population-driven differences in demand, population is not a sufficient statistic for a region’s demand structure. A second key consideration is geography. The spatial distribution of population—whether due to history, amenities, broader economic productivity, or housing costs—determines the number of potential patients located nearby. It is important to control for each region’s proximity to major population centers when estimating home-market effects and inferring their quality of care. We do so by estimating a gravity model.

Adding a stochastic error to the gravity equation (1), taking expectations and then logs,

---

13This application of the Frisch-Waugh-Lovell theorem is only feasible for positive trade volumes.
we obtain the following equation for gross bilateral trade flows:

$$\ln E(\overline{RQ}_{ij}) = \ln \delta_i + \ln \left( \frac{N_j}{\Phi_j} \right) + \gamma X_{ij}.$$  \hspace{1cm} (5)

The outcome variable on the left side of equation (5) is the value of procedures exported from region $i$ to patients residing in $j$. We variously specify the first two right-side regressors as observable demand shifters or fixed effects in different regression specifications described below. $X_{ij}$ is a vector of observed trade-cost shifters, typically log distance and a same-region dummy so that $\gamma X_{ij} = \gamma_1 \ln \text{distance}_{ij} + \gamma_0 1(i = j)$. In other specifications, we add the square of the log distance or replace these continuous distance covariates by dummies indicating distance deciles.

When estimating equation (5) using the total value of bilateral exports as the dependent variable, we aggregate quantities across thousands of distinct medical procedures using the average national Medicare reimbursement rate for each procedure. This produces an expenditure measure that is independent of any spatial variation in reimbursement rates.\footnote{Mechanically, we multiply the quantity of each procedure by the national average price for that procedure and denote the sum across all procedures by $\overline{RQ}_{ij}$.}

We also estimate procedure-level versions of equation (5) for selected procedures, such as LVAD insertion and colonoscopy. In these cases, the dependent variable is simply the procedure count and no aggregation scheme is required. Since observed bilateral trade is zero for many pairs of regions, especially when looking at trade in individual procedures, we estimate equation (5) using the Poisson pseudo-maximum-likelihood (PPML) estimator (Silva and Tenreyro, 2006).

We test for a home-market effect in medical services using market size as an observed demand shifter. Following Costinot et al. (2019), log-linearizing the system of equations (2) and (4) around the symmetric equilibrium yields the local relationship between trade flows and pop-
ulation, allowing us to omit the logit price index, $\Phi_j$. The estimating equation is

$$\ln \mathbb{E}[\overline{RQ}_{ij}] = \lambda_X \ln \text{population}_i + \lambda_M \ln \text{population}_j + \gamma X_{ij}. \tag{6}$$

Relative to equation (5), the right side of equation (6) replaces $\ln \delta_i$ and $\ln \left( \frac{N_i}{\Phi_j} \right)$ by log population in the producing and consuming regions, respectively. As defined in Costinot et al. (2019), a positive coefficient $\lambda_X > 0$ implies a weak home-market effect: gross exports increase with market size. If $\lambda_X > \lambda_M > 0$, there is a strong home-market effect: net exports increase with market size.

One potential concern with estimating equation (6) directly is reverse causality. Suppose current population size is driven by universities and hospitals that emerged as local “anchor institutions.” Historical accident could have led to the formation of leading medical schools and hospitals that persist until the present. Such an accident may be responsible for Johns Hopkins University: The young Johns Hopkins moved to Baltimore to work at his uncle Gerard’s wholesale grocery business (Jacob, 1974). After making his own fortune in trade and railroads, Johns Hopkins pledged his estate to found a research university and hospital, modeled on the leading 19th century German universities. The fact that uncle Gerard lived in Baltimore is likely a major reason Baltimore today exports 16 percent of its healthcare output. If entities such as Hopkins became anchor institutions that support greater contemporary population size, one might fear that our identifying assumption fails: medical exports and contemporary population would both be correlated with this historical factor.

We use two instrumental variables (IVs) 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 perhaps it could today. Since population is persistent over time, 1940 population does predict 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. Levy and Moscona (2020) show that shallower subterranean bedrock makes construction easier, which leads to higher population density. Bedrock depth also predicts population size, so we use this as a second instrument for local demand, again for the importing and exporting regions respectively.\textsuperscript{15}

### 3.4 Results: a strong home-market effect in medical services

Table 1 reports the results of estimating equation (6). The first column shows significant, positive coefficients on both patient and provider market population. The coefficient for provider market population size is two-thirds greater than that for patient market population size. This demonstrates what Costinot et al. (2019) term a \textit{strong home-market effect}. Not only does a 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 -1.6. This is substantially larger than the distance elasticity of -0.95 estimated for trade in manufactures between CBSAs (Dingel, 2017).\textsuperscript{16} 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 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 all of the columns: gross and net exports both increase with market size. The magnitudes are stable in columns 2 and 3, and the magnitude

\textsuperscript{15}At present, this instrument is only available for core-based statistical areas (CBSAs) and not for our main geographic unit, hospital referral regions. We demonstrate that our main results are robust to the use of CBSAs and to the use of both instruments at this geographic unit. Appendix Figure B.1 shows that the first stage is extremely powerful.

\textsuperscript{16}The distance elasticity of medical services trade between core-based statistical areas is -2.2.
of gross (though not net) exports increases when excluding zeros.

The last column of Table 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 B.1 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.

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 cost reduction in larger markets. While doctors do earn less in urban locations (Gottlieb et al., 2020), they are only one part of input costs. 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 nurses’ earnings and real estate costs.\(^{17}\) We confirm that doctors are cheaper in larger markets, but other costs rise with population size. Appendix Figure B.2 shows these relationships. The elasticity of doctors’ earnings with respect to population is -0.01, but that for non-physicians is 0.054. 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 would reinforce these higher labor costs. This spatial variation in costs undercuts the idea that amenities make production cheaper in larger markets.

One final concern is “snowbird” patients who may appear to travel farther than they actually do, as they may need medical care while spending months in a warmer HRR from

\(^{17}\)The Gottlieb et al. (2020) earnings data are only available for 111 commuting zones. Appendix Figure B.3 shows analogous results for many CBSAs, estimated using the American Community Survey (ACS), but this source top-codes income for a substantial share of doctors.
the one listed as their main residence (or the reverse). To demonstrate that our results are not driven by snowbirds, we follow Finkelstein, Gentzkow, and Williams (2016) by re-running our estimates excluding Florida, California, and Arizona.\footnote{These results are available upon request.}

4 Comparing rare and common services

Our theoretical framework predicts larger home-market effects for rare procedures. When returns to scale are decreasing in quantity, the average cost of rare procedures is more sensitive to market size than the average cost of common procedures, and hence the same is true for the quality and quantity of output. Comparing the market-size effects for common and rare procedures is thus a finer test of our theory. Section 4.1 examines how spatial variation in the production and consumption of each procedure relates to market size. Section 4.2 generalizes our gravity-based regression analysis to estimate home-market effects separately for rare and common procedures.

4.1 Spatial variation in production and consumption by frequency

We estimate the population elasticity of production and consumption per Medicare beneficiary for each procedure.\footnote{Davis and Dingel (2020) describe how population elasticities relate to other measures of geographic concentration, such as location quotients, and estimate population elasticities of employment for various skills and sectors.} We find that production rises with market size more than consumption, especially for less common procedures.

4.1.1 Method

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

$$\ln \mathbb{E}\left[ \frac{Q_{pi}}{M_i} \right] = \zeta_p + \beta_p \ln \text{population}_i. \quad (7)$$

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.\footnote{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.}

If the quantity produced were simply proportional to population, $\beta_p$ would be zero.

