Market Size and Trade in Medical Services*

Jonathan I. Dingel
Columbia University,
CEPR, and NBER

Joshua D. Gottlieb
University of Chicago
and NBER

Maya Lozinski
Massachusetts General Hospital

Pauline Mourot
Boston University

June 2026

Abstract

We uncover substantial interregional trade in medical services and investigate the contribution of regional increasing returns. In Medicare, one-fifth of production involves a doctor treating a patient from another region. Larger regions produce greater quantity, quality, and variety of medical services, which they "export" to patients from elsewhere, especially smaller regions. We estimate regional quality based on patients' travel choices, and these revealed-preference estimates align with clinical outcomes. Demographic-based instruments and a case study show that greater local demand increases quality and exports, consistent with economies of scale. We study counterfactual policies that lower travel costs rather than relocating production.

Keywords: healthcare access, market-size effects, Medicare claims data, trade in services
JEL codes: F12, F14, I11, R12

*We thank Scott Blatte, Aidan Buehler, Steve Buschbach, Cedric Elkouh, Aaron Haefner, Feng Lin, Luke Motley, Nick Niers, Noah Sobel-Lewin, Katherine Papen, Vaidehi Parameswaran, Xiaoyang Zhang, and Mingzhe Zhou for excellent research assistance. We are grateful to Rodrigo Adão, Zarek Brot-Goldberg, Darren Bryan, Don Davis, Rebecca Diamond, Liran Einav, Ed Glaeser, Tali Han, Tom Holmes, Loukas Karabarbounis, Wojciech Kopczuk, Doug Miller, Sean Nicholson, Bruce Schakman, Martin Schneider, Bradley Setzler, Jon Skinner, Felix Tintelnot, discussants Jan David Bakker, Barthelemy Bonadio, Jessie Handbury, Naomi Hausman, Tim Layton, Elena Patel, and many seminar participants for helpful feedback. We thank the Becker Friedman Institute at the University of Chicago for funding support and enabling us to access the Medicare claims data. We thank Antoine Levy and Jacob Moscona for sharing CBSA-level bedrock-depth data and Amitabh Chandra, Maurice Dalton, and Doug Staiger for hospital quality data and code. Dingel thanks the Cohen and Keenoy Faculty Research Fund at the University of Chicago Booth School of Business for support. Lozinski thanks the NIH (T32GM007281 and T32AG051146). j.dingel@columbia.edu, jgottlieb@uchicago.edu, mayalozinski@uchicago.edu, pmourot@bu.edu.

Rural Americans have worse health outcomes (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 discussions of whether physicians are geographically "maldistributed" go back decades (Newhouse et al., 1982*a*; Skinner et al., 2019). We consider an alternative explanation: rather than mismatch, do these patterns reflect the benefits of specialization and trade in an industry featuring scale economies?

We find substantial interregional trade in medical services: one-fifth of Medicare production involves a doctor treating a patient from another region. Larger regions produce more medical services than they consume, serving patients from smaller regions. We apply a trade model to these patient flows and find that scale economies explain larger regions' production and export of higher-quality medical care. Smaller regions' residents travel to see specialized and more experienced physicians, who are concentrated in larger markets. After quantifying regional economies of scale and patients' willingness to travel, we use the estimated model to explore counterfactual policies aimed at improving access for patients in smaller regions.

When medical services have increasing returns to scale, there are benefits to geographically concentrating production. Absent trade, however, the only way to serve patients in smaller regions would be to disperse production across space, forgoing the benefits of scale.[1] This is likely necessary for emergency care, which is about 5% of medical spending (Dieleman et al., 2020). But the vast majority of medical services are not emergency care. For example, if patients with cancer can travel across regions in search of the ideal oncologist—one specialized in their particular type of cancer, one with a better reputation, or simply a better personal match—the economic geography of medical care may resemble other trad-

[1]Many economists assume trade costs for medical services are prohibitively high. Hsieh and Rossi-Hansberg (2023): "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." Joskow (1980) recognizes the potential for trade and articulates a closely related tradeoff for hospital scale.

able industries. In that case, society would face a *proximity-concentration tradeoff*: patients who consume medical services produced far from home incur trade costs yet benefit from the higher quality generated by geographically concentrating production.

We investigate these economic forces using Medicare claims data that we introduce in Section 1. Medicare is the federal government's insurance program for the elderly and disabled and spends four percent of US GDP. Medical service providers submit claims that report the specific medical procedure provided, the treatment location, and where the patient lives. Using hundreds of millions of these claims, we find that "imported" medical procedures, defined as a patient receiving a service provided by a physician in a different region, constitute about one-fifth of healthcare consumption. Production is geographically concentrated in larger regions, while consumption is much less so. This contrast means exported medical services are disproportionately produced in larger regions and imports are a larger share of consumption in smaller regions. Although some patients travel thousands of kilometers for care, patient flows decline rapidly with distance between regions.

We show that specialized medicine is concentrated in larger markets and trade expands the population benefiting from specialization: imported care is more likely to be performed by specialists, by appropriate specialists, by more experienced physicians, and using scarce capital equipment, especially for smaller regions' imports. Larger regions also perform a greater diversity of procedures. Rare procedures are traded more and over longer distances. For example, half of the patients having left ventricular assist devices implanted to restore their heart function come from outside the surgeon's region, while only 10% of screening colonoscopies are traded. While these patterns establish an important role for regional specialization and trade, they do not reveal the strength of regional increasing returns or the consequences of healthcare agglomeration.

To guide our investigation of these mechanisms, Section 2 develops a model of trade in medical services. We adapt standard models to a setting in which the government sets prices, so quality variation clears markets. The model delivers equations that we use to

estimate regional service quality, the strength of increasing returns, and market-size effects. Our benchmark exposition assumes external economies of scale, but we show that this model is isomorphic to models featuring division of labor and lumpy capital, which match observed regional variation in the practice of medicine. If there are regional scale economies, larger markets produce higher-quality care and export it. When economies of scale are sufficiently strong, the model predicts a *strong home-market effect*: greater demand makes larger regions *net* exporters of medical services. Because market size matters more at smaller scales, the model predicts less common medical procedures will respond more to differences in market size. The gravity model of trade flows provides a framework to test this hypothesis, as in Costinot et al. (2019).

We recover revealed-preference estimates of regional service quality by estimating patients' willingness to travel to each exporting region for medical services. Regions that attract patients from farther away are revealed to have higher-quality care. These estimates based on trade flows align with conventional measures of quality: they are negatively related to external measures of mortality rates and positively related to other hospital rankings and safety scores. The top-ranked region by revealed preference is Los Angeles, CA, and the top-ranked relative to its size is Rochester, MN, home of the Mayo Clinic.

Service quality is central to numerous healthcare analyses and depends on many dimensions of care, not all of which are captured by clinical outcomes. Revealed preference has the advantage of incorporating whatever broader notion of quality patients care about, including hospital amenities or physician behavior. In many settings, researchers learn from consumers facing quality-price tradeoffs. In healthcare, patients are often insulated from price variation. Our gravity-based revealed-preference approach can be applied whenever patients trade off quality and distance.

We use these revealed-preference estimates to estimate the scale elasticity: how does serving more patients raise a region's quality of medical care? The main threat to identifying scale economies is confounding supply shifters: larger markets may have exogenous advan-

tages in producing medical services. For example, urban amenities might attract doctors at lower wages (Lee, 2010). We use four empirical strategies, detailed in Section 3, to isolate variation in output caused by demand shifters: instruments for regional population; changes over time; regional demographics that predict demand variation across procedures; and a case study of mesothelioma, a rare cancer caused by asbestos exposure.

Section 4 finds substantial regional increasing returns. We estimate the scale elasticity to be about 0.8: a region producing 10% more because of greater demand delivers about 8% higher quality. This result remains when estimated in the cross-section, a panel, or instrumented with historical characteristics that predict current production levels. Using the interaction of regional demographics and procedure-specific elderly demand further supports the presence of scale economies. Our mesothelioma case offers additional evidence, since asbestos exposure is not likely correlated with amenities specifically attracting doctors who treat mesothelioma. It yields a scale elasticity estimate of 0.6, similar to the aggregate result. Consistent with theoretical microfoundations for such scale economies, larger markets have finer division of labor and disproportionately host specialized capital inputs.

The same trade flows show that regional increasing returns are large enough to explain the geographic patterns of production, consumption, and trade documented in Section 1. We find a strong home-market effect in medical services: greater demand induces a greater increase in exports than imports, making larger markets net exporters of medical care. These scale effects cannot be attributed to larger markets having lower input costs or medical production raising population size. Consistent with the model, the home-market effect is stronger for less common procedures: a larger residential population drives a greater increase in net exports for rare procedures.

Rural healthcare is a salient and persistent issue for federal and state healthcare policies, as the $50 billion Rural Health Transformation Program highlights (CMS, 2025). The availability of trade that we have documented expands the set of policies that can be considered for improving healthcare access, and reducing geographic disparities in that access, beyond

the standard approach of increasing payments for rural providers. Section 5 quantitatively explores the consequences of two types of policies: subsidizing healthcare production in rural areas, or subsidizing rural patients' travel to receive care. Per $10 million spent, we find that travel subsidies generate a greater increase in lives saved and in patient market access than production subsidies. This is true whether or not we include scale economies—most of the gains come from trade without any assumption about causal agglomeration effects—but the difference is larger when we incorporate scale economies.

The higher-quality care available in larger markets may not benefit all patients equally. We find that patients with lower socioeconomic status (SES) are less likely to travel farther for high-quality care, even for the same billing code. Thus, the proximity-concentration tradeoff varies by SES; the pattern of aggregate returns masks important heterogeneity. The travel subsidy generates a larger increase in market access and survival for low-income patients. Finally, we show how geography may contribute to the striking relationship between regional income and health in the United States.

This paper builds on research in urban, trade, and health economics. Urban economists have documented skill-biased agglomeration in production as knowledge workers have become more numerous and concentrated in skilled cities (Berry and Glaeser, 2005; Moretti, 2011; Diamond, 2016; Davis and Dingel, 2020; Eckert, Ganapati, and Walsh, 2020). Connecting this to the production and trade of services has been more difficult. Most studies of the geography of services analyze restaurants and retailers (Davis et al., 2019; Agarwal, Jensen, and Monte, 2020; Allen et al., 2021; Burstein, Lein, and Vogel, 2024; Miyauchi, Nakajima, and Redding, 2025). We show that—even in a service-based economy—the sizes of both local and potential export markets influence production and quality. This suggests that healthcare can serve as an export base for large markets (Bartik and Erickcek, 2007).

The trade literature has examined market-size effects in manufacturing but investigated services much less. Davis and Weinstein (2003), Hanson and Xiang (2004), and Bartelme et al. (2025) link manufactures' market size to export patterns, in line with the home-

market effect of Krugman (1980) and Helpman and Krugman (1985). Dingel (2017) shows that market-size effects drive quality specialization across US cities. Market-size effects for pharmaceuticals have been estimated using demographic variation over time (Acemoglu and Linn, 2004) and across countries (Costinot et al., 2019). Services are much less studied, in part because of the paucity of reliable trade data (Lipsey, 2009; Muñoz, 2024). We advance this literature using the detailed procedure and location information in medical claims data.

The importance of medical care for health and welfare generates substantial public-policy interest. Rural locations have worse health outcomes but fewer doctors per capita. Newhouse et al. (1982*a*,*b*,*c*), Newhouse (1990), and Rosenthal, Zaslavsky, and Newhouse (2005) considered this issue and argued against targeting a uniform geographic distribution of physicians. Building on this, we measure interregional trade in medical services, estimate the impact of geography on patient access, and connect this trade to economies of scale. Modern trade theory guides our modeling, estimation strategy, and counterfactual policy analysis.

1 Empirical setting and geographic patterns

This section describes the data and core facts about the geography of healthcare production and consumption. We show that production and consumption volumes vary systematically with market size, trade between regions declines with distance, and the patterns of regional specialization and trade suggest that regional scale influences healthcare consumption. These facts motivate our subsequent model and empirical analysis, and provide core patterns with which any model of geographic patterns—as well as health policy—must contend.

1.1 Empirical setting

Our primary dataset is claims data from Medicare, the largest source of medical spending in the United States.[2] Medicare is the federal government's insurance program for the elderly

[2]Appendix A.1 contains more details about our data and processing.

and disabled. It does not directly employ physicians or run its own hospitals. Instead, it pays bills submitted by independent physicians, physician groups, hospitals, and other medical service providers. These bills—called "claims" in industry terminology—report the specific services provided using 5-digit codes from the Healthcare Common Procedure Coding System (HCPCS). There are over 12,000 distinct HCPCS codes, which identify individual procedures at a granular level.[3] In alternative analyses, we use groupings of patient *diagnoses* to account for potential substitution between treatments.

Federal regulation, not pricing decisions by physicians or hospitals, determines the payment for each claim. For physician payments, Medicare sets procedure-specific "reimbursement rates" largely independent of quality, quantity, or region.[4] We remove regional price differences from our spending measure as described in Appendix A.2. This produces an expenditure measure purged of spatial variation in reimbursement rates. Patients pay a share of these reimbursements through copayments and deductibles, but these cost-sharing rules are constant across regions, and most Medicare patients have a Medigap supplemental or Medicaid insurance that covers most or all of this cost sharing (Cabral and Mahoney, 2019).

Medicare claims report the ZIP codes of both the place of service and the patient's residence.[5] Using these, we construct a trade matrix for medical services. We study all medical care provided by physicians, including both professional and facility charges, outside an emergency room or skilled nursing facility. Because Medicare rarely reimbursed telehealth before

[3]Codes distinguish between, *e.g.*, flu vaccines that protect against three or four strains of flu, whether administration is intramuscular or intranasal, and patient ages. There are distinct codes for chest X-rays based on whether the images are of ribs, the breastbone, or the full chest, both sides or one side of the body, and the number of images taken (1, 2, 3, or 4+). By contrast, all health services are aggregated into one of the nine service sectors available in Canadian data on interprovincial trade (Anderson, Milot, and Yotov, 2014) and the seventeen service sectors in international trade data (Borchert et al., 2021).

[4]While Medicare does have some quality incentive programs, the money at stake is a small share of Medicare's overall spending (Gupta, 2021). Medicare has some spatial variation in physician reimbursements, but it is not very large, and has diminished over time; details on payment rules are in Clemens and Gottlieb (2014). Hospital payments also reflect local wage indices, the hospital's occupational composition, and other factors. Gottlieb et al. (2010) show that price variation does not explain spatial variation in expenditure.

[5]One natural concern is measurement error in Medicare's records of patients' residences. Appendix C.1 demonstrates that our main results are robust to excluding states with large seasonal populations and examines how far dialysis patients appear to travel. We find that residential measurement error is limited and does not drive our results.

2020, this trade involves traveling to receive care delivered in-person.[6] We aggregate the ZIP-code-level information up to 306 hospital referral regions (HRRs), which are geographic units defined by the *Dartmouth Atlas of Health Care* to represent regional markets for tertiary medical care. HRRs are constructed by aggregating residential areas based on where patients were referred for major cardiovascular surgical procedures and for neurosurgery. Each HRR has at least one city where both categories of surgery were performed. Thus, the construction of these geographic units should tend to minimize trade between HRRs, as shown in Appendix Table E.1.[7] We construct HRR-to-HRR trade flows by interpreting the patient's residential HRR as the importing region and the service location's HRR as the exporting region.[8]

Physicians, hospitals, pharmacies, and other healthcare providers submit different types of claims. We use Traditional Medicare (TM) claims data, primarily from 2017, from hospitals, physicians, and outpatient care providers. One year of data includes around 360 million services, representing $230 billion in spending. These claims are not perfectly representative of all US healthcare, because Medicare beneficiaries are elderly or disabled and may choose among more providers than patients whose private insurance imposes geographic restrictions. But the geographic distribution of Medicare beneficiaries is quite similar to the overall population, and Medicare alone finances one-fifth of medical spending. These data likely capture the key features of overall healthcare production and consumption, but we cannot determine whether the geographic patterns we document generalize to the broader population.[9] Ap-

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

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

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

[9] At the level of procedures, we validate procedure frequencies in our 20% sample of physician services with two alternative data sources: more comprehensive Medicare data (based on all Traditional Medicare patients, but without information on patients' locations) and Health Care Cost Institute data for privately insured patients (Appendix A.3). At the level of physicians, we study geographic patterns of production by specialty using the National Plan and Provider Enumeration System (NPPES)—physician registry data containing the ZIP code and specialty of all physicians registered to practice in the United States (Appendix A.1).

pendix Table E.2 reports summary statistics of HRR production, consumption, and trade volumes.

1.2 Spatial variation in production and consumption

Figure 1 shows maps of healthcare production and consumption across regions, based on the place of service and patient's residential address, respectively. The consumption map shows substantial variation that has been well documented by the *Dartmouth Atlas* and related literature on geographic variation in healthcare (Fisher et al., 2003*a*,*b*; Finkelstein, Gentzkow, and Williams, 2016). The production map shows even more pronounced variation: more production in large urban agglomerations and less in rural areas. There is substantial variation in production even between immediate neighbors, while consumption varies more smoothly.

The subsequent panels show patterns of trade. Nationally, 19% of production is exported to a patient in another HRR, and 21% is traded between CBSAs.[10] Panel 1(c) shows the ratio of production to consumption; a value larger than one means an HRR is a net exporter. Net-exporting regions tend to be major urban agglomerations, plus places such as Rochester, Minnesota and Durham, North Carolina that specialize in healthcare. Panel 1(d) shows gross exports as a share of local production for each HRR. Three-quarters of services produced in the Rochester metropolitan area, home to the top-rated Mayo Clinic, are provided to patients from other regions, who travel an average of 545 km to Rochester.[11] As a major healthcare exporter with a population of merely 220,000, Rochester is an outlier: larger regions are responsible for a disproportionate share of medical services production.

[10] 42% of physician production is exported outside the provider's Hospital Service Area (HSA), a smaller *Dartmouth Atlas* geographic unit. A nationally representative survey covering people regardless of insurance provider reports that about 18% of inpatient stays not originating in the emergency department occur outside the patient's home HRR or CBSA, and 10% of outpatient care physician office visits occur outside the patient's home CBSA (Decker, Gourevitch, and Meille, 2024). These shares are based on visits, not spending, and when we construct comparable measures for Traditional Medicare, we find slightly larger magnitudes: 23% of inpatient stays and 12% of outpatient physician office visits occur outside a patient's home region. For comparison, 6% of employees work outside their commuting zone of residence (Monte, Redding, and Rossi-Hansberg, 2018, Appendix Table B.1). For manufactured goods, we obtain a 68% export share for metropolitan areas covered by the Commodity Flow Survey.

[11] These top hospitals tend to have top physicians; Mourot (2025) documents positive assortative matching.

Figure 2 plots the average production and consumption per beneficiary across HRRs of different sizes. Production per beneficiary increases with population, with an elasticity of 0.05. Consumption per beneficiary is virtually uncorrelated with population; in fact, the estimated elasticity is slightly negative. The difference between production and consumption is net trade: larger markets are net exporters and smaller markets are net importers. Gross trade flows exceed net trade flows, with imports comprising about one-third of consumption in the smallest regions. Imports per beneficiary decline with an elasticity of -0.35 with respect to population. Exports per beneficiary are approximately flat, which means total exports are increasing with local population.

1.3 Bilateral trade and bilateral distance

Despite the clear patterns in Figure 2, geographic variation in trade is far from entirely explained by market size. The regions with the lowest export shares are Anchorage, Honolulu, and Yakima, Washington, likely reflecting their remoteness. To account for these geographic patterns, we examine bilateral trade flows.

Figure 3 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 care provided in the patient's home region and in other regions. Within HRRs, there is a narrow distribution of distances that peaks around 17 km. When visiting providers in a different HRR, patients travel a great variety of distances. There is a local plateau between approximately 30–100 km, suggesting a fair amount of travel to nearby HRRs, perhaps indicating regional medical centers. There is substantial long-distance travel for care: many patients travel hundreds or thousands of kilometers.[12]

Panel 3(b) shows that trade declines with distance. The blue curve depicts trade volume

[12]Patients' choices to travel these distances underpin our revealed-preference estimates of regional service quality in Section 4.1. The average patient coming from another region travels 640 km to Chicago and 740 km to New York City, compared with less than 146 km to Urbana-Champaign, Illinois or Charlottesville, Virginia. An older literature cited in Dranove and Satterthwaite (2000) finds that patients who travel farther incur higher hospital costs.

against distance (for pairs of HRRs with positive trade flows) after removing fixed effects for each exporter and each importer.[13] This intensive-margin relationship is roughly log-linear. The red curve shows the extensive margin: the share of pairs with positive trade as a function of distance. This is 100% for nearby pairs and under 80% for the most distant pairs. These patterns motivate the inclusion of distance covariates in our gravity-based analysis.

Patients may vary in their ability or willingness to travel, especially by socioeconomic status. We quantify it here, to the extent feasible in our data, to enrich counterfactual analysis and interpret welfare implications. Panel 3(c) depicts distance elasticities—how fast bilateral trade declines with the distance between regions—estimated separately by neighborhood income decile.[14] We find a strong, nearly monotonic relationship between socioeconomic status and the distance elasticity: patients from the highest neighborhood-income decile exhibit a distance elasticity 25% smaller than those in the lowest decile.[15] This means patients from higher-income neighborhoods are more amenable to traveling farther for medical care. If society faces a proximity-concentration tradeoff, this tradeoff varies by socioeconomic status. This is especially notable given the empirical setting: Medicare insures the near-universe of elderly and disabled Americans.

