Quasi-experimental Evidence of Physician Effects

Peter Tu†

1 November 2017

Please click here for the most recent version. ¹

Abstract

Variation in health care utilization is large and exists both across geographic regions and across doctors in the same region. However, the literature is only just beginning to understand the drivers of such variation. Are doctors prescribing different care for the same types of patients or are patients different in ways unobservable to researchers? This paper estimates the share of variation in physician practice intensity that is due to doctor differences as opposed to patient differences. The quasi-experiment leverages changes in the referral network of primary care physicians resulting from specialist migrations and exits. Using Medicare claims data, I estimate that between 50% and 70% of variation is driven by doctor styles. I replicate the baseline results using an instrumental variables strategy, suggesting that identification is robust to the most-likely endogeneity threats. Higher or lower health care usage is not correlated with mortality or hospitalization rates, but higher initial utilization does lead to greater subsequent utilization. In contrast to past studies, I estimate the treatment effect of doctors, rather than the treatment effect of geographic regions. Since doctors are the key decision-makers in health care provision and since variation across doctors vastly exceeds variation across regions, understanding the causal effect of doctors is vital for understanding the sources of variation and for designing policies to reduce that variation.

JEL Codes: H51, I1, I11

¹Email: petertu@fas.harvard.edu. Department of Economics. Harvard University. I am indebted to the invaluable encouragement, support, and guidance of Nathan Hendren, David Cutler, and Anupam Jena. I also thank Jay Bhattacharya, Edward Glaeser, Brian Chen, David Choi, John Coglianelse, Siddharth George, Heather Sarsons, Alex Segura, Nihar Shah, Chenzi Xu, and seminar participants at Harvard’s Public/Labor lunch and Health Economics seminar for their helpful questions and suggestions.

²http://scholar.harvard.edu/petertu
Contents

1 Introduction ................................................................. 3
  1.1 Relation to previous literature ..................................... 5
2 Empirical Strategy ......................................................... 7
  2.1 Statistical Model ...................................................... 8
  2.2 Identification .......................................................... 11
  2.3 Identification assumptions ......................................... 13
3 Data and Summary Statistics ............................................ 14
  3.1 Defining utilization measures ....................................... 17
  3.2 Identifying referral relationships and specialist separations ... 19
  3.3 Sample restrictions ................................................... 21
  3.4 Summary statistics .................................................... 22
4 Results: Health Care Utilization ........................................ 23
  4.1 Estimating physician effects ......................................... 23
  4.2 Constructing $\delta_{rd}$ ................................................ 24
  4.3 Main results ............................................................ 25
  4.4 Instrumenting for $\delta_{rd}$ ............................................ 30
  4.5 Robustness checks .................................................... 32
  4.6 Estimating break effects directly .................................. 35
  4.7 Components of $\delta_{rd}$ .............................................. 39
  4.8 Variance decomposition .............................................. 40
5 Results: Patient Health ................................................... 41
  5.1 Mortality rates ........................................................ 42
  5.2 Hospitalization rates ................................................. 44
  5.3 Longer-run health care utilization .................................. 44
6 Conclusion ................................................................. 48
Appendices ................................................................. 53
  A Number of episodes by clinical classification ....................... 53
  B Defining referral networks using CCS codes ....................... 56
  C Heterogeneity of treatment effects .................................. 61
1 Introduction

Variation in the consumption of health care in the United States is vast, persistent across many different measures of geography, and for the most part, unexplained. In 2014, for example, Medicare beneficiaries in Miami, Florida received over 60% more care than the average beneficiary in Madison, Wisconsin ($13,000 versus $8,000).² Neither accounting for price differences nor accounting for patient characteristics explains away these differences, and differences in utilization appear to be uncorrelated with patient health (Gottlieb et al. 2010b, Skinner 2012).

Past work on health care variation has focused mostly on such cross-region differences in care. However, the study of geographic variation is limited in its policy importance for two reasons. First, differences in health care utilization are actually larger within regions than across regions (Newhouse et al. 2013). Secondly, decisions about health care are made not at the regional level but at the physician level. Understanding doctor differences is therefore key for any policies aimed at reducing health care variation. In this paper, I study differences in the practice intensity of individual physicians.

At first glance, the existence of practice variation across doctors suggests that similar patients may be receiving dramatically different care. A high-utilization doctor may request several tests and then surgery, whereas one less-intensive may prescribe monitoring and bed rest. On the other hand, variation in treatment may also arise because patients differ in their needs and preferences. Though studies control for patient characteristics and diagnoses, there are many aspects about patients unobservable to researchers but observed by doctors. Plausibly, particularly-sick patients or ones with greater tolerance for invasive care are more likely to be routed to certain physicians. If so, sorting based on these unobservable characteristics would generate variation in health care utilization and explain the lack of correlation with observed patient outcomes. Thus far, however, the literature has been unable to disentangle whether variation in physician practice intensity arises because of patient differences or doctor differences.

In this paper, I decompose variation across doctors by exploiting a quasi-experiment that uses changes in the referral networks for Medicare patients. I estimate that 57.6% is a conservative lower bound for the share of variation caused by doctor differences, and my preferred estimate is that roughly 70% of the variation is driven by provider styles. Hence, unobservable differences in

²These statistics are taken from http://www.dartmouthatlas.org/.
patient health or preferences explain only a small part of why doctors prescribe different amounts of care. Differences between doctors are also large in magnitude even when controlling for price differences and observable patient characteristics. For a single episode of care, the standard deviation in utilization is $127 (in 2012 dollars, against a mean of $259). I also show that variation in the provision of care across individual physicians vastly exceeds variation across both regions and even physician groups.

The natural experiment I use follows primary care physicians (PCPs) who refer patients out to a set of physician specialists. At time $t = 0$, one of these specialists migrates away or otherwise leaves, and the PCP’s network of specialists is forced to change. As a result, patients referred by the same PCP during $t \geq 0$ receive care from a different set of specialists than those referred during $t < 0$. The main assumption is that patients referred during these two time periods are similar, so differences in their care must be physician-driven.

A key advantage of this empirical strategy is that it relies only on variation in specialists within the same PCP and diagnosis network. Thus, my approach is robust to the two main sources of sorting. Needier patients may seek out particular kinds of primary care doctors, and PCPs may choose specialists aligned with their own styles. Neither type of sorting biases my estimates unless the sorting happens to change concurrently with the specialist separation. Using placebo tests studying unaffected referral networks, predicted utilization, and PCP behavior, I assess these potential biases to be minimal. I also repeat the analysis using instrumental variables based only on the pre-break referral network (i.e. the set of specialists during $t < 0$). Estimates using these instrumental variable closely replicate the baseline results, suggesting that classical measurement error and potentially-endogenous PCP reactions pose minor threats to identification.

Next I study the connection between health care usage and patient health. Like many other studies, I find no effect of utilization differences on mortality or hospitalization rates. However, I do find longer-run effects in the form of higher longer-term utilization for patients referred to higher-intensity specialists. Intuitively, receiving more health care today requires more utilization in the future in the form of increased follow-up care.

In Appendix C, I consider heterogeneity in the treatment effect of more-intensive specialists along a number of dimensions including the type of diagnosis and characteristics of the PCP and specialists, such as gender, affiliation with a teaching hospital, and the prestige of the specialist’s training. I do not find strong correlates for the treatment effect. This suggests that the share of
variation due to doctors is similar across these different dimensions.

Next I discuss the existing literature. Section 2 details the empirical strategy. Section 3 describes the data and defines key metrics. Section 4 and Section 5 present results on utilization and patient health respectively. Lastly, Section 6 concludes.

1.1 Relation to previous literature

This project foremost relates to a large literature on variation in health care utilization. Led by the Dartmouth Atlas Project, this strand of research documents widespread variation in U.S. Medicare usage across large geographic areas (Wennberg 1984, Fisher et al. 2003ab, Gawande 2009). These differences across regions persist even after controlling for patient demographics, patient health status, and price differences, suggesting that observably-similar patients are actually receiving different levels of care (Gottlieb et al. 2010b). Furthermore, higher spending regions do not seem to produce better outcomes or healthier patients (Skinner 2012).

Recent work uses natural experiments from patient migrations to decompose inter-region variation into patient-driven “demand” factors and provider-driven “supply” factors. Finkelstein et al. (2016) find that among Medicare beneficiaries who relocate between 1998 and 2008, their usage of health care switches immediately to look more like the destination region compared to the origin region. The authors argue that the patient’s decision to relocate is plausibly-exogenous to the patient’s health care needs, because there is minimal trend in utilization before the move, and because they find symmetric patterns for patients moving to lower-utilization regions and patients moving to higher-utilization regions. Thus, these elderly patients do not appear to be migrating for reasons closely related to their own health. The authors conclude that 40-50% of the variation across regions is driven by patient health and preferences, while the remainder is due to “place-specific supply factors.” Agha et al. (2017) and Laird and Nielsen (2017) likewise use natural experiments from migrating patients to study regional determinants of medical care.

In contrast, Molitor (2016) studies migrating physicians. He estimates that cardiologists who move adopt 2/3 of the habits of the destination region. He concludes that environmental factors such as “financial and legal incentives, hospital capacity, and productivity spillovers” impact physician practice more than any individual physician tendencies. On the other hand, Grytten and Sorensen (2003) find that Norwegian primary care physicians who switch locations maintain their earlier styles.
Compared to these studies, my paper differs on a key aspect. Whereas those works estimate the effect of migration on the migrating patient or physician, I study the impact that a departing physician has on the patients who stay. Looking at patients who do not move holds constant geography and therefore, maintains the place-specific supply factors highlighted by Finkelstein and coauthors. For instance, market structure, hospital capacity, local policies, and any productivity spillovers are likely unchanged for the practitioners who remain. This enables me to isolate the effect of a higher-utilization doctor, rather than the combined effect of a higher-utilization region. Holding fixed the geographic area also mitigates the concerns identified by Song et al. (2010) that diagnosis decisions also differ systematically across regions. Additionally, analyzing non-movers reduces potential confounding due to non-medical factors related to health and tied to geography such as residential segregation and income inequality (Chetty et al. 2014c).

Two other recent studies have looked specifically at within-region variation. Tsugawa et al. (2017) study physicians employed directly by hospitals, known as hospitalists. They make the assumption that patient assignment to hospitalists within the same hospital is quasi-random and find that physicians in the highest-utilization quartile do on average 40% more treatment per-episode than those in the lowest quartile ($1,055 compared to $743). Higher-spending practices, however, do not correlate to differences in either 30-day mortality or 30-day hospital readmission rates.

Chan (2016) likewise studies different physicians within the same hospital, but he looks doctors at the beginning of their careers training at a single institution. He shows that variation across practitioners exists even for first-year practitioners and that the variation does not disappear with increased experience. Instead, as physicians progress through their training, their practice styles become more and more set, but styles mostly do not converge across doctors.

Unlike Tsugawa et al., Chan finds that higher-spending doctors have slightly lower hospital readmission rates. This contrast could be explained by the different patient and physician samples: Tsugawa et al. study Medicare beneficiaries and experienced doctors, whereas Chan’s data contain all hospitalizations at a single institution treated by doctors training in the beginning of their careers.

This paper also relates to a separate line of research that connects health care utilization and patient outcomes by using quasi-random assignment from ambulances transporting patients to hospitals. McClellan et al. (1994), Cutler (2007), and Doyle et al. (2015) exploit how an ambulance’s proximity and affiliation to local hospitals influences whether that patient is treated at a high-
spending or a low-spending hospital. McClellan et al. report minimal benefit to more intensive treatment, whereas Cutler and Doyle et al. find that higher-spending hospitals are cost-effective at extending lives.

Lastly, my work relates to a nascent literature in health economics on physician referral networks. Zeltzer (2016) shows homophily in Medicare referral networks. For example, male physicians are more likely to contract with male specialists, and this has consequences on the wage gap between male and female doctors. But despite a recent call by Song et al. (2014) highlighting the importance of referrals, there is a still relative dearth of research studying the effect of referrals on patients (Carr-Hill et al. 1996, O’Donnell 2000, Sarsons 2017).

2 Empirical Strategy

This project seeks to isolate the causal effect of physicians on health care utilization leveraging breaks in a PCPs network of specialists. While the ideal experiment would randomly assign patients to providers, I rely on changing specialist networks for quasi-random variation. Figure 1 illustrates a stylized example of the natural experiment.

Consider a primary care physician PCP1, who during time \( t < 0 \), refers patients with a given condition to specialists \( s_1, s_2, \) and \( s_3 \). At \( t = 0 \), specialist \( s_3 \) exits the PCP’s network—either due to retirement or by moving to a new location. As a result, PCP1 newly contracts with specialist \( s_4 \) for \( t \geq 0 \). In this example, the PCP’s referral network for this condition expands to include \( s_4 \). Alternatively, the PCP may consolidate referrals to existing links \( s_1 \) and \( s_2 \), or even cease referring patients for this diagnosis altogether.

This project asks whether changing the specialist set affects patient utilization and patient outcomes. Continuing with the example, my empirical strategy works by estimating whether the utilization of patients referred by PCP1 during \( t \geq 0 \) more closely resembles that implied by Figure 1a or Figure 1b. If variation in care is entirely driven by patient characteristics, then changing the specialist network should not alter patient care. At the other extreme, if physician styles fully drive differences in health care, then the utilization will change to reflect the new network’s practice pattern.

A key feature of this empirical strategy is that it uses only within-PCP variation. This greatly reduces the potential endogeneity problems arising from patients sorting to physicians. As long as patients are not switching PCPs at the same time as the specialist separation, any effect of this
Figure 1: Stylized example of the quasi-experiment. For a given diagnosis, (a) and (b) show the referral networks of primary care doctors PCP₁ and PCP₂, before and after the exit of specialist S₃ respectively.

(a) t < 0: before specialist S₃ exits.
(b) t ≥ 0: after specialist S₃ exits.

sorting is held constant. I later show placebo tests in support of the assumption that within a PCP, the distribution of patient health is not changing during the event window.