Our model suggests that scale economies play a larger role for rarer procedures. Following the logic in Panel (c) of Figure 1, we expect less common services to 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:

$$\ln \mathbb{E}\left[ \frac{G_{pi}}{M_i} \right] = \zeta_p^C + \beta_p^C \ln \text{population}_i. \quad (8)$$

If $\beta_p^C \neq \beta_p$, there is net trade in procedure $p$. Our model predicts that procedure frequency influences the pattern of trade, a prediction we test in Section 4.2.

\section*{4.1.2 Results}

Production per beneficiary rises with market size, especially for less common procedures. Panel (a) of Figure 5 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 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 (8), 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. Panel (a) of Figure 5 also plots the population elasticity of consumption per beneficiary $\hat{\beta}^C_p$ for each procedure against its national volume $\ln Q_p$. While this relationship is negative, the slope for consumption is only one third that for production. This difference means that larger markets have higher net exports in less common procedures.

We have thus far modeled patients as having demand for a particular treatment, and providers producing specific service codes. An alternative approach is to think of patients having 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 in different regions might use different treatments for the same care, our estimates thus far could be contaminated by substitution among procedures. We address this by conducting a similar analysis at the level of clinical condition.

Panel (b) shows production and consumption elasticities by diagnosis, rather than by procedure. The key patterns remain similar: production elasticities are higher than consumption, and are more negatively sloped with respect to national patient volume. Both consumption and production elasticities have less steep relationships with national volume than for procedures. This could reflect increased measurement error: the 482 diagnosis categories we use are far coarser than the 8,253 procedures in Panel (a). 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 5 imply trade in medical services between markets of different sizes. In particular, 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. But the population elasticities of production and consumption can only reveal net trade, understating the role of imported consumption to the degree that gross trade exceeds net trade.

4.2 Market-size effects are stronger for less common procedures

We now examine bilateral trade in procedures, emphasizing how trade varies with a procedure’s frequency. Figure 6 Panel (a) shows the distribution of imports as a share of consumption by procedure. This kernel density plot exhibits a spike at just above 20 percent, 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. To start, we divide procedures into two groups, common and rare, based on whether the national quantity of the procedure is above or below the median procedure.

Panel (b) depicts the distribution of import shares across regions for rare and common procedures. The difference is dramatic: rare procedures (those with national frequency below the median procedure) have much higher import shares, while the common procedures are overwhelmingly lower. Nationally, the imported share of consumption is 35\% for below-median-frequency procedures and 22\% for those above the median. This suggests that frequency is an important predictor of import volume. Within both groups of procedures, there is substantial variation in import shares across hospital referral regions. To formally test for differences in home-market effects, we turn again to gravity models.
4.2.1 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 $RQ_{ijc}$ separately for these two categories of care, $c \in \{\text{common}, \text{rare}\}$. We thus have two observations for each $ij$ pair, allowing us to estimate:

$$\ln E[RQ_{ijc}] = \lambda_X \ln \text{population}_i + \lambda_M \ln \text{population}_j + \gamma X_{ij}$$

$$+ (\mu_X \ln \text{population}_i + \mu_M \ln \text{population}_j + \psi X_{ij}) \cdot 1(c = \text{rare}).$$

(9)

An alternate specification introduces $ij$-pair fixed effects, which absorb all the covariates not interacted with $1(c = \text{rare})$. If returns to scale diminish with quantity, we should find stronger market-size effects for rare procedures, $\mu_X > 0$.

4.2.2 Result

Table 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 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 larger for rare services. The coefficient on provider-market population increases by about 50 percent relative to common services. The coefficient on the patient’s market population shrinks by nearly half. Column 4 introduces location-pair fixed effects. The differential effect of population for rare procedures remains stable. 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 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 both population sizes with 1940 log population in both regions. Columns 3 and 4 repeat the exercise using CBSAs as our geographic unit, rather than HRRs, and reveal the same pattern. When instrumenting with depth to bedrock, columns 5 and 6 show that we lose the statistical power to distinguish between estimates for common and rare services, but the home-market effect remains significant in both columns and the point estimates suggest that it is stronger for rare services.

The finding that less common procedures exhibit stronger home-market effects is robust to different ways of defining rare and common care. Table 4 demonstrates that our result holds when we look across diagnoses rather than procedures. As with the production and consumption elasticities in Figure 5b, the magnitude of the difference between rare and common care is reduced. This could reflect substitution across care within a diagnosis or a less precise classification of patients than of procedures. But the qualitative pattern holds and remains significant, consistent with the difference-in-difference prediction from the model.

In Figure 7, we return to categorizing procedures and show that our finding is robust to categorizing procedures differently. We now estimate equation (6) separately for each decile of procedure frequency. 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 increase across the frequency distribution. The difference between the respective coefficients for each decile demonstrates a strong home-market effect, which is stronger the less common the procedure.

For concreteness, Table 5 shows estimates for selected individual procedures. 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.
The potential concern about omitted cost shifters from Section 3.4 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 specialties, 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 shows that this concern does not pan out. If urban amenities drive specialists’ locations, these should be negative, and especially so for rare specialties. Instead, Appendix Figure B.4 shows that the income elasticities are close to zero on average and uncorrelated with the specialty’s national abundance. So whatever labor-supply shocks urban amenities generate do not exhibit the compensating differentials necessary to explain the observed relationship between market size and specialization.

5 Estimating the scale elasticity of quality

This section estimates the strength of local increasing returns. Section 5.1 describes our empirical strategy. Section 5.2 presents the results, showing that revealed-preference quality has an elasticity of around 0.7 with respect to local volume. Section 5.3 studies the division of labor, showing a mechanism behind these results and evidence supporting our interpretation that trade volumes reflect quality of care.

5.1 Empirical approach

We use a two-step procedure, which begins by estimating a fixed-effects version of the gravity equation. In equation (5), the exporting region $i$ component of the bilateral trade flow is its perceived service quality. Thus, we can estimate $\ln \delta_i$ as the origin fixed effect in this gravity equation. Similarly, $\ln \left( \frac{N_i}{\bar{N}_i} \right)$ can be absorbed as a destination fixed effect, denoted
\ln \theta_j. \text{ This yields the following estimating equation:}

\begin{equation}
\ln \mathbb{E}(\overline{R}_{Qij}) = \ln \delta_i + \ln \theta_j + \gamma X_{ij}.
\end{equation}

We interpret the exporter fixed effects \( \hat{\ln} \delta_i \) as a revealed-preference measure of quality, an interpretation we validate using hospital rankings and physician specialization. The importer fixed effects \( \hat{\ln} \theta_j \) enable us to compute \( \Phi_j - 1 = N_j / \hat{\theta}_j \), a measure of patient market access for those who reside in location \( j \).

We use the estimated exporter fixed effects \( \hat{\ln} \delta_i \) to examine the determinants of regional service quality, in particular its scale elasticity, \( \alpha \). In the free-entry condition (3), service quality in region \( i \) is an isoelastic function of the quantity produced, conditional on revenue, cost, and productivity shifters. Taking the log of equation (3) and rearranging terms yields an estimating equation for the quality-quantity relationship across locations:

\begin{equation}
\ln \delta_i = \alpha \ln Q_i + \ln \overline{R} - \ln w_i + \ln A_i.
\end{equation}

Replacing \( \ln \delta_i \) with its estimate \( \hat{\ln} \delta_i \) from (10) yields an estimating equation for \( \hat{\alpha} \).