1.4 Spatial variation in production and consumption by frequency

We next explore differences in outputs and production patterns in larger markets and ask if trade expands the geographic scope of patients benefiting from scale effects.[16] Figure 4(a) shows that larger regions produce a wider variety of procedures. It plots the number of distinct procedures produced in each region in seven procedure categories against regional

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

[14] We estimate these distance elasticities using equation (11) as described in Section 3.

[15] These estimates are consistent with the interaction that Silver and Zhang (2026) estimate between income and distance to care. These differences in distance elasticities are not driven by differences in the composition of procedures. When we estimate elasticities separately for rare and common services—or even for individual procedures (see Appendix Table E.3)—the income gradient of distance elasticities persists.

[16] This section focuses on professional fees for which payments are made at the service level. We observe 210 million claim lines representing $89 billion in spending on professional services.

population. Larger regions produce a greater variety of care in all seven categories.[17]

To understand what types of procedures are provided in larger markets, we estimate each procedure's population elasticity of production per Medicare beneficiary.[18] Let Q_{pi} denote the count of procedure p produced in region i and M_i the number of Medicare beneficiaries residing in i. For each procedure p, we estimate the following relationship across regions:

$$\ln \mathbb{E}\left[\frac{Q_{pi}}{M_i}\right] = \zeta_p + \beta_p \ln \text{population}_i. \tag{1}$$

The estimated elasticity, $\hat{\beta}_p$, describes how production varies with market size, and we estimate it using Poisson pseudo-maximum-likelihood.[19] If the quantity produced were simply proportional to population, β_p would be zero. We relate these population elasticities to national procedure frequency to determine whether scale is more important for rarer procedures.

Production per beneficiary indeed rises with market size, especially for less common procedures. The blue dots in Figure 4(b) relate the population elasticity of production per beneficiary $\hat{\beta}_p$ to the procedure's national volume, $Q_p = \sum_i Q_{pi}$. Across values of Q_p, procedure output per beneficiary increases with market size. Less common procedures have higher elasticities, consistent with stronger economies of scale in rare procedures.

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, we separately estimate the population elasticity of *consumption* per beneficiary for each procedure. Let G_{pi} denote the count of procedure p consumed by patients *residing* in region i. We then estimate an analogue of equation (1) where the dependent variable is $\ln \mathbb{E}\left[G_{pi}/M_i\right]$, and denote by β_p^C

[17] Appendix Figure E.1(a) shows that this is true for consumption, but to a lesser extent: Aggregating across all seven categories, HRRs in the top decile of population produce 3 times, but consume only 2 times, as many unique procedures as those in the bottom decile.

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

[19] 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.

the resulting coefficient. If $\beta_p^C \neq \beta_p$, there is size-predicted net trade in procedure p.

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. The red squares in Figure 4(b) plot the population elasticity of consumption per beneficiary $\hat{\beta}_p^C$ of a procedure against its national volume $\ln Q_p$. While the relationship is negative, the slope for consumption is only one-third that for production. Panel (c) plots the share of consumption that is imported for each region among procedures more and less common than the median frequency. The import shares are notably higher for rare procedures.

Table 1 reports the production, consumption, and trade patterns for two exemplar procedures: screening colonoscopy and left ventricular assist device (LVAD) implantation. Colonoscopies are common and geographically dispersed, while LVAD procedures are rare, geographically concentrated, and traded over longer distances.

We have thus far characterized the geography of healthcare in terms of specific medical services. An alternative view is that patients have a particular medical condition that requires treatment, but the patients may not know what particular care they need; they simply know they require care. As physicians might use different treatments across regions for the same condition, our procedure-level results could reflect substitution across procedures. We address this by conducting a similar analysis across diagnoses.

Figure 4(d) shows production and consumption elasticities by diagnosis, rather than by procedure. The key patterns remain similar: production elasticities are higher, and decline more rapidly with national patient volume, than consumption elasticities. Both sets of elasticities have less steep relationships with national volume than for procedures. This could reflect measurement error within each category: the 482 diagnosis categories we use are far coarser than the 8,210 procedures in Figure 4(b). 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 4 imply trade in medical services between markets of different sizes. Larger markets

tend to 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.

1.5 Travel to access specialized services

We next ask whether the distribution of specialist physicians helps explain trade. Figure 5(a) shows the share of imports and of locally produced consumption that are provided by specialists as a function of regional population.[20] Imports are significantly more specialist-intensive than local production. This difference is especially pronounced in the smallest regions, and it remains true throughout the population distribution.

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

Figure 5(b) shows the share of imports and of locally produced care provided by non-standard specialties. Imports are less likely to come from non-standard specialists than local care, and the distinction is especially pronounced in the smallest regions. The difference is substantial: Local care in the smallest regions is 40% more likely to be provided by a non-standard specialist than in the largest regions (7% vs. 5%). When importing medical services, this share falls to 5%—indistinguishable from the largest regions' locally produced care.

We conduct a similar analysis based on provider experience. Using the public Medicare provider data (based on all Traditional Medicare patients), we count the number of times

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

the physician billed for the specific service in the previous year. We divide this experience measure by the procedure's national mean and average it across all procedures provided to patients in an HRR. We then rescale HRR-level experience by the mean across HRRs. Figure 5(c) shows that, at all population sizes, care imported from other regions is produced by more experienced providers than locally produced care.[21] Patients in larger regions see more experienced providers for both imported and locally produced care.

Specialists are disproportionately located in larger markets, as are physicians with more experience in their procedures. Since imported care is predominantly specialized and provides patients access to higher experience, we conclude that visiting the appropriate specialist based on training or experience is part of the value proposition for trade in medical care.

Figure 5(d) shows analogous patterns for rare capital equipment. We measure the share of care that requires a piece of rare capital equipment and plot these shares separately for local and imported care against regional population. While locally produced care in larger markets is more likely to use rare equipment, imported care has a higher rare-equipment share throughout the population distribution.

This analysis also suggests that trade reflects quality variation. Patients travel to regions with highly-ranked hospitals, which larger markets tend to have—along with rare equipment and the ability to provide rare services. This market-size effect strongly predicts gross and net exports. Together, this suggests that economies of scale play an important role in increasing the quality of care, and trade between regions enables patients from many regions to share the benefits of this agglomeration.

2 Theoretical framework

This section develops a model of regional trade in medical services tailored to our analysis of US healthcare. We develop a competitive model of a market for one medical procedure that

[21] This comparison restricts attention to the 143 procedures performed in all HRRs. Thus, regional variation does not reflect the fact that larger markets produce a greater number of distinct codes (Figure 4(a)).

has a fixed price.[22] With fixed prices, quality variation plays a key role in clearing markets. Patients select quality-differentiated services and face trade costs, so patient flows between regions follow a gravity equation. Regional increasing returns cause the quality-adjusted cost of producing a service to decline with scale. The model delivers equations that allow us to estimate regional service quality, the strength of regional increasing returns, and the home-market effect. We use the estimated model to quantify outcomes in counterfactual policy scenarios.

2.1 Demand

We use a logit model of individuals choosing providers for a given service. Providers and patients are in regions indexed by i or j, with $\mathcal{I}$ denoting the set of regions. All providers in a region are identical. Let N_j denote the number of potential trips by patients residing in region j.[23] A patient in region j choosing a provider in region i for trip k obtains utility

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

The provider-region component δ_i would usually include a product's characteristics and price. With fixed prices, δ_i is simply the quality of care available in region i. The region-pair component ρ_{ij} represents bilateral inverse trade costs (proximity). The idiosyncratic component ϵ_{ik} is independently and identically drawn from a standard Gumbel distribution, so the probability that the patient making trip k selects a provider in region i is

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

[22] For brevity, we present here a competitive model with external economies of scale, but the consequences of regional increasing returns for trade flows in a fixed-price environment do not hinge on this assumption. Appendix B.1 shows that a monopolistic-competition model with one medical provider in each region delivers the same predictions. Appendix B.2 presents four distinct microfoundations that all deliver a regional quality-quantity relationship featuring a constant scale elasticity. As in flexible-price models, many market structures can give rise to a home-market effect (Costinot et al., 2019).

[23] Appendix B.3 extends the model to have multiple patient types.

We allow for an outside option denoted by $i = 0$, which represents individuals choosing to forgo care, and we normalize its common component to zero, $\ln \delta_0 = \ln \rho_{0j(k)} = 0 \ \forall k$.[24]

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

$$Q_{ij} = \delta_i \frac{N_j}{\Phi_j} \rho_{ij}, \tag{2}$$

where $\Phi_j \equiv \sum_{i' \in \{0\} \cup \mathcal{I}} \delta_{i'} \rho_{i'j}$ is the expected value of the choice set for patients in region j. We call this Φ_j "patient market access," as it captures the quality of care available to patients and their costs of accessing it. The share of potential patients in region j forgoing medical care is $\frac{Q_{0j}}{N_j} = \frac{1}{\Phi_j}$. Remote regions are farther from medical providers (lower ρ_{ij}), reducing their patient market access and raising the share of patients that forgo care. Equation (2) is a gravity equation with an origin i component, a destination j component, and an ij pair component. Total demand for care produced in i is

$$Q_i = \delta_i \sum_{j \in \mathcal{I}} \frac{N_j}{\Phi_j} \rho_{ij}. \tag{3}$$

2.2 Production

We assume competitive production of services with free entry and regional increasing returns that are external to individual providers. That is, each price-taking provider chooses its output quality and quantity given total regional production, an exogenous factor price, and an exogenous productivity shifter. A provider in region i that employs L units of the

[24] This formulation of demand is familiar from the hospital competition literature, which has been surveyed by, e.g., Gaynor and Town (2011) and Gaynor, Ho, and Town (2015). Our competitive model does not distinguish between hospitals or physicians within a region.

composite input to produce service of quality δ produces the output quantity

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

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

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

In our setting, output prices are fixed: the Medicare-determined "reimbursement rate" $\overline{R}$ is independent of quality, quantity, and region. Each provider that produces output of the highest quality produced in region i earns revenue $\overline{R}$ per unit.

Provider optimization and free entry require the unit cost to equal the reimbursement rate in each region. Given factor price w_i and productivity shifter A_i, the free-entry condition

$$C(Q_i, \delta_i; w_i, A_i) = \overline{R} \tag{4}$$

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

While our assumptions thus far suffice for qualitative results, we later specify functional forms for additional predictions and empirical quantification; specifically, $K(\delta_i) = \delta_i$ and

$H(Q_i) = Q_i^\alpha$, with a scale elasticity of $\alpha \geq 0$.[25] In this case, the free-entry condition (4) is

$$\overline{R} = \frac{w_i \delta_i}{A_i Q_i^\alpha}. \tag{5}$$

This simple environment features regional economies of scale that are external to providers, but it is isomorphic to models with richer mechanisms generating regional increasing returns. In Appendix B.2, we provide four alternative microfoundations with internal economies of scale that deliver a regional quality-quantity relationship featuring a constant scale elasticity, as in equation (5). These microfoundations feature gains from specialization, fixed costs of entry, and task-level increasing returns.

2.3 Equilibrium

Equilibrium equates supply and demand in each region, $Q_i = \sum_j Q_{ij}$. Given exogenous parameters $\overline{R}$, $\{w_i, A_i, N_i\}_{i \in \mathcal{I}}$, and $\{\rho_{ij}\}_{(i,j) \in (\mathcal{I}, \mathcal{I})}$, an equilibrium is a set of quantities and qualities $\{Q_i, \delta_i\}_{i \in \mathcal{I}}$ that simultaneously satisfy equations (3) and (4).

2.4 Scale effects in autarky

In autarky, patients choose whether to receive care, but they cannot travel between regions ($\rho_{ij} = 0$ for $i \notin \{0, j\}$). In this case, all demand is local and equation (3) simplifies to

$$Q_{jj} = \frac{\delta_j \rho_{jj}}{1 + \delta_j \rho_{jj}} N_j. \tag{6}$$

[25] There are increasing returns to scale if $\alpha > 0$. The four microfoundations presented in Appendix B.2 each deliver a regional production function with scale elasticity α. This formulation is compatible with an equilibrium in which region i does not produce the procedure at all: $Q_i = 0 \implies \delta_i = 0$. The function $K(\delta_i) = \delta_i$ normalizes the cost of quality. Appendix B.4 shows this to be innocuous for all purposes except counterfactual reimbursement rate changes, for which we evaluate a more general functional form $K(\delta_i) = \delta_i^\nu$.

The autarkic equilibrium is at the intersection of the demand curve given by equation (6) and the free-entry isocost curve given by equation (4).[26] An increase in population size, $\Delta N_j > 0$, affects equilibrium outcomes by shifting the demand curve.

Figure 6 illustrates how greater demand affects quality in autarky. Panel (a) shows the role of returns to scale. The vertical axis shows quality δ_i and the horizontal axis shows quantity Q_i (on logarithmic scales). Higher quality attracts more patients, so demand is upward-sloping.[27] We show two cases of the free-entry isocost curve defined by equation (5): the horizontal line depicts constant returns ($\alpha = 0$) and the upward-sloping line depicts increasing returns ($\alpha > 0$). With constant returns, a rightward shift in demand ($\Delta N_j > 0$) causes a proportional increase in quantity produced and no change in output quality. With increasing returns, higher demand moves producers up the isocost curve. This quality improvement causes a more-than-proportional increase in quantity produced.

Panel 6(b) shows that an increase in demand raises quality more as the demand curve is increasingly elastic. The panel depicts two demand curves: the one on the left is more elastic, as we would expect for a less common procedure.[28] Shifting each demand curve to the right raises the equilibrium quality of each procedure because of increasing returns to scale. This market-size effect is larger for the less common procedure with more elastic demand because the demand shift is amplified by a larger increase in quantity demanded.[29]

[26] For the equilibrium to be Marshallian stable, the demand curve must be steeper than the isocost curve at the intersection. There is a stable equilibrium because equation (6) means $Q_{jj} \to N_j$ as $\delta_j \to \infty$.

[27] For visual clarity, we draw a log-linear demand curve. The logit demand function (6) is in fact log-convex, which is consistent with all the comparative statics illustrated in Figure 6.

[28] The demand function (6) is log-convex. This functional form therefore satisfies "Marshall's second law" that demand is more elastic at lower quantities, which has found support in the trade literature (*e.g.,* Mayer, Melitz, and Ottaviano, 2021). We apply the fixed-price analogue here.

[29] Alternatively, one could obtain this prediction by assuming that demand is log-linear and the isocost curve is log-concave, so that rare procedures have stronger scale economies. For example, introducing a fixed cost F would yield the isocost curve $\frac{w_i \delta_i}{A_i Q_i^{\alpha}} + F Q_i^{-1} = \bar{R}$. In this case, the scale elasticity $\frac{\partial \delta_i}{\partial Q_i} \frac{Q_i}{\delta_i} = \frac{\alpha \bar{R} + (1-\alpha) F Q_i^{-1}}{\bar{R} - F Q_i^{-1}}$ is decreasing in quantity produced Q_i for $\alpha \in (0, 1)$. A rightward shift in demand would cause a larger (log) difference in quality for the low-volume procedure on the steeper part of the isocost curve.

2.5 Market-size effects on trade flows

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

Fixing the qualities produced in other regions, an increase in one region's demand affects its trade flows through three mechanisms. First, greater total demand for services proportionally increases a region's demand for imports through the N_j term in equation (2). Second, with increasing returns, an increase in N_i elicits an increase in quality δ_i, which raises region i's *gross* exports to each region. Costinot et al. (2019) call this the "weak home-market effect." Third, if increasing returns are sufficiently strong, quality δ_i improves so much that region i's patient market access Φ_i rises such that $\ln \delta_i$ rises more than $\ln (N_i/\Phi_i)$ does. That is, the increase in region i's gross exports exceeds any increase in its gross imports. This is the "strong home-market effect": an increase in local demand raises a region's *net* exports.

Figures 6(c) and 6(d) introduce trade and illustrate the distinction between weak and strong home-market effects.[30] Panel (c) depicts the quality and quantity produced in one region under two scale elasticities. Comparing points B and C, a given increase in demand elicits a larger quality improvement when increasing returns are stronger. Panel (d) depicts equilibrium exports and imports as a function of the region's demand shifter N_j. The import curves slope up because an increase in local demand raises demand for imports. The export curves slope up because of increasing returns: higher local demand increases quality, which increases gross exports. This is the weak home-market effect. When the scale elasticity α is larger—the free-entry isocost curve in Panel (c) is steeper—greater demand elicits a larger increase in output quality. This steepens the export curve and flattens the import curve

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

in Panel (d). When the export curve is steeper than the import curve, there is a strong home-market effect: the increase in demand raises exports more than imports.

As in the autarkic case, we predict larger effects of market size for less common procedures. An increase in demand raises quality more when demand is more elastic, leading to a stronger home-market effect for rare procedures. If rare procedures also have greater economies of scale (higher α), that would amplify this contrast. This result motivates a research design comparing home-market effects for common and rare procedures.

These results also hold when an increase in demand in one region affects equilibrium outcomes in other regions. To demonstrate this, we examine the home-market effect in the neighborhood of a symmetric equilibrium. If all regions are the same size, $N_i = \bar{N}\ \forall i$, and trade costs are symmetric, $\rho_{ii} = 1$ and $\rho_{ij} = \rho \in (0,1)\ \forall i \notin \{0, j\}$, there is a symmetric equilibrium with quality $\bar{\delta}$ and patient market access $\bar{\Phi}$ in each region. As detailed in Appendix B.5, we totally differentiate the system of equations in terms of $\{\mathrm{d}\delta_i, \mathrm{d}N_i\}_{i=1}^{\mathcal{I}}$ and evaluate a change in one region's demand N_1 at the symmetric equilibrium.

With increasing returns of any magnitude, there is a weak home-market effect; with sufficiently strong increasing returns, there is a strong home-market effect. When $\alpha \in (0,1)$, an increase in region 1's demand $\mathrm{d}\ln N_1 > 0$ elicits a proportional increase in its relative service quality $\mathrm{d}\ln\delta_1 - \mathrm{d}\ln\delta_{j\neq 1} > 0$. This higher quality causes region 1 to export more to every other region: $\frac{\mathrm{d}\ln Q_{1j}}{\mathrm{d}\ln N_1} > 0$. The effect on the region's net exports is

$$\mathrm{d}\ln Q_{1,j\neq 1} - \mathrm{d}\ln Q_{j\neq 1,1} = \left[\frac{1 - \frac{1-\alpha}{\alpha}\frac{1+(\mathcal{I}-1)\rho}{1-\rho}}{\frac{1-\alpha}{\alpha}\frac{(1+(\mathcal{I}-1)\rho)}{(1-\rho)} + \frac{(1-\rho)\bar{\delta}}{1+(1+(\mathcal{I}-1)\rho)\bar{\delta}}}\right]\mathrm{d}\ln N_1. \tag{7}$$

Net exports increase with population size if and only if $\frac{\alpha}{1-\alpha} > \frac{1+(|\mathcal{I}|-1)\rho}{1-\rho}$. This occurs if increasing returns are sufficiently strong (α is large enough) and trade costs are sufficiently large (ρ is small enough). Otherwise, there is a weak home-market effect but not a strong one. Given a strong home-market effect, the effect in equation (7) is diminishing in the population of patients $\bar{N}$, so we predict a stronger home-market effect for rare procedures.

2.6 Regional factor supplies

Thus far, we have assumed that the composite input is supplied at price w_i. Now, suppose that, as microfounded in Appendix B.6, the isoelastic inverse regional factor supply curve is

$$w_i = \chi_i L_i^{\beta}, \tag{8}$$

where χ_i is a regional factor supply shifter and $\beta = 0$ constitutes the perfectly elastic case presented thus far.[31] In the general case, equilibrium must equate supply and demand of both medical services and factor inputs in each region, $Q_i = \sum_j Q_{ij}$ and $w_i L_i = \overline{R} Q_i$. Given exogenous parameters $\overline{R}, \beta$, $\{\chi_i, A_i, N_i\}_{i \in \mathcal{I}}$, and $\{\rho_{ij}\}_{(i,j) \in (\mathcal{I}, \mathcal{I})}$, an equilibrium is now a set of quantities, qualities, and input prices $\{Q_i, \delta_i, L_i, w_i\}_{i \in \mathcal{I}}$ that simultaneously satisfy equations (3), (4), and (8).

Introducing regional factor supplies in equation (8) reveals two points to consider when interpreting the relationship between service quality and market size. First, when increased production raises the cost of hiring inputs, the effective scale elasticity is the technological scale elasticity net of this additional cost. That is, combining equations (5) and (8) yields

$$\delta_i = A_i Q_i^{\alpha - \frac{\beta}{1+\beta}} \chi_i^{\frac{-1}{1+\beta}} \overline{R}^{\frac{1}{1+\beta}}, \tag{9}$$

so the effective scale elasticity is $\tilde{\alpha} \equiv \alpha - \frac{\beta}{1+\beta}$.[32] This effective scale elasticity is the relevant economic notion for most counterfactual scenarios and the parameter that our empirical estimates will recover.

[31] The empirical evidence suggests that healthcare supply is reasonably elastic, though not perfectly elastic as in the $\beta = 0$ case (*e.g.* Finkelstein, 2007; Cabral, Carey, and Miller, 2025; Clemens and Gottlieb, 2014; Geddes and Schnell, 2023; Gottlieb et al., 2025).

[32] Our microfoundation in Appendix B.6 implies upward-sloping factor supply curves ($\beta \geq 0$). In principle, these might be downward-sloping if, for example, physicians prefer to work at large academic medical centers that employ many other doctors. Teaching hospitals tend to be in large regions because they follow healthcare demand: more patients enable physicians to specialize, conduct research, and train medical students. A salient example is Cornell University: after attempting to have medical training in both Ithaca and New York City, the Cornell Trustees quickly closed the Ithaca location in the early 20th century (Flexner, 1910; Gotto and Moon, 2016). Downward-sloping factor supplies would increase the effective scale elasticity: greater medical demand would improve quality through both economies of scale and lower hiring costs.