However, using only within-PCP variation adds a caveat to the interpretation of my results. In addition to referring patients to specialists, PCPs may also recommend the actual treatment, which constrains the specialist’s choice set. If so, my estimated physician share is a lower-bound for the variation in health care utilization driven by specialists.

Lastly, I remain agnostic as to how PCPs refer within their own network. For example, PCP₁ may refer her patients to whichever specialist is soonest-available. Or she may direct all her sickest patients to s₁ and the least-needy to s₃. Hence, Figure 2 is a better representation of my empirical strategy. Aggregating the specialists together permits arbitrarily sorting to specialists within the same PCP network. In Section 2.3, I discuss whether different sorting strategies affects the interpretation of my results.

My empirical strategy resembles that of Chetty et al. (2014b), who estimate the causal effect of high-quality teachers on student test scores. They study high and low-quality teachers exiting and entering schools, whereas I study high and low-utilization specialists leaving and joining referral networks. This project very much lives in the overlap between education and health economics highlighted by Chandra and Staiger (2016).

2.1 Statistical Model

Let \( y_{irdst} \) be the log of health care usage for patient \( i \) referred by PCP \( r \) to specialist \( s \) for diagnosis \( d \). Most generally, total utilization can be expressed as a function of the specialist’s true causal effect
Figure 2: Stylized example of quasi-experimental design. Unlike Figure 1, the PCP decision of which in-network specialist to refer to is treated as black-box to allow for arbitrary sorting of patients to specialists within the same PCP.

(a) $t < 0$: before specialist $s_3$ exits

(b) $t \geq 0$: after specialist $s_3$ exits

$\alpha_s$, and the health and preferences of the patient and PCP. Let $\kappa_{irdst}$ capture all the components not related to the specialist’s causal effect. It is indexed by PCP $r$ and specialist $s$ to allow for the possibility that PCP $r$ chooses the specific specialist $s$ endogenously (i.e. PCP sort patients to specialists based on unobservables).

$$y_{irdst} = H(\alpha_s, \kappa_{irdst})$$

$$= \alpha_s + \kappa_{irdst}$$

This second step assumes that the physician effect and patient sorting are multiplicatively-separable and therefore additively-separable in log specifications. This requires that doctor styles be consistent across patients. For example, an aggressive doctor who does 10% more ($\alpha_s = 0.10$) does 10% more for both very sick and relatively healthy patients. I briefly return to this assumption later when estimating physician effects.

Next I assume that the sorting component $\kappa_{irdst}$ can be decomposed into the sum of patient of observables $X_{it}$ and a sorting effect $\Psi_{rds}$. This second term $\Psi_{rds}$ captures the sorting of patients to PCPs and PCPs to specialists. Lastly, the error term $v_{irdst}$ captures any remaining variation in $\kappa_{irdst}$ orthogonal (by assumption) to $\Psi_{rds}$ and $X_{it}$.

$$y_{irdst} = \alpha_s + [\Psi_{rds} + X_{it}\gamma] + v_{irdst}$$
A key constraint is that $\alpha_s$—the true causal effect of physician $s$—is never observed, not even in the population since there is always some sorting $\kappa$ present. Instead, let $A_s$ be the observational estimate of specialist $s$’s effect. It combines both the true causal effect $\alpha_s$ and average sorting $\kappa_s$ to that physician, $A_s = g(\alpha_s, \kappa_s)$.

$$A_s = \alpha_s + \kappa_s$$ (4)

This project investigates how much of the variation in $A$ is causal (related to $\alpha$) versus sorting.

Since I observe only $A_s$ and not $\alpha_s$, my model is instead:

$$y_{irdst} = \beta A_s + [\Psi_{rds} + X_{it} \gamma] + \nu_{irdst}$$ (5)

For simplicity here, I temporarily ignore $X_{it}$ (i.e. setting $\gamma = 0$),

$$\beta = \frac{\text{Cov}(A_s, y_{irdst})}{\text{Var}(A_s)}$$ (6)

$$= \frac{\text{Cov}(\alpha_s, y_{irdst}) + \text{Cov}(\kappa_s, y_{irdst})}{\text{Var}(A_s)}$$ (7)

$$= \frac{\text{Var}(\alpha_s) + \text{Cov}(\alpha_s, \Psi_{rds}) + \text{Cov}(\kappa_s, \alpha_s) + \text{Cov}(\kappa_s, \Psi_{rds})}{\text{Var}(A_s)}$$ (8)

$$\beta = \frac{\text{Var}(\alpha_s) + \text{Cov}(\alpha_s, \kappa_s)}{\text{Var}(A_s)} \text{ when } \Psi_{rds} \text{ is included as a control}$$ (9)

Thus, $\beta$ combines the variance from the true causal effect $\text{Var}(\alpha_s)$ and a selection component $\text{Cov}(\alpha_s, \kappa_s)$ that measures the degree to which sicker patients sort to less-aggressive or more-aggressive doctors. If unobservably-sicker patients ($\kappa > 0$) sort to more aggressive doctors ($\alpha > 0$), then the covariance term would be positive.\(^3\)

While I cannot directly estimate the selection term, I discuss in Section 4.6 how $\beta^2$ is a lower-bound for $\text{Var}(\alpha_s)/\text{Var}(A_s)$ and I attempt to estimate $\text{Var}(\alpha_s)/\text{Var}(A_s)$ directly. I find no evidence that this covariance term is positive, so my preferred interpretation is that $\beta$ is a reasonable estimate for causal share of variation in observed physician effects.

\(^3\)This covariance component capturing selection is common across the “movers” literature. For example, Chetty and Hendren (2017a) also cope with the challenge of removing this selection term.
2.2 Identification

In Equation 5’s current form, the coefficient of interest $\beta$ cannot be estimated, because $A_s$ and $\Psi_{rds}$ are collinear. In this section, I transform the model to implement the empirical design motivated by the quasi-experiment.

First, I remove the conditioning on specialist $s$:

$$
y_{irdt} = \beta \times \mathbb{E}_{rd}[A_s] + \mathbb{E}_{rd}[\Psi_{rds}] + X_{it}\gamma + \epsilon_{irdt}
$$

(10)

$$
y_{irdt} = \Psi_{rd} + \beta \times \mathbb{E}_{rd}[A_s] + X_{it}\gamma + \epsilon_{irdt}
$$

(11)

Identification requires changes in $\mathbb{E}_{rd}[A_s]$ within the same PCP-diagnosis network. In keeping with the quasi-experimental design, I consider the observational network effect $\mathbb{E}_{rd}[A_s]$ before and after the specialist separation at $t = 0$.

$$
y_{irdt} = \Psi_{rd} + \beta \left( \mathbb{E}_{rd}[A_s|t \geq 0] \times \mathbbm{1}\{t \geq 0\} + \mathbb{E}_{rd}[A_s|t < 0] \times \mathbbm{1}\{t < 0\} \right) + X_{it}\gamma + \epsilon_{irdt}
$$

The key assumption is conditional mean independence of the term labeled “observational network effects.” To make clearer what this exogenous object is, I add $(\beta \mathbb{E}_{rd}[A_s|t < 0] - \beta \mathbb{E}_{rd}[A_s|t < 0])$ to the right-hand side.

$$
y_{irdt} = \Psi_{rd} + \ldots
$$

$$
\beta \left( \mathbb{E}_{rd}[A_s|t \geq 0] \times \mathbbm{1}\{t \geq 0\} + \mathbb{E}_{rd}[A_s|t < 0] \times \mathbbm{1}\{t < 0\} - \mathbb{E}_{rd}[A_s|t < 0] \right) + \ldots
$$

$$
\beta \times \mathbb{E}_{rd}[A_s|t < 0] + X_{it}\gamma + \epsilon_{irdt}
$$

$$
= (\Psi_{rd} + \beta \times \mathbb{E}_{rd}[A_s|t < 0]) + \beta \underbrace{(\mathbb{E}_{rd}[A_s|t \geq 0] - \mathbb{E}_{rd}[A_s|t < 0]) \times \mathbbm{1}\{t \geq 0\}}_{\delta_{rd}} + X_{it}\gamma + \epsilon_{irdt}
$$

Combining the first two terms into a new per-event intercept $\bar{\Psi}_{rd}$ yields an estimable equation:

$$
y_{irdt} = \bar{\Psi}_{rd} + \beta \times \left[ \delta_{rd} \times \mathbbm{1}\{t \geq 0\} \right] + X_{it}\gamma + \epsilon_{irdt}
$$

(12)

where $\delta_{rd}$ is the exogenous object and captures the change in network effects before and after the network separation. More specifically, $\delta_{rd}$ is the difference in expected utilization for a patient that
PCP $r$ refers out for condition $d$ after the specialist separation ($t \geq 0$) compared to before ($t < 0$). If the new network is dramatically different than the old network, then the magnitude of $\delta_{rd}$ should be large. On the other hand, if the separating specialist played only a minor role or performed similarly to other specialists, then $\delta_{rd}$ should be close to zero.

For $\delta_{rd}$ to be exogenous, PCP behavior and the health distribution of the referred patients must remain steady around $t = 0$. This forms the key identifying assumption. I enumerate its potential pitfalls and argue its plausibility in Section 2.3.

I now describe the empirical construction of $\delta_{rd}$:

$$
\delta_{rd} = \mathsf{E}_{rd}[A_s|t \geq 0] - \mathsf{E}_{rd}[A_s|t < 0]
$$

Let $S_{rdt}$ be the set of specialists to which PCP $r$ refers for diagnosis $d$ during time period $t$, and $f_{rds} \in (0, 1]$ is the share of referrals to $s$. Then I can construct the weighted average:

$$
\mathsf{E}_{rd}[A_s|t] = \sum_{s \in S_{rdt}} f_{rds} \times A_s
$$

Henceforth, I also refer to $\mathsf{E}_{rd}[A_s|t]$ as the observational network effect for PCP $r$ and referred-diagnosis $d$ during time period $t$. Using the sample analog of Equation 14, I can construct $\hat{\delta}_{rd}$:

$$
\hat{\delta}_{rd} = \left[ \sum_{s \in S_{rd}, t \geq 0} \hat{f}_{rds,t \geq 0} \times \hat{A}_s \right] - \left[ \sum_{s \in S_{rd}, t < 0} \hat{f}_{rds,t < 0} \times \hat{A}_s \right]
$$

where

- $\hat{f}_{rds,t} =$ fraction of observed diagnosis $d$ claims that referrer $r$ sends to specialist $s \in S$ during time period $t$,
- $\hat{A}_s =$ physician $s$’s effect estimated on a different data sample.

I discuss both identifying $S_{rdt}$ and estimating the physician effect $A_s$ in the following section.

**Equation 16** gives my main specification in event study form, where $i$ is a patient-episode, $r$ is the PCP, $d$ the referred-diagnosis, and $q$ is the quarter relative to the specialist separation. $\hat{\delta}_{rd}$ is the expected change in utilization under the null that the observational physician effects are fully
causal, $A_s = \alpha_s$. Thus, $\beta_q$ are the coefficients of interest.

\[
y_{irdt} = \psi_{rd} + \sum_{q} \beta_q [1\{t = q\} \times \hat{\delta}_{rd}] + \tau_t + X_{it} \gamma + \epsilon_{irdt} \tag{16}
\]

$\psi_{rd}$ represents a PCP-diagnosis fixed effect, so the coefficients are identified off changes within the same referrer and referred diagnosis over time. $\tau_t$ contains year-month fixed effects to capture potential seasonality in patient health and patient visits. Lastly, $X_{it}$ is a vector of patient information including 5-year age bins, race, various comorbidities, and the Charlson Comorbidity Index, a composite of health factors used to predict patient mortality. If changes in network structure $\hat{\delta}_{rd}$ are quasi-exogenous, then $\beta_q$ should be zero for $q < 0$.

This specification parallels that of Finkelstein et al. (2016). Their approach uses the natural experiment of patients migrating across regions, so they use a patient fixed effect instead of a PCP-diagnosis one. Analogously, their $\delta$ is constructed as the difference between average utilization in the destination region and average utilization in the origin region.

### 2.3 Identification assumptions

The key identifying assumption is that the change in network effect $\delta_{rd}$ is orthogonal to changes in the PCP’s practice style, patient health, and patient preferences.

The foremost threat to this assumption is patients switching PCPs as a result of the separation. I hypothesize that this violation is unlikely though. Plausibly, patients choose their own PCPs, but likely only a small minority chooses according to the PCP’s specialist network. Even the existence of such a minority would only violate orthogonality if those patients left their chosen PCP in reaction to a specialist departure.

My empirical strategy is also robust to more subtle changes in PCP behavior. Consider different models of how PCPs sort patients to specialists. One possibility is that within a diagnosis category, the PCP segments specialists by severity tier: the sickest patients are referred to $s_1$, the moderately-sick to $s_2$, and so on. I control for various covariates, but this sorting could occur on unobservables (i.e. attributes of the patient that PCPs observe but I do not). However, in the case that the new specialist simply replaces the departing specialist and treats the same “tier” of patients, my orthogonality assumption clearly holds.

In general, changes in how the PCP refers to specialists would only threaten my identification
assumption if $\delta_{rd}$ was correlated with the error term $\epsilon_{irdt}$ as a result. However, as long as the distribution of patients referred out by the PCP is stable around $t = 0$, then this correlation should be zero. This claim does rely on the assumption made earlier that physicians have consistent styles across patients.

My identifying assumption could be violated if the PCP changes behavior in reaction to losing a specialist. For example, suppose the PCP ceased referring out for a condition. In fact, referrals on average do decline slightly after $t = 0$ but only slightly and not enough to drive my findings. I later show placebo tests in support of this claim that PCPs are not changing behavior.

Furthermore, measurement error in $A_s$ and therefore in $\hat{\delta}_{rd}$ would attenuate my estimates. I show instrumental variable estimates that closely replicate the baseline estimates, so I claim classical measurement error poses only a mild concern. The same instrumental variable strategy also shows that my identification is not biased by potentially endogenous actions by the PCPs.