One potential concern with estimating equation (11) by ordinary least squares is reverse causality. The same “anchor institutions” concern discussed in Section 3.3 is even more applicable here, as the main regressor is healthcare output rather than population.

We use three instruments to address this: contemporaneous population, along with the same historical population and bedrock depth instruments from Section 3.3. Current population is a natural instrument for healthcare output and is valid if it does not affect healthcare quality other than through its impact on local demand. The other two instruments operate through current population. Historical population remains valid if the anchor institutions concern is a problem in 2017 but was not in 1940. Bedrock depth remains valid if correlates of geology don’t affect healthcare demand other than through population.
Despite our instruments, other channels related to population size could generate the 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 moving to larger regions because of preferable jobs available there. This could arise for a variety of reasons: more patients allow the physicians to specialize in their specific interests, to conduct research, to teach medical students or trainees, or are associated with hospitals that have these amenities. This would not be a problem for our interpretation, as 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 a type of agglomeration externality, and ought to be included in our estimates.

The challenge to our interpretation arises if physicians prefer larger markets for non-professional reasons, and this drives up 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 (11). The analysis of local costs in Section 3.4 and Appendix Figures B.2 and B.3 mitigates this concern.

### 5.2 Scale improves quality

Figure 9 plots \( \ln \delta_i \) against the log of total output in region \( i \), and Table 6 reports the regression results. The estimated scale elasticity is around 0.7 and remains stable if we instrument for output, control for spatial variation in reimbursements, or omit \( Q_{ii} \) observations when estimating the gravity equation (10). Across twelve estimates, the lowest elasticity is 0.537 and the highest is 0.872. 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 core-based statistical areas, reported in Appendix Table B.2, are also stable across specifications and when using the alternative bedrock instrument. These estimates of
\( \hat{\alpha} \) will be key for our counterfactual calculations in Section 6.\(^ {21} \)

To validate our interpretation of \( \ln \delta_i \) as reflecting quality, we first 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.\(^ {22} \) We also obtain Hospital Safety Grades from the Leapfrog Group and average them by HRR. We estimate equation (10) for rare and common procedures separately, yielding \( \hat{\ln \delta_i}^{\text{rare}} \) and \( \hat{\ln \delta_i}^{\text{common}} \).

Figure 8 shows the exporter fixed effects on the vertical axis and the external measure on the horizontal axis. The significant positive slopes in both Panels (a) and (b) 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 from the Leapfrog Group, shown in Panels (c) and (d).\(^ {23} \) The positive relationships with both measures suggest that our estimates capture a meaningful measure of hospital quality.

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 we see a stronger relationship between the U.S. News rankings and exporter fixed effects estimated from rare services; the coefficient in Panel (b) is twice as large as that for common services in Panel (a). In contrast, safety grades are not differentially relevant for rare services. Indeed, Panels (c) and (d) show virtually identical slopes.

5.3 Scale facilitates the division of labor

One source of increasing returns could be division of labor among physicians. In particular, the specialized labor required to produce rare services could drive the patterns we found in

\(^{21}\) 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, Hornbeck, and Moretti (2010) report an analogous elasticity above 1 (a 12 percent increase in total factor productivity in response to adding a plant representing 8.6 percent of the county’s prior output).

\(^{22}\) U.S. News 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. 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.

\(^{23}\) 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.
Section 4 across treatments and diagnoses. Specialized services may require physicians with specific training, and low demand in smaller HRRs may not support specialized physicians.

5.3.1 Specialization as a source of local increasing returns

To study this mechanism, we estimate the population elasticity of physicians and relate it to physician frequency. Let $Y_{si}$ be the number of doctors of specialty $s$ in location $i$, and $Y_s = \sum_h Y_{si}$ be the specialty’s national frequency.\(^{24}\) We estimate a Poisson model,

$$\ln \mathbb{E}[Y_{si}] = \kappa_s^S + \beta_s^S \ln \text{population}_i,$$

for each specialty $s$ by maximum likelihood.

Figure 10 Panel (a) shows the relationship between the estimated population elasticity $\hat{\beta}_s^S$ and each specialty’s national frequency, $\ln Y_s$. There is a negative relationship between a specialty’s population elasticity and the national number of physicians in that specialization. The similarity of these population elasticity-frequency relationships between medical procedures and physician specializations suggests that whatever scale phenomenon is relevant for physicians could help explain the pattern of production of procedures.

Consistent with this idea, Appendix Figure B.5 shows that larger markets produce a greater variety of procedures. 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. The elasticity of unique procedure count with respect to population is 0.37.

---

\(^{24}\)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 (NPIs) 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.
5.3.2 Imports are specialist-intensive

We next ask whether the distribution of specialties helps explain trade. Panel (b) of Figure 10 shows the share of imports and of domestic consumption that are provided by specialists as a function of regional population.\textsuperscript{25} Imports are significantly more specialist-intensive than domestic 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 (billing code) 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.

Panel (c) of Figure 10 shows the share of imports and of domestic consumption provided by the standard specialties. We see a similar pattern, with imports more likely to come from the standard specialist than domestic care, and the distinction especially pronounced in the smallest regions. The magnitude of this difference is substantial: Domestic care in the smallest regions is 50 percent more likely to be provided by a non-standard specialist than in the largest regions (7.5 percent vs. 5 percent). When importing medical services, this share falls to 5 percent—indistinguishable from domestic consumption in the largest regions.

Since imported care is predominantly specialty care—and specialists are disproportionately located in larger markets—we conclude that visiting the appropriate specialist is part of the value proposition for trade in medical care. Trade enables patients to visit the appropriate specialist for the care they need, providing a second validation of our interpretation that it reflects quality. Furthermore, they travel to regions with highly-ranked hospitals. Larger markets tend to have these high-quality hospitals and rare services available, and this home-market size 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

\textsuperscript{25}We define “specialist” to mean all physicians except those whose primary specialty is internal medicine, general practice, or family practice.
enables patients from many different regions to share the benefits of this agglomeration.

6 Tradeoffs and counterfactual scenarios

Given the estimated strength of local increasing returns, the geographic concentration of healthcare production generates substantial benefits. Larger regions employ a greater variety of specialized physicians and produce meaningful volumes of more specialized procedures. But this geographic concentration implies that patients in smaller regions may suffer from limited access to physicians. This perceived maldistribution has led to numerous federal and state policies designed to encourage physicians to locate in “under-served regions.” These policies include higher payments to hospitals and physicians in rural locations and bonuses and loan forgiveness for physicians who locate in Health Professional Shortage Areas.26 The effectiveness of these policies is not clear (Khoury, Leganza, and Masucci, 2022), and our analysis offers a reason why: it may not just be a matter of physicians’ location preferences. Instead, less populous regions may be fundamentally constrained by producing at smaller scales.

These differences may not affect all patients equally. Section 6.1 estimates distance sensitivity separately by income group. We find that patients in low-income neighborhoods are more sensitive to travel distance than those who live in high-income neighborhoods.

In Section 6.2, we use our estimates of the scale elasticity $\alpha$, distance elasticity $\gamma$, and region-specific quality $\delta_i$ to compute counterfactual values of patient market access $\Phi_j$. At baseline, access is considerably higher in more populous markets: the value of patient market access has a population elasticity of 0.60. We then examine two counterfactual policy scenarios. First, we reallocate quantities produced from larger markets to smaller markets, thereby reducing the scale-related gap in quality of service. Second, we compute how much more willing to travel patients in each region would need to be in order to equalize the

26 Examples of higher payments in rural locations include the Critical Access Hospital program and rural subsidies in the Geographic Practice Cost Indexes.
value of patient market access across regions. We show that socioeconomic status interacts with these equity-efficiency considerations and quantify the changes to distance sensitivity that different income groups would require to eliminate size-related geographic disparities in patient market access.