Second, equation (9) illustrates the possibility that urban scale might raise output quality not only through increased demand but also by shifting regional factor supplies. Suppose that regional amenities have an exogenous component and a population-driven component: $\chi_i^{-1} = \tilde{B}_i \tilde{N}_i^{\varsigma}$, where $\tilde{N}_i$ is the region's population size. A market with greater population $\tilde{N}_i$ has greater medical demand N_i, and thus its service quality would increase as a consequence of both increased medical scale ($\alpha > 0$) and improved amenities ($\varsigma > 0$). As described in Appendix B.6, better amenities might improve output quality in two ways. If doctors were homogeneous, regions with better amenities would enjoy lower factor prices. If doctors varied in their ability and better doctors valued amenities more, a region with better amenities would produce higher quality not only via lower factor prices but also via selection on doctor ability that improves A_i. This conceptual distinction between market-size effects in the sense of greater medical demand with regional increasing returns and in the consequences of urban scale for regional factor supplies informs our empirical approach.

3 Empirical strategies

This section describes the empirical strategies we use to estimate regional quality δ_i, estimate the scale elasticity α, and detect home-market effects.

3.1 Revealed-preference measure of regional quality δ_i

Equation (2) shows that we can obtain a revealed-preference measure of each region's quality, δ_i, by looking at patient flows. Letting S_{ij} denote the value of procedures exported from region i to patients residing in j, we take expectations and then logs of equation (2) to obtain

$$\ln \mathbb{E}\left[S_{ij}\right] = \ln \delta_i + \ln\left(\frac{N_j}{\Phi_j}\right) + \gamma X_{ij}, \tag{10}$$

where we parameterize the region-pair component of preferences as $\ln \rho_{ij} = \gamma X_{ij} + \upsilon_{ij}$, with X_{ij} a vector of observed trade-cost shifters and υ_{ij} an orthogonal unobserved component.[33] The right side contains three terms: exporter quality $\ln \delta_i$, importer demand shifter $\ln\left(\frac{N_j}{\Phi_j}\right)$, and observed trade costs γX_{ij}. These trade costs may reflect the opportunity cost of patients' time, fiscal costs of travel, adverse health consequences of travel, and difficulties of coordinating medical care farther from home. Our estimates of the γ coefficients will capture all these distance-correlated costs.

To estimate each region's service quality, we replace the first two terms with exporter fixed effects and importer fixed effects, respectively:

$$\ln \mathbb{E}\left[S_{ij}\right] = \underbrace{\ln \delta_i}_{\text{exporter FE}} + \underbrace{\ln \theta_j}_{\text{importer FE}} + \gamma X_{ij}. \tag{11}$$

The exporter fixed effects are a revealed-preference measure of regional quality, which we validate using external quality metrics. The importer fixed effects, combined with an assumption about the number of potential patients (detailed in Appendix D.2), enable us to compute $\Phi_j = N_j/\theta_j$, a measure of patient market access for those who reside in location j.

To estimate equation (11), we must parameterize observed trade costs γX_{ij}. We first use log distance and a same-region dummy, *i.e.* $\gamma X_{ij} = \gamma_1 \ln \text{distance}_{ij} + \gamma_0 \mathbf{1}(i = j)$. Alternative specifications add $(\ln \text{distance}_{ij})^2$ or replace these continuous distance covariates with indicators for distance deciles. Since observed bilateral trade is zero for many pairs of regions, especially when looking at trade in individual procedures, we estimate (11) using the Poisson pseudo-maximum-likelihood (PPML) estimator (Santos Silva and Tenreyro, 2006).

[33] For much of the analysis, we define S_{ij} as total value across all care, computed as described in Section 1.1. When we study groups of procedures, such as terciles of elderly-specific care, or a specific disease, such as mesothelioma, we aggregate spending over the relevant procedures. When analyzing individual procedures, the dependent variable is the procedure count Q_{ij} rather than spending S_{ij} because no aggregation is required.

3.2 Estimating the scale elasticity α

The relationship between quality and quantity reveals the scale elasticity of the regional production function, α. Per the free-entry condition (5), quality is an isoelastic function of the quantity produced, conditional on price, cost, and productivity shifters. Rearranging equation (5) and taking logs yields $\ln \delta_i = \alpha \ln Q_i + \ln\left(\frac{\overline{R}}{w_i}\right) + \ln(A_i)$. To convert this to an estimating equation, we replace quality with its empirical estimate from equation (11), replace $\ln\left(\frac{\overline{R}}{w_i}\right)$ with a more flexible set of controls $\mathbf{X}_i$, add state fixed effects $\Xi_{s(i)}$, and define a residual u_i that incorporates $\ln(A_i)$ plus measurement error:

$$\widehat{\ln \delta_i} = \alpha \ln Q_i + \Xi_{s(i)} + \Upsilon \mathbf{X}_i + u_i. \tag{12}$$

Because our data only include Traditional Medicare patients, we do not observe the total production of medical services in each region. We thus use two different proxies for Q_i in equation (12). The first proxy we use is the Traditional Medicare production in each region, Q_i^{TM}. Traditional Medicare beneficiaries are a smaller share of the population in larger regions, and Appendix C.2 defines the resulting "observability bias" in our estimate of the scale elasticity α. We gauge the magnitude of this bias using a second proxy, the labor supply in each region L_i, which is based on all medical providers in an area and thus covers all insurance segments. We define it using a CES aggregator of doctors and nurse practitioners, as detailed in Appendix C.3. Our estimates are not sensitive to the assumed elasticity of substitution between types of medical providers.

Our goal is to estimate the slope of the isocost curve, α, but Q_i (or its proxy) may be endogenous in equation (12) for three reasons. First, supply shifters (higher A_i) would raise both the region's production volume, the regressor of interest, and its output quality, the dependent variable. Second, a region's success in exporting medical services, with local hospitals serving as major "anchor institutions," may increase local labor demand and raise current population size. Third, larger markets might have both higher demand and lower

input costs, for instance because physicians prefer to live in large markets due to urban amenities as outlined in Section 2.6, which could bias the estimated scale elasticity upward.

To address these concerns, we use a range of controls and then introduce four empirical strategies beyond OLS estimation of equation (12). First, we include state fixed effects to account for state policies, such as certificate of need laws and licensure requirements, that may affect both quality and quantity of care in a state. Second, we introduce controls intended to capture R_i and w_i to mitigate remaining concerns about population being correlated with supply shifters. Section 4 will also show that input costs don't appear to fall with market size. We also present results produced using a version of $\widehat{\ln \delta_i}$ estimated while excluding the diagonal elements of the trade matrix.

Regional-variation instruments. To address the first concern—supply shifters that impact both quality and quantity—we instrument for local production with local population. Population predicts healthcare production, since larger populations require more care.[34]

Second, we address reverse causality by using either historical (1940) population or bedrock depth as instruments. In 1940, medicine was a far smaller industry and did not drive local population in the way it might today. Because local populations persist, population in 1940 predicts current population and thus current healthcare demand. Bedrock depth uses local geology, which is arguably not influenced by local medicine industry success, to predict healthcare demand using its effect on construction costs, ultimately influencing current population density (Rosenthal and Strange, 2008; Levy and Moscona, 2022).[35]

As Section 2.6 and Appendix B.6 demonstrate, physician sorting on amenities correlated with regional population is a key potential confounder when estimating α. To tackle this concern, we use the elderly share of a region conditional on regional population as an instru-

[34]The variation in demand predicted by population is overwhelmingly driven by headcount, not income per capita: the population elasticity of income per capita is below 0.1, and the income elasticity of health spending is likely below 1.0 (Acemoglu, Finkelstein, and Notowidigdo, 2013).

[35]This instrument is currently only available for CBSAs and CZs, not HRRs. Our main results hold when defining regions as CBSAs and CZs. Levy and Moscona (2022) show that the bedrock instrument has ample first-stage power for predicting CBSA population density; the same is true for our endogenous variables (population levels).

ment for demand for healthcare services in a region. We define the elderly share of a region as the share of the population over 65. Since the elderly tend to consume more healthcare, a higher share of elderly residents in a region will increase the total demand for healthcare services in that region. The share of the population over 65, conditional on total population, is orthogonal to scale-driven supply shifters, such as scale-dependent regional amenities that would attract better physicians.

This approach assumes that amenities that disproportionately attract the elderly are orthogonal to the amenities that disproportionately attract doctors conditional on population. This assumption is plausible for two main reasons. First, prior literature on elderly sorting and mobility suggests that seniors and the working-age population move in opposite directions on average along the population density dimension (Badilla Maroto et al., 2024). High-human-capital workers such as physicians are even more likely to have different sorting patterns from seniors (Chen and Rosenthal, 2008), suggesting that, if anything, this mobility would tend to bias our estimated scale economies downward. Second, we empirically test for a relationship between the elderly share and doctor ability. We use MCAT scores of the medical schools attended by doctors practicing in a region as a proxy for doctor ability. While Appendix Figure E.2(a) shows that average MCAT scores increase with regional population, Appendix Figure E.2(b) shows no relationship between regional MCAT scores and the elderly share conditional on population level, strengthening the idea that the demographic instrument addresses physician sorting.

Procedure- and regional-variation instrument. To further tackle the concern of physician sorting across regions, we use the interaction between each region's elderly population share and procedures' Medicare share as a shifter for local demand for the procedure. Demand for elderly-specific procedures will be higher in regions with a larger share of elderly residents. In this empirical strategy, physicians can sort—even on amenities correlated with regional elderly share, conditional on population—as long as physicians who tend to perform elderly-specific procedures do not differentially sort on these amenities.

We define a procedure as "elderly-specific" using the share Medicare represents in the total procedure frequency. Details on how we construct this variable are in Appendix A.1. We aggregate procedures p by whether they are above the median elderly specificity, labeling these broad aggregates as procedure groups g. For each procedure group, we estimate the region $\times$ procedure-group quality $\ln \delta_i^g$ using

$$\ln \mathbb{E}\left[S_{ij}^g\right] = \ln \delta_i^g + \ln \theta_j^g + \gamma^g X_{ij}. \tag{13}$$

We then estimate the reduced-form effect of scale on quality using

$$\widehat{\ln \delta_i^g} = \Lambda(\text{above-median elderly specificity}_g \times \ln \text{elderly share}_i) + \eta_g + \eta_i + u_{ig}, \tag{14}$$

where η_g and η_i are procedure group and region fixed effects, respectively. We estimate another version where g indexes terciles of Medicare share rather than halves. With this empirical strategy, we recover the reduced-form estimate $\hat{\Lambda}$ because we only observe Traditional Medicare production. This precludes us from estimating the first-stage impact on quantities produced inside and outside the TM population necessary to estimate α with this strategy.

Panel variation. Physicians may sort across regions based on amenities correlated with population size, which would bias our estimates of α upward when using the population instruments. The predominant concerns related to amenities valued by physicians correlated with city size are likely stable over a modest time frame since population sizes are persistent. If these are time-invariant, we can difference them out using panel variation. Using Δ to denote differences from 2013 to 2017, we estimate

$$\widehat{\Delta \ln \delta_i} = \alpha \Delta \ln Q_i + \Xi_{s(i)} + \Upsilon \Delta \mathbf{X}_i + \Delta u_i. \tag{15}$$

Mesothelioma. Finally, we use variation in mesothelioma, a rare type of cancer caused by past exposure to asbestos. Mesothelioma offers an ideal case study to identify scale economies

for two key reasons. First, it is caused almost exclusively by past exposure to asbestos. Asbestos use peaked in the 1970s, so Medicare beneficiaries exposed to asbestos in those years would be consuming medical care during our period of study (Attanoos et al., 2018). These beneficiaries comprise a substantial share of mesothelioma care consumption: 82% of mesothelioma deaths are from people aged 65 and older. Most importantly, asbestos exposure decades ago is plausibly orthogonal to amenities that might drive sorting of skilled physicians today conditional on population. Second, mesothelioma is an aggressive cancer with extremely low survival rates. This means that patients with mesothelioma are very unlikely to be undiagnosed and that we can proxy for past exposure to asbestos using death rates from mesothelioma. Mesothelioma is rare, so to increase power we pool 2016–2019 data, as detailed in Appendix A.1.

We leverage residual variation in demand for mesothelioma care conditional on regional population to address concerns about physician sorting related to market size. Appendix Figure E.3(a) shows that mesothelioma death rates are unrelated to our quality estimates for aggregate medical services, conditional on population. To violate our main assumption, a supply shifter would need to be correlated with mesothelioma death rates conditional on population to raise mesothelioma quality but not the average quality of all other medical services. This seems unlikely, and is therefore reassuring for this empirical strategy. Appendix Figure E.3(b) shows that the variation in mesothelioma death rates we leverage is relatively geographically dispersed after controlling for population and our regional controls.

We estimate the scale elasticity α^{meso} using

$$\widehat{\ln \delta_i^{\text{meso}}} = \alpha^{\text{meso}} \ln Q_i^{\text{meso}} + \tau \ln \text{population}_i + \Xi_{s(i)} + \Upsilon \mathbf{X}_i + u_i. \tag{16}$$

We estimate exporter fixed effects $\widehat{\ln \delta_i^{\text{meso}}}$ using a version of equation (11), with the sample restricted to mesothelioma care. We instrument for the log quantity of mesothelioma care produced in a region $\ln Q_i^{\text{meso}}$ with the log of mesothelioma death rates.

3.3 Test for the home-market effect

Having estimated increasing returns, we next ask whether they explain the pattern of trade described in Section 1. We do this using a more parsimonious gravity regression, treating each region's population as a demand shifter. Following Costinot et al. (2019), we differentiate the system of equations (3) and (4) around the symmetric equilibrium and replace $\ln \delta_i$ and $\ln\left(\frac{N_j}{\Phi_j}\right)$ in (10) with log population in the producing and consuming regions, respectively:

$$\ln \mathbb{E}\left[S_{ij}\right] = \lambda_X \ln \text{pop}_i + \lambda_M \ln \text{pop}_j + \gamma X_{ij}. \tag{17}$$

Estimating $\lambda_X > 0$ implies a weak home-market effect as defined in Costinot et al. (2019): *gross* exports increase with market size. If $\lambda_X > \lambda_M$, the home-market effect is strong: *net* exports increase with market size. The estimated λ_M is biased downward by the same observability bias discussed above. Appendix C.4 calculates the correction applicable here.

To address the same identification challenges discussed in Section 3.2, we estimate slight variants of these specifications. First, we use the share of the population over 65 as a demand shifter conditional on population:

$$\ln \mathbb{E}\left[S_{ij}\right] = \lambda_X \ln \text{over } 65_i + \lambda_M \ln \text{over } 65_j + \Gamma_X \ln \text{pop}_i + \Gamma_M \ln \text{pop}_j + \gamma X_{ij}. \tag{18}$$

As above, Appendix C.4 provides an observability-bias correction.

Second, we let population growth proxy for *changes* in local demand using a panel regression with a region-pair fixed effect:

$$\ln \mathbb{E}\left[S_{ijt}\right] = \lambda_X \ln \text{pop}_{it} + \lambda_M \ln \text{pop}_{jt} + \phi_{ij} + \gamma_t X_{ij}, \tag{19}$$

where S_{ijt} denotes the gross bilateral trade flow in year t. In this specification, changes in population across time periods t identify λ_X and λ_M.

Third, we contrast common and rare services by splitting procedures based on whether

their national volume exceeds that of the median procedure.[36] The model predicts a stronger home-market effect for rare procedures, so we estimate the following specification:

$$\ln \mathbb{E}\left[S_{ijg}\right] = \lambda_X \ln \text{pop}_i + \lambda_M \ln \text{pop}_j + \gamma X_{ij} + \left(\mu_X \ln \text{pop}_i + \mu_M \ln \text{pop}_j + \psi X_{ij}\right) \cdot \mathbf{1}(g = \text{rare}). \tag{20}$$

Given a strong home-market effect overall, theory predicts $\mu_X > \mu_M$. An alternate specification introduces ij-pair fixed effects, which absorb all covariates not interacted with the group indicator $\mathbf{1}(g = \text{rare})$. We estimate similar models with other types of heterogeneity, such as procedure intensity and frequency of patient engagement, and for specific procedures.

4 Regional increasing returns in medical services

This section tests for increasing returns in healthcare and the implications for trade flows. Section 4.1 estimates regional quality of medical care and shows that external quality indicators line up with revealed-preference measures based on trade flows. Using these measures and the demand shifters described above, Section 4.2 finds substantial regional increasing returns in healthcare production. In line with these estimates, Section 4.3 shows that larger markets have finer division of labor and disproportionately host specialized capital inputs. Section 4.4 demonstrates a strong home-market effect, which Section 4.5 shows to be even stronger for rarer services. These results imply that scale economies indeed generate the observed patterns of production and trade.

4.1 Revealed-preference quality estimates reflect clinical quality

Estimating the scale elasticity requires a measure of regional service quality δ_i. We rely on revealed-preference measures, namely the exporter fixed effects $\widehat{\ln \delta_i}$ from estimating equa-

[36] Figure 4(c) plots the distribution of import shares across regions for the two groups.

tion (11).[37] To validate these quality estimates, Figure 7 relates them to external measures of regional hospital quality. We aggregate estimated hospital mortality from Chandra, Dalton, and Staiger (2023) and the Centers for Medicare and Medicaid Services (CMS) by HRR. We also tally hospitals' appearances in the *U.S. News* Best Hospitals rankings.[38] Panels (a)–(c) establish that patients travel farther to obtain care from regions with lower hospital mortality rates or a higher share of *U.S. News* rankings. Appendix Table E.4 reports the corresponding regressions. It also shows a regression relating the exporter fixed effects simultaneously to a variety of hospital safety, mortality, accreditation, and other quality measures. This regression has an adjusted R^2 of 0.32. While our revealed-preference estimates are intended to capture all attributes that influence patient choices, their relationship to these other quality measures suggests that our estimates capture many characteristics of clinical interest.

4.2 Scale improves quality

Estimated service quality $\widehat{\ln \delta_i}$ rises substantially with the regional production volume $\ln Q_i$. Figure 7(d) depicts this relationship, which appears consistent with our isoelastic specification. Table 2 reports 20 regression estimates of the scale elasticity α, with each row corresponding to a different empirical strategy and each column a different specification. Across all 20 estimates, the lowest elasticity is 0.55 and the highest is 1.61.[39]

The baseline OLS estimate shown in the first column is 0.8. The second column omits the diagonal S_{ii} observations when estimating the gravity equation (11). The third column con-

[37] Other empirical strategies require quality estimates for specific groups of procedures. We estimate equation (13) to obtain $\widehat{\ln \delta_i}^{g}$ for procedure groups g determined by the elderly population's share of consumption. We estimate $\widehat{\ln \delta_i}^{\text{meso}}$ using a version of equation (11) with the sample restricted to mesothelioma care.

[38] Compared with the CMS measures, Chandra, Dalton, and Staiger (2023) use empirical Bayes estimation to account for differences in hospital volume and quality drift over time. Appendix A.1 details the *U.S. News* rankings. Appendix Figure E.4 presents results using the Leapfrog Group's Hospital Safety Grades.

[39] Appendix Figure E.5 shows an alternative version of Figure 7(d) that uses observed Traditional Medicare production instead of our labor-based proxy; it is extremely similar. The estimates in Table 2 overlap with the range of prior estimates of 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% increase in total factor productivity caused by adding a plant representing 8.6% of the county's prior output).

trols for regional variation in the MCAT proxy for doctor ability, Medicare reimbursements, mean two-bedroom property value, and mean annual earnings for non-healthcare workers. The final column replaces the quantity regressor with a proxy based on regional employment of doctors and NPs to account for non-Medicare production.[40]

The first three rows rely on cross-sectional variation in market size. All scale elasticity estimates are highly significant and reasonably similar, although excluding the diagonal of the trade matrix when estimating quality yields a slightly higher estimate. The contemporaneous and historical population instruments have strong first-stage regressions, as shown by the F statistics reported in square brackets under each estimate. For both instruments, the estimate in each column is similar to its OLS counterpart.

The main challenge to our interpretation arises if physicians prefer larger markets for non-medical reasons, and this labor supply shift increases quality or reduces costs in larger markets. We investigate whether this mechanism is strong enough to reduce net costs in larger markets. Doctors are cheaper in larger markets (Gottlieb et al., 2025), but other costs rise with population size. Appendix Figure E.7 shows that the population elasticity of doctors' earnings is -0.01, but that for non-physicians is 0.043.[41] To compute the population elasticity of overall labor costs, note that non-physician labor's share of healthcare production is twice the physician labor's share (Appendix Table E.5). The population elasticity of labor costs is thus positive. The higher cost of real estate in larger markets reinforces these higher labor costs. This spatial variation undermines the idea that amenities make production cheaper in larger markets. Figure 8(a) further shows that specialty-specific income elasticities are close to zero on average and unrelated to the number of specialists.

Another way to address the threat of doctor sorting is to use the population share above age 65 as an instrument, while controlling for total population. Recall from Section 3.2 that our measure of doctor quality is conditionally independent of the elderly share and that

[40] Appendix C.5 shows how these estimates should be scaled by a factor of $\frac{1}{1+\beta}$ to recover the target scale elasticity when $\beta > 0$. The typical value of $\beta \approx 0.4$ in the literature implies an estimated effective scale elasticity ranging from 0.47 to 1.16.

[41] Appendix A.1 discusses subtleties of the income data.

prime-age high-human-capital workers generally have different geographic sorting patterns from the elderly. The over-65 variation is thus distinct from the first three rows of Table 2 and orthogonal to any size-related amenities. The fourth row shows that this instrument is weaker than raw population but strong enough to not raise weak-instrument concerns. The resulting estimates are somewhat less precise but still show substantial scale economies, and the differences from zero are highly statistically significant.

