NBER WORKING PAPER SERIES

TEAM FORMATION AND PERFORMANCE: EVIDENCE FROM HEALTHCARE REFERRAL NETWORKS

Leila Agha
Keith Marzilli Ericson
Kimberley H. Geissler
James B. Rebitzer

Working Paper 24338
http://www.nber.org/papers/w24338

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2018

We acknowledge funding from the Boston University Questrom School of Business and the National Institute of Health Care Management. We are especially grateful to Benjamin Lubin for his thoughtful contributions to the project. Buqu Gao provided excellent research assistance. We thank Jay Bhattacharya, Jon Skinner, Doug Staiger and Jay Bhattacharya for helpful comments, along with seminar participants at Boston University, Dartmouth College, Emory University, RAND, University of Toronto, and Yale School of Public Health, and conference attendees at the Caribbean Health Economics Symposium, iHEA, Midwest Health Economics Conference, ASSA Annual Meeting, and the NBER Health Care meetings. We are also grateful to Jean Roth and Mohan Ramanujan for assistance obtaining and managing the Medicare claims data. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Team Formation and Performance: Evidence from Healthcare Referral Networks
Leila Agha, Keith Marzilli Ericson, Kimberley H. Geissler, and James B. Rebitzer
NBER Working Paper No. 24338
February 2018
JEL No. D85,I10,I11,L2,M5

ABSTRACT

How does team structure affect productivity? We address this question with an application to healthcare by examining the teams that primary care physicians (PCPs) assemble when they refer patients to specialists. Our theoretical model analyzes how PCPs trade off costly coordination against beneficial specialization, predicting that coordination improves when PCPs concentrate their referrals within a smaller set of specialists. Empirically we find that patients of PCPs with concentrated referrals have lower healthcare costs. This effect exists for commercially insured and Medicare populations; is statistically and economically significant; and holds under identification strategies that account for unobserved patient and physician characteristics.

Leila Agha
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
leila.agha@dartmouth.edu

Kimberley H. Geissler
University of Massachusetts at Amherst
School of Public Health and Health Sciences
715 N Pleasant St
325 Arnold Hall
Amherst, MA 01003
kgeissler@umass.edu

Keith Marzilli Ericson
Boston University
Questrom School of Business
595 Commonwealth Avenue
Boston, MA 02215
and NBER
kericson@bu.edu

James B. Rebitzer
Professor of Management, Economics, Public Policy
Markets, Public Policy and Law Department
Boston University School of Management
595 Commonwealth Ave.
Boston, MA 02215
and NBER
rebitzer@bu.edu

1. Introduction

Teams are pervasive in economic organizations, so the performance of teams matters a great deal for organizational efficiency. Potential obstacles to optimal team performance include free-riding (Holmstrom 1982) and incomplete contracting (Williamson 1985; Hart 2017), but even in the absence of these problems, coordinating complex tasks and sharing information within teams is difficult (Marschak and Radner 1972, Becker and Murphy 1992, Dessein and Santos 2006).

The efficiency of teams is especially important and challenging in healthcare because a wide array of specialist and primary care clinicians must interact to treat patients who have complex health problems. These clinical teams are especially vulnerable to sub-optimal performance because they are often comprised of members from different firms, which complicates coordination and exacerbates incomplete contracting.[1] Prior research suggests that repeated interactions among team members may improve performance by encouraging investments in team relationships (Crawford 1990) and team human capital (Chillemi and Gui 1997; Mailath and Postlewaite 1990). In this paper, we build on these ideas by investigating how team structure affects the costs of healthcare delivery.

Primary care physicians (PCPs) regularly refer patients to specialists, but researchers know little about how the structure of these referrals influences the cost or quality of care that patients receive. Our analysis rests on the assumption that repeated interactions between PCPs and specialists facilitate investments in team coordination.[2] When PCPs concentrate their patient referrals within a narrower group of providers within a specialty, e.g. refer their cardiology patients to a smaller set of cardiologists, this encourages repeat interactions and so incentivizes greater investments in team coordination. This enhanced coordination, however, comes at a cost: the loss of the quality gains from a better patient-specialist match when a larger set of diverse specialists is available. We formalize this tradeoff between coordination and specialization in a model of team formation. The model shows that the concentration of specialist referrals within a PCP's team is a meaningful proxy for coordination effort. In our empirical work, we use this insight to develop a novel measure of PCP *team referral concentration*. We apply this measure to study the effects of PCP-specialist team structure on patient costs.