6.1 Heterogeneity in distance elasticities by socioeconomic status

The large and robust scale economies we have estimated mean that patients who live in, or can travel to access, regions with abundant healthcare production have access to higher-quality care. This may not be equally true for all patients. We consider one key dimension of patient heterogeneity: differences in socioeconomic status. Our data do not contain patients’ wealth, so we classify them based on residential ZIP code. We split ZIP codes into deciles based on median household income, and estimate equation (10) separately by decile.

Figure 11 plots the estimated distance elasticities by income decile. We find a strong, nearly monotonic relationship between income and the distance elasticity: patients from the highest neighborhood-income decile exhibit a distance elasticity 25% lower than those in the lowest decile. While we do not have exogenous variation in income, our estimates are consistent with the interaction that Silver and Zhang (2022) estimate between income and distance to care. We can directly rule out an obvious concern: 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 B.3)—the income gradient of distance elasticities persists.

In summary, patients who reside in higher-income neighborhoods are more likely to travel longer distances for medical care. This means that the benefits of agglomeration—availability of higher-quality rare care in major centers—may not be distributed evenly. This is especially notable in this setting, since Medicare is a nearly-universal insurance program for elderly and disabled Americans.
6.2 Proximity-concentration tradeoffs

Our first counterfactual scenario forcibly reallocates production levels to reduce spatial inequity. Using primes to denote counterfactual values, we simply impose counterfactual production levels $Q'_i = Q_i \left( \frac{\text{population}_i}{\text{population}} \right)^{-0.15}$. That is, we reduce the population elasticity of output by 0.15, while keeping aggregate output unchanged ($\sum_i Q'_i = \sum_i Q_i$). We use our estimate of the quality elasticity, $\alpha$, to translate this reallocation into quality changes: $\delta'_i = \delta_i \left( \frac{Q'_i}{Q_i} \right)^\alpha$. We then compute each hospital referral region’s patient market access in this counterfactual scenario, $\Phi'_j - \delta_{0j} = \sum_{i \in I} \phi_{ij} \delta'_i$.\(^{27}\)

Figure 12 depicts the consequences of this reallocation. The scale elasticity $\alpha$ translates the reduction of the size-based gap in quantities into a reduction of the size-based gap in service quality. As shown by the red series, quality falls nearly 30% in the largest markets, while it increase almost 15% in the smallest markets. The size-based changes in patient market access, depicted by the blue series, exhibit a similar pattern with more modest magnitudes, because patient market access is a distance-weighted sum of service quality. This counterfactual reallocation of production improves patient market access in small markets. But it comes at a high cost: the majority of the Medicare population resides in larger markets, as depicted by the gray series, and thus the population-weighted average value of all patient market access falls by 9%.

Our second counterfactual scenario examines how much more travel would be required to equalize patient market access across regions of different sizes. Distance-related bilateral trade costs incorporate fiscal costs, information costs, preferences, and other frictions. For each region, we compute the increase in willingness to travel (the percentage shrinkage in the coefficients on distance covariates from $\gamma$ to $\gamma'$) that would cause the counterfactual value of patient market access in that region, $\Phi'_j - \delta_{0j} = \sum_{i \in I} \exp(\gamma' X_{ij}) \delta_i$, to be equal to patient market access in the largest region, $\Phi_{\text{largest HRR}} - \delta_{0,\text{largest HRR}}$. We compute these changes holding equilibrium outcomes in all other regions, including service quality $\delta_i$, fixed.

\(^{27}\)Recall $\phi_{ij} = \exp(\gamma X_{ij})$, where $X_{ij}$ contains log distance, log distance squared, and a same-HRR dummy.
Figure 13 panel (a) depicts the reductions in distance sensitivity necessary to equalize patient market access across regions of different sizes. Since less populous markets have worse patient market access in the initial equilibrium, patients from smaller markets must increase their willingness to travel more. To obtain patient market access equal to that enjoyed by patients in the largest markets, the distance sensitivity of patients in the smallest markets would have to be reduced by 40%. This counterfactual scenario would involve substantially more travel by these patients. Panel (b) repeats this analysis separately based on the distance sensitivities of patients residing in the highest- and lowest-income zip codes.\(^{28}\) We see that lower-income patients in small and mid-sized markets need larger subsidies than higher-income ones. This reflects the extra reduction in market access they experience when residing in areas with less access while having a higher sensitivity to distance.

These counterfactual outcomes come with a number of caveats. First, both counterfactual scenarios, moving physicians and subsidizing travel, would involve fiscal costs that we have not estimated.\(^{29}\) Both could be refined to improve targeting rather than using average outcomes by market size. Even so, these simple counterfactual scenarios highlight the central tradeoff we have identified: healthcare production has substantial local increasing returns, and patient travel plays a meaningful role in enabling access to higher-quality care.

7 Conclusion

Rural and smaller markets are home to fewer specialized physicians, produce fewer procedures per capita, and have worse health outcomes than larger markets. But thanks to ample trade in medical services, these facts do not translate one-for-one into substantially lower consumption of medical services in these markets. Instead, trade affords patients in smaller markets access to higher-quality care. This higher quality comes in part from consuming

\(^{28}\)The only difference we use between the two groups is the distance sensitivity, from Figure 11. We use common area quality estimates \(\hat{\delta}_i\) between the two groups.

\(^{29}\)Carroll (2019) finds some evidence that the Critical Access Hospital program has health benefits, but concludes that it is cost-ineffective.
services that would otherwise be unavailable, seeing specialized physicians, and having care provided by the appropriate specialist.

This trade amplifies the scale advantages of large markets, and hence the quality of care they produce. We identify a strong home-market effect, meaning that medical care exhibits increasing returns to scale in a region, and both gross and net exports of medical services increase with local population. 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 space may impact the quality of care available. We simulate policies that aim to improve care access in “under-served” markets, and find that they have very different welfare and distributional consequences. Policy efforts to improve access must account for the proximity-concentration tradeoff, as some policies create tension between facilitating access and quality of care.
References


Figure 1: Illustrative model diagrams

(a) Equilibrium in autarky

Isocost Curve = \( R \)

Demand in Small City
Demand in Big City

\( Q_S \)
\( 
\delta_S \)
\( Q_B \)
\( 
\delta_B \)

Quantity (\( Q \))

(b) Equilibrium with trade

Isocost Curve = \( R \)

Net Imports
Net Exports

\( \delta'_S \)
\( \delta'_B \)

(c) Rare vs. common procedures: Autarky

Isocost Curve = \( R \)

Demand for Rare Procedure
Demand for Common Procedure

\( \delta'^S \)
\( \delta'^B \)

\( Q'_S \)
\( Q'_B \)

Quantity (\( Q \))

(d) Rare vs. common procedures with trade

Isocost Curve = \( R \)

Net Exports

\( \delta'^R \)
\( \delta'^B \)

\( Q'_R \)
\( Q'_B \)