Table 2's last row uses panel variation, thus removing any time-invariant regional supply shifters. As described in Section 3.2, this strategy addresses the threat of doctor sorting by using variation distinct from all of the cross-sectional approaches. This removes persistent unobserved amenities as a potential confounder. Scale elasticities remain precise and significant, and if anything are larger than those estimated in the cross section.[42]

Table 3 provides further evidence for scale economies in our context. It shows estimates of equation (14), which regresses quality at the region-by-procedure-group level on the interaction of the region's elderly share and the procedure group's elderly demand share. This regression includes region and procedure group fixed effects, so a positive coefficient on the interaction means that regions with a higher elderly share produce disproportionately higher quality for care that the elderly disproportionately demand. Unlike Table 2, this is a reduced-form regression; we do not observe the non-elderly's procedure-by-region production that would be necessary to estimate the scale elasticity itself with this strategy.

Table 3 shows results for two splits of services (at the median, and into terciles), and two specifications (including and excluding the diagonal elements of the trade matrix when estimating quality). Columns 1 and 2 show that procedures with above-median elderly demand have significantly higher quality when provided in elderly-abundant regions. Columns 3 and 4 split procedures into terciles, and show a monotonic relationship: the highest tercile of elderly intensity has a larger positive coefficient on the elderly share interaction than the middle tercile. These are not sensitive to the inclusion of the diagonal of the trade matrix

[42] Appendix Tables E.6 and E.7 report results for CBSAs and CZs, which are also stable across specifications and when using the bedrock instrument.

when estimating quality.

These estimates are further evidence for scale economies. For supply shifters to explain these results, factors used to produce elderly-oriented care would have to be relatively cheaper in elderly-abundant regions. We think this is unlikely. Doctors who specialize in care disproportionately consumed by the elderly are not likely to have systematically different preferences for regional amenities from other doctors.

Our final strategy for estimating α is to focus on care for mesothelioma. Table 4 shows estimates of the scale elasticity for mesothelioma care, using variation in demand predicted by mesothelioma death rates, after conditioning on state fixed effects and the region's overall population.[43] First-stage F statistics exceed 40 in all specifications. The baseline scale elasticity estimate in column 1 is 0.6, similar to the scale elasticity for aggregate services. Column 2 removes diagonal elements of the trade matrix when estimating quality, column 3 adds controls, and column 4 adds a control for the ratio of Medicare Advantage to Traditional Medicare beneficiaries to address observability bias.[44] Results are similar across specifications and highly significant.

These regressions using demand variation caused by past asbestos exposure suggest scale economies similar in magnitude to our aggregate estimates. Because mesothelioma death rates are orthogonal to aggregate service quality when controlling for log population and state fixed effects (Appendix Figure E.3), a peculiar alignment of supply shifters would be required to explain these estimates without scale economies. Namely, the doctors who treat mesothelioma would need to find regions with past asbestos exposure particularly attractive, conditional on state fixed effects and total population.

[43] Appendix Figure E.8 shows that our mesothelioma care quality estimates $\widehat{\ln \delta_i^{\text{meso}}}$ are positively correlated with survival times after diagnosis, for both raw and risk-adjusted survival times.

[44] Since mesothelioma primarily affects the elderly (CDC, 2024), unobserved production is concentrated among Medicare Advantage beneficiaries. We can thus address the observability bias by controlling for the ratio of Medicare Advantage to Traditional Medicare beneficiaries in each region, as a proxy for the (unobserved) ratio of mesothelioma care production.

4.3 Mechanisms for regional increasing returns

This section shows that the division of labor and lumpy capital, which are among the theoretical microfoundations for regional returns to scale modeled in Appendix B.2, vary systematically with market size. This evidence does not preclude other agglomeration mechanisms from also playing a role. Knowledge diffusion (Baicker and Chandra, 2010), thicker input markets (Friedson and Li, 2015), and more output varieties (Figure 4(a)) could also be important benefits of scale. We focus here on specialization and equipment use because they can be measured using claims data and are closely related to procedure-level returns to scale.

4.3.1 Scale facilitates the division of labor

One source of increasing returns could be division of labor among physicians (Arrow, 1963; Baumgardner, 1988). In particular, the specialized labor required to produce rare services could explain rare treatments' concentration in larger markets and higher traded share. Specialized services may require physicians with specific training, whose presence may require high demand (Dranove, Shanley, and Simon, 1992).

To study this mechanism, we estimate the population elasticity of physicians per capita for each specialty and relate it to the number of physicians in the specialty. Let Y_{si} be the number of doctors of specialty s in location i. We estimate a separate Poisson model,

$$\ln \mathbb{E}\left[\frac{Y_{si}}{\text{population}_i}\right] = \zeta_s^S + \beta_s^S \ln \text{population}_i, \tag{21}$$

for each specialty s by maximum likelihood.

Figure 8(b) shows a clear negative relationship between a specialty's per capita population elasticity $\hat{\beta}_s^S$ and the national number of physicians in that specialty.[45] This parallels the patterns for procedures and diagnoses from Figure 4. A natural explanation for the

[45] This pattern is not attributable to spatial sorting driven by rare specialties commanding higher earnings. In fact, a specialty's number of physicians and mean earnings are uncorrelated. Appendix Table E.8 shows that controlling for a specialty's earnings has no effect on the negative relationship between population elasticity and number of physicians across specialties.

geographic concentration of both rare medical care and rare specialties in larger regions is that the size of the market limits the division of labor. To the extent that providing rare treatments requires specialized physicians, increased patient volume makes production economically viable.

A large medical literature about specific procedures suggests that this division of labor directly impacts the clinical quality of care, not only its availability. This "volume-outcome" research finds that providers with higher procedure volume achieve better patient outcomes (Birkmeyer et al., 2003, in 8 cardiovascular and cancer surgeries; Gaynor, Seider, and Vogt, 2005, in coronary artery bypass graft surgery; Domenella, 2026, in hip and knee replacements; Chowdhury, Dagash, and Pierro, 2007, in a meta-analysis). Our estimates speak to the same forces at a regional level. Because rare services and the specialized physicians who provide them are concentrated in larger regions, patients in those markets are treated by higher-volume, more specialized providers. The combination of this volume-outcome literature and the fact that physicians specialize more in larger regions suggests a very natural mechanism that likely underlies the regional scale elasticities we estimate.

4.3.2 Lumpy capital as a source of regional increasing returns

Another potential source of increasing returns is lumpy capital, namely expensive medical equipment. For instance, if a piece of medical equipment is rarely used in procedures, small markets may not use it enough to justify the investment.

We provide evidence consistent with this hypothesis using a CMS dataset that enumerates the types of equipment used by each medical procedure (HCPCS code).[46] We measure equipment use by linking this dataset to procedure-level production data. The frequency of equipment use H_{di} for a piece of equipment d in region i is defined as the sum of the procedure volume in region i across all procedures that use equipment d. We use a Poisson

[46] See Appendix A.1 for details.

model to estimate the population elasticity of per-capita equipment use:

$$\ln \mathbb{E}\left[\frac{H_{di}}{\text{population}_i}\right] = \zeta_d^K + \beta_d^K \ln \text{population}_i. \tag{22}$$

Figure 8(c) plots the population elasticity of per-capita equipment use β_d^K against log national use, $\ln H_d$. The strong negative correlation indicates that equipment used less often tends to concentrate in larger markets. Similar to the patterns for physician specializations, rare capital equipment is disproportionately employed in larger regions.

4.4 A strong home-market effect in medical services

While we have established there are regional increasing returns, we have yet to see if they are sufficiently strong to explain the fact that larger markets are net exporters. This question determines whether market size enables larger regions to specialize in healthcare production. We test for a strong home-market effect in medical services by estimating equation (17).

Table 5 presents the results. The first column shows significant, positive coefficients on both provider- and patient-market population.[47] The coefficient on provider-market population is substantially greater than that on patient-market population. This demonstrates a *strong* home-market effect (Costinot et al., 2019). 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).[48] 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.

[47] In all estimates using trade matrices, we two-way cluster standard errors by patient and provider market.

[48] We find a distance elasticity of medical services trade between CBSAs of -2.4. The analogous elasticity of health care and social assistance services trade between Canadian provinces is -1.42 (Anderson, Milot, and Yotov, 2014). The distance elasticity of international trade is typically near -0.9 (Disdier and Head, 2008).

Columns 2 and 3 of Table 5 demonstrate that more flexible distance-covariate specifications—including (log distance)^2 or using dummies for deciles of distance—deliver the same result. Column 4 reports the same result when using the historical population instrument: both gross and net exports increase with market size. Results are similar when excluding states with large seasonal populations (Appendix C.1).

The final two columns of Table 5 use demand shifters that are distinct from regional variation in total population to estimate the home-market effect. Column 5 uses regional variation in the share of residents over 65, while controlling for total (log) population (equation 18). This eliminates size-related confounders, such as physicians sorting on urban amenities. These point estimates align with those in the other columns, but they have large standard errors. Column 6 uses changes in population over time and includes market-pair fixed effects that absorb all cross-sectional variation (equation 19). This eliminates any time-invariant regional supply shifters. This specification shows a strong home-market effect that is similar in magnitude to the cross-sectional estimates.

4.5 How market-size effects vary with procedure characteristics

Given the presence of a strong home-market effect overall, the model predicts that this should be especially true for rare services. We have already seen in Figure 4 that trade is indeed more prevalent in rare services. We now use the specific pattern of these trade flows to test the model's prediction that scale economies drive the production and trade of rare services.

Table 6 reports estimates of equation (20), gravity regressions in which each pair of locations has two observations: one for rare services and one for common. In all specifications, the coefficient on the interaction of provider-market population and the rare-service indicator is positive and highly significant, while the coefficient on the interaction of patient-market population and the rare-service indicator is negative. These estimates allow us to conclude that, consistent with theory, the home-market effect is stronger for rare services. When

examining diagnoses rather than procedures, we find the same result.[49]

The model predicts larger home-market effects for rare services when they have more elastic demand or larger scale economies. Separate estimates of equation (12) by procedure frequency yield a larger scale elasticity for rare services in some but not all specifications (Appendix Table E.11). Finding broadly similar scale elasticities suggests that more elastic demand may explain the stronger home-market effects for rare services.

Some procedures require more capital and specialized knowledge, and the consequences of market size for these resource-intensive procedures may illuminate sources of increasing returns. Relative value units (RVUs) capture the extent of physician work, expertise, and additional resources required to perform the procedure and determine the procedure's Medicare payment. We divide procedures into terciles by intensity as measured by the total RVUs CMS assigns to a procedure. We find that the home-market effect is increasing in procedure intensity (Appendix Table E.12). This result is consistent with the patterns of physician specialization and equipment usage from Section 4.3.

5 Health policy with trade and increasing returns

The trade we have documented implies that patients' choice sets for healthcare extend beyond their home regions. This observation expands the set of policies that can be considered for improving healthcare access and reducing geographic disparities in care availability, beyond the standard approach of increasing payments to rural providers. Rural healthcare is a salient and persistent issue for federal and state healthcare policies, most notably illustrated by the Rural Health Transformation Program (RHTP) launched in 2025 (CMS, 2025).

This section first uses our trade measures and model estimates to study two ways to sub-

[49]This result holds when using finer definitions of rare care (Appendix Figure E.9). Among illustrative procedures, Appendix Table E.9 shows a larger home-market effect for rare services. Appendix Table E.10 shows that the rare-common distinction is robust to controlling for how often an individual patient receives the same procedure. Patients are more sensitive to distance for high-engagement procedures, but engagement is not very correlated with frequency and controlling for this heterogeneity does not meaningfully alter the estimated differential impacts of population size for rare procedures.

sidize rural healthcare access: subsidizing healthcare production in rural areas, or subsidizing rural patients' travel to receive care.[50] For both policies, we consider the impacts with and without scale economies. Even absent scale economies, patients' travel decisions shape the consequences and incidence of these policies. We then introduce scale economies and show how they generate spillovers from subsidizing one region onto its neighbors, amplify the differences between the two policies, and shape how each policy's mortality and market-access effects vary across the income distribution. Finally, we show how geography may contribute to the striking relationship between regional income and mortality in the US.

5.1 Method to compute counterfactual outcomes

We compute counterfactual equilibrium outcomes relative to the baseline equilibrium. For the baseline equilibrium, define export shares $x_{ij} \equiv \frac{Q_{ij}}{\sum_{j'} Q_{ij'}}$ and import shares $m_{ij} \equiv \frac{Q_{ij}}{N_j}$. The import share denominator is based on all Traditional Medicare beneficiaries; we use the Medicare Current Beneficiary Survey to estimate the patients in each region choosing the outside option.[51] For every variable or parameter y, denote the ratio of its counterfactual value y' to its baseline value y by $\hat{y} \equiv \frac{y'}{y}$. Appendix D.1 shows how we solve for the relative counterfactual endogenous qualities $(\hat{\delta})$ using baseline equilibrium shares (x_{ij}, m_{ij}), the scale elasticity (α), and relative counterfactual parameters $(\hat{A}, \hat{R}, \hat{w}, \hat{\rho}, \hat{N})$. Counterfactual

[50] We calibrate the model using only the Traditional Medicare trade matrix, so our counterfactuals are for the Traditional Medicare population. Because our estimation sample excludes emergency services, our estimates and counterfactual scenarios omit any complementarities between emergency and non-emergency care. Our partial-equilibrium model assumes elastic regional factor supplies. Our model and estimates represent long-run elasticities, omitting any adjustment costs or short-run diseconomies of scale due to crowding or queuing.

[51] To compute import shares, we need to know the share of potential patients in each region choosing the outside option. We calibrate this number using our gravity estimates and the Medicare Current Beneficiary Survey, which asks patients whether they have forgone any healthcare in the past year. See Appendix D.2. The qualitative and spatial patterns of counterfactual outcomes are robust to a range of different assumptions about this outside option share. Appendices D.3 and D.4 generalize this method of computing counterfactual outcomes to the model with multiple types of patients introduced in Appendix B.3.

qualities are given by a system of $\mathcal{I}$ equations with unknowns $\{\hat{\delta}_i\}_{i=1}^{\mathcal{I}}$:

$$\hat{\delta}_i = \left(\frac{\hat{R}_i \hat{A}_i}{\hat{w}_i}\right)^{\frac{1}{1-\alpha}} \left(\sum_{j \in \mathcal{I}} \frac{x_{ij} \hat{\rho}_{ij} \hat{N}_j}{m_{0j} + \sum_{i' \in \mathcal{I}} m_{i'j} \hat{\delta}_{i'} \hat{\rho}_{i'j}}\right)^{\frac{\alpha}{1-\alpha}}. \tag{23}$$

The first term of this expression, $\left(\hat{R}_i \hat{A}_i / \hat{w}_i\right)^{\frac{1}{1-\alpha}}$, shows that the scale elasticity α governs the effect of exogenous supplier shifters, including reimbursements $\hat{R}_i$, on quality produced in a region. Reimbursement rates shift the scale of production, and stronger scale economies (higher α) amplify these shifts. The second term shows how changes in other regions influence local outcomes through trade, combined with scale. Thus, counterfactual scenarios with $\alpha > 0$ rely on both our estimates of the scale elasticity α and observed trade patterns. Absent scale economies ($\alpha = 0$), equation (23) depends only on the exogenous shifters: $\hat{\delta}_i = \hat{R}_i \hat{A}_i / \hat{w}_i$.

We use our estimated model to compute two counterfactual outcomes patients and policymakers might care about: patient market access and mortality. Market access Φ_j is a proximity-weighted sum of regional qualities, so captures all dimensions of quality that patients value according to revealed preference. We compute mortality impacts based on the empirical relationship between risk-adjusted mortality and δ_i, as described in Appendix D.6.

Patient market access can also be interpreted as the flip side of forgone care. Recall from Section 2.1 that the share of potential patients in region j choosing the outside option is the inverse of patient market access: $m_{0j} = {\Phi_j}^{-1}$. Because our counterfactual policies do not change the outside option, the share of patients forgoing care moves inversely with changes in market access: $\hat{m}_{0j} = \hat{\Phi}_j^{-1}$. Results about $\hat{\Phi}_j$ are thus equivalently (inverse) results about the change in forgone care: increases in market access capture reductions in forgone care.

5.2 Subsidizing output vs. subsidizing travel

Consider a policymaker whose goal is to improve patient market access for rural patients, for example by spending the RHTP's $50 billion over five years. We analyze counterfactual policies that target rural patients' access to uncommon procedures, because these are more

likely to be provided by specialists not available near rural patients (Figure 8(b)). We present two versions: the first subsidizes Coronary Artery Bypass Grafting (CABG), a major surgery that treats coronary artery disease, and the second subsidizes all uncommon procedures. We define "uncommon" to make the counterfactual policy's fiscal cost similar to the funding going to Traditional Medicare patients if the discretionary half of RHTP were allocated proportionally to baseline medical spending. We order procedure codes by frequency and label the bottom nine deciles "uncommon," so fiscal costs are roughly $850 million per year.

We compare two policies for improving rural access to these types of care. The first is a *reimbursement increase* for uncommon procedures produced in rural regions.[52] These policies are common in practice (MedPAC, 2012, 2024; Arredondo et al., 2023), and the literature has shown that healthcare supply is responsive to such reimbursement changes (fn. 31).[53] The second policy is a *travel subsidy* for rural patients receiving uncommon procedures. We model it on travel subsidy policies in Montana and South Dakota's Medicaid programs (details in Appendix D.5). We apply both policies to all rural HRRs, defined as HRRs below the median population density. The reimbursement increase is set such that, when applied to all rural HRRs, it has the same fiscal cost as the Montana-style travel subsidy.

5.2.1 Policy impacts without scale economies ($\alpha = 0$)

To isolate the role of trade patterns, Table 7 compares the gains from each policy in the absence of scale economies ($\alpha = 0$). For each policy and each outcome, columns 1 and 2 show the impact when the subsidies apply to all uncommon procedures, whereas in columns 3 and 4 they apply to CABG. Appendix Figure E.11 shows the geographic distribution of travel subsidy spending and lives saved for both CABG and uncommon procedure subsidies.

Focusing on lives saved per $10 million spent, first notice in column 3 of row 1 that the

[52] Given our regional production function, output subsidies are equivalent to input subsidies. A policy that raises the output price from $\overline{R}$ to $\overline{R}(1+s_i)$ is equivalent to a policy that lowers the input price from w_i to $w_i/(1+s_i)$ in terms of the equilibrium quantities and qualities that satisfy free-entry condition (5).

[53] As shown in Appendix B.4, our assumption of a unit cost elasticity of quality, $K(\delta) = \delta$, is conservative in that our counterfactual outcomes overstate the quality response to reimbursement changes.

output subsidy for CABG *costs* lives. The increase in local quality induces some patients to switch from importing care at baseline to consuming the improved local option. Because rural regions have substantially higher CABG mortality rates, these switches cost lives. When applying the subsidy to all uncommon procedures in column 1, the reimbursement increase saves lives, but the gains remain small. For both CABG and uncommon procedures, raising reimbursements for rural providers improves patient market access for rural residents.

Most importantly, comparing the reimbursement and travel-subsidy rows within each panel shows that the travel subsidy has a much larger positive impact on each outcome than the reimbursement increase. Unlike the reimbursement increase, the travel subsidy encourages patients to substitute towards *higher*-quality providers in other regions, which tends to reduce their mortality. Some patients who did not travel at baseline start traveling and benefit from lower-mortality care, and some patients who traveled to a certain region now travel to a different region with even lower mortality. Finally, Table 7 shows that a travel subsidy can become less cost-effective at saving lives if it becomes too generous: an expanded subsidy for all uncommon care increases patient market access more (row 6), but saves fewer lives (row 3), compared with our baseline subsidy.

These counterfactual results do not rely on any assumptions about patients' preferences over survival rates; mortality results reflect the empirical relationship between revealed preferences and mortality rates. Thus regions with the largest mortality gains are not necessarily those with the largest gains in market access. The travel subsidy changes mortality outcomes when it induces patients to change which region treats them. The policy also makes subsidy payments to inframarginal patients who do not change their decisions; this increases market access without affecting mortality outcomes. The correlations between market access improvements and lives saved are 0.21 for CABG and 0.19 for broader travel subsidies.

Figure 9 compares the two policies for patients in different income terciles. Willingness to travel for care varies with income (Section 1.3 and Figure 3(c)), so gains from the two policies differ across the income distribution. Still assuming away scale economies, Panels (a) and (b)

of Figure 9 show that the reimbursement increases generate very modest improvements in market access (dashed lines) for all income terciles, while the travel subsidy generates gains that are especially large for lower-income patients.

For both types of care, the travel subsidy increases lower-income patients' survival more (solid lines). Because lower-income patients are more sensitive to travel costs, more of them start importing care from higher-quality regions, implying a negative relationship between income and lives saved. When subsidizing uncommon-procedure reimbursements, lower-income patients again gain more survival, because they are more likely to consume services produced in rural regions and so benefit from rural quality improvements. The gradient flips when subsidizing CABG reimbursement because lower-income patients are more likely to switch to local care when local quality improves, and rural regions have high CABG mortality. Overall, the travel subsidy is more progressive than the reimbursement increase regardless of whether one evaluates it by lives saved or market access gains.

5.2.2 Policy impacts with scale economies ($\alpha > 0$)

The policy analysis thus far solely reflects trade patterns. We now introduce scale economies to see how agglomeration affects these results. To build intuition, we first examine policies targeting particular regions, then return to potential widespread policy interventions.