Our empirical results can be briefly summarized: patients of PCPs with more concentrated specialist referrals have lower total healthcare costs. This association exists for both commercially insured and Medicare populations; is statistically and economically significant; and holds under various identification

[1] The "Stark law" effectively prohibits physicians from being financially compensated for referrals. As a result, a referring physician cannot write a principal-agent incentive contract. Kolber (2006) provides more detail.

[2] Barnett et al. (2012b) survey PCPs about the reasons they choose a referral, and find that patient experience and ease of communication are important reasons. Geissler et al. (2017) find that practice sites and medical groups (but not physician contracting networks) are important determinants of referrals.

strategies. For commercially insured, chronically-ill patients in Massachusetts, those treated by PCPs with the average below-median team referral concentration have 12% lower spending and 6.3% lower utilization, compared to those treated by PCPs with the average above-median team referral concentration after controlling for detailed patient, physician, and insurer characteristics. The effect of team referral concentration on spending persists even using within-PCP, cross-specialty variation in team referral concentration, or when comparing patients who consult the same specialist but are referred from PCPs with different levels of team referral concentration. For Medicare beneficiaries, we use an additional identification strategy to study those who switch doctors as the result of a move across geographic regions. In these analyses, we use patient fixed effects and an instrumental variables strategy based on regression to the mean to remove potential endogeneity in choice of PCP. In the Medicare population, we also find that an increase in the team referral concentration of a patient's PCP is associated with lowered spending.

Our study contributes to a growing empirical literature on the economics of team organization and productivity.[3] In work that examines the role of coordination of expertise within software development teams, Faraj and Sproull (2000) highlight the role of coordination on team performance. Two prior studies have demonstrated a link between team familiarity (i.e., repeated interactions between team members) and team performance for software development teams (Huckman et al. 2009) and surgical teams (Reagans et al. 2005).

In addition to contributing to the general economics literature on team performance, our paper also relates to a large body of literature arguing that better care coordination may reduce healthcare costs and improve quality.[4] This prior research has helped motivate important policy initiatives that aim to improve care coordination, including policies designed to alter organizational form and incentives (e.g. Accountable Care Organizations and Patient Centered Medical Homes) and public subsidies to health information technology such as electronic medical records.

[3]See Bloom and Van Reenan (2011) for a detailed discussion of how team management practices are related to productivity. Other research has studied team formation. Hamilton, Nickerson, and Owan (2003) examine the impact of team versus individual work for productivity in a garment plant and examine who chooses to join a team. Experimental economists have examined the formation of teams in the lab (e.g. Weber 2006, Feri, Irlenbusch, and Sutter 2010, and Grosse, Putterman, and Rockenback 2011). There is also a literature in psychology on the performance of teams, reviewed in Kozlowski and Ilgen (2006).

[4] For example, see Agha et al. (2017); Hussey et. al. (2014); Romano et. al (2015); and Milstein and Gilbertson (2009). In economics, care coordination is often referred to by the obverse term, "care fragmentation". For a discussion and review of the literature on fragmentation see Cebul et. al. (2008); Frandsen and Rebitzer (2014); and Rebitzer and Votruba (2011). For a contrary view on the role of care coordination in lowering costs, see McWilliams (2016). The problem of building organizations and institutions that coordinate the activities of specialized providers has implications that extend beyond healthcare costs. See, for example, Meltzer (2001) and Meltzer and Chung (2010) on hospitalists.

A little discussed limitation of the preceding care coordination literature is that it has focused on the relationship an individual patient has with their set of providers rather than on relationships within teams of providers.[5] To see the significance of this distinction, imagine a PCP who refers each of her patients with diabetes to two specialists: a cardiologist and an endocrinologist. The PCP can choose to make each of her patient's referrals to the same cardiologist and the same endocrinologist, or could refer to a different cardiologist and different endocrinologist for each patient. In both cases, the distribution of an individual patient's care across providers is the same: each patient sees her PCP, cardiologist, and endocrinologist. But in the former case, the provider team works together more frequently, thus facilitating improved team relationships and coordination.

An advantage of studying teams in healthcare as opposed to other industries is that we have micro data on team formation, tasks and performance across a large number of PCP-specialist teams. Specifically, our empirical analysis relies primarily on the Massachusetts All Payer Claims Database (APCD).[6] The APCD data contains almost all commercially insured patients in the state.

The APCD is uniquely suited for our study design because it allows us to observe a larger fraction of a given provider's links (via shared patients) to other providers than has been available in other data sets from Medicare or commercial insurers. Comprehensiveness matters because team referral concentration reflects the entire network of physician referrals. With this type of network variable, we show that omitting substantial numbers of patient referrals – as is routine in more commonly used Medicare or commercial insurer data sets – introduces measurement error and attenuation bias. We also replicate our main findings using a 20% sample of Medicare fee-for-service enrollees. This data set provides broader geographic coverage and the ability to conduct longitudinal analyses, at the expense of more statistical noise in our measure of team referral concentration.