Lastly, I note that my empirical approach does not require that PCPs be referring patients to specific specialists. A PCP may just be directing patients to a specialty group, and that group assigns patients to particular physicians, either quasi-randomly by availability or strategically through some kind of matching. A member of that specialty group leaving generates quasi-random variation in the PCP’s network in the same manner that a specific specialist’s departure does.

3 Data and Summary Statistics

This project uses a 20% random sample of fee-for-service Medicare beneficiaries and claim-level data. This same dataset has been used in a number of research studies including Finkelstein et al. (2016), Molitor (2016), and Agha et al. (2017). I restrict to data over 2008–2013 to avoid needing crosswalks to map the different provider identification codes. Data from 2009–2013 are used for the event study estimation, while data from 2008 are withheld to estimate physician fixed-effects.

I focus on physician providers and omit referrals made by or to physician assistants, clinical social workers, nurse practitioners, and institutional providers such as laboratories. However, procedures reimbursed to these non-physician entities are included in the measures of utilization discussed in Section 3.1.

Specifically, my analysis restricts to providers labeled “Allopathic & Osteopathic Physicians” by the Health Care Provider Taxonomy Code Set available through the National Uniform Claim Committee (http://www.nucc.org/).
Physician identifiers

Beginning in 2007, health care providers are all assigned a unique ten-digit National Provider Identifier (NPI). The NPI is reported to all payers, including Medicare, for reimbursement. NPIs are permanent and stay with the provider even after job and/or location changes. The CMS also maintains a public dataset of NPI registrations, including basic information about the provider, such as name, gender, and specialty.

Diagnosis categories and referral networks

Each claim in my dataset includes a primary diagnosis in the form of an International Statistical Classification of Diseases (ICD-9) code. These ICD-9 diagnosis codes are very specific, so I rely on the Clinical Classifications Software (CCS), developed by the Agency for Healthcare Research and Quality, to group ICD-9 codes into “clinically meaningful categories.” It collapses over 14,000 ICD-9 codes into 18 system categories.

Table 1 presents a list of the system categories along with the number of episodes and number of specialist separations per category. Appendix A splits the four larger categories into more narrow CCS categories. The count of episodes are the number of observations in my dataset that a PCP referred a patient out for that diagnosis category. I define my measure of a episode in Section 3.3.

Diseases of the circulatory system (7) include chest pain, heart diseases, and hypertension. Lung issues such as pneumonia and bronchitis fall within the respiratory category (8), while cancers fall under Neoplasms (2).

I use these same disease categories to group specialists. Recall that in constructing the PCP’s network effect $E_{rd}[A_s]$ in Equation 14, a key step is identifying $S_{rd}$, the set of specialists to whom PCP $r$ sends patients for diagnosis $d$.

One concern may be that these system categories may be too broad. That is, the set of physicians grouped by these 18 systems associates non-substitutable specialists. This actually does not threaten my empirical strategy. The expected change in utilization $\delta_{rd}$ will be smaller, because the departing specialist was not relevant to a portion of the network. But the consequent effect of the specialist separation should also be smaller, so that specialist share $\beta$ should still be accurate.

Instead, greater risk lies in defining the set of specialists too narrowly. One alternative is group-

---

The CCS code mapping is available for download at [https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp](https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp).
Table 1: Number of episodes and events by disease system.

<table>
<thead>
<tr>
<th>System</th>
<th># episodes</th>
<th># events</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Infectious and parasitic diseases</td>
<td>181</td>
<td>7</td>
<td>0.0%</td>
</tr>
<tr>
<td>2. Neoplasms</td>
<td>63,323</td>
<td>1,043</td>
<td>10.6%</td>
</tr>
<tr>
<td>3. Endocrine; nutritional; and metabolic diseases and immunity disorders</td>
<td>2,527</td>
<td>80</td>
<td>0.4%</td>
</tr>
<tr>
<td>4. Diseases of the blood and blood-forming organs</td>
<td>215</td>
<td>10</td>
<td>0.0%</td>
</tr>
<tr>
<td>5. Mental illness</td>
<td>1,784</td>
<td>30</td>
<td>0.3%</td>
</tr>
<tr>
<td>6. Diseases of the nervous system and sense organs</td>
<td>40,824</td>
<td>498</td>
<td>6.9%</td>
</tr>
<tr>
<td>7. Diseases of the circulatory system</td>
<td>178,196</td>
<td>2,779</td>
<td>30.0%</td>
</tr>
<tr>
<td>8. Diseases of the respiratory system</td>
<td>83,633</td>
<td>1,621</td>
<td>14.1%</td>
</tr>
<tr>
<td>9. Diseases of the digestive system</td>
<td>35,347</td>
<td>589</td>
<td>5.9%</td>
</tr>
<tr>
<td>10. Diseases of the genitourinary system</td>
<td>32,221</td>
<td>525</td>
<td>5.4%</td>
</tr>
<tr>
<td>11. Diseases of the skin and subcutaneous tissue</td>
<td>5,474</td>
<td>96</td>
<td>0.9%</td>
</tr>
<tr>
<td>12. Diseases of the musculoskeletal system and connective tissue</td>
<td>85,234</td>
<td>1,486</td>
<td>14.3%</td>
</tr>
<tr>
<td>13. Congenital anomalies</td>
<td>37</td>
<td>1</td>
<td>0.0%</td>
</tr>
<tr>
<td>14. Injury and poisoning</td>
<td>6,812</td>
<td>214</td>
<td>1.1%</td>
</tr>
<tr>
<td>15. Symptoms; signs; and ill-defined conditions</td>
<td>58,717</td>
<td>1,196</td>
<td>9.9%</td>
</tr>
<tr>
<td>16. Residual codes; unclassified; all e codes [259. and 260.]</td>
<td>314</td>
<td>12</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

Table 2: Number of disease system and CCS types referred to the departing specialists.

<table>
<thead>
<tr>
<th>Diagnosis Category</th>
<th># of Diagnoses Type</th>
<th>Mean</th>
<th>25%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems</td>
<td>1.13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CCS</td>
<td>2.53</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

The fact that specialists treat multiple CCS categories is not a problem in of itself. Instead, the worry arises because in my data, while the PCP makes the referral decision, the specialist chooses actual referral diagnosis. When the PCP sends a patient to the specialist, the data records the condition that the specialist assesses the patient as having. Variation may arise precisely because different specialists vary in their diagnosis of similar patients. When the delineations between CCS codes are fuzzy, different specialists may code similar patients differently. If so, identifying the set
of specialists $S_{sd}$ by CCS instead of system understates the specialist share of variation. Indeed, defining networks by CCS reduces the specialist share estimate from 70% to 60%, but does not change the other main findings (see Appendix B).

3.1 Defining utilization measures

A single episode of health care often involves many providers. For example, a sequence of care might begin with one doctor conducting the initial exam, requesting blood tests from a laboratory, sending the patient for surgery, and then having the patient stay overnight in a hospital. In this example, Medicare reimburses at least four different parties for the same episode of care (the doctor, the laboratory, the surgeon, and the hospital). However, all the procedures stemmed from decisions made by first doctor, but that doctor likely would not be explicitly listed on all the downstream claims.

Thus, the data do not allow me to conclusively attribute utilization to physicians (Pham et al. 2007). To address this limitation, I define multiple different utilization measures and repeat my analysis for these different outcomes. Table 3 summarizes each of these measures.

The first row (category A) presents the most expansive definition: it includes all utilization for that patient related to the systems-level diagnosis. For example, if the original referral was for a heart-related condition, then “All system-level” utilization captures all heart-related utilization conducted on that patient regardless of the provider claiming the reimbursement. Category B (all CCS-level) is slightly more narrow in that it includes all utilization related to one of 284 CCS codes (rather than 18 systems definitions). Both these categories include non-physician utilization as well (e.g. payments made to hospitals, laboratories, and nurse practitioners), though the results hold when restricted to just physician reimbursements.

Categories C restricts the measure of utilization to reimbursements explicitly claimed by the referred specialist. For claims to count under this measure, the referred specialist’s NPI must be listed somewhere on the claim.

An example may clarify these different measures of utilization. Suppose a patient is referred to specialist $s_1$, who conducts an exam, diagnoses the patient with pneumonia, and charges Medicare $P_E$. $s_1$ also sends the patient for a chest X-ray, done by a nurse practitioner who charges Medicare $P_X$. After examining the X-ray, $s_1$ notices a separate problem in the patient’s lungs and then refers to the patient to specialist $s_2$, who then conducts a procedure costing $P_P$. Concurrently, the patient sees
another specialist for his regular diabetes check-up, costing Medicare $P_D$. The different utilization measures would be:

\[
\begin{align*}
\text{A} &= P_E + P_X + P_P \\
\text{B} &= P_E + P_X \\
\text{C} &= P_E
\end{align*}
\]

Measure A is the most-inclusive, whereas measure C is the least. Critically, reimbursement for the patient’s diabetes check-up $P_D$ is unrelated to the referral and not included in any of the measures.

I also consider the different outcome measures across different time periods to better account for the fact that a single health incident can require multiple days of care and follow-up visits. Specifically, I consider three different episode lengths: the actual referral day, within 30 days of the referral, and within 90 days of the referral.

The empirical construction of each utilization measure adopts the procedure from Gottlieb et al. (2010a) and used by Finkelstein et al. (2016) to aggregate outpatient, inpatient, and physician reimbursement while removing regional adjustments. Medicare reimburses providers based on federally-administered prices, and price differences across regions are relatively small. Nevertheless, this procedure purges those price differences to build a price-independent measure of health care quantity that I call utilization.

Table 3 gives the mean and standard deviation of the three outcome measures across the three time horizons. The units of utilization can be interpreted in 2012 dollars. Across all measures, the standard deviations are multiples of the means, consistent with evidence of long tails in health care usage.

Table 4a presents the correlation across different 30-day measures. While all are positively correlated, measure C that restricts to the referred specialist clearly is only a loose proxy for total utilization. The correlation coefficient for categories A and C is 31.3%. Undoubtedly, category C understates the effect of the specialist on the patient’s care, whereas the broadest category A likely overstates it. For this reason, I show results for both outcomes A and C. I also focus on 30-day episode lengths; extending to 90 days risks capturing multiple episodes. Moreover, the strong correlation between 30-day and 90-day utilization in Table 4b supports the reliability of the 30-day results. I study utilization after the 30-day mark further in Section 5.3.
Table 3: Mean and standard deviation of different utilization measures

<table>
<thead>
<tr>
<th></th>
<th>Episode Length</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 day</td>
<td>30 days</td>
<td>90 days</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>$1,119.86</td>
<td>$2,173.65</td>
<td>$3,145.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4,534.9)</td>
<td>(6,991.4)</td>
<td>(8,838.0)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$714.72</td>
<td>$1,181.55</td>
<td>$1,555.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3,211.6)</td>
<td>(4,488.9)</td>
<td>(5,395.5)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$215.92</td>
<td>$345.13</td>
<td>$427.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1,717.4)</td>
<td>(2,452.1)</td>
<td>(2,818.6)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Correlation between different measures of utilization

(a) Correlation between different 30-day utilization measures

<table>
<thead>
<tr>
<th>30-day utilization measure</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100%</td>
<td>65.5%</td>
<td>31.3%</td>
</tr>
<tr>
<td>B</td>
<td>100%</td>
<td>31.7%</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Correlation between different time windows for utilization A.

<table>
<thead>
<tr>
<th>Episode Length</th>
<th>1 day</th>
<th>30 days</th>
<th>90 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td>100%</td>
<td>70.2%</td>
<td>57.6%</td>
</tr>
<tr>
<td>30 days</td>
<td>100%</td>
<td></td>
<td>84.4%</td>
</tr>
<tr>
<td>90 days</td>
<td></td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

3.2 Identifying referral relationships and specialist separations

My empirical strategy requires identifying referral networks and networks breaks. For certain claims, Medicare requires that the specialist report a referral by a Medicare-recognized referrer. This is a gatekeeping method to deter fraud, and Medicare denies reimbursement for certain claims missing a valid referral (CMS 2016).

Two factors prevent me from identifying the full referral network. First, my dataset is a random sample of traditional Medicare beneficiaries, so I only observe PCPs and specialists for those patients. But since the beneficiaries are randomly sampled, the providers I observe and the links I infer are also a random sample. Secondly, Medicare only requires explicit referrals for certain cases, so I miss referrals in which the specialist does not record the PCP.

Despite these limitations, the data are sufficiently rich to build an approximation of referral networks. Zeltzer (2016) uses the same data when studying homophily among physician referral networks. In 2012, just under 40% of physician payments in my dataset included an explicit referral.
Figure 3: Average fraction of physician reimbursement that is referred by total annual physician reimbursement.

Notes: To construct this graph, I first divide each physician’s annual referred pay by total pay, where pay is Medicare physician reimbursement from my sample. Then I bin physicians into 30 groups based on their total annual pay. $y$ is then the average fraction referred for physicians in each bin.

This fraction is a lower bound for the amount of care that is actually referred, both because of the earlier recording issue and because specialists may continue treating the patient after that initial referral.

Figure 3 graphs the average fraction of Medicare annual reimbursement paid to physician $s$ that is referred against the total annual reimbursement on the $x$-axis. The portion of pay that is referred is not vastly different for specialists treating many Medicare patients versus few patients.

Next I identify specialist separations. Specifically, I look at physician NPIs with Medicare claims who either drop out the claims database altogether or move HRRs. I define moves as instances where a NPI switches both hospital referral region and state of reimbursed claims. Table 5 lists the number of specialist separations between 2010 and 2013 used in my sample, and Figure 4a shows the the year and month of those separations. Unsurprisingly, more exits occur in July and December, but there are separations during other months as well.

I construct $\hat{\delta}_{it}$ as the difference between the network one-year before the separation and one-year after the separation. Hence, I require at least a year of data before and after the break, so my
main specification uses only separations occurring from 2010–2012.

Figure 4b presents the number of episodes by quarter relative to the specialist separation at $q = 0$. Note that the panel is only balanced for eight quarters around $q = 0$, which explains why there are fewer episodes for quarters $-6, -5, 4,$ and $5$. Somewhat surprising, there is no evidence of any tapering to the separating specialist in anticipation of the exit. This suggests that I should expect minimal pre-trend in the event study.