Local reimbursement increase. Figure 10 contrasts the consequences of raising reimbursements for uncommon procedures by 8% in either Boston or Paducah, Kentucky. This is the reimbursement increase used in Table 7 to match the fiscal cost of the Montana-style travel subsidy in the presence of scale economies ($\alpha = 0.6$). Panel (a) shows that higher reimbursements in Boston raise the quality of care produced there, while quality *declines* in the rest of New England as patients substitute away and scale economies translate lower volumes into lower quality (an "agglomeration shadow"; Fujita and Krugman, 1995). These spillovers, driven by our estimated agglomeration effects, diminish with distance to Boston.

Yet patient market access moves nearly opposite to local output quality (Panel (b)). Patients in Boston benefit most from the higher reimbursement of their local production. Nearby regions import enough from Boston that its improved quality outweighs the decline in their local production, so their market access rises—and it rises most where local quality falls most, since those are the regions whose residents substitute toward Boston most strongly. In more distant regions, the welfare impacts are virtually zero.

Raising reimbursements by 8% in Paducah generates very different spillovers (Panel (c)). Unlike in the Boston scenario, Panel (d) shows that most neighboring regions also experience declines in patient market access, because Paducah's improved quality is not significant enough to offset the neighbors' reduced scale. Appendix Figure E.12 shows that this contrast generalizes: the spillovers from a local reimbursement increase are governed by the region's net trade in medical care, and changes in regional output quality need not align with changes in patient market access.

Local reduction in travel costs. We next consider a policy that applies the travel subsidy analyzed in Section 5.2.1 for Paducahans traveling for care. Unlike a Paducah reimbursement increase, this has positive spillovers on neighboring regions (Figure 11): neighbors increase exports, raising their own scale, quality, and residents' market access. Paducahans benefit even though facilitating travel reduces the quantity—and thus the quality—produced in Paducah itself. An analyst focused on care provided within Paducah would reach a very different conclusion than one focused on Paducahans' market access.

Output vs. travel subsidies with economies of scale. With this intuition in mind, we return to the comparison of nationwide rural output and travel subsidies, now allowing for economies of scale. Columns 2 and 4 of Table 7 show that scale economies amplify the travel subsidy's advantage over the reimbursement increase. Under a reimbursement increase, rural quality gains are larger because more patients consume locally when quality improves; but patients in the regions from which rural patients no longer import are hurt

as those regions lose agglomeration and quality. This second effect worsens the mortality losses in the CABG case and shrinks the mortality gains for uncommon procedures. The travel subsidy instead amplifies both mortality and market access gains: regions that receive more patients improve in quality through scale, propagating gains from rural to non-rural patients. The geographic distribution of subsidy spending and lives saved is nearly identical to the no-scale case (Appendix Figure E.11).

Allowing for economies of scale leaves the income gradient little changed (Figures 9(c) and 9(d)): gains from the reimbursement increase resemble the $\alpha = 0$ case, and the travel subsidy remains similarly progressive in both mortality and market access. The policy benefits bottom-income-tercile patients more both because they are more sensitive to travel costs and because they disproportionately reside in rural regions.

These scenarios carry significant caveats, and we do not attempt to identify the optimal policy. Even so, they compare two feasible policies for improving rural healthcare access on a dollar-for-dollar basis and illustrate the economic mechanisms in our model. The two policies have very different consequences for patient market access and mortality even without scale economies, with the differences driven by trade patterns; economies of scale magnify these differences and generate spillovers that likewise depend on trade patterns.

5.2.3 Geography and healthcare inequality

The US has a steeper relationship between regional income and mortality than other rich economies (Appendix Figure E.14(a)). Our framework suggests that the US's greater geographic spread—and, in particular, the distances between low-income populations and urban centers—may contribute to this gradient. Appendix Figure E.14(b) shows the observed and three counterfactual CABG-mortality-income gradients across HRRs: autarky, the more generous travel subsidy studied above, and that subsidy applied to all regions. Panel (c) plots the difference between each region's log mortality under the travel-subsidy counterfactual and at baseline against income. Lower-income regions improve more: the counterfactual

mortality-income gradient is about one-quarter lower than its baseline value.

This illustrates the importance of travel for healthcare inequality. Trade reduces the gradient relative to autarky, and subsidies to travel for care would reduce it further. This implies that distance from high-quality care is a widespread obstacle, and hence that unequal health outcomes partly reflect the country's economic geography.

6 Conclusion

Smaller markets have fewer specialized physicians, produce less medical care per capita, and have worse health outcomes than larger markets. Thanks to trade in medical services, less production does not translate one for one into less consumption of medical services. Instead, trade affords patients who live in smaller markets access to higher-quality care. This quality comes in part from consuming services that would otherwise be unavailable, visiting appropriate specialists, and accessing experienced physicians. This trade amplifies the scale advantages of large markets and hence the quality of care they produce. This means the healthcare industry can serve as an export base for large cities.

Because patients can and do travel for medical care, policymakers can consider trade-based policies when aiming to improve medical access. They must also account for interregional spillovers when considering a wide range of policy decisions. Understanding the role of scale is crucial for evaluating policy decisions, and insights from economic geography more generally can enrich the study of healthcare markets. Conversely, as the service economy grows, healthcare provides a rich setting to inform the study of economic geography.

References

AAPC. 2021. "What Is CPT?" Available online at https://www.aapc.com/resources/medical-coding/cpt.aspx.

Acemoglu, Daron, Amy Finkelstein, and Matthew J. Notowidigdo. 2013. "Income and Health Spending: Evidence from Oil Price Shocks." *The Review of Economics and Statistics*, 95(4): 1079–1095.

Acemoglu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *The Quarterly Journal of Economics*, 119(3): 1049–1090.

Agarwal, Sumit, J. Bradford Jensen, and Ferdinando Monte. 2020. "Consumer Mobility and the Local Structure of Consumption Industries." NBER Working Paper 23616.

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

Anderson, James E., Catherine A. Milot, and Yoto V. Yotov. 2014. "How much does geography deflect services trade? Canadian answers." *Int'l Econ. Rev.*, 55(3): 791–818.

Arora, Abhishek, and Melissa Dell. 2023. "LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models."

Arredondo, Kelley, Hilary N Touchett, Sundas Khan, Matthew Vincenti, and Bradley V Watts. 2023. "Current programs and incentives to overcome rural physician shortages in the United States: a narrative review." *Journal of General Internal Medicine*, 38(Suppl 3): 916–922.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review*, 53(5): 941–973.

Attanoos, Richard L., Andrew Churg, Francoise Galateau-Salle, Allen R. Gibbs, and Victor L. Roggli. 2018. "Malignant Mesothelioma and Its Non-Asbestos Causes." *Archives of Pathology & Laboratory Medicine*, 142(6): 753–760.

Badilla Maroto, Marco, Benjamin Faber, Antoine Levy, and Mathilde Munoz. 2024. "Senior Migration, Local Economic Development and Spatial Inequality."

Baicker, Katherine, and Amitabh Chandra. 2010. "Understanding Agglomerations in Health Care." *Agglomeration Economics*, 211–236. University of Chicago Press.

Bartelme, Dominick G., Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare. 2025. "The Textbook Case for Industrial Policy: Theory Meets Data." *Journal of Political Economy*, 133(5): 1527–1573.

Bartik, Timothy, and George Erickcek. 2007. "Higher Education, the Health Care Industry, and Metropolitan Regional Economic Development: What Can 'Eds & Meds' Do for the Economic Fortunes of a Metro Area's Residents?" Upjohn Inst. Working Paper 08-140.

Baumgardner, James R. 1988. "Physicians' Services and the Division of Labor across Local Markets." *Journal of Political Economy*, 96(5): 948–982.

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

Berry, Christopher R., and Edward L. Glaeser. 2005. "The Divergence of Human Capital Levels across Cities." *Papers in Regional Science*, 84(3): 407–444.

Birkmeyer, John D, Therese A Stukel, Andrea E Siewers, Philip P Goodney, David E Wennberg, and F Lee Lucas. 2003. "Surgeon volume and operative mortality in the United States." *New England Journal of Medicine*, 349(22): 2117–2127.

Borchert, Ingo, Mario Larch, Serge Shikher, and Yoto V. Yotov. 2021. "The International Trade and Production Database for Estimation (ITPD-E)." *International Economics*, 166(C): 140–166.

Bound, John, and Harry J. Holzer. 2000. "Demand Shifts, Population Adjustments, and Labor Market Outcomes During the 1980s." *Journal of Labor Economics*, 18(1): 20–54.

Bradley, Cathy J, and David Neumark. 2017. "Small cash incentives can encourage primary care visits by low-income people with new health care coverage." *Health Affairs*, 36(8): 1376–1384.

Burstein, Ariel, Sarah Lein, and Jonathan Vogel. 2024. "Cross-border shopping: evidence and welfare implications for Switzerland." *Journal of International Economics*, 152: 104015.

Cabral, Marika, and Neale Mahoney. 2019. "Externalities and taxation of supplemental insurance: A study of Medicare and Medigap." *American Economic Journal: Applied Economics*, 11(2): 37–73.

Cabral, Marika, Colleen Carey, and Sarah Miller. 2025. "The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid." *American Economic Journal: Economic Policy*, 17(1): 106–43.

CDC. 2024. "U.S. Cancer Statistics: Malignant Mesothelioma." Centers for Disease Control and Prevention. Available online at https://www.cdc.gov/united-states-cancer-statistics/publications/mesothelioma.html. Accessed May 1, 2025.

Centers for Medicare and Medicaid Services. 2024. "National Health Expenditure Data."

Chan, Alex. 2024. "Discrimination Against Doctors: A Field Experiment." Working Paper.

Chandra, Amitabh, Maurice Dalton, and Douglas O Staiger. 2023. "Are Hospital Quality Indicators Causal?" NBER Working Paper 31789.

Chaney, Thomas, and Ralph Ossa. 2013. "Market size, division of labor, and firm productivity." *Journal of International Economics*, 90(1): 177–180.

Chen, Yong, and Stuart S Rosenthal. 2008. "Local amenities and life-cycle migration: Do people move for jobs or fun?" *Journal of Urban Economics*, 64(3): 519–537.

Chipman, John S. 1970. "External Economies of Scale and Competitive Equilibrium." *The Quarterly Journal of Economics*, 84(3): 347–385.

Chowdhury, MM, H Dagash, and A Pierro. 2007. "A systematic review of the impact of volume of surgery and specialization on patient outcome." *Journal of British Surgery*, 94(2): 145–161.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Treatment Patterns and Patient Health?" *American Economic Review*, 104(4): 1320–1349.

CMS. 2025. "Rural Health Transformation Program." Center for Medicare & Medicaid Services Notice of Funding Opportunity CMS-RHT-26-001. Available online at https://grants.gov/search-results-detail/360442#:~:text=CMS%2DRHT%2D26%2D001%2DFull%20Announcement%20%2D%20RHT%20Program.zip.

Costinot, Arnaud, Dave Donaldson, Margaret Kyle, and Heidi Williams. 2019. "The More We Die, The More We Sell? A Simple Test of the Home-Market Effect." *The Quarterly Journal of Economics*, 134(2): 843–894.

Davis, Donald R., and David E. Weinstein. 2003. "Market access, economic geography and comparative advantage: an empirical test." *J. of International Economics*, 59(1): 1–23.

Davis, Donald R., and Jonathan I. Dingel. 2020. "The comparative advantage of cities." *Journal of International Economics*, 123(C).

Davis, Donald R., Jonathan I. Dingel, Joan Monras, and Eduardo Morales. 2019. "How Segregated Is Urban Consumption?" *Journal of Political Economy*, 127(4): 1684–1738.

Decker, Sandra L., Rebecca A. Gourevitch, and Giacomo Meille. 2024. "Patient Travel Patterns and Concordance With Geographic Market Boundaries." *Annals of Internal Medicine*, 177(12): 1732–1734. PMID: 39496184.

Diamond, Rebecca. 2016. "The Determinants and Welfare Implications of U.S. Workers' Diverging Location Choices by Skill: 1980–2000." *American Economic Review*, 106(3): 479–524.

Dieleman, Joseph L., Jackie Cao, Abby Chapin, Carina Chen, Zhiyin Li, Angela Liu, Cody Horst, Alexander Kaldjian, Taylor Matyasz, Kirstin Woody Scott, Anthony L. Bui, Madeline Campbell, Herbert C. Duber, Abe C. Dunn, Abraham D. Flaxman, Christina Fitzmaurice, Mohsen Naghavi, Nafis Sadat, Peter Shieh, Ellen Squires, Kai Yeung, and Christopher J. L. Murray. 2020. "US Health Care Spending by Payer and Health Condition, 1996-2016." *JAMA*, 323(9): 863–884.

Dingel, Jonathan I. 2017. "The Determinants of Quality Specialization." *Review of Economic Studies*, 84(4): 1551–1582.

Dingel, Jonathan I., and Felix Tintelnot. 2026. "Spatial Economics for Granular Settings." *Econometrica*, 94(2): 407–464.

Disdier, Anne-Célia, and Keith Head. 2008. "The Puzzling Persistence of the Distance Effect on Bilateral Trade." *The Review of Economics and Statistics*, 90(1): 37–48.

Domenella, Dante. 2026. "Market Size and the Returns to Surgeon Volume: Evidence from Joint Replacements." Stanford University, mimeo. Available online at https://ddomenella.github.io/job_market/jmp_domenella.pdf.

Dorsey, E. Ray, and Eric J. Topol. 2016. "State of Telehealth." *New England Journal of Medicine*, 375(2): 154–161. PMID: 27410924.

Dranove, David, and Mark A. Satterthwaite. 2000. "The Industrial Organization of Health Care Markets." *Handbook of Health Economics*, 1: 1093–1139.

Dranove, David, Mark Shanley, and Carol Simon. 1992. "Is Hospital Competition Wasteful?" *The RAND Journal of Economics*, 23(2): 247–262.

Duranton, Gilles, and Diego Puga. 2004. "Micro-foundations of Urban Agglomeration Economies." In *Handbook of regional and urban economics*. Vol. 4, 2063–2117. Elsevier.

Eckert, Fabian, Sharat Ganapati, and Conor Walsh. 2020. "Skilled Scalable Services: The New Urban Bias in Economic Growth." CESifo Working Paper Series 8705.

Fajgelbaum, Pablo D., Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. 2019. "State Taxes and Spatial Misallocation." *The Review of Economic Studies*, 86(1): 333–376.

Finkelstein, Amy. 2007. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *The Quarterly Journal of Economics*, 122(1): 1–37.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence from Patient Migration." *The Quarterly Journal of Economics*, 131(4): 1681–1726.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2021. "Place-Based Drivers of Mortality: Evidence from Migration." *American Economic Review*, 111(8): 2697–2735.

Fisher, Elliott S., David E. Wennberg, Thérèse A. Stukel, Daniel J. Gottlieb, F. Lee Lucas, and Etoile L. Pinder. 2003*a*. "The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care." *Ann. Intern. Med.*, 138(4): 273–287.

Fisher, Elliott S., David E. Wennberg, Thérèse A. Stukel, Daniel J. Gottlieb, F. Lee Lucas, and Etoile L. Pinder. 2003*b*. "The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care." *Ann. Intern. Med.*, 138(4): 288–298.

Flexner, Abraham. 1910. "Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching." Carnegie Fnd. Bulletin 4.

Fowler, Christopher S, and Leif Jensen. 2020. "Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA." *Environment and Planning A: Economy and Space*, 52(7): 1395–1414.

Friedson, Andrew I., and Jing Li. 2015. "The impact of agglomeration economies on hospital input prices." *Health Economics Review*, 5(38).

Fujita, Masahisa, and Paul Krugman. 1995. "When is the economy monocentric?: von Thünen and Chamberlin unified." *Regional Science and Urban Econ.*, 25(4): 505–528.

Gaynor, Martin, and Robert J Town. 2011. "Competition in health care markets." *Handbook of Health Economics* Vol. 2, 499–637. Elsevier.

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

Gaynor, Martin, Kate Ho, and Robert J Town. 2015. "The industrial organization of health-care markets." *Journal of Economic Literature*, 53(2): 235–284.

Geddes, Eilidh, and Molly Schnell. 2023. "The expansionary and contractionary supply-side effects of health insurance." National Bureau of Economic Research Working Paper 31483.

Gottlieb, Daniel J, Weiping Zhou, Yunjie Song, Kathryn Gilman Andrews, Jonathan S Skinner, and Jason M Sutherland. 2010. "Prices don't drive regional Medicare spending variations." *Health Affairs*, 29(3): 537–543.

Gottlieb, Joshua D., and Sean Nicholson. 2026. "Physician Competition: Entry and Substitution." *Journal of Economic Perspectives*, 40(2): 117–142.

Gottlieb, Joshua D., Maria Polyakova, Kevin Rinz, Hugh Shiplett, and Victoria Udalova. 2020. "Who Values Human Capitalists' Human Capital? The Earnings and Labor Supply of U.S. Physicians." U.S. Census Bureau Working Paper CES 20-23.

Gottlieb, Joshua D., Maria Polyakova, Kevin Rinz, Hugh Shiplett, and Victoria Udalova. 2025. "The Earnings and Labor Supply of U.S. Physicians." *The Quarterly Journal of Economics*, 140(2): 1243–1298.

Gotto, Antonio M., and Jennifer Moon. 2016. *Weill Cornell Medicine: A History of Cornell's Medical School.* Ithaca: Cornell University Press.

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

Gupta, Atul. 2021. "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program." *American Economic Review*, 111(4): 1241–83.

Hanson, Gordon H., and Chong Xiang. 2004. "The Home-Market Effect and Bilateral Trade Patterns." *American Economic Review*, 94(4): 1108–1129.

Helpman, Elhanan, and Paul R. Krugman. 1985. *Market Structure and Foreign Trade.* MIT Press.

Hsieh, Chang-Tai, and Esteban Rossi-Hansberg. 2023. "The Industrial Revolution in Services." *Journal of Political Economy Macroeconomics*, 1(1): 3–42.

Jensen, J. Bradford, and Lori G. Kletzer. 2005. "Tradable Services: Understanding the Scope and Impact of Services Offshoring." *Brookings Trade Forum*, 75–116.

Joskow, Paul L. 1980. "The Effects of Competition and Regulation on Hospital Bed Supply and the Reservation Quality of the Hospital." *The Bell Journal of Economics*, 11(2): 421–447.

Kline, Patrick, and Enrico Moretti. 2013. "Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority." *The Quarterly Journal of Economics*, 129(1): 275–331.

Krugman, Paul. 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade." *American Economic Review*, 70(5): 950–59.

Lee, Sanghoon. 2010. "Ability Sorting and Consumer City." *Journal of Urban Economics*, 68(1): 20–33.

Levy, Antoine, and Jacob Moscona. 2022. "Specializing in Density: Spatial Sorting and the Pattern of Trade." MIT, mimeo. Available online at https://economics.mit.edu/sites/default/files/inline-files/specializing_in_cities_website.pdf.

Lipsey, Robert E. 2009. "Measuring International Trade in Services." In *International Trade in Services and Intangibles in the Era of Globalization.* , ed. Marshall B. Reinsdorf and Matthew J. Slaughter, 27–74. University of Chicago Press.

Lopez, Eric, and Gretchen Jacobson. 2020. "How Much More Than Medicare Do Private Insurers Pay? A Review of the Literature."

Mayer, Thierry, Marc J. Melitz, and Gianmarco I. P. Ottaviano. 2021. "Product Mix and Firm Productivity Responses to Trade Competition." *The Review of Economics and Statistics*, 103(5): 874–891.

MedPAC. 2012. "Serving Rural Medicare Beneficiaries." In *Report to the Congress: Medicare and the Health Care Delivery System.* Chapter 5. Washington, DC:Medicare Payment Advisory Commission.

MedPAC. 2024. "Mandated Report: Rural Emergency Hospitals." In *Report to the Congress: Medicare Payment Policy.* Chapter 15. Washington, DC:Medicare Payment Advisory Commission.

Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J. Redding. 2025. "The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data." *The Quarterly Journal of Economics*, 140(4): 2507–2570.

Monte, Ferdinando, Stephen J Redding, and Esteban Rossi-Hansberg. 2018. "Commuting, Migration, and Local Employment Elasticities." *American Economic Review*, 108(12): 3855–3890.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

Moretti, Enrico. 2011. "Local Labor Markets." In *Handbook of Labor Economics*. Vol. 4, 1237–1313. Elsevier.

Mourot, Pauline. 2025. "Should Top Surgeons Practice at Top Hospitals? Sorting and Complementarities in Healthcare." Boston University, mimeo.

Muñoz, Mathilde. 2024. "Trading Nontradables: The Implications of Europe's Job-Posting Policy." *The Quarterly Journal of Economics*, 139(1): 235–304.

Neufeld, Jonathan D., and Charles R. Doarn. 2015. "Telemedicine Spending by Medicare: A Snapshot from 2012." *Telemed J E Health*, 21(8): 686–693.

Newhouse, Joseph P. 1990. "Geographic Access to Physician Services." *Annual Review of Public Health*, 11(1): 207–230.

Newhouse, Joseph P., Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. 1982*a*. "Does the Geographical Distribution of Physicians Reflect Market Failure?" *The Bell Journal of Economics*, 13(2): 493–505.

Newhouse, Joseph P., Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. 1982*b*. "The Geographic Distribution of Physicians: Is the Conventional Wisdom Correct?" Santa Monica: RAND Corp. Publ. No. R-2734.

Newhouse, Joseph P., Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. 1982*c*. "Where Have All the Doctors Gone?" *JAMA*, 247(17): 2392–2396.

Notowidigdo, Matthew J. 2020. "The Incidence of Local Labor Demand Shocks." *Journal of Labor Economics*, 38(3): 687–725.