The consistency of our findings across both a commercially insured, working age population and a Medicare population is noteworthy for two reasons. First, because it suggests that insurance characteristics are not likely to be driving the observed relationship between team referral concentration and healthcare spending. Secondly, this pattern reduces concerns that unique features of healthcare delivery in one state drive our findings. A limitation of our empirical analyses is that we do not assess the impact of team referral concentration on care quality, since quality is multidimensional and difficult to observe. This matters because, as our theoretical model of team formation emphasizes, there may be quality gains available from enhanced specialization and improved patient-specialist

[5] There is also another literature that maps out patient-sharing networks among physicians (see e.g. Barnett et al. 2012a). This literature uses characteristics of the social network inferred from claims data that are distinct from our measure of referral concentration. Jackson, Rogers, and Zenou (2017) review the underlying concepts of network structure (e.g. clustering, centrality).

[6] See Ericson and Starc (2015) for a more detailed description of the Massachusetts APCD.

matching when team referral concentration is low. From this perspective, it may be helpful to interpret our cost results as suggesting a bound on the value of any offsetting quality improvements required to make low levels of team referral concentration worthwhile.

The paper proceeds as follows. Section 2 lays out a theory of coordination, specialization, and team formation, providing a theoretical motivation for our measure of team referral concentration. Section 3 introduces our empirical measure and describes the empirical approach. Section 4 describes the data. Section 5 presents results from the Massachusetts APCD. Section 6 extends our work to the Medicare sample and provides an alternative approach to identification. Section 7 concludes.

2. Coordination, Specialization, and Team Formation

In forming teams to maximize productivity, there is a tradeoff between coordination and specialization. In our context, when PCPs concentrate their patient referrals within a narrower group of providers within a specialty, e.g. refer their cardiology patients to a smaller set of cardiologists, this encourages repeat interactions and so incentivizes greater investments in team coordination. This enhanced coordination, however, comes at the cost of the quality gains that can occur from referring to a more diverse set of specialists with an improved patient-specialist match.

We formalize this tradeoff in the following model of team formation. The model shows that team referral concentration is a useful indicator of coordination within teams. It also suggests factors that may cause team referral concentration to vary across physicians independent of the clinical characteristics of the PCP's patients. Both of these results will prove useful in interpreting the subsequent empirical work.

A. Model Set-Up

In our model, PCPs all have the same number of chronically ill patients, normalized to measure one. Each patient is referred to a specialist and there is only one specialty.[7] The PCP chooses how many specialists *N* to work with and how much effort r_s she puts into coordinating with each specialist *s*. Then, the PCP selects a specialist for each patient referral.

The PCP's utility from a patient's referral depends on the quality and cost of care the patient receives when referred to a specialist. Thus, a PCP's utility for referring patient *i* to specialist *s* is given by

$$U_{is} = Q_s - \theta c_s + \varepsilon_{is},$$

[7] Another interesting dimension that we discuss in the empirical section is the PCP's decision of whether to refer to the specialist at all. To highlight the coordination/specialization tradeoff in team formation, our model abstracts away from that decision and assumes a fixed number of referrals.

where Q_s is a measure of the quality of care provided by the specialist *s* and c_s is the (overall) healthcare costs the patient will incur if they are referred to the specialist. Note that c_s includes all healthcare costs generated by both PCP and specialist actions.[8] The relative weight the PCP places on costs versus quality is given by $\theta > 0$. Finally, ε_{is} is the idiosyncratic match value between the patient *i* and specialist *s.* We assume that ε_{is} is distributed i.i.d. Type 1 Extreme Value.

Given these assumptions, the PCP's total expected utility from all her referrals is:

$$E[\max_{s \in N}(U_{is})] = \ln\left[\sum_{s=1}^{N} e^{V_{is}}\right]$$

where V_{is} is the deterministic component of a referral's utility ($V_{is} = Q_s - \theta c_s$). This is the standard log-sum formula for welfare with differentiated products (Small and Rosen 1981).[9] We see that without any effort costs of working with a specialist, the benefits of matching each patient to the specialist that is best for them creates an incentive to work with as many specialists as possible.

We now turn to the PCP's effort costs, which include a set-up cost to work with each specialist and a coordination cost that improves quality and reduces spending. We assume that PCPs must pay a fixed set-up cost of effort $\varphi > 0$ the first time they refer a patient to a particular specialist. For example, this could be due to the search cost of identifying the specialist or the startup cost of establishing a communication channel with the specialist. Once paid, the PCP does not need to pay the setup cost for subsequent patients referred to that specialist.