Figure 4

(a) Number of specialist separations by year-month.

(b) Number of episodes by quarter.

3.3 Sample restrictions

Having identified referral relationships and specialist separations, I next restrict to referral networks “affected” by the specialist separation. For example, a PCP who refers to a retiring specialist just once before the retirement is unlikely to be affected by the separation. Therefore, I only consider PCPs who referred patients to the separating specialist for the relevant diagnosis at least four times in the year prior to the separation. Four referrals in my 20% random sample means on average at least 20 referrals of traditional Medicare patients over the year. Increasing this threshold reduces my sample size but does not affect the main takeaways.

Next, I want to look at new referrals, so I require that the patient did not already visited the
specialist in the two quarters before the referral. This ensures the referral is not part of a continuation of care by the specialist but more likely to be a distinct episode.

Similarly, if a patient is referred to different specialists within 90-days for the same system-level condition, I only keep the first referral in my sample. For these cases, I run the danger of mis-attributing utilization to the first referred specialist, but dropping subsequent referrals means I avoid having overlapping episodes and double-counting reimbursements for the same patient across multiple specialists. Similarly, I drop rare cases that the patient is referred out for multiple different conditions or to multiple specialists on the same day. Table 5 presents the count of physicians and episodes in my main estimation sample.

<table>
<thead>
<tr>
<th>Separating Specialists</th>
<th>Affected PCPs</th>
<th>PCP-Diagnoses</th>
<th>Episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exit</td>
<td>2,433</td>
<td>6,510</td>
<td>7,376</td>
</tr>
<tr>
<td>Move</td>
<td>1,027</td>
<td>2,537</td>
<td>2,811</td>
</tr>
<tr>
<td>Total</td>
<td>3,460</td>
<td>9,047</td>
<td>10,187</td>
</tr>
</tbody>
</table>

3.4 Summary statistics

Table 6 gives summary statistics of specialists and referrers in my sample. The number of links gives the average number of distinct physicians with whom the doctor has explicit referral relationships in a calendar year. The number does not include non-doctor entities such as laboratories and physician assistants, but it does not filter by the sample restrictions described earlier to identify plausibly-new episodes. Therefore, the number of links overestimates the number of actual referral contacts. Lastly, the number of links counts referrals for all diagnoses. I discuss diagnosis-specific referral networks in Section 4.5.
Table 6: Average annual utilization performed and referred by specialists and PCPs.

<table>
<thead>
<tr>
<th></th>
<th>Specialists</th>
<th></th>
<th>PCPs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>std.dev</td>
<td>Mean</td>
<td>std.dev</td>
</tr>
<tr>
<td>totalpay</td>
<td>49,741.18</td>
<td>(82,204.27)</td>
<td>65,099.84</td>
<td>(97,140.14)</td>
</tr>
<tr>
<td>Own utilization, referred</td>
<td>30,905.38</td>
<td>(56,056.70)</td>
<td>16,638.24</td>
<td>(54,213.97)</td>
</tr>
<tr>
<td># links inbound</td>
<td>89.12</td>
<td>(88.11)</td>
<td>24.56</td>
<td>(36.48)</td>
</tr>
<tr>
<td>Referred utilization</td>
<td>8,368.58</td>
<td>(21,866.62)</td>
<td>40,600.04</td>
<td>(47,363.66)</td>
</tr>
<tr>
<td># links outbound</td>
<td>16.24</td>
<td>(23.38)</td>
<td>53.81</td>
<td>(35.92)</td>
</tr>
<tr>
<td>Physician effect, (\hat{A})</td>
<td>5.30</td>
<td>(0.71)</td>
<td>5.02</td>
<td>(0.60)</td>
</tr>
<tr>
<td>(N)</td>
<td>82,204</td>
<td></td>
<td>9,047</td>
<td></td>
</tr>
</tbody>
</table>

4 Results: Health Care Utilization

4.1 Estimating physician effects

To estimate the physician fixed-effects \(\hat{A}_p\) used in constructing \(\hat{\delta}_{rd}\), I use a log-utilization specification with fixed effects for diagnosis \(\zeta_d\) and month-year \(\tau_t\) as well as patient demographic and comorbidity controls \(X_{it}\). Specifically, \(X_{it}\) is a matrix of patient gender, patient race, 5-year-age bins, and dummies for patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions.\(^6\)

\[
y_{ipdt} = A_p + \zeta_d + \tau_t + X_{it}' \Phi + v_{ipdt} \quad (17)
\]

This manner of estimating physician fixed effects follows from the statistical model in Section 2.1 and likewise assumes that the physician maintains his style across different diagnoses. This claim of consistent styles has some suggestive support from the literature (Phelps et al. 1994, Cutler et al. 2015, Currie et al. 2016).

To account for possible drift in physician effects, I adopt the approach that Chetty et al. (2014b) use to estimate teacher value-added. I estimate separate \(\hat{A}_{py}\) for different years \(y\) and combine these estimates to predict \(\hat{A}_p\). Consistent with the finding from Epstein and Nicholson (2009) that physician styles are relatively static (among physician obstetricians), the amount of drift in \(A_p\) is

\(^6\)I use the R implementation from Gaure (2013) to implement the strategy outlined in Abowd et al. (2002) for estimating regressions with two fixed effects.
Critically, I avoid using the same patient episodes in the estimation of $A_p$ and the main event study by only using non-event years for $A_p$. For instance, for a specialist separation occurring in 2010, I use the estimated specialist’s effect $\hat{A}_{s,y}$ for years $y = \{2008, 2012, 2013\}$ to predict $\hat{A}_{s,10}$. Similarly, for a separation in 2012, I predict $\hat{A}_{s,12}$ using the fixed effects from 2008, 2009, and 2010. Withholding data from the event years avoids generating mechanical correlation from using the same data for both independent and dependent variables (Chetty et al. 2014a). Henceforth, for expositional ease, I use $\hat{A}_p$ to indicate the predicted specialist effect $\hat{A}_{p,y}$ in the relevant event year $y$. Figure 5 plots the density of $\hat{A}_p$ for two different measures of utilization.

4.2 Constructing $\hat{\delta}_{rd}$

As presented in Section 2.1, I construct $\hat{\delta}_{rd}$ following Equation 15 and using the physician effects $\hat{A}_p$ described in Section 4.1. Figure 6 shows density plots of $\hat{\delta}_{rd}$ for 30-day utilization measures A and C. Recall from Table 3 that measure A sums the reimbursement from all treatment for the referred
Figure 6: Density of $\hat{\delta}_{rd}$ for utilization measures A and C.

(a) in percentages

(b) in levels

Notes: The left-hand graph plots the density of $\hat{\delta}_{rd}$, the percent change in the observational physician effects $A_s$ in the specialist network of PCP $r$ and diagnosis $d$ (Equation 13). The right-hand graph shows the implied change in levels. “System-related utilization” includes all utilization related to the initial diagnosis system in the 30-days following the initial referral, whereas “utilization by referred specialist” includes just utilization by the referred-to specialist. Table 3 describes these different utilization measures.

patient broadly related to the original referred diagnosis $d$. Measure C is far more restrictive and counts only the treatment conducted on the patient by the referred specialist. While $\delta_{rd}$ is more dispersed for C in percent terms, the levels are much lower, consistent with the claim that measure C captures only a subset of the care the patient receives.

4.3 Main results

Figure 7 plots the coefficients from the event study specification in Equation 16. Since $q = -1$ is the last quarter in which the departing specialist $s^*$ is still present, $\beta_{-1}$ is normalized to zero as the baseline. Importantly, though the specialist $s^*$ claims Medicare reimbursements during $q = -1$, I do not constrain that the PCP $r$ be referring out to specialist $s^*$ during $q = -1$. In other words, I allow for the PCP and $s^*$ to have concluded their relationship before $q = -1$.

These results are for 30-day episode lengths. Hence, I omit any episode without a full 30-days of data. I also omit events starting within 30-days of the specialist separation (i.e. within 30-days
**Figure 7:** Event study coefficients $\beta_q$ from Equation 16 for $Y = \log$ of 30-day utilization.

(a) Utilization Measure A: all disease-system related utilization

(b) Utilization Measure C: All utilization by the referred specialist

Notes: Graphs plot $\hat{\beta}_q$, the estimated coefficients on $[\delta_{tq} \times 1\{t = q\}]$, the change in network effect interacted with relative-quarter indicator variables. Specifications with FE include a PCP-diagnosis fixed effect $\psi_{rd}$. Specifications with a linear pre-FE omit the PCP-diagnosis fixed effect and instead include an estimate for the baseline network effect during $t < 0$, $E_{rd}[^{\text{PRE}}_{\text{FE}}(A_t|t < 0)]$. Controls include year-month dummies, patient gender, race, Charlson Comorbidity Index, 5-year-age bins, and dummies for patient health history of cancer, diabetes, Alzheimer’s disease, heart disease, chronic obstructive pulmonary disease, and chronic kidney disease. Bars represent 95% confidence intervals constructed using standard errors from 50 bootstrapped samples.

before $q = 0$). Doing so prevents inadvertent censoring for the utilization measures that restrict to the referred specialist (i.e. measure C). I carry over the same restriction for other utilization measures to maintain a consistent dataset, but this restriction does not impact my results, which suggests that PCPs only refer to soon-departing specialists for very-short-term procedures.

After including the PCP-diagnosis fixed effect $\psi_{rd}$, adding in control variables for patient demographics and health status along with year-month time effects does not change the main coefficients at all. This evidence that observable health measures are uncorrelated to network changes offers initial support for the orthogonality assumption discussed in Section 2.3. The last specification labeled “linear pre-FE control” replaces the many PCP-diagnosis fixed effects $\psi_{rd}$ with a single linear
control for the network effect \( \mathbb{E}_{rd}[A_s|t < 0] \) before the separation. With this specification, the coefficients lose precision but change only slightly. For the remaining figures, I use the PCP-diagnosis fixed effect and the full suite of controls unless otherwise noted. Throughout, standard errors are calculated using 50 bootstrapped samples. For each bootstrap iteration, the entire estimation sequence beginning with the construction of \( \hat{A}_p \) is bootstrapped. I block-bootstrap at the physician level when constructing \( \hat{A}_p \) and at the PCP level when estimating regressions.

Although the graphs show the event studies six quarters on both sides of \( t = 0 \), I calculate \( \delta_{rd} \) based on four-quarter differences. That is, I construct the pre-break network effect by using specialists from \( q = -4 \) to \( q = -1 \) and the post-break network effect from \( q = 0 \) and \( q = 3 \), inclusive. Since I require at minimum a year of data before and after the separation, calculating \( \delta_{rd} \) this way ensures balance across all events.

Nevertheless, I report the coefficients outside the two-year bandwidth. These coefficients \( \hat{\beta}_{-6}, \hat{\beta}_{-5}, \hat{\beta}_{4}, \) and \( \hat{\beta}_{5} \) are similar in magnitude to the coefficients within the two-year window. The differences do, however, suggest there is some natural flux in referral networks as well. I address the potential endogeneity worries this raises in my instrumental variable strategy.

Figure 7a suggests that patients who are referred to a 1% observationally-more intensive specialist network receive on average 0.7% more care. On the other hand, Figure 7b shows that this effect is more powerful when utilization is defined as just the treatment done by the referred specialist. This second measure of specialist effect being larger is consistent with intuition that specialists have more control over their own treatment of the patient compared to the overall utilization conducted on the patient. The gap also has to do with the different institutional structures that physicians oversee. Some specialists manage (and charge for) their own imaging services, while others outsource those services to other providers (Clemens and Gottlieb 2014).

Results with utilization measure A (total utilization related to the diagnosis), thus, provide a more complete description of the care patients actually receive. However, finding a strong effect with utilization measure C corroborates the mechanism claim that decisions made by the referred specialist are driving the changes in total utilization.

Recall from Equation 9 that the coefficients \( \beta \) combine the causal share of variation and a selection term. Thus, these estimates give the treatment effect of observationally-higher intensity specialists but are not the causal share of variation (i.e. not Var(\( a_s \))/Var(\( A_s \))). I attempt to directly estimate the causal share in Section 4.6.
Figure 7a and Figure 7b show that the change in utilization is immediate—the coefficients jump right at $q = 0$, and there is minimal trend afterwards. The absence of strong pre-trend and post-trend both reinforces the credibility of the event study design and allows me to collapse the event study into pre and post-break periods (Equation 18). I maintain the PCP-diagnosis fixed effect and the same controls. Figure 8 shows the bin scatter. The approximate linearity suggests that the response to $\hat{\delta}_{rd}$ is on average the same across events with different size $\hat{\delta}_{rd}$. In other words, separations generating large referral network changes (i.e. large $|\hat{\delta}_{rd}|$) have similar treatment effects as separations generating small deltas.

$$y_{irdt} = \Psi_{rd} + \beta[1\{t > 0\} \times \hat{\delta}_{rd}] + \theta[1\{t > 0\}] + \tau_t + X_{it}\gamma + \epsilon_{irdt}$$  (18)

Next I investigate possible asymmetries in the event study. First, I allow for differential responses for positive and negative $\hat{\delta}_{rd}$ events. This tests the hypothesis that health care costs more readily increase than decrease. Such an asymmetry could arise, for instance, if habit formation (on the part of the PCP) attenuates the effect of negative $\hat{\delta}_{rd}$ events because the PCP is accustomed to higher-touch service for her patients. Lowering the amount of services may worry the PCP about liability so that she encourages specialists to practice more “defensive medicine” as a result. The bin scatter in Figure 8 suggests such an asymmetry is unlikely, but differences may occur in timing rather than overall effect. Figure 9a shows that I cannot reject differences in either average treatment effect or timing between negative and positive $\hat{\delta}_{rd}$ events.

Figure 9b allows for differential responses for separations due to specialist exit or specialist migration. Again, I fail to reject a statistically-significant difference between the event types. Separations due to specialists leaving Medicare and specialists migrating on average have similar effects on how the remaining network responds.