Rosenblatt, R. A., and L. G. Hart. 2000. "Physicians and Rural America." *The Western Journal of Medicine*, 173(5): 348–351.

Rosenthal, Meredith B., Alan Zaslavsky, and Joseph P. Newhouse. 2005. "The Geographic Distribution of Physicians Revisited." *Health Services Research*, 40(6p1): 1931–1952.

Rosenthal, Stuart S., and William C. Strange. 2008. "The Attenuation of Human Capital Spillovers." *Journal of Urban Economics*, 64(2): 373–389.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Megan Schouweiler, and Matthew Sobek. 2022. "IPUMS USA: Version 12.0." Minneapolis, MN: IPUMS.

Salop, Steven C. 1979. "Monopolistic Competition With Outside Goods." *The Bell Journal of Economics*, 141–156.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics*, 88(4): 641–658.

Silver, David, and Jonathan Zhang. 2026. "Invisible Wounds: How Mental Disability Benefits Shape Veteran Well-Being." *American Economic Journal: Economic Policy*, 18(1): 282–317.

Skinner, Lucy, Douglas O. Staiger, David I. Auerbach, and Peter I. Buerhaus. 2019. "Implications of an Aging Rural Physician Workforce." *NEJM*, 381(4): 299–301.

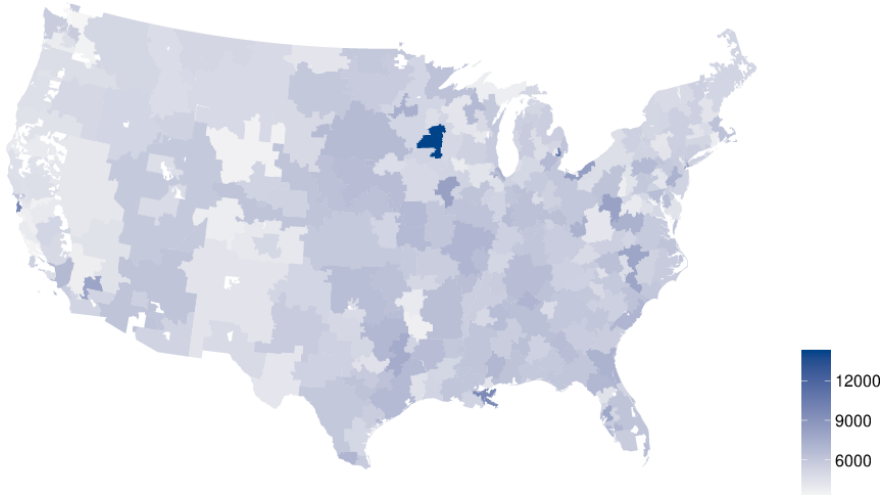
Suárez Serrato, Juan Carlos, and Philippe Wingender. 2014. "Estimating the Incidence of Government Spending." Working Paper, Duke University and the International Monetary Fund.

Zhao, Yafu, and William Encinosa. 2009. "The Cost of End-of-Life Hospitalizations, 2007." Agency for Healthcare Research and Quality 81.

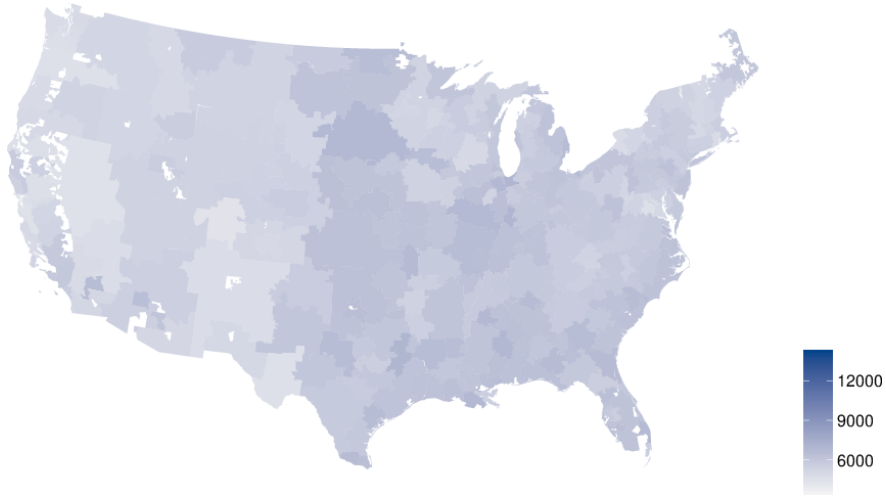
Zuckerman, Stephen, Laura Skopec, and Joshua Aarons. 2021. "Medicaid Physician Fees Remained Substantially Below Fees Paid By Medicare In 2019." *Health Affairs*, 40(2): 343–348.

Figure 1: Production, consumption, and trade across regions

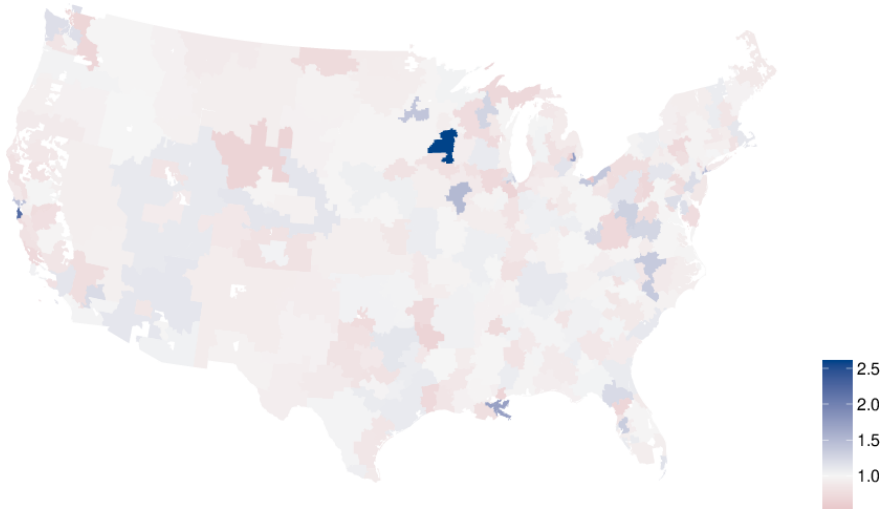
(a) Production per capita



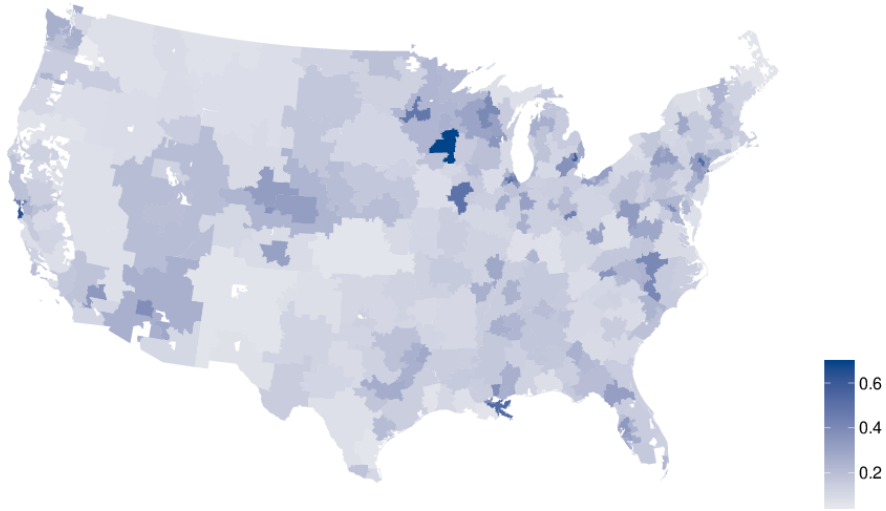
(b) Consumption per capita



(c) Production divided by consumption

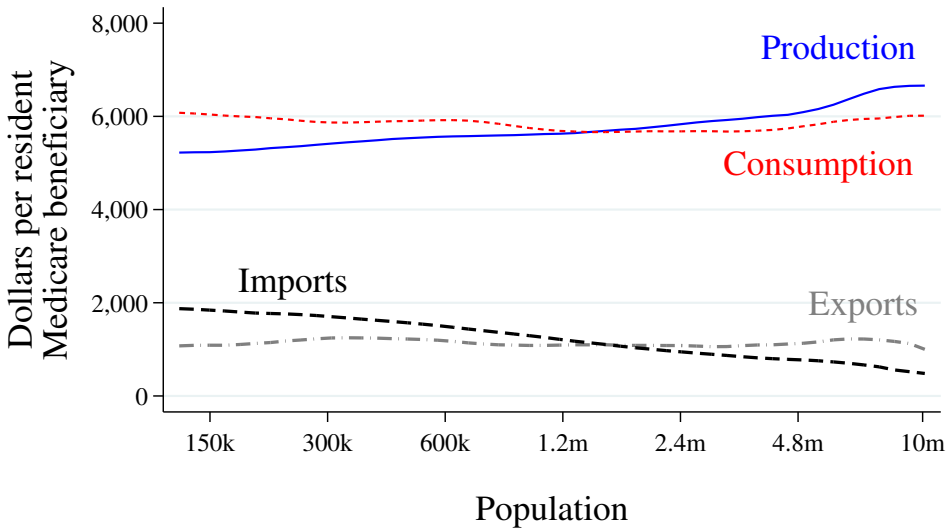


(d) Gross exports relative to production



Notes: These maps depict production, consumption, and trade by hospital referral region (HRR). Panel (a) shows dollars of production per capita. The HRR of production is the location where the service is provided. Panel (b) shows dollars of consumption per capita. The HRR of consumption is based on the patient's residential address. Panel (c) shows the ratio of production per capita to consumption per capita. Panel (d) shows gross exports as a share of production. Section 1.1 and Appendices A.1 and A.2 detail how we compute trade flows in physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. The Anchorage and Honolulu HRRs are not depicted.

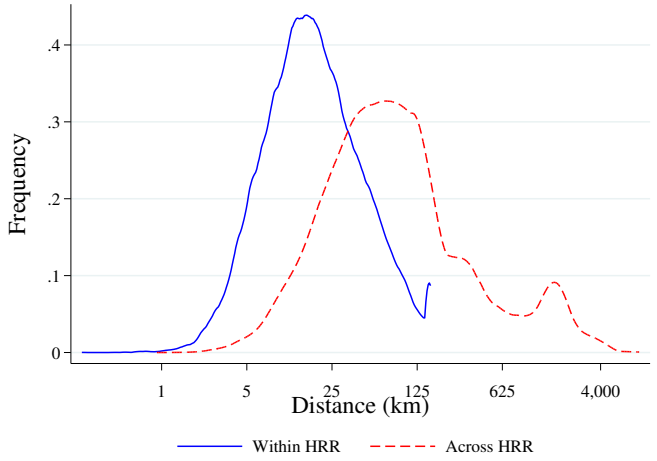
Figure 2: Production and consumption of medical care across regions



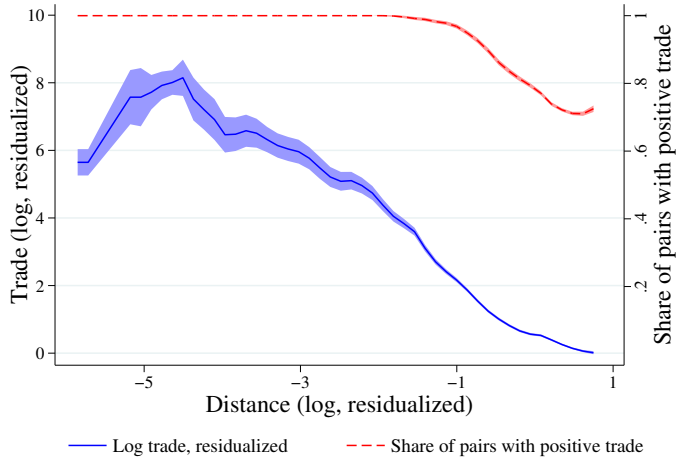
Notes: This figure shows production, consumption, and trade per capita of Medicare services across hospital referral regions (HRRs) of different sizes, all smoothed via local averages. The blue series shows production of medical care per Medicare beneficiary residing in the HRR of production. The red series shows consumption of medical care per Medicare beneficiary residing in the HRR of consumption. The dashed gray series shows interregional "exports" of medical care and the dashed black series shows interregional "imports" of medical care, again per Medicare beneficiary. Section 1.1 and Appendices A.1 and A.2 detail how we compute trade flows in physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project.

Figure 3: Patients travel between regions and trade declines with distance, more so for lower-income patients

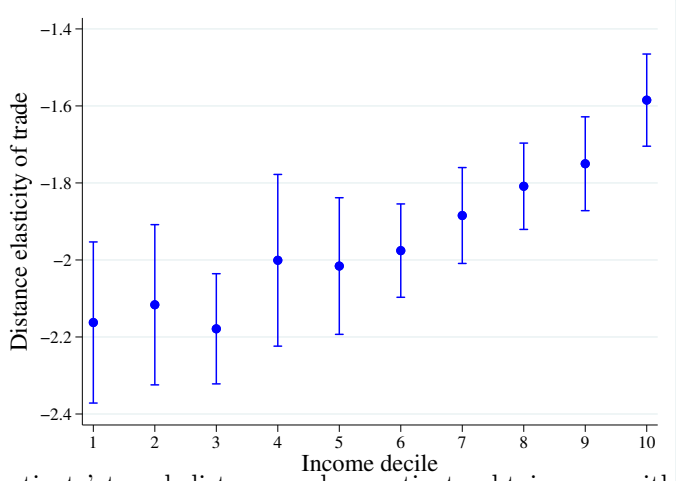
(a) Distribution of travel distances within and across HRRs



(b) Trade volume and extensive margin by distance



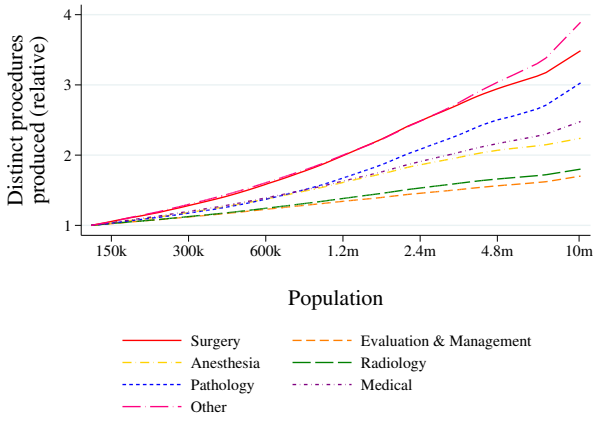
(c) Higher-income patients are less sensitive to distance



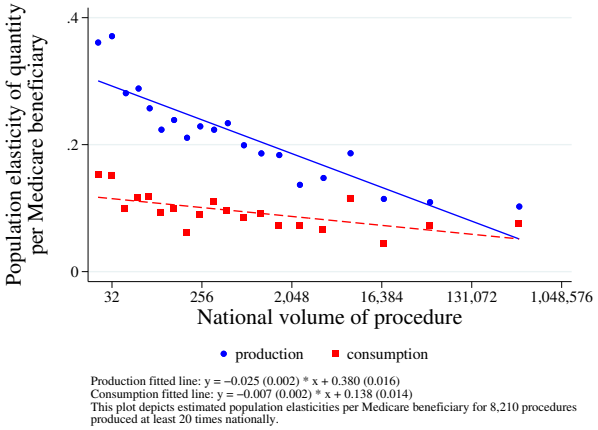
Notes: Panel (a) shows the distribution of patients' travel distances when patients obtain care within their home HRR (blue distribution) and when they travel across HRRs (red distribution). Travel distances measure the distance between home and treatment locations. For travel within a hospital referral region, the distance measure reflects the distance between the centroid of the patient's residential ZIP code and the ZIP code of the service location. We use ZCTA-to-ZCTA distances downloaded from the National Bureau of Economic Research; those exceeding 160 kilometers are winsorized at 160 kilometers. For travel across HRRs, we use ZCTA-to-ZCTA distances when they are within 160 kilometers and (for computational ease) use HRR-to-HRR distances beyond 160 kilometers. In Panel (b), the blue series depicts the volume of trade against distance, after conditioning out the fixed effects in equation (11), for positive-trade pairs of locations. The red series shows the observed share of HRR pairs with positive trade as a function of residualized distance from conditioning out the same set of fixed effects as in the blue series. Panel (c) depicts the coefficient on log distance obtained by estimating equation (11) separately for each decile of the national ZIP-level median-household-income distribution. The 95% confidence intervals are computed using standard errors two-way clustered by both patient HRR and provider HRR. Patients from higher-income ZIP codes are less sensitive to distance.

Figure 4: Production and consumption of medical care by market size and service

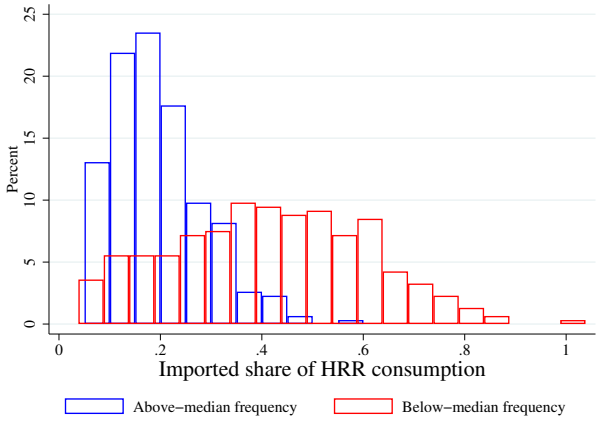
(a) Variety of procedures produced by market size



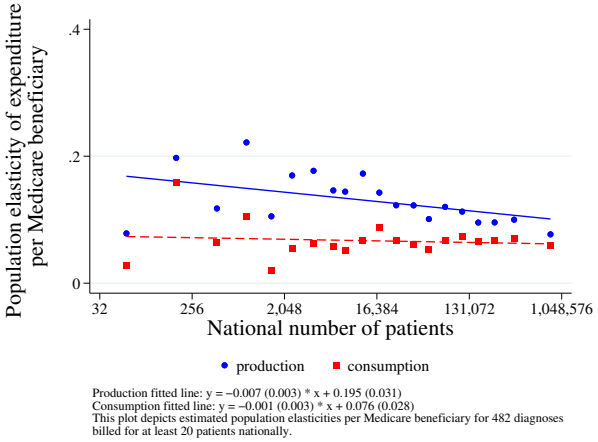
(b) Population elasticities by procedure frequency



(c) Import share distributions: common and rare



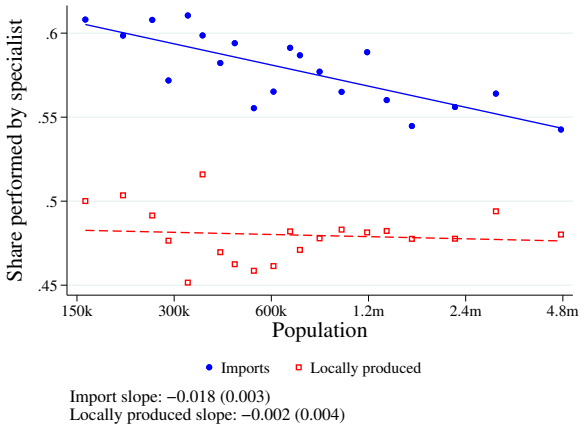
(d) Population elasticities by diagnosis frequency



Notes: Panel (a) depicts the number of distinct services produced in an HRR as a function of population. We use procedure classifications from the American Academy of Professional Coders, which groups codes into surgeries, anesthesia, radiology, pathology, medical, and evaluation & management services (AAPC, 2021). We combine Category II codes, Category III codes and Multianalyte Assays into "other." Within each group, the procedure count is relative to that of the smallest regions. More populous HRRs produce a greater variety of services. In Panels (b) and (d), the vertical axes are the population elasticities of quantity of medical care produced (blue dots) and consumed (red squares) per local Medicare beneficiary. The elasticities are computed using the Poisson model in equation (1) based on place of service and patients' residential location, respectively. Panel (b) estimates these elasticities for each of the procedures provided at least 20 times nationally in the Medicare data. Panel (d) estimates the elasticities for care provided to treat each of the Clinical Classifications Software Refined diagnoses billed for at least 20 patients nationally in the Medicare data. Each panel depicts these population elasticities as a function of the national volume (of procedures and diagnoses, respectively). Section 1.1 and Appendices A.1 and A.2 detail how we compute production and consumption of physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files. The contrasting population elasticities of production and consumption imply trade in medical services between markets of different sizes, with more net trade for rare procedures and rare diseases. Panel (c) shows the distribution of the imported consumption share across regions for two groups of procedures, split based on how often they are performed nationally. Imports are defined as care provided to a patient who lives in one HRR at a service location in a different HRR. Those performed less often than the median are shown in red, and those performed more often than the median are shown in blue. Import shares are substantially higher for the rare services.