Note: These diagrams illustrate demand and isocost curves in quantity and quality space. Prices are fixed and supply exhibits local increasing returns to scale. Panel (a) shows the autarkic (no trade) equilibrium in two regions. Greater demand in the big region means that in equilibrium higher quality is produced for the same average cost. Panel (b) introduces trade. Some small-region patients import higher-quality care from the big region. This further increases the quality gap between the locations. Panel (c) contrasts rare and common procedures. Because the isocost curve is concave, market-size effects are more pronounced for rare than for common procedures. In the case illustrated, the small region does not even produce the rare procedure in autarky. Panel (d) introduces trade in both procedures. Trade flows are larger in the rare procedure, as small-region patients import higher-quality care.
Figure 2: Production, consumption, and exports across space

(a) Production per capita

(b) Consumption per capita

(c) Gross exports relative to production

(d) Net exporter vs net importer

Note: Panel (a) shows production per capita, including professional and facility fees. The HRR of production corresponds to the location where the service was provided. Panel (b) shows consumption per capita, including professional and facility fees. The HRR of consumption corresponds to the patient’s HRR. Colors depict deciles of production per capita in both panels (a) and (b). Panel (c) shows gross exports as a share of total production by HRR for professional fees. Panel (d) contrasts HRRs that are net exporters with HRRs that are net importers for professional fees. All calculations exclude emergency-room services and Medicare Advantage patients.
Figure 3: Production and consumption of medical care across regions and services

Note: This figure shows production, consumption, and trade per capita of Medicare services across HRRs of different sizes, all smoothed via local averages. The dollar value of physician services is computed using national average prices and excluding emergency-room care. 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 exports of medical care and the dashed light-blue series shows imports of medical care, again per Medicare beneficiary. The orange series at the bottom of the graph reports the distribution of HRR populations.
Figure 4: Patients travel between regions and trade declines with distance

(a) Distribution of travel distances within and across HRRs

Within-HRR ZCTA-to-ZCTA distances more than 160km censored at 160km. Across-HRR ZCTA-to-ZCTA distances more than 160km replaced by HRR-to-HRR distances.

(b) Trade volume and extensive margin by distance

Note: 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 distances 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 obtain geographic centroids from the Census Bureau’s corresponding ZIP code tabulation area (ZCTA). 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 (10), 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 (10).
Figure 5: Population elasticities of production and consumption

(a) Population elasticities by procedure frequency

(b) Population elasticities by diagnosis frequency

Note: 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 (7) and (8) 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. 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. The horizontal axis shows the total national volume of care for the service or diagnosis. 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. We see 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 6: 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

Note: Panel (a) shows the distribution of the imported consumption share across procedures, for all procedures with at least $5,000 in annual spending (in our 20% sample of Medicare claims). Imports are defined as care provided to a patient who lives in one HRR at a place of service 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 7: The home-market effect is stronger for rarer procedures

Note: This figure groups all services in the Medicare claims data into deciles based on the national frequency of each procedure. For each decile, we estimate equation (6), testing for a home market effect, and plot the estimated coefficients on provider and patient market log population with their 95-percent 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 8: Estimated quality is positively correlated with 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. quality for common services

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

Note: These panels show the relationship between our exporter fixed effects and external quality measures. The vertical axis shows the exporter fixed effects for each HRR estimated from equation (10), in Panels (a) and (c) using trade in common services, and in Panels (b) and (d) using trade in rare services. 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 Panels (c) and (d) 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 take 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 9: Quality is higher in regions producing more output

![Graph showing the relationship between production and quality across HRRs. Production, on the horizontal axis, is measured as log Medicare output produced in the HRR (in millions US dollars). The revealed-preference measure of quality, on the vertical axis, is the exporter fixed effect in equation (10). The production elasticity of quality is 0.778.]

Note This figure shows the relationship between production and quality across HRRs. Production, on the horizontal axis, is measured as log Medicare output produced in the HRR (in millions US dollars). The revealed-preference measure of quality, on the vertical axis, is the exporter fixed effect in equation (10). The production elasticity of quality is 0.778.
Figure 10: Imports are specialist-intensive, especially in smaller regions

(a) Population elasticities of physician specializations

(b) Specialty care imports

(c) “Standard” specialty care

Note: Panel (a) plots the elasticities of quantity of physicians in an HRR with respect to population for each specialty in the NPPES data on the vertical axis. The elasticities are computed using the Poisson model in equation (12). The horizontal axis shows the number of physicians nationally in each specialty in NPPES. 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 procedures performed by a specialist and smaller markets’ imports are more likely to be performed by specialists. Panel (c) examines procedures that are typically performed by specialists, and classifies the “standard” specialists as the top two specialties performing the procedure nationally. Panel (c) 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.
Figure 11: Higher-income patients are more willing to travel

Distance elasticity of trade by income decile

Note: This figure depicts the coefficient on log distance obtained by estimating equation (10) separately for each decile of the national ZIP-level median-household-income distribution. The 95% confidence intervals are computed using standard errors clustered by both patient HRR and provider HRR.

Figure 12: Patient market access in production reallocation scenario

Note: This figure depicts the consequences of forcibly reallocating production from larger to smaller markets by reducing population elasticity of quantities produced by 0.15. The gray line depicts the distribution of the Medicare population across hospital referral regions by population size. The red line depicts the change in the service quality produced by market size. The blue line depicts the change in the value of patient market access by market size. The exercise is described in detail in Section 6.
Figure 13: Reduced distance sensitivity required to equalize patient market access

(a) Aggregate

(b) By income group

Note: Panel (a) depicts the percentage reduction in the distance elasticity required to raise patient market access to the value in the largest markets by market size. The distance elasticity would need to be reduced the most in the smallest markets to equalize patient access. Counterfactual patient market access in region \( i \) with distance sensitivity \( \gamma' \) is \( \sum_j \exp(\gamma'X_{ij})\delta_j \). We compute the percentage reduction in the coefficients on distance covariates from \( \gamma \) to \( \gamma' \) such that the size-predicted value of patient market access is equal to estimated patient market access in the largest region. Panel (b) repeats the same exercise based on the distance sensitivity of patients residing in the highest- and lowest-income decile zip codes separately.
Table 1: Gravity regression: aggregate medical services exhibit a strong home-market effect

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation method:</td>
<td>PPML</td>
<td>PPML</td>
<td>PPML</td>
<td>IV</td>
</tr>
<tr>
<td>Provider-market population (log)</td>
<td>0.636</td>
<td>0.641</td>
<td>0.643</td>
<td>0.594</td>
</tr>
<tr>
<td></td>
<td>(0.0627)</td>
<td>(0.0603)</td>
<td>(0.0448)</td>
<td>(0.0719)</td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.378</td>
<td>0.376</td>
<td>0.405</td>
<td>0.365</td>
</tr>
<tr>
<td></td>
<td>(0.0608)</td>
<td>(0.0580)</td>
<td>(0.0417)</td>
<td>(0.0515)</td>
</tr>
<tr>
<td>Distance (log)</td>
<td>-1.656</td>
<td>0.0550</td>
<td>0.0362</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0498)</td>
<td>(0.305)</td>
<td>(0.268)</td>
<td></td>
</tr>
<tr>
<td>Distance (log, squared)</td>
<td>-0.173</td>
<td>-0.171</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0296)</td>
<td>(0.0262)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
</tr>
<tr>
<td>Distance elasticity at mean</td>
<td></td>
<td></td>
<td>-2.42</td>
<td>-2.42</td>
</tr>
<tr>
<td>Distance deciles</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Two-way clustered standard errors in parentheses