Lastly, Figure 10 shows how the number of referred claims, unique specialists, and the number of referred claims changes by quarter. Specifically, I aggregate observations into counts by quarter and estimate $\hat{\beta}_q^*$ in Equation 19, while still maintaining a PCP-diagnosis fixed effect $\Psi_{rd}^*$. One caveat is that these numbers refer to plausibly first-time referrals; that is, patients who have not already visited that specialist before and have not been recently referred to other specialists for the same
Figure 8: Bin scatter of Equation 18 for $Y = \log$(30-day utilization). The dotted line indicates the 45° line.

(a) Utilization Measure A: all disease-system related utilization

(b) Utilization Measure C: All utilization by the referred specialist

Notes: These figures present binned scatter plots of the relationship between log utilization and the network change interacted with a dummy for after the specialist separation $[1\{t \geq 0\} \times \hat{\delta}_{rd}]$. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. The coefficient $\hat{\beta}$ is from the OLS regression of log utilization residualized against controls against $\hat{\delta}_{rd}$ residualized against the same controls, and the standard error is from 50 bootstrapped samples.

There appears to be a slight decrease in the number of referrals out for the affected diagnosis (Figure 10a), but there are still outbound referrals. The drop at $q = 0$ does suggest that there exist marginal referral cases (Choudhry et al. 2014). The results also suggest that over 60% of PCPs in my sample contract with a new specialist following the break, whereas others consolidate referrals to their remaining network.

Lastly, Figure 10c shows the number of distinct specialist organizations that the PCPs refers
Figure 9: Event study allowing for effect asymmetries for outcome A, $Y = \log(30$-day utilization by disease system).

(a) Positive v. negative $\delta_{rd}$ events.

(b) Events due to specialist exits v. specialist moves.

Notes: Bars represent 95% confidence intervals constructed with bootstrapped standard errors. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.

to within each quarter. One hypothesis posits that when specialist $s^*$ departs, the PCP redirects referrals to the “protégé” of $s^*$ or to another specialist working within the same specialty group. Figure 10 indicates that is not the necessarily the case. The number of distinct organizations drops about half the amount as the number of specialists. However, PCPs were already referring to, on average, two specialists within each organization, so the number of distinct organizations should only drop in some of the cases when $s^*$ leaves. Instead, it seems more likely that PCPs were already referring patients to the separating specialists’ protégé and groupmates before the departure.

4.4 Instrumenting for $\tilde{\delta}_{rd}$

My quasi-experiment bases its exogeneity claim on the assumption that $\delta_{rd}$ is driven by a specialist leaving. However, simultaneous to the specialist separation at $t = 0$, other network changes may also be occurring. If so, $\delta_{rd}$ captures more than just the specialist separation. It could also include
**Figure 10**: Plot of coefficients $\beta_q^\#$ from Equation 19 with different dependent variables.

(a) $Y_{rdt} =$ number of referred claims by quarter.  
(b) $Y_{rdt} =$ number of unique specialists by quarter.  
(c) $Y_{rdt} =$ number of distinct specialist organizations by quarter.

Notes: Bars represent 95% confidence intervals using analytical standard errors clustered at the PCP level. The mean of the dependent variable and standard deviation during the pre-period ($t < 0$) are reported below each respective figure.

endogenous switches in the referral network. For example, as discussed in Section 2.3, PCPs might be strategically adjusting other referral decisions in response to unobservable (to the econometrician) shifts in patient health or their own evolving styles.

To allay these endogeneity concerns, I instrument for $\hat{\delta}_{rd}$ using just features of the exiting specialist and the pre-break network. By blinding itself to post-break ($t \geq 0$) information, this estimation strategy avoids using variation arising from potentially-endogenous PCP actions. I construct two instruments $Z_M$ and $Z_P$, each motivated by a different thought experiment.

\begin{align}
Z_{M,rd} &\equiv -\hat{f}_{rd,s^*,t<0} \times (\hat{A}_{z,s^*} - \overline{A}_z) \\
Z_{P,rd} &\equiv -\hat{f}_{rd,s^*,t<0} \times (\hat{A}_{z,s^*} - \hat{A}_{z,rd-s^*})
\end{align}

Both instruments stem from the question: how is a PCP likely to adjust following the loss of a specialist? The PCP could contract with a new specialist or she could consolidate patients to her remaining network. $Z_M$ models the first scenario. It is the expected change in network effect if the PCP replaces $s^*$ by taking a new draw from available specialists (i.e. one with mean physician-effect $\overline{A}$ in expectation).
For the second scenario, $Z_P$ gives the expected change if the PCP reallocates the patients she would have referred to $s^*$ among her remaining network ($rd - s^*$) in the same proportions. Throughout for notation, the subscript-$z$ marks physician effects $A_z$ estimated on a different sample than the physician effects used to construct the potentially-endogenous regressor $\hat{\delta}_{rd}$.

First, I study the collapsed event study in Equation 18. The actual instruments used are $Z_M$ and $Z_P$ interacted with a post-period indicator:

$$Z_{k,rdt} \equiv 1\{t \geq 0\} \times Z_{k,rd} \quad \text{for } k \in \{M,P\}$$

(22)

Table 8a reports the OLS, IV, and first-stage coefficients for the broadest utilization measure $A$. Both instruments produce coefficient estimates comparable to the OLS, suggesting minimal bias from endogeneity threats and from classical measurement error.

Conversely, Table 8b shows that the IV coefficients are roughly 40pp lower than the OLS for the narrowest utilization measure $C$, the specialist’s own utilization. The finding is consistent with the specialist’s own utilization being idiosyncratic to which services the referred specialists conduct themselves as opposed to contracting out to third parties. As a result, features of the old referral network are less correlated with the reimbursements charged to the new network (i.e. lower reduced form coefficients).

In the event study, the potentially-endogenous regressors are quarter dummies interacted with $\hat{\delta}_{rd}$, so the analogous interactions serve as the actual instruments (Equation 23). Figure 11 shows the IV coefficients moving in tandem with the OLS event study.

$$Z_{k,rdqt} \equiv 1\{t = q\} \times Z_{k,rd} \quad \text{for } k \in \{M,P\}$$

(23)

4.5 Robustness checks

As a first robustness check, I consider the same PCP and the same window but untreated referral networks. This placebo test works because the specialist separation should only affect certain referral networks. For example, a cardiologist leaving should not impact how the PCP refers patients with respiratory ailments. On the other hand, if the underlying patient health is shifting, then I should
Table 7: OLS and IV coefficients for Equation 18.

(a) \( y = \log(30\text{-day system-related utilization}) \)

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV-( Z_M )</th>
<th>IV-( Z_P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\delta}_{id} \times 1{ t \geq 0 } )</td>
<td>0.729***</td>
<td>0.758***</td>
<td>0.622***</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.083)</td>
<td>(0.118)</td>
</tr>
<tr>
<td>( 1{ t \geq 0 } )</td>
<td>−0.014</td>
<td>−0.015</td>
<td>−0.013</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>( \hat{Y}_{irdt} )</td>
<td>0.374***</td>
<td>0.422***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.026)</td>
<td></td>
</tr>
</tbody>
</table>

(b) \( y = \log(30\text{-day utilization by referred specialist}) \)

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV-( Z_M )</th>
<th>IV-( Z_P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\delta}_{id} \times 1{ t \geq 0 } )</td>
<td>0.846***</td>
<td>0.612***</td>
<td>0.572***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.040)</td>
<td>(0.053)</td>
</tr>
<tr>
<td>( 1{ t \geq 0 } )</td>
<td>−0.011</td>
<td>−0.006</td>
<td>−0.005</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>( \hat{Y}_{irdt} )</td>
<td>0.340***</td>
<td>0.520***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.018)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors are from 50 bootstrapped samples. The first-stage coefficients gives the coefficients on \( Z \times 1\{ t \geq 0 \} \) in a regression of \( Z \times 1\{ t \geq 0 \} \) and controls on \( \hat{\delta}_{id} \times 1\{ t \geq 0 \} \). The instruments \( Z_M \) and \( Z_P \) (defined in Equation 20 and 21) are respectively, the expected change in network effect if the PCP replaced the separating specialist with an average new specialist or did not contract with a new specialist. Controls include the PCP-diagnosis fixed effect \( \Psi_{rd} \) and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.

see utilization changes for both affected and unaffected referral networks. Thus, I repeat my main specification for the same PCPs but on patients referred for conditions different than those treated by leaving specialist. I term these networks “untreated” since the specialist separation should have no direct impact on patients referred to these networks.

Figure 12a plots the coefficients for both the treated and untreated referral networks. The placebo coefficients stay flat at \( q = 0 \), supporting the identifying assumption that the change in utilization for affected diagnoses is not driven by changing underlying patient health.

The previous placebo test would not detect the particular case in which only the health of patients with the “treated” diagnosis is changing, while that of other patients stays steady. A second placebo test repeats the main specification for affected networks with \textit{predicted} utilization instead of actual utilization as the outcome variable. In a separate sample, I regress utilization against patient covariates and CCS diagnoses but omit physician fixed effects. Using these coefficients, I construct predicted utilization \( \hat{Y}_{irdt} \) and repeat the main specification (omitting the covariates used to predict
Figure 11: Coefficients from OLS and IV event studies for \( y = \log(30\text{-day system-related utilization}) \).

Notes: These graphs plot the OLS and IV event study coefficients for Equation 16, where the instruments \( Z_M \) and \( Z_P \) are interacted with quarterly dummies (Equation 23). The instruments are defined in Equation 20 and 21 and capture the expected change in network effect if the PCP replaced the separating specialist with a new specialist or did not contract with a new specialist. Controls include the PCP-diagnosis fixed effect \( \Psi_{rd} \) and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. Bars represents the 95% confidence interval constructed using bootstrapped standard errors.

\( \hat{Y} \).

Figure 12b plots the event study coefficients with actual utilization and predicted utilization. The coefficients for predicted utilization are small across all quarters, but there is a statistically-significant but small increase for \( q \geq 0 \). I interpret this as evidence that the new specialist set is coding diagnoses differently than the old set. Since CCS diagnosis codes are included in predicting \( \hat{Y} \), I observe a slightly positive \( \beta_{q \geq 0} \) for predicted utilization. However, the results still suggest the vast share of the treatment effect occurs within CCS diagnosis.

As one last robustness check, I examine the behavior of the PCP during the window of the network change. The worry here is the PCPs adjust their own behavior to compensate for losing
Figure 12: Placebo tests.

(a) Placebo test comparing the event study for affected v. unaffected diagnosis networks.
(b) Placebo test comparing the event study for actual utilization v. predicted utilization.

Notes: These figures repeat the main event study coefficients (labeled “Treated” and “Actual Utilization” respectively) along with coefficients from two different placebo tests. The placebo test in (1) replaces the dependent variable with 30-day utilization for patients referred by the same PCP to unaffected specialist networks. The placebo test in (b) replaces the dependent variable with predicted utilization. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. Bars represent 95% confidence intervals constructed using bootstrapped standard errors.

a specialist. Figure 13 suggests that any adjustments are minimal. The left-hand graph shows the number of visits the PCP sees during the window. Patients do not seem to leaving the PCP. The right-hand figure shows the coefficients when I repeat my event study but using the PCP’s own utilization (rather than the specialists’) as the dependent variable. The coefficients are small and statistically insignificant, suggesting that actions by the PCP are not driving my key results.

4.6 Estimating break effects directly

Thus far, my approach has been to calculate a predicted change in utilization $\hat{\delta}_{rd}$ and estimate the actual change as a fraction $\beta$ of that $\hat{\delta}_{rd}$. A limitation of this approach, shown in Section 2.1, is that $\beta$
Figure 13: PCP behavior in window surrounding specialist separation.

(a) $Y_{rt} =$ number of episodes beginning in quarter $t$ for PCP $r$.

(b) $y_{irt} = \log(30$-day utilization reimbursed to PCP $r$ for patient $i$ and episode beginning at date $t$).

Notes: Figure (a) plots coefficients $\beta_q$ from the regression $Y_{rt} = \Psi_r + \sum \beta_q \mathbb{1}\{t = q\} + u_{rt}$ where $\Psi_r$ is a PCP fixed effect. Figure (b) plots $\beta_q$ from the regression $y_{irt} = \Psi_r + \sum \beta_q [\mathbb{1}\{t = q\} \times \delta_{rd}] + X_{it} \gamma + \tau_t + u_{irt}$.

Bars represent 95% confidence intervals constructed using analytical standard errors.

conflates the causal variation and a selection covariance component. Abstracting from the controls, Equation 18 estimates:

$$\beta = \frac{\text{Var}(\Delta \alpha_s | r, d) + \text{Cov}(\Delta \alpha_s, \Delta \kappa_s | r, d)}{\text{Var}(\Delta \alpha_s | r, d)}$$

(24)

In this section, I directly estimate the effect of the network change at $t = 0$ and use my estimates as an attempt to bound the causal share of variation in health care utilization.

Equation 25 presents the specification for estimating the break effect $\Omega_{rd}$. The dependent variables are the same measures of log-utilization as before. I retain the PCP-diagnosis fixed effect $\Psi_{rd}$ but also include another PCP-diagnosis-specific effect $\Omega_{rd}$ interacted with a post-break indicator variable. This interaction captures the change in utilization after $t = 0$ for PCP $r$ and referred-diagnosis $d$.

$$y_{irdt} = \Psi_{rd} + [\Omega_{rd} \times \mathbb{1}\{t \geq 0\}] + \theta [\mathbb{1}\{t \geq 0\}] + \tau_t + X_{it} \gamma + u_{irdt}$$

(25)
Table 9: Coefficients for Equation 26.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\delta}_{rd})</td>
<td>0.759***</td>
<td>0.745***</td>
<td>0.747***</td>
<td>0.738***</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.030)</td>
<td>(0.034)</td>
<td>(0.030)</td>
</tr>
</tbody>
</table>

Controls: Y Y
Weighted: Y Y

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>10,187</td>
<td>10,187</td>
<td>10,187</td>
<td>10,187</td>
</tr>
<tr>
<td>(R^2)</td>
<td>6.4%</td>
<td>7.1%</td>
<td>7.2%</td>
<td>7.9%</td>
</tr>
</tbody>
</table>

*\(p<0.1; *\(p<0.05; **p<0.01\)

Notes: Regressions (2) and (4) weight events by the number of episodes, which matches the approach taken in the main event study (Table 7). Standard errors are from 50 bootstrapped samples. Controls include the event year, diagnosis, and the specialization of the separating specialist.