Next, the PCP can put effort into coordinating care with a specialist. We assume that heterogeneity in specialist characteristics is limited to the idiosyncratic match value, ε_{is}. That is, all specialists have the same average healthcare cost c and quality Q ex ante, before PCPs exert coordination effort. We make the following assumption for how coordination affects healthcare cost and quality:

> **Assumption 1**: "Coordination Spillovers". If a PCP invests effort r_s in coordination with a specialist *s,* this benefits all of that PCP's patients who see that specialist via lower healthcare costs *c* and higher quality *Q* as follows:
>
> $$Q_s = Q + \omega r_s$$

[8] PCPs may in fact care differentially about costs generated by their own care versus costs generated by the specialist, but we abstract away from this issue.

[9] The assumptions underlying the log-sum formula can be restrictive. For our purposes, however, it offers a parsimonious expression of the welfare gain from increasing the number of specialists, holding fixed costs of coordination.

$$c_s = c - r_s$$

where $\omega > 0$ scales the relative effect coordination effort has on quality relative to cost

We also assume that the cost of effort is quadratic. Now, we can write the PCP's objective function as the sum of her utility of patient referrals and disutility of effort:

$$\max_{N,\{r_s\}_{s\in N}} \ln\left[\sum_{s=1}^{N} e^{V_{is}}\right] - \varphi N - \sum_{s=1}^{N}\left[\frac{1}{2}r_s^{\,2}\right]$$

We also make a technical assumption to guarantee that the optimal choice of coordination effort is symmetric across specialists: $\omega + \theta < \sqrt{2}$.

B. Model Results: Team Referral Concentration and Investments in Care Coordination

Solving the model outlined above, we have the following results (proofs in Appendix).

Result 1: The PCP's optimal team structure has a number of specialists N*. The PCP invests the same amount of coordination effort r_s^* with each specialist and refers to each specialist with the probability $\frac{1}{N*}$.

Next, we investigate how variation in the number of specialists a PCP works with translates into cost and quality.

Result 2: Coordination effort r_s^* is inversely proportional to the number of specialists in the team, *N. A PCP with a higher fixed cost φ of working with an additional specialist will work with fewer specialists, invest more coordination effort with each specialist, and have lower expected healthcare costs for their patients.**

To see this, from the first order condition for the choice of effort, we find that

$$r_s^* = \frac{1}{N^*}(\omega + \theta),$$

where $\frac{1}{N^*}$ is the probability the PCP refers to that specialist. Higher fixed costs φ of adding additional specialists to the referral pool lowers the optimal number of specialists and increases r_s^*. Note also that, holding fixed the number of specialists in the team, a higher value of θ (PCP disutility from healthcare costs) raises coordination effort. Effort also increases when ω is higher—that is, when effort has a stronger effect on quality.

C. Discussion

The model analyzes how the tradeoff between specialization and coordination influences team formation. PCPs who work with a smaller team of specialists have more concentrated specialist referrals, increasing coordination within the team, but their patients experience a worse idiosyncratic match with their specialists. The model does not specify the precise source of specialist match value but it plausibly includes things like the specialist's ability or experience with the patient's specific disease, the patient's travel time, or appointment schedules. In forming a team, the PCP balances this loss of quality against the gains in care coordination that are enabled by repeat interactions with a smaller set of specialists. The model is also agnostic about the source of gains from coordination but the medical literature suggests this likely includes better personal relationships between the PCP and the specialist; improved communication; or even establishing interoperable electronic medical records and e-referral systems. [10]

Our model identifies factors that would cause $\frac{1}{N^*}$ (probability of referral to a given specialist) to vary independent of the clinical characteristics of a PCP's patient panel. Specifically it highlights the importance of the fixed cost of adding another specialist to the team, φ, as a source of variation across providers. PCPs who have worked with specialists in the past have already paid the start-up cost of establishing a relationship. This would effectively lower φ. Variation in PCP knowledge about specialist options in the area could also lead to variation in φ, as could variation in the size and scope of local multi-specialty physician practices. The parameter φ could also respond to regional variation in the demand and supply of specialists. If, for example, a PCP is operating in an area where demand for specialists greatly exceeds supply, this may increase the cost to a PCP of establishing a relationship with another specialist.[11]

If, as we assume in our model, improved care coordination reduces costs, then $\frac{1}{N^*}$ will also vary with θ, the weight PCPs place on reducing total care costs. This assumption is reasonable given prior research that stresses cost savings from eliminating duplicative diagnostic studies; avoiding cascades of testing and low-value therapeutic interventions; and reducing polypharmacy (the use of a large number of