Note: This table reports estimates of equation (6), which estimates the presence of gross and/or net home market effects. The sample is all HRR pairs \( (N = 306^2) \), and 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 data are computed from the Medicare 20 percent carrier Research Identifiable File. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.
Table 2: The home-market effect is stronger for rare procedures

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider-market population (log)</td>
<td>0.638</td>
<td>0.624</td>
<td>0.623</td>
<td>0.630</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0634)</td>
<td>(0.0613)</td>
<td>(0.0614)</td>
<td>(0.0598)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.377</td>
<td>0.379</td>
<td>0.380</td>
<td>0.379</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0615)</td>
<td>(0.0590)</td>
<td>(0.0591)</td>
<td>(0.0572)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider-market population (log) × rare</td>
<td>0.306</td>
<td>0.291</td>
<td>0.316</td>
<td>0.287</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0472)</td>
<td>(0.0455)</td>
<td>(0.0480)</td>
<td>(0.0458)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-market population (log) × rare</td>
<td>-0.229</td>
<td>-0.219</td>
<td>-0.232</td>
<td>-0.211</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0698)</td>
<td>(0.0671)</td>
<td>(0.0704)</td>
<td>(0.0658)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>187,272</td>
<td>113,468</td>
<td>113,468</td>
<td>113,468</td>
<td>113,468</td>
<td>113,468</td>
</tr>
<tr>
<td>Distance controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance [quadratic] controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Patient-provider-market-pair FEs</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

Two-way clustered standard errors in parentheses

Note: This table reports estimates of equation (9), 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 $2 \times 306^2$ in column 1. 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 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 is especially true for rare services. Trade data are computed from the Medicare 20 percent carrier Research Identifiable File. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.
Table 3: The stronger home-market effect for rare procedures is robust to instrumenting for population

<table>
<thead>
<tr>
<th>Geography:</th>
<th>Instrument:</th>
<th>Procedure Sample:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HRR</td>
<td>HRR</td>
<td>(1940 pop)</td>
<td>(1940 pop)</td>
<td>(1940 pop)</td>
<td>(1940 pop)</td>
<td>Bedrock</td>
<td>Bedrock</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CBSA</td>
<td>CBSA</td>
<td>CBSA</td>
<td>CBSA</td>
<td>CBSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider-market population (log)</td>
<td>0.595</td>
<td>1.080</td>
<td>0.716</td>
<td>0.895</td>
<td>1.157</td>
<td>1.753</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0733)</td>
<td>(0.0913)</td>
<td>(0.0249)</td>
<td>(0.0388)</td>
<td>(0.307)</td>
<td>(0.524)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.361</td>
<td>0.0476</td>
<td>0.396</td>
<td>0.328</td>
<td>0.182</td>
<td>-0.582</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0522)</td>
<td>(0.114)</td>
<td>(0.0261)</td>
<td>(0.0344)</td>
<td>(0.373)</td>
<td>(0.580)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance (log)</td>
<td>0.0756</td>
<td>0.973</td>
<td>-3.412</td>
<td>-1.378</td>
<td>-4.678</td>
<td>-4.631</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.270)</td>
<td>(0.449)</td>
<td>(0.294)</td>
<td>(0.989)</td>
<td>(1.049)</td>
<td>(2.520)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance (log, squared)</td>
<td>-0.177</td>
<td>-0.261</td>
<td>0.105</td>
<td>-0.0742</td>
<td>0.210</td>
<td>0.181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0265)</td>
<td>(0.0503)</td>
<td>(0.0287)</td>
<td>(0.0935)</td>
<td>(0.0845)</td>
<td>(0.199)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>93,636</td>
<td>93,636</td>
<td>857,476</td>
<td>857,476</td>
<td>781,456</td>
<td>781,456</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance elasticity at mean</td>
<td>-2.45</td>
<td>-2.76</td>
<td>-1.91</td>
<td>-2.43</td>
<td>-1.68</td>
<td>-2.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: This table reports estimates of equation (6), 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 computed by 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. Standard errors (in parentheses) are two-way clustered by patient market and provider market.
Table 4: The home-market effect is stronger for rarer diagnoses

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider-market population (log)</td>
<td>0.624</td>
<td>0.611</td>
<td>0.606</td>
<td>0.612</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0619)</td>
<td>(0.0597)</td>
<td>(0.0592)</td>
<td>(0.0576)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.380</td>
<td>0.382</td>
<td>0.385</td>
<td>0.384</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0607)</td>
<td>(0.0581)</td>
<td>(0.0578)</td>
<td>(0.0560)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider-market population (log) × rare</td>
<td>0.117</td>
<td>0.111</td>
<td>0.122</td>
<td>0.110</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0438)</td>
<td>(0.0403)</td>
<td>(0.0444)</td>
<td>(0.0390)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-market population (log) × rare</td>
<td>-0.0786</td>
<td>-0.0759</td>
<td>-0.0792</td>
<td>-0.0719</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0336)</td>
<td>(0.0315)</td>
<td>(0.0346)</td>
<td>(0.0307)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>187,272</td>
<td>113,534</td>
<td>113,534</td>
<td>113,534</td>
<td>113,534</td>
<td>113,534</td>
</tr>
<tr>
<td>Distance controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Distance [quadratic] controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient-provider-market-pair FEs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

[Note:] This table augments equation (6) 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 {exporting HRR, importing HRR, rare/common indicator} so the number of observations is $2 \times 306^2$ in column 1. 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 × 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 × rare across all columns indicates that the strong home-market effect is especially true for rare diagnoses. Trade data are computed from the Medicare 20 percent carrier Research Identifiable File. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.
Table 5: Gravity regression by procedure: individual procedures exhibit a strong home-market effect

<table>
<thead>
<tr>
<th>Procedure:</th>
<th>(1) Colonoscopy</th>
<th>(2) Cataract surgery</th>
<th>(3) Brain radiosurgery</th>
<th>(4) Brain tumor removal</th>
<th>(5) LVAD implantation</th>
<th>(6) Colon removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCPCS code:</td>
<td>G0121</td>
<td>66982</td>
<td>61798</td>
<td>61510</td>
<td>33979</td>
<td>44155</td>
</tr>
<tr>
<td>Provider-market population (log)</td>
<td>0.504 (0.0647)</td>
<td>0.550 (0.0616)</td>
<td>1.164 (0.118)</td>
<td>0.940 (0.0779)</td>
<td>1.187 (0.155)</td>
<td>0.998 (0.184)</td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.334 (0.0643)</td>
<td>0.408 (0.0557)</td>
<td>0.172 (0.0744)</td>
<td>0.201 (0.0692)</td>
<td>0.223 (0.133)</td>
<td>-0.0859 (0.163)</td>
</tr>
<tr>
<td>Distance (log)</td>
<td>0.239 (0.393)</td>
<td>-0.00287 (0.344)</td>
<td>1.453 (0.729)</td>
<td>0.862 (0.518)</td>
<td>1.882 (1.091)</td>
<td>2.477 (1.674)</td>
</tr>
<tr>
<td>Distance (log, squared)</td>
<td>-0.198 (0.0394)</td>
<td>-0.177 (0.0352)</td>
<td>-0.303 (0.0726)</td>
<td>-0.257 (0.0541)</td>
<td>-0.342 (0.111)</td>
<td>-0.443 (0.173)</td>
</tr>
<tr>
<td>Observations</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
<td>93,636</td>
</tr>
<tr>
<td>Distance elasticity at mean</td>
<td>-2.60</td>
<td>-2.54</td>
<td>-2.88</td>
<td>-2.82</td>
<td>-3.02</td>
<td>-3.87</td>
</tr>
<tr>
<td>Total count</td>
<td>58,798</td>
<td>43,604</td>
<td>752</td>
<td>1,922</td>
<td>333</td>
<td>112</td>
</tr>
</tbody>
</table>