Figure 14a shows the density of \(\hat{\Omega}_{rd,s}\), and Figure 14b shows the bin scatter of the relationship between \(\hat{\Omega}_{rd}\) and \(\hat{\delta}_{rd}\) from Equation 26.

\[
\hat{\Omega}_{rd} = \Lambda_0 + \Lambda_1 \hat{\delta}_{rd} + v_{rd}
\] (26)

Taking the variance of Equation 26 lower-bounds the causal share of variance:

\[
\text{Var}(\hat{\Omega}_{rd}) = \text{Var}(\Lambda_1 \hat{\delta}_{rd}) + \text{Var}(v_{rd}) + 2\text{Cov}(\Lambda_1 \hat{\delta}_{rd}, v_{rd})
\] (27)

\[
\text{Var}(\hat{\Omega}_{rd}) = \Lambda_1^2 \text{Var}(\hat{\delta}_{rd}) + \text{Var}(v_{rd})
\] (28)

\[
\frac{\text{Var}(\hat{\Omega}_{rd})}{\text{Var}(\hat{\delta}_{rd})} \geq \Lambda_1^2
\] (29)

Table 9 reports my estimates of \(\hat{\Lambda}_1\). Unsurprisingly, \(\hat{\Lambda}_1\) is near-identical to \(\hat{\beta}\), since \(\Omega_{rd}\) replaces \((\hat{\beta}\hat{\delta}_{rd})\) in going from Equation 18 to Equation 25. \(\Lambda_1^2 = 0.759^2 = 57.6\%\) provides a lower-bound for the causal share of variation in observational physician effects \(A_p\).

Alternatively, I can use the same approach to estimate \(\text{Var}(\Delta \alpha_{i|r,d})\) directly. The estimated break effects \(\hat{\Omega}_{rd}\) can be expressed as the sum of the true break effect \(\Omega_{rd}\) and an orthogonal sam-
Notes: The right hand graph is a bin scatter of the relationship between \( \hat{\delta}_{rd} \) and \( \hat{\Omega}_{rd} \). I also show the best-fit line from the OLS regression and report the bootstrapped standard error. The displayed OLS line matches Regression (1) in Table 9.

pling error component \( \eta_{rd} \).

\[
\hat{\Omega}_{rd} = \Omega_{rd} + \eta_{rd} \tag{30}
\]

\[
\text{Var}(\hat{\Omega}_{rd}) = \text{Var}(\Omega_{rd}) + \text{Var}(\eta_{rd}) \tag{31}
\]

Following Chetty and Hendren (2017b), I can estimate \( \text{Var}(\Omega_{rd}) \) as the difference between the observed variance \( \text{Var}(\hat{\Omega}_{rd}) \) and an estimate of the sampling variance \( \mathbb{E}[\hat{\sigma}^2_{\Omega}] \), the average squared standard errors of \( \hat{\Omega}_{rd} \). Throughout, I use precision weights \( 1/\hat{\sigma}^2_{\Omega} \), though equal weighting does not alter the main takeaway.

\[
\hat{\text{Var}}(\Omega_{rd}) = \text{Var}(\hat{\Omega}_{rd}) - \mathbb{E}[\hat{\sigma}^2_{\Omega}] \tag{32}
\]

I use an analogous procedure to estimate \( \hat{\text{Var}}(\delta_{rd}) \), except sampling error in \( \delta_{rd} \) is estimated as the
variance of $\hat{\delta}_{rd}$ from 50 bootstrapped samples.

**Table 10: Estimated variances of $\delta_{rd}$ and $\Omega_{rd}$.**

<table>
<thead>
<tr>
<th></th>
<th>$\text{Var}(\hat{\Omega}_{rd}) = 0.131$</th>
<th>$\text{Var}(\hat{\delta}_{rd}) = 0.0271$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{E}[\hat{\sigma}_{\Omega}^2]$</td>
<td>$0.094$</td>
<td>$0.0022$</td>
</tr>
<tr>
<td>$\text{Var}(\Omega_{rd}) = 0.0364$</td>
<td>$\text{Var}(\delta_{rd}) = 0.0249$</td>
<td></td>
</tr>
</tbody>
</table>

The estimates in Table 10 show that the variance in true causal effects $\text{Var}(\Omega_{rd})$ is actually larger than the variance in observational physician effects $\text{Var}(\delta_{rd})$. This suggests that unobservably-sicker patients sort to lower-utilization doctors, $\text{Cov}(\Delta\alpha_s, \Delta\kappa_s | r, d) < 0$, and that this covariance is highly negative.

However, my estimates of $\Omega_{rd}$ are very noisy: over 70% of the observed variance in $\hat{\Omega}_{rd}$ arises from sampling noise (i.e. $\mathbb{E}[\sigma_{\Omega}^2]$). Hence, I take this as only suggestive evidence that selection term is not positive, so my preferred interpretation is that $\Lambda_1 = 75.9\%$ is a reasonable estimate for the share of variation in health care utilization driven by specialist styles.

### 4.7 Components of $\delta_{rd}$

In this section, I seek to understand what types of utilization differ between high and low-intensity doctors. Are the high-intensity networks merely ordering more imaging exams and laboratory tests? Or are they also prescribing more expensive procedures?

To answer this question, I rely the Berenson-Eggers Type of Service (BETOS) codes, which categorize Medicare-reimbursed actions into seven broad bins: (1) Evaluation and Management; (2) Procedures; (3) Imaging; (4) Testing; (5) Durable Medical Equipment; (6) Other; and (7) Exceptions/Unclassified. To understand which margins are changing with $\delta_{rd}$, I repeat my main event study specification in Equation 16 using the dependent variable:

$$y_{irdst, b} = \log(1 + 30\text{-day utilization for disease-system } d \text{ by BETOS } b)$$

(33)

Figure 15 plots the original event study coefficients, along with the coefficients for utilization on procedures, testing, imaging, and evaluation & management. Clearly, the action lies entirely along the procedures margin, as the coefficients for the other BETOS categories budge only slightly
around $q = 0$. These results suggest that the variation in health care usage is due to costly procedures, not just incremental testing and imaging services, consistent with survey evidence that differences between doctors arise because some prescribe intensive care and others recommend more palliative treatment (Cutler et al. 2015).

**Figure 15:** Event study coefficients decomposing overall utilization into types of utilization by BETOS category.

![Event study coefficients](image)

Notes: This figure shows the coefficients $\beta_q$ from Equation 16 where the dependent variable is the log utilization by BETOS (Berenson-Eggers Type of Service) category (Equation 33). Bars show the 95% confidence interval of coefficients using analytical standard errors. I report summary statistics in the table below. The means are lower than those in Table 3, because they omit inpatient reimbursements, which do not have BETOS classifications.

<table>
<thead>
<tr>
<th>Type</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedures</td>
<td>$337.79$</td>
<td>1,762.95</td>
</tr>
<tr>
<td>Testing</td>
<td>$41.68$</td>
<td>146.41</td>
</tr>
<tr>
<td>Imaging</td>
<td>$125.48$</td>
<td>380.98</td>
</tr>
<tr>
<td>Evaluation &amp; Management</td>
<td>$167.4$</td>
<td>461.71</td>
</tr>
</tbody>
</table>

4.8 Variance decomposition

Newhouse et al. (2013) and Tsugawa et al. (2017) show evidence that even though most of the past literature starting with the Dartmouth Atlas focus on variation across large regions known as hospital referral regions (HRRs),\(^7\) variation is actually larger within HRRs (Fisher et al. 2003ab). Here I quantify the variation at different levels using a hierarchical model for observational physician

\(^7\)There are 307 HRRs in the United States.
Table 11: Variance decomposition of log utilization from Equation 34.

<table>
<thead>
<tr>
<th>N</th>
<th>Share Variance</th>
<th>Variance of Variance</th>
<th>Variance in Levels</th>
<th>St. Dev. in Levels ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Referral Region</td>
<td>306</td>
<td>0.006</td>
<td>1.6%</td>
<td>308.00</td>
</tr>
<tr>
<td>Practice Group</td>
<td>62,744</td>
<td>0.106</td>
<td>29.5%</td>
<td>6,749.36</td>
</tr>
<tr>
<td>Physician</td>
<td>228,412</td>
<td>0.217</td>
<td>60.2%</td>
<td>16,301.15</td>
</tr>
</tbody>
</table>

\[ \hat{A}_p = 5.45 \]

Table 11 reports the variance at each level when \( y \) is the log of 30-day utilization by disease system. Only a sliver of the total variation across physicians lies at the HRR level. Comparable to Tsugawa et al. (2017), I find that variation within the same practice group is about half the variation across physicians within the same group. In comparison, the standard deviation in annual Medicare utilization per beneficiary across HRRs is $779 (Finkelstein et al. 2016).

5 Results: Patient Health

Having shown that health care usage does indeed change as a result of changing referral networks, I next investigate whether higher or lower utilization impacts patient health. Specifically, I consider three different outcomes up through 180-days after the initial 30-day episode: mortality, hospitalization, and longer-run utilization.

---

To estimate the hierarchical model, I require that physicians are fully nested within groups, so I omit the \( \approx 15\% \) of physicians who I observe practicing in multiple groups.
5.1 Mortality rates

Mortality is a drastic outcome and only captures the most severe deteriorations in health. Aside from exceptional outliers, an one-time specialist referral seems unlikely to either trigger an immediate plummet in health or rescue someone severely ill, so I hypothesize minimal impact on mortality rates. Nonetheless, as the ultimate outcome, mortality merits studying.

First, I repeat the main specification as a linear probability model using a binary dependent variable for whether the patient died within 180-days after the initial 30-day referral episode. In addition, I show the results using categorical independent variables. Equation 37 presents the alternative model. Same as before, I include a PCP-diagnosis fixed effect as well as month-year and patient controls. However, rather than interact the relative quarter with $\hat{\delta}_{rt}$, I interact the relative quarter with binaries for whether the network effect $E_{rt}[A_s]$ increased or decreased; that is, whether specialists became on average higher utilization or lower utilization. In order to evaluate whether the magnitude of $\hat{\delta}_{rt}$ also matters, I re-run the regression restricting to different thresholds of $|\hat{\delta}_{rt}|$.

The coefficients of interest are $\beta_{q}^+$ and $\beta_{q}^-$ for specialist networks that became higher-utilization and lower-utilization respectively.

$$y_{irdt} = \psi_{rt} + \sum_q \left\{ \beta_q^+ \left[ 1\{t = q\}1\{\hat{\delta}_{rt} > 0\} \right] + \beta_q^- \left[ 1\{t = q\}1\{\hat{\delta}_{rt} < 0\} \right] \right\} + \tau_t + X_{it}\gamma + \epsilon_{irdq} \quad (37)$$

Figure 16b plots the key coefficients, $\beta_{q}^+$ and $\beta_{q}^-$, for all events and for events with larger utilization differences ($|\hat{\delta}_{rt}| \geq 5\%$ and $|\hat{\delta}_{rt}| \geq 10\%$). The plots suggest that changes in utilization have minimal effect on mortality rates. Table 16a presents the coefficients from the collapsed event study. None of the coefficients are statistically significant at the 5% level. For the full sample, I can reject mortality rate changes greater than 0.4pp (or 5%) on a base of 7.39%. For networks with $|\hat{\delta}_{rd}| \geq 10\%$, my estimates become noisier, and I can only reject mortality changes greater than 0.5pp or 6.8%.

While mortality differences may not arise in aggregate, potentially higher or lower health care usage could lead to divergences for more severe conditions. My empirical design lacks power to study just these severe conditions, so I defer such work to future study.
Figure 16: Coefficients for $Y = 1$ if patient died within 210-days after the initial referral.

(a) Coefficients for collapsed event study.

| All Diagnoses | All Events | All Events | $|\hat{\delta}_{rd}| \geq 5\%$ | $|\hat{\delta}_{rd}| \geq 10\%$ |
|---------------|------------|------------|-----------------------------|-----------------------------|
| $\hat{\delta}_{rd} \times 1\{t \geq 0\}$ | 0.0029 | 0.0044 | \textbf{0.0007} | \textbf{0.0018} |
| $1\{\hat{\delta}_{rd} < 0\} \times 1\{t \geq 0\}$ | -0.0020* | -0.0004 | \textbf{0.0007} | \textbf{0.0018} |
| $1\{\hat{\delta}_{rd} > 0\} \times 1\{t \geq 0\}$ | -0.0004 | -0.0004 | 0.0015 | 0.0015 |

| N | 571,974 | 571,974 | 361,240 | 228,152 |
| # Events | 10,187 | 10,187 | 6,751 | 4,453 |
| $\hat{\gamma}$ | 7.39% | 7.39% | 7.1% | 6.73% |

95% C.I. for $(\beta^{+} - \beta^{-})$:[-0.002, 0.005] [-0.004, 0.004] [-0.003, 0.008]

Notes: Standard errors are from 50 bootstrapped samples. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.

(b) 180-day mortality rates for all events.

(b1) All events

(b2) $|\hat{\delta}_{rd}| \geq 5\%$

(b3) $|\hat{\delta}_{rd}| \geq 10\%$

Notes: Bars represents 95% confidence intervals constructed using bootstrapped standard errors.
5.2 Hospitalization rates

The sickest patients with higher mortality risks likely receive the same treatment regardless of provider, so not finding overall effects of $\delta_{rd}$ on mortality is unsurprising. Next, I look at hospitalization rates as a more intermediate health outcome. I define a patient as visiting a hospital if Medicare makes an inpatient reimbursement to a hospital for that patient.

Table 17a reports the coefficients. As was the case with mortality, I do not find any statistically significant effect on patient hospitalization rates within a half year after the initial referred episode. Figure 17b plots the event study coefficients and likewise does not suggest any discernible impacts on hospitalization rates. I can reject more than a 0.5pp (2.6%) change in hospitalization rates.