[10] In a recent article in the *New England Journal of Medicine,* Press (2014) described the significance of these relationships as follows: "… one ingredient that's essential to effective teamwork across care settings [is] relationships. Having a relationship with another clinician makes it easier to communicate because the social barrier is lower and opportunities to communicate are more frequent." Press's anecdotal description of physician teams aligns with economic theory suggesting that relationship-specific investment is important to productive team relationships (cf. Crawford 1990), Stille (2005), and Bodenheimer (1999) describe coordination between primary care and specialists as a central goal of primary care and document widespread failures of communication when specialist referrals are made. Stille (2005) identifies a successful general practice model as referring within a "tight web of consultants in which physicians know one another well and can share work effectively".

[11] This sort of congestion may also influence wait times for specialists, but exogenously determined specialist wait time does not alter the results of our model. It is isomorphic to reducing the quality benefit of specialist referral, thus leading to increased team referral concentration. We thank Jay Bhattacharya for sharing this insight into modeling specialist congestion.

medications).[12] We note, however, that recent research on ACOs has cast some doubt on the relationship between improved care coordination and cost reductions (McWilliams, Chernew, and Landon 2017). The empirics presented in the next section will subject the assumption that lower team referral concentration (and thus improved care coordination) lowers costs to rigorous empirical testing.

Although we do not explicitly model the decision to refer to a specialist at all, our model suggests that cross PCP variation in $\frac{1}{N^*}$ likely correlates with variation in the number of specialists seen by an *individual* patient. PCPs who put greater weight on reducing total care costs (i.e. with PCPs with higher θ) will be also tend to be to be judicious in their use of specialists, since this practice pattern is one way to directly reduce healthcare costs. In addition, as the clinical literature suggests, it is also likely that enhanced coordination between PCPs and specialists will lead to fewer, and more appropriate, specialist referrals (Bodenheimer 1999). The correlation between team referral concentration and the number of specialists an individual patient sees is an important empirical issue that we take up in section three below.

Finally, it is worth noting that in some cases the PCP may not chose team structure optimally. This can happen, for example, when organizational rules or externally imposed narrow networks limit the set of specialists to whom a PCP can practically refer. In this case, however, the logic of our model suggests that optimal choice of coordination effort will still be inversely proportional to team size.

3. Empirical Implementation of Team Referral Concentration

In this section, we introduce an empirical measure of team-based referrals that generalizes from the preceding model. To build intuition for our team referral concentration measure, consider Figure 1. The left hand side of the figure depicts four patients, each of whom sees one PCP and two specialists. The differences across patients in their chosen specialists are due to the referral practices of their PCPs. PCP A, who treats patients 1 and 2, refers each of them to a different set of specialists while PCP B refers patients 3 and 4 to the same set of specialists. These referral patterns give rise to very different levels of repeat interactions for the two PCPs. As depicted on the right hand side of the figure, PCP A interacts with each of four specialists only once while PCP B interacts with each of two specialists twice. PCP B's referrals are more concentrated within a smaller set of specialists than PCP A's.

If repeat interactions ease coordination challenges, it follows that the patients of PCP B will have superior coordination of care than the patients of PCP A – even though the care of each individual patient involves the same number of providers. In our empirical specifications we compare otherwise

[12] Bodenheimer (1999) outlines the cost saving mechanisms highlighted above.

similar patients whose PCPs have different levels of team referral concentration and measure the relationship to healthcare costs and utilization.

Our theory motivates a measure of team referral concentration, 1/N, using the special case where PCPs refer equally to each specialist in their team. We generalize this measure by calculating a PCP's Herfindahl–Hirschman Index (HHI) of shared patients within each specialty. A PCP has a *shared patient* with a specialist physician if that patient visits both the PCP and the specialist during our sample window (calendar year 2012). For each PCP *d*, for each specialist *s* in each specialty *j* (e.g. endocrinology), we calculate the number of shared patients m_{ds}. Then, to translate these shared patients into "market shares", we calculate the PCP's total number of patient-specialist links for that specialty,[13] or $M_{dj} = \sum_{s \in j} m_{ds}$. For each PCP-specialist pair *ds*, we then calculate that specialist's share of total PCP referrals in that specialty: $share_{ds} = \frac{m_{ds}}{M_{dj}}$.

Our PCP-level measure of team referral concentration within each specialty *j* is then the HHI: $ReferralCon_{dj} = \sum_{s \in j} (share_{ds})^2$. When PCPs refer equally to each specialist, this measure reduces to 1/N, the same measure of team referral concentration that emerged in our model.