Two-way clustered standard errors in parentheses

Note: This table reports estimates of equation (6) 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 data are computed from the Medicare 20 percent carrier Research Identifiable File. 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 colonoscopy (1) and cataract surgery (2). In a colonoscopy, the physician visualizes the large bowel with a camera to aid diagnosis. In a cataract surgery, the surgeon removes a cloudy lens from the eye to improve vision. Relatively rare procedures include brain radiosurgery (3), brain tumor removal (4), left ventricular assist devices, e.g. LVADs, (5) and colon removals (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 aforementioned structures.
<table>
<thead>
<tr>
<th></th>
<th>No Controls</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Diag</td>
<td>Diag</td>
</tr>
<tr>
<td>OLS</td>
<td>0.804</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>2SLS: pop</td>
<td>0.799</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>2SLS: pop1940</td>
<td>0.660</td>
<td>0.550</td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.069)</td>
</tr>
</tbody>
</table>

**Note:** This table reports estimates of $\alpha$ from ordinary least squares (OLS) or two-stage least squares (2SLS) regressions of the form $\hat{\ln} \delta_i = \alpha \ln Q_i + \ln R_i + u_i$, where the $R_i$ covariate is Medicare’s Geographic Adjustment Factor, $u_i$ is an error term, and we instrument for $\ln Q_i$ using the designated instrumental variables. The “no controls” designation indicates the $R_i$ covariate is omitted. The “no diag” designation indicates that $Q_{ii}$ observations were omitted when estimating $\hat{\ln} \delta_i$ using the gravity equation. Standard errors are robust to heteroskedasticity.
A Theory appendix

A.1 Derivations of results in Section 1.6

Starting from the $I$ equations with the unknowns $\{\delta_i\}_{i=1}^I$:

$$\delta_i = \left(\frac{RA_i}{w_i}\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in I} \sum_{i' \in 0 \cup I} \frac{\phi_{ij}}{\delta_{i'}} \phi_{i'j} N_{i'}\right)^{\frac{\alpha}{1-\alpha}}$$

For brevity, assume $\frac{RA_i}{w_i} = 1 \forall i$.

Totally differentiating the above system of equations in terms of $\{d\delta_i, dN_i\}_{i=1}^I$ and evaluating it at the symmetric equilibrium $(\bar{Q}_1 = \frac{1+(I-1)\phi}{\Phi} \bar{N})$ yields the following expression:

$$\Phi \frac{1}{N} \frac{1-\alpha}{\alpha} \bar{N}^{1-2\alpha} \frac{d\bar{N}}{\alpha} = -\left[d\delta_i + \phi \sum_{i' \neq i} d\delta_{i'}\right] + \Phi \frac{dN_i}{N} + \sum_{j \neq i} -\phi \left[d\delta_j + \phi \sum_{i' \neq j} d\delta_{i'}\right] + \sum_{j \neq i} \phi \Phi \frac{dN_j}{N}.$$\

Given $\alpha > 0$, $dN_1 > 0$, and $dN_j = 0 \forall j \neq 1$, tedious algebra delivers the expression for quality changes reported in Section 1.6:

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

The weak home-market effect is derived as follows:

$$\ln Q_{1,j \neq 1} = \alpha \ln Q_1 + \ln \phi - \ln \Phi_j + \ln N_j$$

$$\frac{d \ln Q_{1,j \neq 1}}{d \ln N_1} = \alpha \frac{d \ln Q_1}{d \ln N_1} - \frac{\alpha}{\Phi_j} \left(\phi Q_1^{\alpha-1} \frac{dQ_1}{d \ln N_1} + Q_j^{\alpha-1} \frac{dQ_j}{d \ln N_1} + \phi \sum_{i' \neq 1,j} Q_{i'}^{\alpha-1} \frac{dQ_{i'}}{d \ln N_1}\right)$$

$$= \frac{d \ln \delta_1}{d \ln N_1} - \frac{1}{\Phi_j} \left(\phi \delta_1 \frac{d \ln \delta_1}{d \ln N_1} + \phi \frac{d \ln \delta_j}{d \ln N_1} + \phi \sum_{i' \neq 1,j} \delta_{i'} \frac{d \ln \delta_{i'}}{d \ln N_1}\right)$$

$$= \left(\frac{\bar{N} - Q_{1j}}{N}\right) \frac{d \ln \delta_1}{d \ln N_1} - \left(\frac{\bar{N} - Q_{0j} - Q_{1j}}{N}\right) \frac{d \ln \delta_j}{d \ln N_1}$$

$$= \left(\frac{\bar{N} - Q_{1j}}{N}\right) \left[\frac{d \ln \delta_1}{d \ln N_1} - \frac{d \ln \delta_j}{d \ln N_1}\right] + \frac{Q_{0j}}{N} \frac{d \ln \delta_j}{d \ln N_1} > 0$$
The condition for the strong home-market effect is derived as follows:

\[ Q_{1,j\neq 1,1} = \frac{Q_1^\alpha \phi}{1 + Q_1^\alpha \phi + Q_j^\alpha \phi + \sum_{i\neq 1,j} Q_i^\alpha \phi} N_j - \frac{Q_j^\alpha \phi}{1 + Q_1^\alpha \phi + Q_j^\alpha \phi + \sum_{i\neq 1,j} Q_i^\alpha \phi} N_1 \]

\[
d\ln Q_{1,j\neq 1} - d\ln Q_{j\neq 1,1} = d\ln N_j - d\ln N_1 + \frac{\alpha Q_1^\alpha \phi}{\phi} \left[ \frac{1 + (I - 1)\phi}{Q} + (1 - \phi) \right] (d\ln Q_1 - d\ln Q_j) \]

\[
= -d\ln N_1 + \frac{1}{Q^{-\alpha} + 1 + (I - 1)\phi} \left[ \frac{1 + (I - 1)\phi}{Q} + (1 - \phi) \right] (d\ln \delta_1 - d\ln \delta_j) \]

\[
= \frac{\frac{\alpha}{N} - (1 - \alpha) \frac{1+(I-1)\phi}{1-\phi}}{\frac{1}{1+(I-1)\phi} N + (1 - \alpha) \frac{1+(I-1)\phi}{1-\phi}} d\ln N_1 \]

The values of this expression at \( \alpha = 1 \) and \( \alpha = 0 \) are

\[
\lim_{\alpha \to 1} (d\ln Q_{1,j\neq 1} - d\ln Q_{j\neq 1,1}) = \frac{1 + (I - 1)\phi}{(1 - \phi)Q} d\ln N_1 > 0 \]

\[
\lim_{\alpha \to 0} (d\ln Q_{1,j\neq 1} - d\ln Q_{j\neq 1,1}) = -d\ln N_1 < 0 \]

There is a strong home-market effect if

\[
d\ln Q_{1,j\neq 1} - d\ln Q_{j\neq 1,1} = \frac{\frac{\alpha}{N} - (1 - \alpha) \frac{1+(I-1)\phi}{1-\phi}}{\frac{1}{1+(I-1)\phi} N + (1 - \alpha) \frac{1+(I-1)\phi}{1-\phi}} d\ln N_1 > 0 \]

\[
\iff \frac{\alpha}{1 - \alpha} > \frac{1 + (I - 1)\phi}{1 - \phi} N \]

This is true if \( \alpha \) is large enough and \( N \) is small enough.