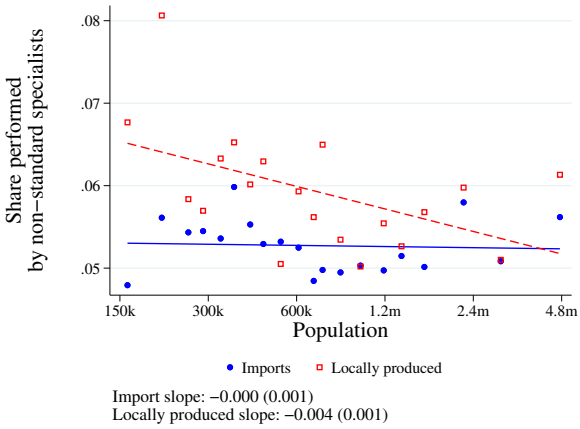
Figure 5: Imports are specialist-intensive, especially in smaller regions

(a) Specialty care imports



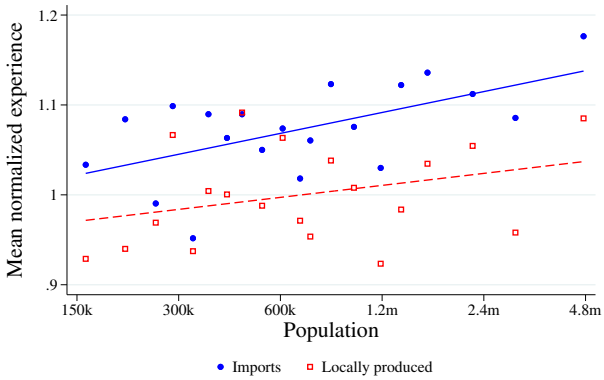
Import slope: −0.018 (0.003)
Locally produced slope: −0.002 (0.004)

(b) "Non-standard" specialty care



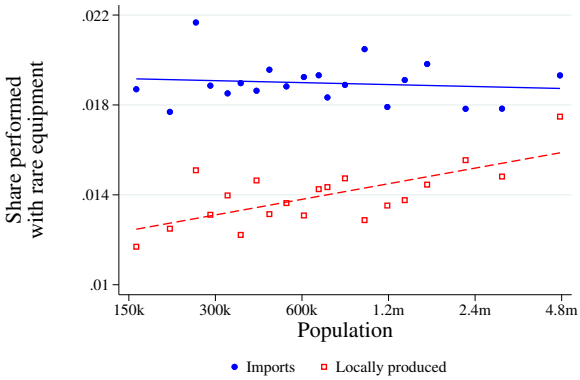
Import slope: −0.000 (0.001)
Locally produced slope: −0.004 (0.001)

(c) Provider experience



Imports slope: 0.034 (0.012)
Locally produced slope: 0.021 (0.018)

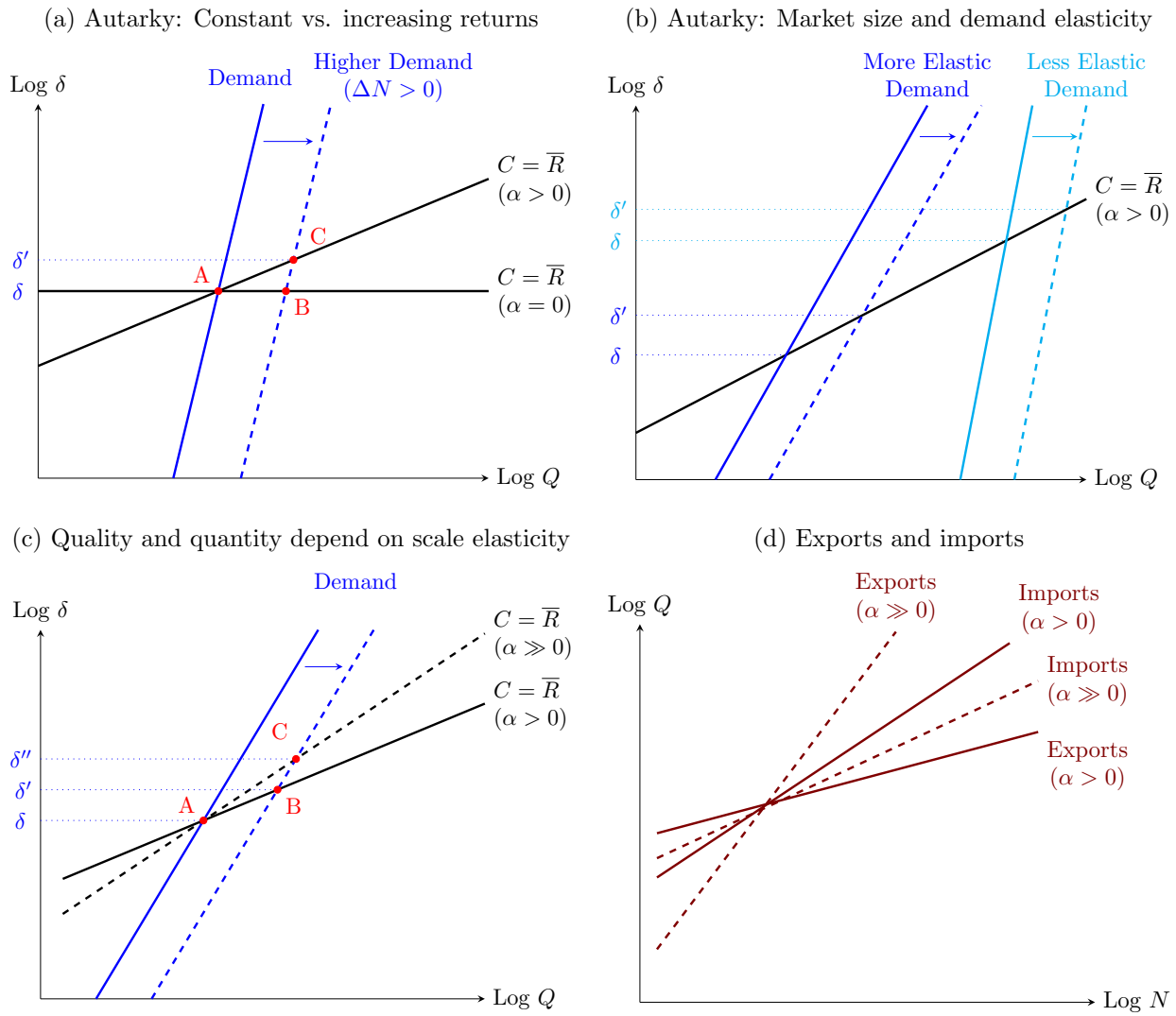
(d) Rare capital equipment imports



Import slope: −0.000 (0.000)
Locally produced slope: 0.001 (0.000)

Notes: Panel (a) shows the share of procedures that are performed by a specialist, for imports and locally produced procedures, by market size. We define generalists as internal-medicine, general-practice, and family-practice physicians and define specialists as all other physicians. Imports are more likely to be performed by a specialist, and smaller markets' imports especially so. Panel (b) examines procedures that are typically performed by specialists, and classifies the "standard" specialists as the top two specialties performing the procedure nationally. It shows the shares of procedures performed by the "non-standard" specialties in imported specialty care and locally produced specialty care as a function of local population size. Imports are less likely to be performed by "non-standard" specialties, especially for smaller regions. Panel (c) shows the mean relative experience of providers for care produced locally and imported by population size of the patient's region. This panel describes only procedures that are performed in all hospital referral regions (143 procedures). In public-use Medicare data, we define a provider's experience for a given procedure as the number of times they performed the procedure for Traditional Medicare patients in the prior calendar year. Before aggregating to the regional level, we rescale experience in each procedure so that its mean is one. We further normalize at the regional level so that the experience of the average HRR is one. On average, patients in larger markets obtain treatment from more experienced providers. At all population levels, imported care is produced by more experienced providers than local care. Panel (d) shows the share of procedures that use rare capital equipment in imported and locally produced care by market size. Rare equipment is defined as those pieces of equipment with below-median use in the Medicare data, defined as in Appendix A.1. We see that locally produced care in larger markets is more likely to use rare equipment, and imported care has a higher rare-equipment share throughout the population distribution.

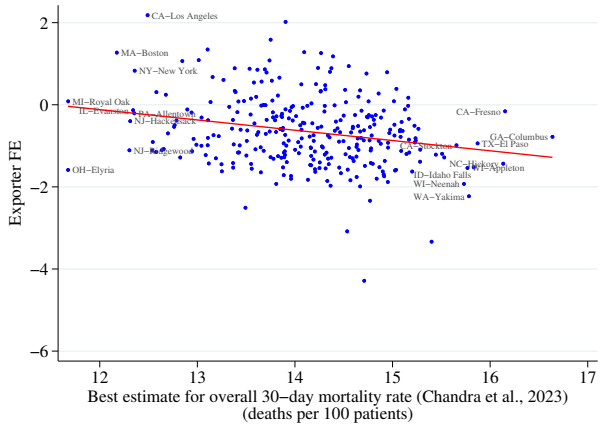
Figure 6: Illustrative model diagrams



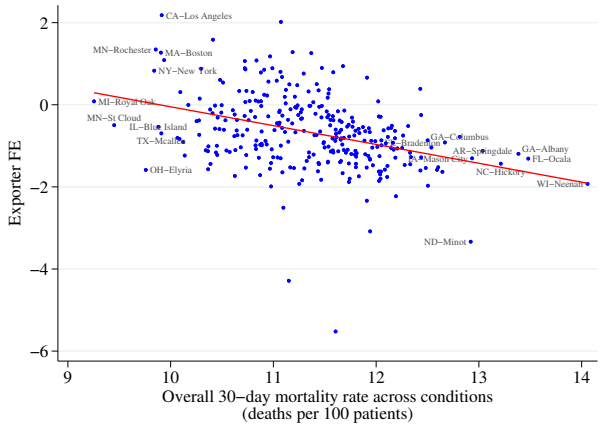
Notes: This figure depicts how increasing demand in one region affects its equilibrium outcomes. In Panels (a)–(c), quantity produced Q is on the horizontal axis and service quality δ is on the vertical axis. The black lines depict the free-entry isocost curve, $C = \overline{R}$, given by equation (4). The blue and cyan lines depict demand for the region's service, which we depict as log-linear for visual clarity. (The logit demand function is actually log-convex, which is consistent with all the depicted comparative statics.) Equilibrium is the intersection of the demand and isocost curves. An increase in demand is the rightward shift from the solid to the dashed demand curve. This shift increases equilibrium quality from δ to δ'. Panel (a) shows that higher demand elicits higher quality if there are increasing returns to scale. Panel (b) shows that this quality improvement is larger when demand is more elastic. Panels (c) and (d) introduce trade and compare the extent of quality improvement under two different magnitudes of increasing returns ($\alpha > 0$ and $\alpha \gg 0$). These magnitudes govern the patterns of interregional trade, shown in Panel (d) as a function of the number of potential patients N. Imports from other regions rise with N. With increasing returns to scale ($\alpha > 0$), exports to other regions also rise with N (a weak home-market effect). When the scale elasticity α is larger ($\alpha \gg 0$), the import curve is flatter and the export curve is steeper. With sufficiently strong increasing returns, an increase in local demand causes a greater increase in exports than imports (a strong home-market effect).

Figure 7: Estimated quality is positively correlated with total output and external quality metrics

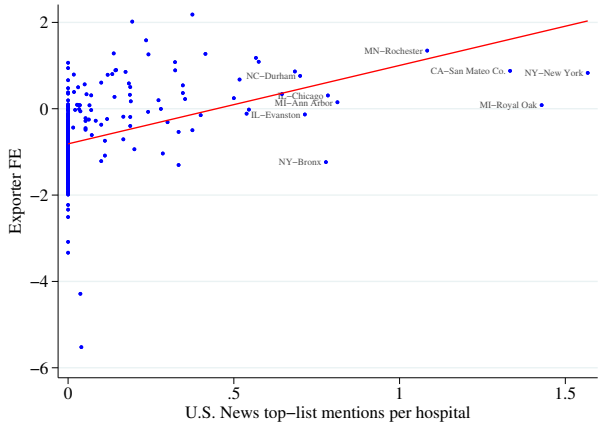
(a) Quality estimates vs. mortality rate (Chandra et al.)



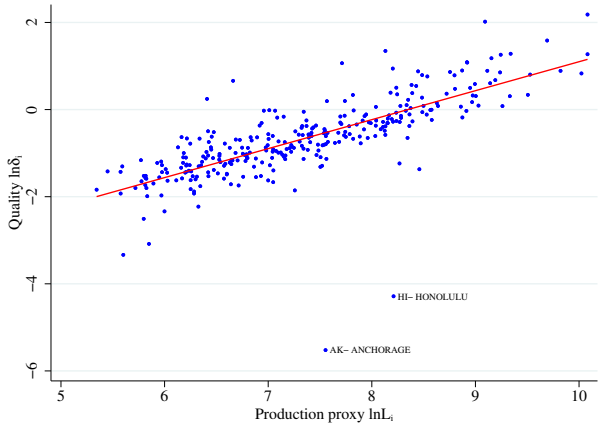
(b) Quality estimates vs. mortality rate (CMS)



(c) Quality estimates vs. *U.S. News*-ranked hospitals



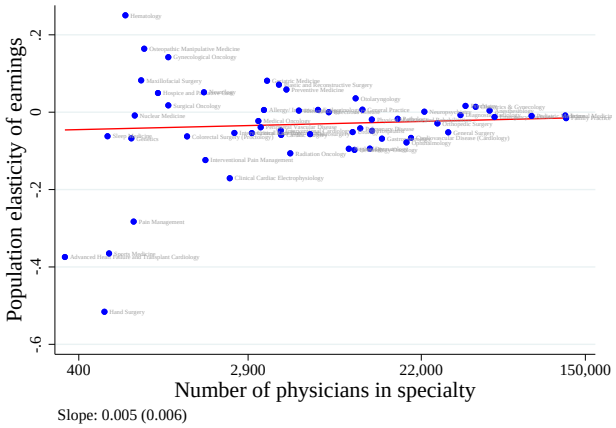
(d) Quality is higher in regions producing more output



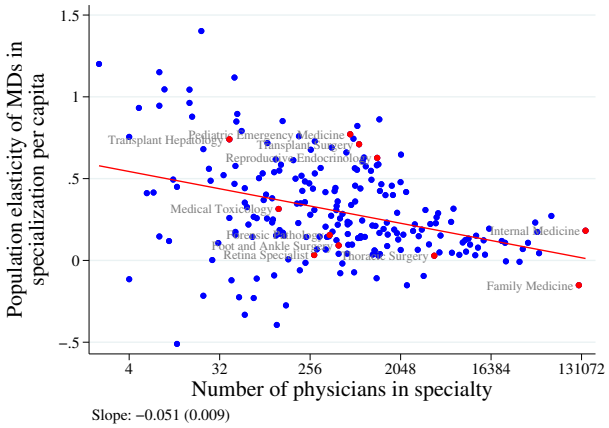
Notes: The first three panels show the relationship between the exporter fixed effects (our revealed-preference measure of quality) and external quality measures. The vertical axis shows the exporter fixed effects for each HRR estimated from equation (11). The horizontal axis in Panel (a) is an average of the hospital mortality rates estimated by Chandra, Dalton, and Staiger (2023) in a region, and in Panel (b) an average of those estimated by CMS. Chandra, Dalton, and Staiger (2023) do not provide a mortality rate for Anchorage, Alaska. The negative correlations indicate patients travel farther to obtain care from regions with lower hospital mortality rates. The horizontal axis in Panel (c) is the number of times each region's hospitals appear on the *U.S. News* list of best hospitals, divided by the number of hospitals in the region. *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 and normalize by the region's number of hospitals. The relationship is positive, indicating that patients travel farther to obtain care from regions highly ranked by *U.S. News*. The horizontal axis in Panel (d) is the labor-based proxy for production described in Appendix C.3. The scatterplot using Traditional Medicare production on the horizontal axis is extremely similar (Appendix Figure E.5). Section 1.1 and Appendices A.1 and A.2 detail how we compute production and trade of physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files.

Figure 8: Regional patterns of specialized physicians and rare capital equipment

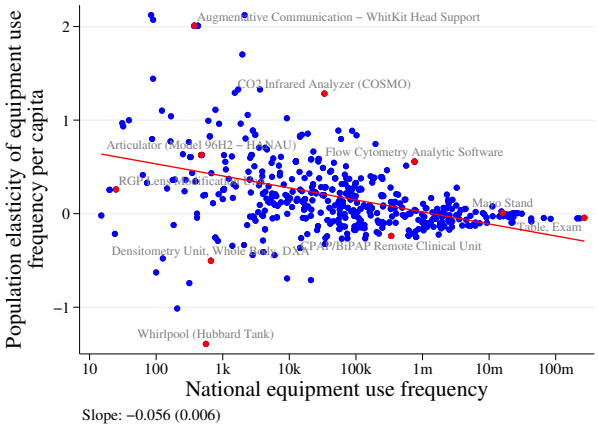
(a) Population elasticities of specialties' earnings



(b) Population elasticities of physician specializations

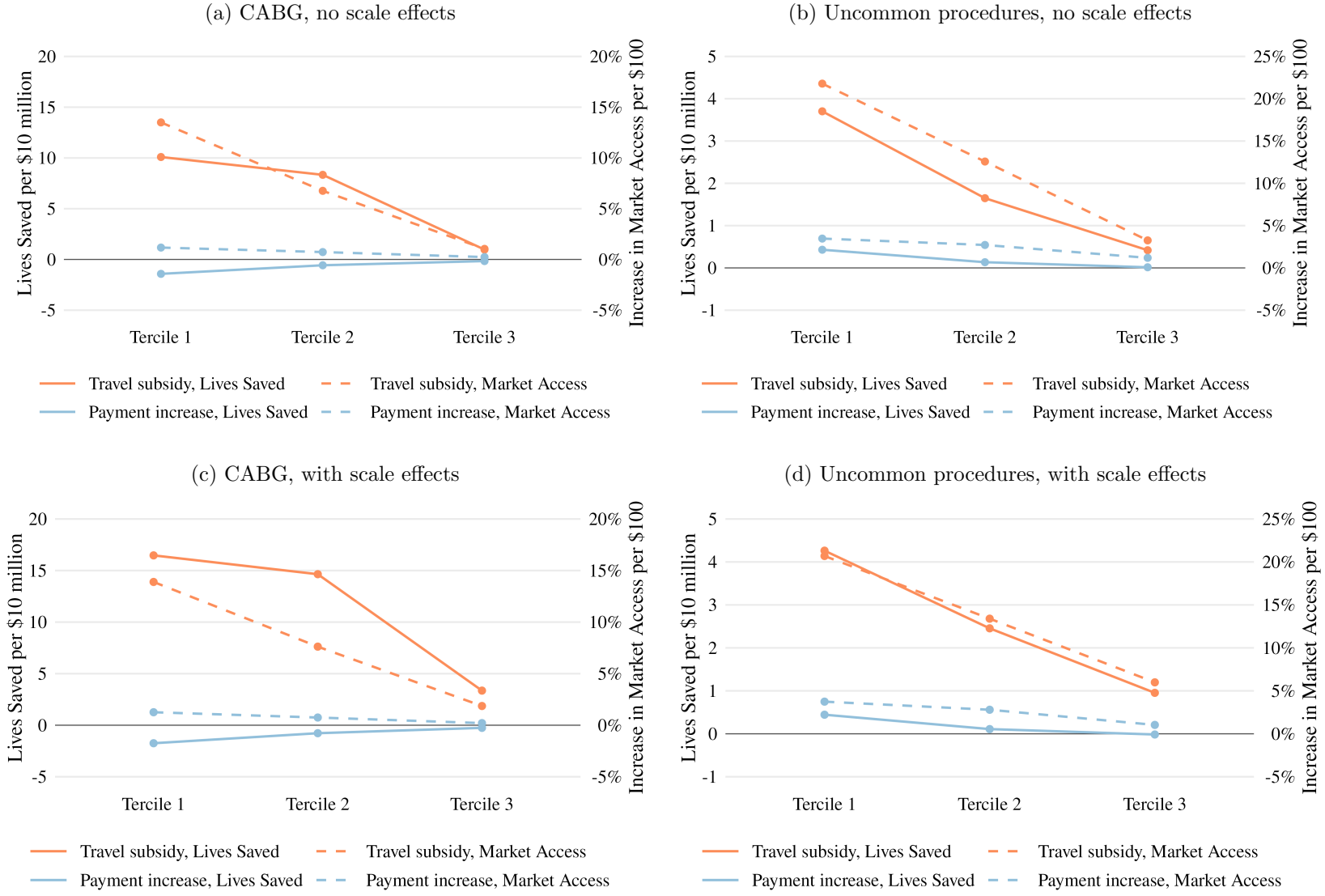


(c) Population elasticities of equipment use



Notes: Panel (a) 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 regression line weights each specialty by the number of commuting zones used to estimate its income elasticity. The graph shows that these elasticities are unrelated to the total national count of physicians in those specialties. The vertical axis of Panel (b) depicts the population elasticities of quantity of physicians in an HRR. The population elasticities are computed for each specialty using the Poisson model in equation (21). The horizontal axis shows the nationwide number of physicians in each specialty. The negative relationship indicates that rare specialties are disproportionately concentrated in high-population regions. Panel (c) plots the population elasticity of equipment use frequency per capita against the national frequency of equipment use, for each piece of equipment. Pieces of equipment with elasticities above the 98th percentile are excluded. The graph shows that rarely used equipment is used more intensively in large markets.

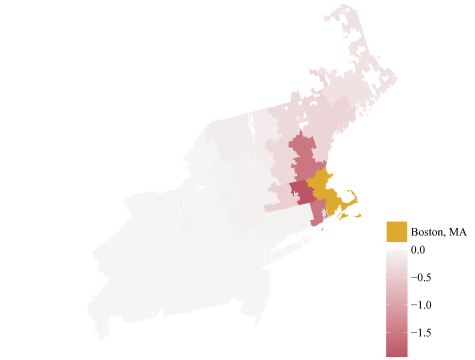
Figure 9: Gains from counterfactual rural reimbursement or travel subsidies by income tercile



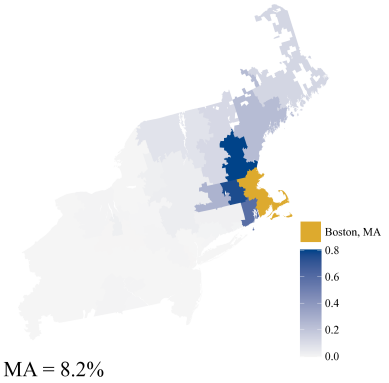
Notes: Panels (a) and (b) compare changes in mortality and patient market access for a reimbursement subsidy in rural regions versus a travel subsidy for patients in rural regions across income terciles, under the assumption of no scale effects ($\alpha = 0$). Panels (c) and (d) show the same comparison across income terciles but with scale effects ($\alpha = 0.6$). We model the travel subsidy based on the actual travel subsidy policy in Montana's Medicaid program as detailed in Appendix D.5. Rural HRRs are defined as HRRs with below-median population density. CABG surgeries are identified using HCPCS codes 33510–33536 in the carrier file. The exercise is described in detail in Section 5 and Appendices D.5 and D.6.