Our theoretical model also considered only a single specialty. To accommodate the reality of referrals across many specialties, our empirical measure of team referral concentration averages across the various specialties weighting each specialty equally.

The resulting measure of team referral concentration, $ReferralCon_d$, describes the *network* of connections among providers where connections are defined by patients shared between PCPs and specialists. Like other network measures, the closer the sample of patients in the data is to the underlying population, the more accurate is our measure of team referral concentration. Because the sample of patients included in the Massachusetts APCD includes a large fraction of the underlying population, our results in that sample are less vulnerable to measurement error. Measurement error is a much bigger concern in our Medicare results where we rely on a 20% sample of Medicare beneficiaries. We discuss the complications this creates in Section 6.

Figure 1 demonstrates that our team-based referral measure is conceptually distinct from measures of the number of distinct providers an individual patient sees. As we have already noted in the theory section, however, these two features of care delivery are unlikely to be independent of each other. Concentrating an individual patient's visits in a smaller number of providers likely increases the

[13] Note that this is not the number of patients seeing a specialist, but the number of patient-specialist links. We define it this way because a PCP sharing a patient with a specialist is a referral. Sharing the patient with two specialists then counts as two referrals. The market share we use is the market share of "referrals" within a given specialty.

continuity of care a patient receives and is a plausible channel through which teams can achieve more cost-effective team coordination. Alternatively, reducing individual patient visits to distinct providers may reduce costs entirely independently of any differences in team coordination. Because concentrating patient visits in a smaller number of providers can be either a mechanism for enhanced team coordination or a potential confounder, we report results with and without controls for patient care continuity.

Following prior practice in the empirical literature, we measure patient care continuity using an HHI that summarizes the concentration of a patient's visits across different providers.[14] Defining n_{ip}, the number of visits during the year by patient *i* to each provider p (who may be a PCP or a specialist), we construct the patient care continuity HHI as $\sum_{p \in physicians} \left(\frac{n_{ip}}{N_i}\right)^2$ where $N_i = \sum_{p \in physicians} n_{ip}$ is the total number of visits by patient *i* to all physicians p.

4. Data on the Commercially Insured (Mass. APCD)

Our primary data comes from the 2012 Massachusetts All Payer Claims Database (APCD), version 2.1; in section 6, we replicate and extend our findings in a national sample of Medicare beneficiaries. We create two extracts from the APCD:[15] a broad sample that allows us to characterize PCP's referral patterns, and an analysis sample on which we run our regressions relating team referral concentration to total spending.

A. Data on PCP team referral concentration within teams

We use a *broad sample* to construct our PCP-level measure of team referral concentration. In the broad sample, we limited claims to evaluation and management visits for patients aged 21 and older with primary health insurance available in the APCD. This includes patients enrolled in commercial health insurance, self-insured employers, Medicaid fee-for-service, Medicaid managed care, and Medicare Advantage whose claims are processed by the 12 largest payers. Using the National Provider Identifier (NPI) associated with each insurance claim, we link claims to physician specialty and demographic information in the National Plan and Provider Enumeration System data.[16] We analyzed five common specialties: cardiology, dermatology, endocrinology, obstetrics and gynecology (OB/GYN), and orthopedics. We categorized each physician as a PCP or as one of these five specialist types; if the

[14] Prior empirical work has used either patient level Herfindahl -Hirschman indices of patient visits across providers or the closely related Bice-Boxerman continuity of care index. See Pollack et al. (2016) for a review.
[15] In all cases, we limit to 12 large payers with complete claims versioning information.
[16] Our extract of the National Plan and Provider Enumeration System data was downloaded in February 2014. We used the physician's primary specialty if available. If multiple specialties were listed, but none were indicated as primary, we used the most specialized as their classification.

physician did not fall into one of these categories, they were excluded from the construction of team referral concentration. Physicians outside of these specialty categories were included in the construction of other visit and cost measures

Team referral concentration is constructed based on physician links. Each link represents a PCP-specialist pair who share at least one patient, with the strength of the link determined by the number of patients in common. We calculate team referral concentration using these links with the method described in Section 3 above. Referral concentration for each PCP is first calculated separately by specialty for each of the five specialties, and then averaged equally across specialties to define a single PCP-level measure of team referral concentration. Subsequent analyses also exploit within-PCP variation in referral concentration by specialty.

B. Data on patient outcomes

Our *analysis sample* limits to chronically ill patients residing in Massachusetts, aged 21-64, who are continuously enrolled with the same commercial insurer or self-insured employer for all of 2012.[17] We focus on chronically ill patients because we expect coordination of care to matter most for patients with complex conditions that often require the care of specialists. The restriction to continuous enrollment helps remove noise or confounds associated with insurance churn, and facilitates calculation of annual spending and utilization.