**B Additional exhibits**
Figure B.1: Bedrock depth predicts population across CBSAs

Note: This figure shows the relationship across CBSAs between log population (on the vertical axis) and median depth to bedrock (Levy and Moscona, 2020). There is a strong negative relationship: the F-statistic is 37.3. Bedrock depth is thus a relevant instrumental variable for current log population.
Figure B.2: Population elasticities of input costs in 2017 (commuting zones)

(a) Physicians’ earnings

(b) Other Healthcare Workers’ earnings

(c) Median House Value

Note: This figure shows the relationships between input costs and local population. Panel (a) shows physicians’ earnings, using data from Gottlieb et al. (2020). Panel (b) shows other healthcare workers’ earnings, using data from the 2015–2019 American Community Survey (ACS). Panel (c) shows median house value, also using ACS data, as a proxy for real estate and other locally priced inputs.
Figure B.3: Population elasticities of input costs in 2017 (CBSA)

(a) Physicians’ earnings

\[ Y = 12.9246 - 0.0518 \times [0.0158] X + \varepsilon \]
\[ N = 251; R^2 = 0.0416 \]

(b) Other Healthcare Workers’ earnings

\[ Y = 10.1368 + 0.0473 \times [0.0065] X + \varepsilon \]
\[ N = 251; R^2 = 0.1738 \]

(c) Median House Value

\[ Y = 10.3770 + 0.1437 \times [0.0213] X + \varepsilon \]
\[ N = 251; R^2 = 0.1544 \]

Note: This figure shows the relationships between input costs and local population. Panel (a) shows physicians’ earnings, using data from the American Community Survey (2015–2019) (ACS). Panel (b) shows other healthcare workers’ earnings, also using ACS data. Panel (c) shows median house value, also using ACS data, as a proxy for real estate and other locally priced inputs.
Figure B.4: Specialists’ income patterns do not explain the output-population gradient

Note: 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 B.5: Larger markets produce greater variety of procedures

Note: 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: the procedure count has a population elasticity of 0.37.
Table B.1: Estimates of a strong home-market effect by CBSA

<table>
<thead>
<tr>
<th>Estimation method:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument:</td>
<td>PPML IV</td>
<td>PPML IV</td>
<td></td>
<td>IV</td>
</tr>
<tr>
<td>Provider-market population (log)</td>
<td>0.734 (0.0229)</td>
<td>0.712 (0.0241)</td>
<td>0.734 (0.0254)</td>
<td>1.134 (0.292)</td>
</tr>
<tr>
<td>Patient-market population (log)</td>
<td>0.395 (0.0286)</td>
<td>0.398 (0.0255)</td>
<td>0.395 (0.0304)</td>
<td>0.199 (0.360)</td>
</tr>
<tr>
<td>Distance (log)</td>
<td>-3.509 (0.314)</td>
<td>-3.451 (0.284)</td>
<td>-3.449 (0.337)</td>
<td>-4.631 (0.969)</td>
</tr>
<tr>
<td>Distance (log, squared)</td>
<td>0.115 (0.0313)</td>
<td>0.110 (0.0277)</td>
<td>0.111 (0.0336)</td>
<td>0.208 (0.0783)</td>
</tr>
</tbody>
</table>

| Observations | 857,476 | 857,476 | 781,456 | 781,456 |
| Sample:      | All CBSAs | All CBSAs | Bedrock data | Bedrock data |
| Distance elasticity at mean | -1.87 | -1.88 | -1.86 | -1.67 |

Two-way clustered standard errors in parentheses

Note: This table reports estimates of equation (6), which estimates the presence of gross and/or net home market effects. The sample is all CBSA pairs (N = 926²), and 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 1 reports the baseline PPML estimate when using CBSAs rather than HRRs as the geographic unit. Column 2 uses the provider-market and patient-market log populations in 1940 as instruments for the contemporaneous log populations. Column 3 reports the PPML estimate on the subsample of regions for which we have data on depth to bedrock available (N = 884²). Column 4 uses depth to bedrock in the importing and exporting regions as instruments for current log population in those regions, respectively. Trade data are computed from the Medicare 20 percent carrier Research Identifiable File. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.
Table B.2: Scale elasticity estimates for CBSAs

<table>
<thead>
<tr>
<th></th>
<th>No Controls</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Diag</td>
<td>Diag</td>
</tr>
<tr>
<td>OLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>2SLS: pop</td>
<td>1.023</td>
<td>0.845</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>2SLS: pop1940</td>
<td>0.928</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>2SLS: bedrock</td>
<td>0.762</td>
<td>0.810</td>
</tr>
<tr>
<td></td>
<td>(0.099)</td>
<td>(0.038)</td>
</tr>
</tbody>
</table>

Note: This table reports estimates of $\alpha$ from OLS or 2SLS regressions of the form $\ln \delta_i = \alpha \ln Q_i + \ln R_i + \ln w_i + u_i$ using core-based statistical areas (CBSAs) as the geographic units. The $R_i$ covariate is Medicare’s Geographic Adjustment Factor, the $w_i$ covariate include mean average property value (two bedrooms) and mean annual earnings for non-healthcare workers, and we instrument for $\ln Q_i$ using the designated instrumental variables. The “no controls” designation indicates the $R_i$ and $w_i$ covariates are omitted. The “no diag” designation indicates that $Q_{ii}$ observations were omitted when estimating $\ln \delta_i$ using the gravity equation. Standard errors are robust.
Table B.3: Higher-income patients are more willing to travel: Procedure-level estimates

<table>
<thead>
<tr>
<th>Procedure</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25min visit</td>
<td>cataract removal</td>
<td>knee joint repair</td>
<td>heart artery bypass</td>
<td>gallblader removal</td>
</tr>
<tr>
<td>Distance (log)</td>
<td>-2.060</td>
<td>-2.235</td>
<td>-2.217</td>
<td>-2.253</td>
<td>-2.116</td>
</tr>
<tr>
<td></td>
<td>(0.0764)</td>
<td>(0.0796)</td>
<td>(0.0903)</td>
<td>(0.0872)</td>
<td>(0.0853)</td>
</tr>
<tr>
<td>Distance (log) × income tercile 2</td>
<td>0.0912</td>
<td>0.157</td>
<td>0.141</td>
<td>0.120</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>(0.0613)</td>
<td>(0.0682)</td>
<td>(0.0728)</td>
<td>(0.0847)</td>
<td>(0.0712)</td>
</tr>
<tr>
<td>Distance (log) × income tercile 3</td>
<td>0.197</td>
<td>0.296</td>
<td>0.199</td>
<td>0.405</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>(0.0759)</td>
<td>(0.0835)</td>
<td>(0.0916)</td>
<td>(0.0925)</td>
<td>(0.0900)</td>
</tr>
<tr>
<td>Observations</td>
<td>271,728</td>
<td>268,400</td>
<td>262,352</td>
<td>239,760</td>
<td>250,800</td>
</tr>
<tr>
<td>Patient market-income FE &amp; Provider market FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Two-way clustered standard errors in parentheses

Note: This table reports the coefficient on log distance for each income tercile from gravity regressions estimated separately for five procedures: 25 min office visit (HCPCS 99214), cataract removal (66984), knee joint repair (27447), heart artery bypass (33533), and gallblader removal (47562). Each model 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 data are computed from the Medicare 20 percent carrier Research Identifiable File. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.