Figure 10: Counterfactual outcomes for higher reimbursements in one region

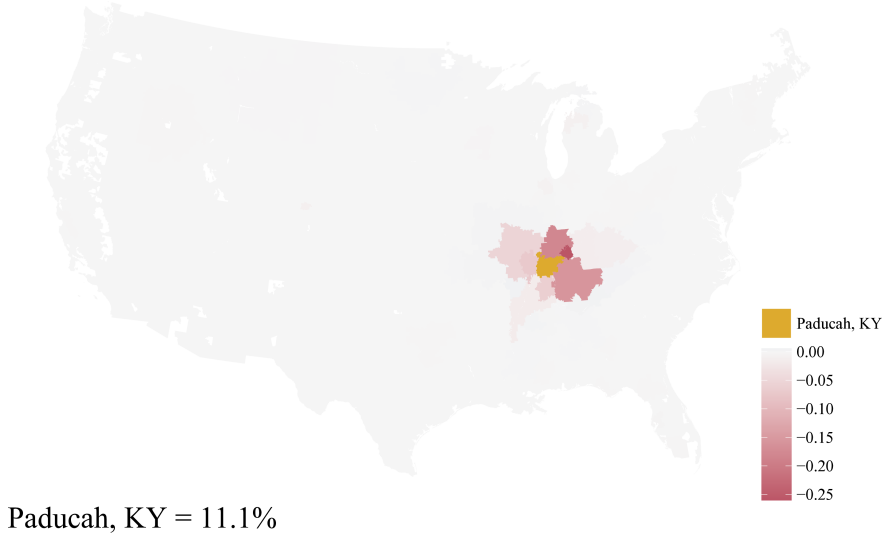
(a) Change (%) in output quality δ_i: higher reimbursement in Boston, MA



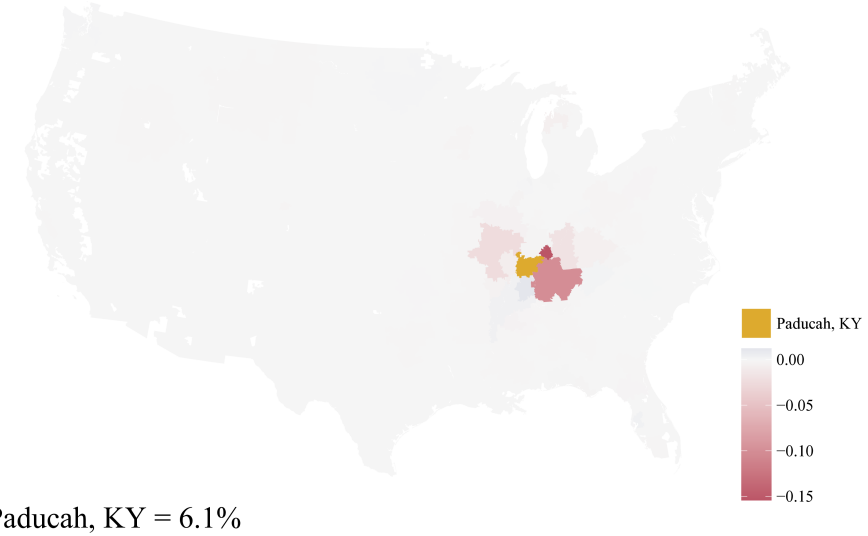
(b) Change (%) in market access Φ_i: higher reimbursement in Boston, MA



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

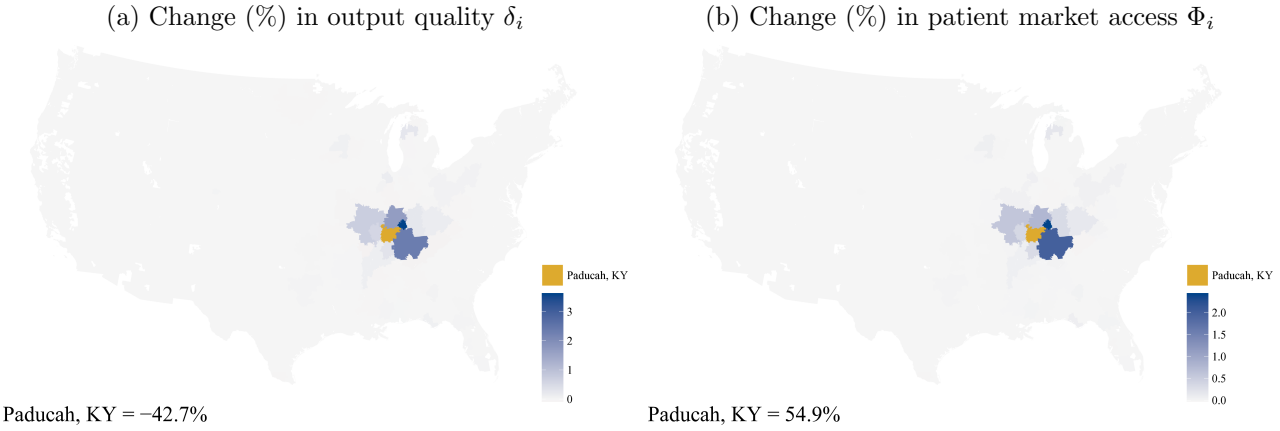


(d) Change (%) in market access Φ_i: higher reimbursement in Paducah, KY



Notes: Panels (a) and (b) show the impacts of increasing reimbursements for uncommon procedures by 8% in the Boston, MA HRR ($\hat{R}_i = 1.08$) based on our estimated model. Panel (a) illustrates the percentage change in quality of care δ_i provided in each region. Panel (b) illustrates the percentage change in the value of market access Φ_i for patients who live in a region. Panels (c) and (d) are analogous, but for an 8% increase in reimbursements in Paducah, KY, a net importer. In all panels, the predicted change for the region whose reimbursement changes ("treated region") is listed on the map itself. In both cases, the quality produced in neighboring regions declines (Panels (a) and (c)). Patients in regions near Boston benefit from increased access to the treated region (Panel (b)), so there is a negative relationship between the percentage changes in δ and Φ across regions. In contrast, patients in regions near Paducah suffer a decrease in access (Panel (d)). The contrasting outcomes stem from Boston being a net exporter and Paducah being a net importer in the baseline equilibrium. The exercise is described in detail in Section 5.

Figure 11: Counterfactual outcomes when changing travel costs for Paducah, KY residents



Notes: This figure shows the impacts of a counterfactual standard travel subsidy for Paducah residents only. We model the travel subsidy based on the actual travel subsidy policy in Montana's Medicaid program as detailed in Appendix D.5 and apply it to travel for uncommon procedures. Panel (a) depicts the percentage change in quality of care δ_i produced in each region. Panel (b) depicts the percentage change in the value of market access Φ_i for patients who live in a region. The notes report the changes for Paducah.

Table 1: Contrasting geographies of colonoscopies and LVAD implantation

	Colonoscopy	LVAD Implant
Code	G0121	33979
N	58,785	333
Physicians	13,469	177
$\hat{\beta}_p^{\text{production}}$	-0.01	0.87
$\hat{\beta}_p^{\text{consumption}}$	-0.01	0.03
Share traded (HRR)	0.10	0.48
Share traded (CBSA)	0.11	0.48
Median distance traveled (km)	13.83	65.27
Share > 100km	0.04	0.37

Notes: This table reports statistics for two HCPCS codes: screening colonoscopy (G0121) and LVAD implantation (33979). We report the number of times the procedure is performed in 2017 in our 20% sample of Medicare patients and the number of distinct physicians performing it. The population elasticities of production and consumption are estimated using the Poisson models in equation (1) based on production HRR and patients' residential HRR, respectively. We also report the shares of procedures in which the patient and service location are in different HRRs or CBSAs, the median distance traveled for all care, and the share in which the patient and service location are more than 100 kilometers apart.

Table 2: Scale elasticity estimates

All services	Baseline	No Diagonal	Controls	Labor proxy
OLS: 2017	0.797	1.102	0.723	0.690
	(0.025)	(0.050)	(0.026)	(0.024)
2SLS: log population	0.775	1.017	0.693	0.659
	(0.027)	(0.055)	(0.028)	(0.024)
	[2163]	[2163]	[1361]	[5226]
2SLS: log 1940 population	0.766	1.113	0.698	0.680
	(0.030)	(0.053)	(0.031)	(0.028)
	[391]	[391]	[362]	[353]
2SLS: log share 65+	0.553	1.438	0.673	1.607
	(0.175)	(0.472)	(0.162)	(0.508)
	[90.1]	[90.1]	[76.7]	[12.4]
2013–2017 difference	1.244	1.233	1.223	0.794
	(0.100)	(0.109)	(0.101)	(0.223)

Notes: This table reports ordinary least squares (OLS) or two-stage least squares (2SLS) estimates of α from equation (12), $\widehat{\ln \delta}_i = \alpha \ln Q_i + \Upsilon \mathbf{X}_i + \Xi_{s(i)} + u_i$, using hospital referral regions (HRRs) in multi-HRR states as the geographic units (so $N = 298$ in all regressions). The dependent variable $\widehat{\ln \delta}_i$ is estimated in equation (11) using a same-region dummy and a quadratic function of log distance. Q_i is region i's total production for Medicare beneficiaries, $\mathbf{X}_i$ is a vector of optional controls explained below, $\Xi_{s(i)}$ is a state fixed effect, and u_i is an error term. In the "no diagonal" column, S_{ii} observations were omitted when estimating $\widehat{\ln \delta}_i$ in equation (11). The third column adds controls for log housing prices, log non-healthcare wages, Medicare's Geographic Adjustment Factor, and our proxy for doctor ability (coefficients not reported). In the "labor proxy" column, $\ln Q_i$ is replaced by the labor proxy $\ln L_i$ (Appendix C.3). The 2013–2017 difference specification estimates equation (15), $\Delta\widehat{\ln \delta}_i = \alpha \Delta \ln Q_i + \Xi_{s(i)} + \Upsilon \Delta \mathbf{X}_i + \Delta u_i$. In the rows labeled "2SLS," we instrument for $\ln Q_i$ using the specified instruments. The standard errors in parentheses are robust to heteroskedasticity. For 2SLS estimates, first-stage effective F-statistics (Montiel Olea and Pflueger, 2013) are reported in square brackets. All estimates reveal substantial scale economies.

Table 3: Regions with more elderly residents produce higher-quality elderly-specific care

	Above median		Terciles	
	Diag	No Diag	Diag	No Diag
Elderly specificity above median × regional log elderly share	0.167	0.218		
	(0.0666)	(0.0652)		
Elderly specificity tercile 2 × regional log elderly share			0.141	0.130
			(0.0932)	(0.0940)
Elderly specificity tercile 3 × regional log elderly share			0.325	0.337
			(0.125)	(0.109)
Observations	612	612	918	918
Region fixed effects	Yes	Yes	Yes	Yes
Procedure group fixed effects	Yes	Yes	Yes	Yes

Notes: This table reports estimates of equation (14), $\widehat{\ln \delta_i^g} = \Lambda(\text{above-median elderly specificity}_g \times \ln \text{elderly share}_i) + \eta_g + \eta_i + u_{ig}$, where $\widehat{\ln \delta_i^g}$ is the estimated quality of region i's production in procedure group g, obtained from a gravity equation estimated separately for each procedure group. A procedure's elderly specificity is its share of services provided to Medicare patients out of total services provided to Medicare and HCCI (private-insurance) patients. Columns 1–2 split procedures into two groups based on whether their Medicare share is above or below the median; columns 3–4 split procedures into terciles of Medicare share, with the omitted group being the lowest tercile. The elderly share in region i is the share of its population aged 65 and older. The "Diag" columns use all HRR-pair observations to estimate the outcome variable, procedure-group-by-region quality $\widehat{\ln \delta_i^g}$, while the "No Diag" columns omit the within-region (diagonal) observations. All specifications include region fixed effects η_i and procedure-group fixed effects η_g. The positive coefficients on the interaction with the elderly share indicate that regions with a larger share of elderly residents produce higher-quality care for elderly-specific procedures. The standard errors in parentheses are clustered by HRR.

Table 4: Scale elasticity for mesothelioma care

	Baseline	No Diagonal	Controls	MA share
Log production	0.583	0.683	0.559	0.582
	(0.097)	(0.249)	(0.105)	(0.096)
Log population	0.286	0.662	0.204	0.286
	(0.140)	(0.331)	(0.165)	(0.142)
$\text{Log}(1 + M_i^{\text{MA}}/M_i^{\text{TM}})$				0.038
				(0.320)
Observations	298	269	298	298
State fixed effects	Yes	Yes	Yes	Yes
First-stage F statistic	49	48	42	59

Notes: This table reports two-stage least squares (2SLS) estimates of the scale elasticity of mesothelioma care from equation (16), $\widehat{\ln \delta_i}^{\text{meso}} = \alpha^{\text{meso}} \ln Q_i^{\text{meso}} + \tau \ln \text{population}_i + \Xi_{s(i)} + \Upsilon \mathbf{X}_i + u_i$ using hospital referral regions (HRRs) as the geographic units. The dependent variable $\widehat{\ln \delta_i}^{\text{meso}}$ is estimated from a gravity equation for mesothelioma care using a same-region dummy and a quadratic function of log distance. Log production $\ln Q_i^{\text{meso}}$ is instrumented using the HRR-level mesothelioma death rate, calculated using population estimates from the 2015–2019 ACS and mesothelioma death counts over 1999–2020 from CDC WONDER. To increase statistical power when studying this rare disease, all regressions pool data over 2016–2019. Column 1 reports the baseline 2SLS specification with log population and state fixed effects as controls. Column 2 omits the diagonal observations (S_{ii}) when estimating $\widehat{\ln \delta_i}^{\text{meso}}$. This column has fewer observations because around 10% of regions have zero exports of mesothelioma care, so no quality estimate is available once the diagonal of the trade matrix is omitted. Column 3 adds controls for log housing prices, log non-healthcare wages, Medicare's Geographic Adjustment Factor, and our proxy for doctor ability (coefficients not reported). Column 4 adds a control for the ratio of Medicare Advantage to Traditional Medicare beneficiaries, $\ln\big(1 + M_i^{\text{MA}}/M_i^{\text{TM}}\big)$, which proxies for measurement error in $\ln Q_i$ due to only observing Traditional Medicare patients. All regressions control for state fixed effects $\Xi_{s(i)}$. The standard errors in parentheses are robust to heteroskedasticity. The first-stage effective F-statistics (Montiel Olea and Pflueger, 2013) are reported in the bottom row. All estimates indicate substantial scale economies in mesothelioma care.

Table 5: Aggregate medical services exhibit a strong home-market effect

	Cross-sectional PPML			IV: 1940 population	Cross-sec 65+ share	2013–2017 panel
Provider-market log population	0.671	0.681	0.671	0.757	0.715	1.150
	(0.0543)	(0.0505)	(0.0366)	(0.0547)	(0.0634)	(0.199)
Patient-market log population	0.312	0.304	0.338	0.336	0.299	-0.028
	(0.0547)	(0.0501)	(0.0346)	(0.0467)	(0.0616)	(0.199)
Provider-market log share over 65					0.425	
					(0.301)	
Patient-market log share over 65					-0.097	
					(0.295)	
Log distance	-1.627	0.344		0.377	0.318	
	(0.0489)	(0.304)		(0.250)	(0.291)	
(Log distance)^2		-0.199		-0.201	-0.197	
		(0.0305)		(0.0247)	(0.0293)	
p-value: no strong HME	<0.001	<0.001	<0.001	<0.001	0.188	<0.001
Observations	93,636	93,636	93,636	93,636	93,636	162,678
Fixed effects						ij
Distance elasticity at mean		-1.59		-1.57	-1.60	
Distance deciles			Yes			

Notes: This table reports estimates of equations (17), (18), and (19), which evaluate the presence of weak or strong home-market effects. The sample is all HRR pairs ($N = 306^2$), and both 2013 and 2017 observations for the last column. The dependent variable in all regressions is the value of trade when including professional and facility (inpatient and outpatient) fees at national average prices. Columns 1 through 5 use 2017 data. In column 1, 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. Column 5 adds the patient- and provider-market log shares of the population aged 65 and older as demand shifters, conditional on log population. Column 6 presents estimates of equation (19) including ij fixed effects using data from 2013 and 2017. Section 1.1 and Appendices A.1 and A.2 detail how we compute trade flows in physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files. The coefficients on patient-market population and 65+ share are adjusted for observability bias following Appendix C.4. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 6: The home-market effect is stronger for rare procedures and diagnoses

	Procedure		Procedure		Procedure		Diagnosis	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
λ_X Provider-market log population	0.618	0.605	0.603		0.606		0.601	
	(0.0516)	(0.0493)	(0.0493)		(0.0488)		(0.0489)	
λ_M Patient-market log population	0.360	0.364	0.366		0.364		0.371	
	(0.0519)	(0.0492)	(0.0492)		(0.0486)		(0.0486)	
μ_X Provider-market log population × rare			0.344	0.329	0.362	0.317	0.120	0.110
			(0.0447)	(0.0405)	(0.0452)	(0.0392)	(0.0232)	(0.0206)
μ_M Patient-market log population × rare			-0.241	-0.239	-0.250	-0.220	-0.0986	-0.0915
			(0.0606)	(0.0587)	(0.0612)	(0.0564)	(0.0186)	(0.0172)
p-value for H_0: $\lambda_X \leq \lambda_M$	0.005	0.006	0.007		0.005		0.008	
p-value for H_0: $\mu_X \leq \mu_M$			<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Observations	187,272	110,402	110,402	110,402	110,402	110,402	109,658	109,658
Distance [linear] controls	Yes	Yes	Yes	Yes				
Distance [quadratic] controls					Yes	Yes	Yes	Yes
Patient-provider-market-pair FEs				Yes		Yes		Yes

Notes: This table reports estimates of equation (20), which introduces interactions with an indicator for whether a procedure or diagnosis is "rare" (provided less often than the median, 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/diagnoses. The unit of observation is {rare indicator, exporting HRR, importing HRR} so the number of observations is 2×306^2 in column 1, and the dependent variable in all regressions is the value of trade. Columns 1–6 reflect estimates by procedure and Columns 7 and 8 reflect estimates by diagnosis. Columns 2 onwards drop HRR pairs with zero trade in both procedure groups, and column 2 shows that this restriction has a negligible impact on the estimated log population coefficients. Columns 3 onwards include the rare indicator interacted with patient- and provider-market populations and distance covariates. Columns 1–4 control for distance using the log of distance between HRRs. Columns 5–8 add a control for the square of log distance. Columns 4, 6, and 8 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 and negative coefficient on patient-market population × rare across all columns indicate that the strong home-market effect is greater for rare services. This result is robust to defining rare procedures using all-payer frequency instead of Traditional-Medicare frequency. Valid primary diagnoses observed in 1,000 distinct claims or more nationally in the professional fees 20% sample are included. Section 1.1 and Appendices A.1 and A.2 detail how we compute these trade flows for physician services (excluding emergency-room care and skilled nursing facilities) at standardized prices from the Medicare 20% carrier, 100% MedPAR, and 100% outpatient claims Research Identifiable Files. HRR definitions are from the Dartmouth Atlas Project. Standard errors (in parentheses) are two-way clustered by patient market and provider market.

Table 7: Counterfactual gains from reimbursement vs. travel subsidies for rural care

	(1)	(2)	(3)	(4)
Eligible services	Uncommon		CABG	
Scale elasticity	$\alpha = 0$	$\alpha = 0.6$	$\alpha = 0$	$\alpha = 0.6$
Panel A. Lives saved				
(1) Rural reimbursements	0.5	0.4	-2.2	-2.9
(2) Travel subsidies	5.4	7.4	24.4	41.1
(3) Generous travel subsidies	4.8	7.3	28.9	49.9
Panel B. Market access				
(4) Rural reimbursements	7.5%	7.7%	2.1%	2.2%
(5) Travel subsidies	41.9%	43.4%	22.2%	24.2%
(6) Generous travel subsidies	67.9%	69.0%	37.0%	38.9%

Notes: This table summarizes outcomes for three counterfactual policies subsidizing rural healthcare: a reimbursement increase, a standard travel subsidy, and a more generous travel subsidy. Panel A reports lives saved per $10 million spent (rows 1–3); Panel B reports the percentage change in patient market access per $100 spent (rows 4–6). Columns 1 and 2 consider all uncommon procedures; columns 3 and 4 consider Coronary Artery Bypass Grafting (CABG). Columns 1 and 3 report results without scale effects ($\alpha = 0$); columns 2 and 4 report results with scale effects ($\alpha = 0.6$). We model the travel subsidy based on the actual travel subsidy policy in Montana's and South Dakota's Medicaid programs as detailed in Appendix D.5. Rural HRRs are defined as HRRs with below-median population density. CABG surgeries are identified using HCPCS codes 33510–33536 in the Medicare Carrier file. The exercise is described in detail in Section 5 and Appendices D.5 and D.6.