Following Frandsen et al. (2015), we defined chronically ill enrollees as those with at least one claim with an ICD-9 diagnosis code indicating one or more of the following conditions: coronary artery disease, cerebrovascular disease, peripheral arterial disease, mesenteric vascular disease, other ischemic vascular disease or conduction disorders, heart failure, migraine and cluster headache, hypertension, hyperlipidemia, diabetes mellitus, asthma, chronic obstructive pulmonary disease, hypercoagulability disorders, osteoarthritis, and/or rheumatoid arthritis.[18]

We assigned each patient to a PCP based on the "plurality primary care physician algorithm" developed by Pham et al. (2007), which assigns each enrollee to the PCP with the highest number of evaluation and

[17] Medicaid Fee-for-Service, Medicaid Managed Care, and Medicare Advantage enrollees are included in the calculation of PCP team referral concentration, but they are not included in the analysis sample due to data limitations. Medicaid Fee-for-Service and Medicaid Managed Care patients (aged 21-64) are included in the number of patients treated by the PCP.
[18] Specific ICD-9 codes for identifying these individuals are included in Appendix B.

management visits, with ties broken by assignment to the PCP with the highest total billed claims. We drop patients from the sample who cannot be assigned to a plurality PCP with this algorithm.[19]

We capture health status by hierarchical condition category (HCC) risk scores and binary condition categories calculated using the Massachusetts "Market-wide Risk Adjustment" calculator and an individual's claims during the year.[20] These HCC risk scores are calculated using a diagnosis-based algorithm that assigns individuals binary indicators for each condition category if they have claims that indicate a given condition (e.g., diabetes without complications).[21]

We calculate total spending at the enrollee level for all inpatient and outpatient claims in 2012. Spending outcomes are based on the insurer allowed amount, which consists of the insurer paid amount and any patient cost-sharing. Higher annual patient spending can correspond to more procedures being performed, or to the same number of procedures being performed but by a higher price provider or in a higher price setting (e.g., hospital vs. physician's office).

To distinguish the contribution of price and quantity changes in our aggregate spending outcome, we also create a measure of utilization using standardized prices. Standardized prices are defined as the mean price per CPT code, procedure modifier, and quantity of procedure units.[22] These standardized prices are constant for each service across insurers and providers.[23] After applying standardized prices to each claim, we aggregate these amounts to the patient level to create a measure of annual care utilization for 2012.

[19] We calculated the modal ZIP code for each physician as the location where they practiced most days. In order to exclude physicians who may be treating many out-of-state patients who are not in our data set, we exclude physicians for whom this ZIP code was not in Massachusetts. Patients matching to these PCPs are in turn excluded from our analysis.

[20] In line with software instructions, we limited to inpatient hospital, outpatient facility, and professional claims (Kautter et al. 2014).

[21] These indicator variables, along with demographic characteristics, are used to assign individuals an HCC risk score given their plan's metal level (based on actuarial value). We assumed all individuals to be in "gold" plans, indicating an actuarial value of 80%, without cost sharing reductions.

[22] Before calculating standardized prices, we winsorized the payment data, rounding all non-zero payments in the bottom 1% up to the 1st percentile prices and all payments in the top 1% down to the 99th percentile price.

[23] We standardized prices based on procedure codes as described. For some procedures, the CPT code is different for a service provided in an inpatient versus outpatient setting, for which our method does not correct.

5. Team Referral Concentration and Spending for the Commercially Insured in Massachusetts

A. Descriptive Statistics

Team referral concentration has a mean value of 0.131 and varies widely across PCPs, with a median value of 0.119; it has a standard deviation of 0.064 and the distribution is displayed in Figure 2. As reported in Table 1, PCPs with above median concentration have an average referral concentration of 0.18, compared to 0.08 among below median physicians. The differences between these are equivalent to a PCP increasing the number of specialists in each specialty category they refer to from 5.6 to 12.5 specialists, if they referred with equal frequency to each specialist within a specialty.

Figure 3 reports a PCP-level scatterplot of referral concentration and the average of log per patient total spending. The graph also displays the corresponding lowess smoother. At the PCP level, higher levels of referral concentration are associated with lower average spending throughout the entire observed distribution of concentration levels. The figure suggests a negative relationship that is strongest in the lower part of the team referral concentration distribution, where most of the data lies. This would be consistent with very uncoordinated care being more expensive. An alternative interpretation is that PCPs with the sickest (i.e. most costly) patients refer to many different specialists within a specialty (e.g., two cardiologists with different sub-specialty expertise) due to clinical need.