5.3 Longer-run health care utilization

Lastly, I consider whether patients use more care in the half-year after the episode. Potentially, patients who receive more care initially require less care later on. The opposite hypothesis is also plausible: more intensive care initially requires more follow-up monitoring and greater risk of complications.

I implement the same regression as before but using dependent variable:

$$y_{irdt} = \log(1 + \text{utilization during 180-days after initial referred episode})$$

The dependent variable is now the log of a continuous variable, so the coefficients $\beta_q^-$ and $\beta_q^+$ have percent-change interpretations. Table 18a reports the key coefficients.

My estimates offer greater support for the second hypothesis. Patients referred into a higher-utilization network have on average 6.97% higher utilization in the half-year after the referred episode. This effect is more dramatic for specialist separations generating larger $\hat{\delta}_{rd}$ as well. As shown in Figure 18b, the effect seems to manifest immediately at $q = 0$, though the estimates are noisy.

Furthermore, the results depict an asymmetry in the effect on longer-run utilization: lower initial utilization ($\hat{\delta}_{rd} < 0$) does not seem to lead to lower utilization subsequently. In contrast, there was no asymmetry for the original referral episode (Figure 9a). This suggests that when PCPs

---

9Previous studies studying this hypothesis have found large effects but over longer time horizons and for larger treatment sizes. Wherry et al. (2017) consider the effect of having Medicaid during childhood on adult outcomes, and Fang and Gavazza (2011) look at the effect of having employer-sponsored health insurance on health care utilization after retirement.
Figure 17: Coefficients for $Y = 1$ if patient was hospitalized between 30-days and 210-days after the initial referral.

(a) Coefficients for collapsed event study.

| All Diagnoses | All Events | All Events | $|\hat{\delta}_{rd}| \geq 5\%$ | $|\hat{\delta}_{rd}| \geq 10\%$ |
|---------------|------------|------------|-----------------|-----------------|
| $\hat{\delta}_{rd} \times 1\{t \geq 0\}$ | 0.0017     |            | (0.0067)        |                 |
| $1\{\hat{\delta}_{rd} < 0\} \times 1\{t \geq 0\}$ | 0.0015     | 0.0018     | (0.0020)        | (0.0025)        | 0.0045*        | (0.0030)        |
| $1\{\hat{\delta}_{rd} > 0\} \times 1\{t \geq 0\}$ | 0.0024     | 0.0012     | (0.0019)        | (0.0021)        | 0.0038        | (0.0024)        |

| N            | 571,974    | 571,974    | 361,240         | 228,152         |
| # Events     | 10,187     | 10,187     | 6,751           | 4,453           |
| $\hat{\Psi}$ | 22\%       | 22\%       | 22\%            | 21.8\%          |

95\% C.I. for

$(\beta^+ - \beta^-)$

[-0.004, 0.006] [-0.007, 0.006] [-0.008, 0.007]

Notes: Standard errors are from 50 bootstrapped samples. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.

(b) 180-day hospitalization rate.

(b1) All events

(b2) $|\hat{\delta}_{rd}| \geq 5\%$

(b3) $|\hat{\delta}_{rd}| \geq 10\%$

Notes: Bars represents 95\% confidence intervals constructed using bootstrapped standard errors.
switch to less-intensive specialists for the initial referral, they may be taking other measures to maintain the previous level of care. Otherwise, I would expect $\beta_q^- < 0$ for $q \geq 0$. These results cannot identify the drivers of the asymmetry, but the takeaway that longer-run utilization more readily increases than decreases warrants further investigation.

To summarize, like a large portion of previous literature (Skinner 2012, Tsugawa et al. 2017), I do not find a strong link between health care utilization and patient health. However, I suffer similar limitations as this other work: a shortage of measures more-reliably correlated with actual health than long-run utilization, and more nuanced than mortality or hospitalization. Hence, revisiting this setting will be interesting following the development of better patient outcome measures (Center for Clinical Standards and Quality 2016).
**Figure 18:** Coefficients for $y = \log(1 + \text{utilization during 180-days after initial referred episode})$.

(a) Coefficients for collapsed event study.

| All Diagnoses | All Events | All Events | $|\hat{\delta}_{rd}| \geq 5\%$ | $|\hat{\delta}_{rd}| \geq 10\%$ |
|---------------|------------|------------|-----------------|-----------------|
| $\hat{\delta}_{rd} \times 1\{t \geq 0\}$ | 0.131** | (0.058) | | |
| $1\{\hat{\delta}_{rd} < 0\} \times 1\{t \geq 0\}$ | 0.0330** | (0.0137) | 0.0295* | (0.0206) |
| $1\{\hat{\delta}_{rd} > 0\} \times 1\{t \geq 0\}$ | 0.0697*** | (0.0135) | 0.0727*** | (0.0176) |

N = 571,974 571,974 361,240 228,152

# Events = 10,187 10,187 6,751 4,453

Y = 2,172.47 2,172.47 2,210.09 2,213.23

95% C.I. for $(\beta^+ - \beta^-)$ = [-0.005, 0.079] [-0.013, 0.100] [-0.002, 0.119]

Notes: Standard errors are from 50 bootstrapped samples. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.

(b) Event study coefficients

(b1) All events

Notes: Plotted points are coefficients $\beta_q$ from Equation 37, and the dependent variable is defined in Equation 38. Bars represent 95% confidence intervals constructed using bootstrapped standard errors.
6 Conclusion

In 2015, almost 18% of U.S. GDP was spent toward health care expenditures, including $646.2 billion or 3.6% of GDP on Medicare alone (Martin et al. 2016). Facing mounting costs, policymakers have increasingly sought ways to bend the health care cost curve. One potential avenue for cost containment begins with the observation that some providers appear much less costly than others without producing worse patient outcomes. Thus, cost reductions may be achieved by pushing high-intensity suppliers toward emulating their less-expensive counterparts.

Such a policy approach makes two key assumptions. First, variation in health care utilization is supply-driven and not the result of underlying differences in patient needs or preferences. The second assumption is that forcing doctors to alter their styles will not harm patients.

This project justifies the first assumption. Using a natural experiment with quasi-exogenous changes in referral networks, I find that as a conservative lower bound, about 50% of the variation in health care utilization across specialists is physician-driven. My preferred estimate is that roughly 70% of the variation is attributable to physicians rather than patients.

These estimates are robust to the inclusion of controls and stricter sample restrictions, as well as different definitions of utilization and referral networks. I also replicate the estimates with instrumental variables, suggesting minimal bias from the most-likely endogeneity threats as well as from classical measurement error.

The differences in utilization are not correlated with patient mortality and hospitalization in the following 180 days. However, there do appear to be longer-term effects from receiving more initial care. I find that patients quasi-randomly referred to higher-intensity specialists for a given episode are more likely to receive more care in the future as well. Intuitively, more aggressive care (such as surgery) often requires more follow-up care as a result of continued monitoring and treatment of potential complications.

This evidence suggests that physicians do indeed have varying styles. However, it still may be that individual physicians are all performing at their own personal efficiency frontiers. If so, forcing doctors to alter their styles may still harm patients. More study is needed on both the origins of physician styles and the effect of policies encouraging different practice habits.

Lastly, this paper hints at one additional policy lever: the relationship between PCPs and specialists. My results suggest that PCPs allow specialists to determine the course of their patients’
care, and thus, changing referral networks also changes patient care. Tightening the connection between PCPs and their specialists may narrow the variation in health care utilization if PCPs constrain specialist behavior. Indeed, one component of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) begins to tie physician reimbursement to specialist actions and patient outcomes (Kocher and Chigurupati 2016). This gives PCPs an additional financial incentive to more closely monitor the follow-up care referred patients receive. The consequences of MACRA on PCP and specialist practices will make for fascinating future study.
References


Appendix

A Number of episodes by clinical classification

This section reports the number of patient referral episodes by CCS codes within the largest four disease system categories discussed in Section 3.

Table A1: Number of episodes within disease systems.

(a) Diseases of the circulatory system

<table>
<thead>
<tr>
<th>Clinical Classification</th>
<th># episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonspecific chest pain</td>
<td>49,527</td>
</tr>
<tr>
<td>Cardiac dysrhythmias</td>
<td>38,142</td>
</tr>
<tr>
<td>Coronary atherosclerosis and other heart disease</td>
<td>33,550</td>
</tr>
<tr>
<td>Other circulatory disease</td>
<td>17,274</td>
</tr>
<tr>
<td>Essential hypertension</td>
<td>17,062</td>
</tr>
<tr>
<td>Heart valve disorders</td>
<td>15,200</td>
</tr>
<tr>
<td>Occlusion or stenosis of precerebral arteries</td>
<td>14,546</td>
</tr>
<tr>
<td>Congestive heart failure; nonhypertensive</td>
<td>10,693</td>
</tr>
<tr>
<td>Peripheral and visceral atherosclerosis</td>
<td>9,585</td>
</tr>
<tr>
<td>Conduction disorders</td>
<td>9,477</td>
</tr>
<tr>
<td>Aortic; peripheral; and visceral artery aneurysms</td>
<td>7,543</td>
</tr>
<tr>
<td>Other and ill-defined heart disease</td>
<td>6,816</td>
</tr>
<tr>
<td>Acute cerebrovascular disease</td>
<td>6,340</td>
</tr>
<tr>
<td>Hypertension with complications</td>
<td>3,285</td>
</tr>
<tr>
<td>Peri-; endo-; and myocarditis; cardiomyopathy</td>
<td>3,273</td>
</tr>
<tr>
<td>Other and ill-defined cerebrovascular disease</td>
<td>3,106</td>
</tr>
<tr>
<td>Transient cerebral ischemia</td>
<td>2,879</td>
</tr>
<tr>
<td>Phlebitis; thrombophlebitis and thromboembolism</td>
<td>2,655</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>2,441</td>
</tr>
</tbody>
</table>
Table A1: Number of episodes within disease systems (continued...).

(b) Diseases of the respiratory system

<table>
<thead>
<tr>
<th>Clinical Classification</th>
<th># episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other lower respiratory disease</td>
<td>73,097</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>15,198</td>
</tr>
<tr>
<td>Pleurisy; pneumothorax; pulmonary collapse</td>
<td>14,448</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>7,208</td>
</tr>
<tr>
<td>Respiratory failure; insufficiency; arrest</td>
<td>3,980</td>
</tr>
<tr>
<td>Other upper respiratory disease</td>
<td>2,774</td>
</tr>
<tr>
<td>Asthma</td>
<td>2,336</td>
</tr>
<tr>
<td>Other upper respiratory infections</td>
<td>2,086</td>
</tr>
</tbody>
</table>

(c) Neoplasms

<table>
<thead>
<tr>
<th>Clinical Classification</th>
<th># episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other and unspecified benign neoplasm</td>
<td>26,365</td>
</tr>
<tr>
<td>Other non-epithelial cancer of skin</td>
<td>14,638</td>
</tr>
<tr>
<td>Cancer of breast</td>
<td>12,626</td>
</tr>
<tr>
<td>Cancer of prostate</td>
<td>5,263</td>
</tr>
<tr>
<td>Neoplasms of unspecified nature</td>
<td>4,990</td>
</tr>
<tr>
<td>Cancer of bronchus; lung</td>
<td>4,848</td>
</tr>
<tr>
<td>Cancer of bladder</td>
<td>3,753</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>3,047</td>
</tr>
<tr>
<td>Cancer of colon</td>
<td>2,525</td>
</tr>
</tbody>
</table>
Table A1: Number of episodes within disease systems (continued...).

(d) Diseases of the musculoskeletal system and connective tissue

<table>
<thead>
<tr>
<th>Clinical Classification</th>
<th># episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spondylosis; intervertebral disc disorders; other back problems</td>
<td>43,891</td>
</tr>
<tr>
<td>Other non-traumatic joint disorders</td>
<td>28,892</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>21,339</td>
</tr>
<tr>
<td>Other connective tissue disease</td>
<td>16,663</td>
</tr>
<tr>
<td>Other bone disease and musculoskeletal deformities</td>
<td>5,584</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>3,202</td>
</tr>
</tbody>
</table>

(e) Symptoms; signs; and ill-defined conditions

<table>
<thead>
<tr>
<th>Clinical Classification</th>
<th># episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other screening for suspected conditions</td>
<td>49,712</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>10,438</td>
</tr>
<tr>
<td>Other aftercare</td>
<td>8,758</td>
</tr>
<tr>
<td>Medical examination/evaluation</td>
<td>7,785</td>
</tr>
<tr>
<td>Syncope</td>
<td>2,303</td>
</tr>
<tr>
<td>Malaise and fatigue</td>
<td>2,089</td>
</tr>
</tbody>
</table>
B Defining referral networks using CCS codes

This section repeats the main empirical specification using a more narrow definition of referral networks. My preferred approach groups a PCP’s specialists into networks by 18 different diagnosis categories as explained in Section 3. An alternative design classifies the 14,000+ ICD-9 diagnoses into 284 CCS codes and groups specialists using those codes.\(^{10}\) This section adopts that alternative approach but otherwise replicates the main empirical strategy.

Figure B1 shows the coefficients from the main event study (Equation 16) using this more narrow definition of referral networks, and Figure B2 shows the corresponding bin scatter for the same measures of utilization as Figure 7 and Figure 8. Similarly, Table B1 and Figure B3 replicate the instrumental variable approach presented in Section 4.4.

The patterns are broadly the same—coefficients jump immediately at \(q = 0\) with minimal pre-trend or post-trend. However, the coefficients are roughly 10pp lower than those in the main specification. This difference is consistent with CCS codes being an overly-narrow definition of referral networks. In my data sample, specialists code the referred diagnosis. If the specialist effect includes the tendency to classify patients as certain CCS codes over others, then this approach should understate the causal effect. In particular, this way of coding referral networks may miss some affected patients that would have been referred to the separating specialist \(s^*\) but are instead referred elsewhere and diagnosed under a related but different CCS. The systems-level diagnosis categories used in the main specification are broader and thus, are less likely to mis-categorize affected patients.

\(^{10}\)Appendix A gives examples of CCS categories.
Figure B1: Event study coefficients $\beta_q$ from Equation 16 for $Y = \log$ of 30-day utilization using CCS codes to identify PCP referral networks.

(a) Utilization Measure A: all disease-system related utilization

(b) Utilization Measure C: All utilization by the referred specialist

Notes: Graphs plot $\hat{\beta}_q$, the estimated coefficients on $\hat{\delta}_{rd}$, the change in network effect before and after the specialist separation. Specifications with FE include a PCP-diagnosis fixed effect $\Psi_{rd}$. Specifications with a linear pre-FE omit the PCP-diagnosis and instead include an estimate for the baseline network effect during $t < 0$, $E_{rd}[A_i | t < 0]$. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. Bars represent 95% confidence intervals constructed using standard errors from 50 bootstrapped samples.
Figure B2: Bin scatter of Equation 18 for $Y = \log(30$-day utilization $)$ using CCS codes to identify PCP referral networks. The dotted line indicates the 45° line.

(a) Utilization Measure A: all disease-system related utilization

(b) Utilization Measure C: All utilization by the referred specialist

Notes: These figures present binned scatter plots of the relationship between log utilization and the network change interacted with a dummy for after the specialist separation $[1(t \geq 0) \times \delta_{rd}]$. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. The coefficient $\hat{\beta}$ is from the OLS regression of log utilization residualized against controls against $\delta_{rd}$ residualized against the same controls, and the standard error is from 50 bootstrapped samples.
### Table B1: OLS and IV coefficients for Equation 18 using CCS codes to identify PCP referral networks.

(a) \( y = \log(30\text{-day system-related utilization}) \)

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV-(Z_M)</th>
<th>IV-(Z_P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_d \times 1{t \geq 0} )</td>
<td>0.646***</td>
<td>0.594***</td>
<td>0.493***</td>
</tr>
<tr>
<td>( 1{t \geq 0} )</td>
<td>-0.014</td>
<td>-0.013</td>
<td>-0.012</td>
</tr>
<tr>
<td>( N )</td>
<td>505,545</td>
<td>505,545</td>
<td>505,545</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>19.6%</td>
<td>19.6%</td>
<td>19.6%</td>
</tr>
<tr>
<td>Controls</td>
<td>( Y )</td>
<td>( Y )</td>
<td>( Y )</td>
</tr>
<tr>
<td>First-stage</td>
<td>0.317***</td>
<td>0.438***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.102)</td>
<td>(0.108)</td>
</tr>
</tbody>
</table>

(b) \( y = \log(30\text{-day utilization by referred specialist}) \)

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV-(Z_M)</th>
<th>IV-(Z_P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_d \times 1{t \geq 0} )</td>
<td>0.769***</td>
<td>0.535***</td>
<td>0.610***</td>
</tr>
<tr>
<td>( 1{t \geq 0} )</td>
<td>-0.013</td>
<td>-0.007</td>
<td>-0.009</td>
</tr>
<tr>
<td>( N )</td>
<td>505,545</td>
<td>505,545</td>
<td>505,545</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>37.7%</td>
<td>37.6%</td>
<td>37.6%</td>
</tr>
<tr>
<td>Controls</td>
<td>( Y )</td>
<td>( Y )</td>
<td>( Y )</td>
</tr>
<tr>
<td>First-stage</td>
<td>0.342***</td>
<td>0.573***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.023)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors are from 50 bootstrapped samples. The first-stage coefficients are the coefficients on \( Z \times 1\{t \geq 0\} \) in a regression of \( Z \times 1\{t \geq 0\} \) and controls on \( \delta_d \times 1\{t \geq 0\} \). The instruments \( Z_M \) and \( Z_P \) (defined in Equation 20 and 21) are respectively, the expected change in network effect if the PCP replaced the separating specialist with the average new specialist or did not contract with a new specialist. Controls include the PCP-diagnosis fixed effect \( \Psi_{rd} \) and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies.
Figure B3: Coefficients from OLS and IV event studies for $y = \log(30$-day system-related utilization) using CCS codes to identify PCP referral networks.

Notes: These graphs plot the OLS and IV event study coefficients for Equation 16, where the instruments $Z_M$ and $Z_P$ are interacted with quarterly dummies (Equation 23). The instruments are defined in Equation 20 and 21 and capture the expected change in network effect if the PCP replaced the separating specialist with a new specialist or did not contract with a new specialist. Controls include the PCP-diagnosis fixed effect $\Psi_{rd}$ and the same list of controls used in the main event study: patient gender, race, Charlson Comorbidity Index, 5-year-age bins, patient health history of cancer, diabetes, Alzheimer’s disease, heart conditions, kidney conditions, and year-month dummies. Bars represents the 95% confidence interval constructed using bootstrapped standard errors.
C Heterogeneity of treatment effects

In this section, I explore the possibility that the treatment effect of network changes differs by observables. Recall that $\beta$ in my main specification combines the true share of causal variation and a selection term.

$$\beta = \frac{\text{Var}(\alpha_s) + \text{Cov}(\alpha_s, \kappa_s)}{\text{Var}(A_s)}$$  \hspace{1cm} (39)

Plausibly, the extent of unobservable patient sorting varies by diagnosis or by PCP and specialist characteristics. Thus, there may exist heterogeneity in $\beta$ by group.

Alternatively, heterogeneity may also arise due to varying treatment effects. $\beta$ also has the interpretation: a patient referred to a specialist network that is observably 1% more utilization-intensive receives on average $\beta\%$ more care. There are various hypotheses for why this effect could differ by group. For example, suppose that after specialist $s^*$ retires, the PCP contracts with a new specialist $s^\dagger$ trained at a less-prestigious residency program. Wary about the abilities of $s^\dagger$, the PCP might more closely monitor the new specialist’s actions and suggest treatment regimes. This particular story suggests that $\beta$ is increasing in the prestige of the new specialist’s residency program.

My analysis of heterogeneity in $\beta$ cannot distinguish between these two sources. It could be that there exists more unobservable patient sorting among certain specialist types. Or it could be that the treatment effect of changing specialist differs systematically. Furthermore, these two sources of heterogeneity could be reinforcing or countervailing.

Equation 40 presents the random coefficients specification. I amend $\beta$ from the main specification to be a per-event $\beta_{rd}$ drawn from a common distribution (Equation 41).

$$\log y_{irdt} = \Psi_{rd} + \beta_{rd} \{t > 0 \} \times \hat{\delta}_{rd} + \theta \{t > 0 \} + \tau_t + X_{it} \gamma + \epsilon_{irdt}$$ \hspace{1cm} (40)

$$\begin{pmatrix} \Psi_{rd} \\ \beta_{rd} \end{pmatrix} \sim \mathcal{N} \begin{pmatrix} \Psi_0 & \sigma_{\Psi}^2 \\ \beta_0 & \sigma_{\beta}^2 \end{pmatrix}$$ \hspace{1cm} (41)

Figure C1 shows the interquartile range of $\hat{\beta}_{rd}$ and $\hat{\delta}_{rd}$ for different systems-level diagnoses for which I have at least $N \geq 100$ distinct events. Categories marked by “$$>>$$” are ones that have nonzero coefficients in a LASSO regression (see Figure C3). The center column $p$ indicates the
within-group correlation between $\hat{\beta}_{rd}$ and $|\hat{\delta}_{rd}|$. Large positive $\rho$ indicates larger treatment effects for larger separation events.

While there does appear to be some differences across diagnoses, the extent of heterogeneity is not large. Within-group means range from 61% and 79%. The specialist effect appears smallest for “Symptoms & ill-defined conditions,” suggesting that there exists more sorting on patient unobservables for patients referred in this category.

**Figure C1:** Heterogeneity in specialist effect by diagnosis system.

![Figure C1: Heterogeneity in specialist effect by diagnosis system.](image)

**Notes:** Horizontal bars show the interquartile range of $\hat{\beta}_{rd}$ and $|\hat{\delta}_{rd}|$ by group, where $\hat{\beta}_{rd}$ is estimated using Equation 40. Points indicate the within-group means, and the center column $\rho$ reports $\text{Corr}(\hat{\beta}_{rd}, |\hat{\delta}_{rd}|)$.

Next I test if the specialist effect varies by the prestige of the specialist’s training. In this case, heterogeneity more likely arises due to differences in actual treatment effects, rather than the extent of patient sorting. PCPs may be more willing to defer to the medical decisions of specialists they perceive to be well-trained. If so, then $\beta_{rd}$ should be increasing with the prestige of the specialist’s training program.

In this case, I study heterogeneity in $\beta_{rd}$ by the perceived quality of the physician’s residency program. After medical school, physicians must complete a multi-year residency program before
they can practice without supervision. I test residency programs rather than medical schools, because rankings are more complete for residency programs and because even international medical school graduates must complete an accredited residency program to practice in the United States.

I gather specialists’ residency information from U.S. Health News and 2017-18 residency ranking information from Doximity. The residency names from U.S. Health News sometimes differ slightly from those from Doximity, so I use a fuzzy matching algorithm to merge them. I lack historical rankings information, so I cannot study heterogeneity based on the residency prestige during that specialist’s actual training. Instead, this limitation forces me to assume relative stationarity in residency rankings, or that present rankings are more salient to PCPs than historical ones.

**Figure C2a** presents heterogeneity by the percentile of the new specialist’s residency program. I use percentiles to account for the fact that some specialties have more programs than other specialties. For cases with more than one new specialist, I take the average of rankings, weighted by the number of referrals. **Figure C2b** shows heterogeneity by the new specialist’s experience across the broad bins reported by U.S. Health News.

Strikingly, neither figure displays heterogeneity along their respective margins. The average treatment effect is similar across the residency rank and experience of the new specialist. Potentially, the lack of heterogeneity is due to the fact that PCPs have already agreed to refer to these specialists. To the extent that either residency prestige or experience matters, those factors might deter PCPs from forming a relationship with that specialist initially. Testing that hypothesis is beyond the scope of this particular project.
Figure C2: Heterogeneity of $\beta_{rd}$ estimated in Equation 40 by features of the new specialist.

(a) Heterogeneity in $\beta_{rd}$ by percentile ranking of new specialist’s residency program. “N/A” indicates that the PCP did not contract with a new specialist or I am unable to match the specialist’s residency program to a ranking.

(b) Heterogeneity of $\beta_{rd}$ by new specialist’s experience.

Notes: Horizontal bars show the interquartile range of $\hat{\beta}_{rd}$ and $\hat{\delta}_{rd}$ by group, where $\hat{\beta}_{rd}$ is estimated using Equation 40. Points indicate the means within group, and $\rho$ reports $\text{Corr}(\hat{\beta}_{rd}, |\hat{\delta}_{rd}|)$. 
**Figure C3**: Coefficients from a 20-fold cross-validation LASSO regression of features on $\hat{\beta}_{rd}$.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
<th>$b$</th>
<th>Interquartile $\delta$ range</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta (in levels)</td>
<td>$[1.58, 1.67]$</td>
<td>0.126</td>
<td></td>
<td>2,037</td>
</tr>
<tr>
<td>Delta (in levels)</td>
<td>$[1.62, 1.79]$</td>
<td>0.084</td>
<td></td>
<td>2,037</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Injury and poisoning</td>
<td>0.077</td>
<td></td>
<td>214</td>
</tr>
<tr>
<td>Delta (in levels)</td>
<td>$[1.72, 1.85]$</td>
<td>0.050</td>
<td></td>
<td>2,038</td>
</tr>
<tr>
<td>Class of $s^*$</td>
<td>Surgery</td>
<td>0.035</td>
<td></td>
<td>153</td>
</tr>
<tr>
<td># of claims during pre-period</td>
<td>$[51, 209]$</td>
<td>0.021</td>
<td></td>
<td>1,969</td>
</tr>
<tr>
<td>Charlson Comorbidity Index</td>
<td>[0.66, 3.09]</td>
<td>0.020</td>
<td></td>
<td>2,021</td>
</tr>
<tr>
<td>Delta (in levels)</td>
<td>$[503, 773]$</td>
<td>0.014</td>
<td></td>
<td>2,037</td>
</tr>
<tr>
<td># of claims during pre-period</td>
<td>$[37, 51]$</td>
<td>0.013</td>
<td></td>
<td>1,966</td>
</tr>
<tr>
<td>New specialist organization?</td>
<td>TRUE</td>
<td>0.012</td>
<td></td>
<td>2,952</td>
</tr>
<tr>
<td>Was $s^*$ at a teaching hospital?</td>
<td>TRUE</td>
<td>-0.012</td>
<td></td>
<td>3,387</td>
</tr>
<tr>
<td>% of claims to $s^*$</td>
<td>(50%, 100%)</td>
<td>-0.013</td>
<td></td>
<td>415</td>
</tr>
<tr>
<td>New specialists’ residency rank percentile</td>
<td>(75, 100]</td>
<td>-0.015</td>
<td></td>
<td>1,116</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Digestive system</td>
<td>-0.022</td>
<td></td>
<td>589</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Respiratory system</td>
<td>-0.032</td>
<td></td>
<td>1,621</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Symptoms &amp; Ill-defined conditions</td>
<td>-0.052</td>
<td></td>
<td>1,196</td>
</tr>
</tbody>
</table>

**Notes**: Features used in the LASSO include $\hat{\delta}_{rd}$; $|\hat{\delta}_{rd}|$; 5-bins of $\hat{\delta}_{rd}$; 5-bins of $|\hat{\delta}_{rd}|$; 5-bins of $\delta_{rd}$ in levels; 5-bins of $f_{rd}$, the share of patients referred to the separating specialist $s^*$; PCP’s gender; fraction of patients hispanic, fraction black; 5-bins of the total count of pre-period claims by PCP; 5-bins of the Charlson Comorbidity Index; the new specialist’s gender; the residency rank of the new specialist; whether the majority of the PCP’s claims are at a teaching hospital; whether a majority of the new specialist claims are at a teaching hospital; whether the new specialist works in a new organization; 5-bins of the new specialist’s observed-physician effect $A_{pd}$; and diagnosis categories.