While the uncontrolled, cross-sectional PCP-level comparison shown in Figure 3 raises significant concerns of omitted variable bias, it is worth noting the distribution of disease categories, gender composition of patients, and rates of hospitalization are quite similar across patients seeing PCPs with high versus low team referral concentration. Table 1 reports these comparisons at the patient level.

Table 1 does not show large differences in patient care continuity HHI between PCPs above and below the median team referral concentration, though patients treated by high referral concentration PCPs also tend to have their patient visits concentrated among a slightly smaller number of providers. A more concentrated patient care continuity HHI may in fact be one of the channels through which team referral concentration has an effect: a more coordinated PCP-specialist team could reduce the number of unique providers seen by a patient through improved coordination. However, PCPs with high referral concentration may be more careful about referral decisions generally, and have both a higher threshold for whether they refer a patient to a specialist at all and thus have a more concentrated pattern of referrals. In the next section, we use a regression approach to isolate variation in referral concentration from patient care continuity HHI and other possible confounders.

To examine differences in team referral concentration by organizational affiliation, we link the physician NPIs to the 2010 Massachusetts Provider Database (MPD) maintained by Massachusetts Health Quality Partners. The MPD has information on the organizations and physician contracting networks to which the PCP belongs.[24] Appendix Table A1 lists the mean and standard deviation of team referral concentration by physician contracting network. Team referral concentration varies substantially across different physician contracting networks. However, team referral concentration also varies substantially across different PCPs within a physician contracting network. Also of note, the largest, highest price hospital system in Massachusetts (Partners Community Health Care, see e.g. Seltz et al. 2016) has an average team referral concentration of 0.11, near the average team referral concentration in our analysis sample.

B. Empirical Approach and Identification

We now investigate the relationship between referral concentration within teams and spending. We pursue three identification strategies, beginning with a simple controlled regression. Baseline regressions take the following form:

$$\log y_i = \alpha ReferralCon_{-i} + \beta X_i + \gamma Z_i + \varepsilon_i$$

where y_i is patient *i's* spending[25], X_i is a set of patient characteristics and Z_i is a set of the assigned PCP's characteristics. $ReferralCon_{-i}$ denotes the team referral concentration of patient i's PCP. In the regression analyses, we use a jackknifed calculation of the PCP's referral concentration, $ReferralCon_{-i}$, that omits the contribution of the current patient *i* to the doctor's team referral concentration. The jackknifing procedure overcomes an important endogeneity threat: that a *patient's* own severe health status necessitates more unusual referrals, thus reducing the *physician's* team referral concentration and driving up the patient's own spending.

All regressions include a rich vector of patient and insurer controls including: patient sex, 5-knot splines for both age and HCC risk score, and patient ZIP code fixed effects to capture local heterogeneity in patient demand for care. Insurer controls include a fixed effect for each payer and an indicator for each of the 13 types of insurance plans defined by the APCD (i.e. Health Maintenance Organization [HMO], Preferred Provider Organization [PPO], Exclusive Provider Organization [EPO], etc.). We then augment

[24] Physician Contracting Network is defined by the Massachusetts Provider Database as "An organization of medical groups and/or practice sites with an integrated approach to quality improvement that enters into contracts with payers on behalf of its provider members." See Massachusetts Health Quality Partners (2016).

[25] We exclude any individuals with zero or negative net spending which can result from, for instance, reversed claims. All individuals in our sample must have had healthcare spending during the year in order to pass our chronic illness sample screen and to be assigned a PCP.

this baseline specification with a series of additional controls for patient and physician characteristics. Given the inclusion of these rich controls, the baseline specification is identified by variation in PCP referral concentration among patients who reside in the same ZIP code, and have similar health status, demographics, insurer, and insurance type.

There are two main threats to identification in these baseline controlled regression specifications. First, PCPs with varying referral concentration may also differ in their practice style along other dimensions. If, for example, physician taste for more intensive care correlates with low referral concentration, this could bias our estimates. To account for this possibility, we run additional specification checks that directly control for PCP fixed effects, exploiting differences in PCP referral concentration across different specialties. For example, if a PCP is more concentrated in her cardiology referrals than her endocrinology referrals, her cardiology patients should have relatively higher costs, compared to a peer physician with the inverse pattern. We describe this approach in more detail and report results in Section 5D below.

A second major threat to interpretation in the baseline specifications is the possibility that patients seeing low referral concentration physicians are in worse health. While observed patient characteristics such as age and HCC risk score do not suggest obvious differences in health status across low and high referral concentration PCPs, there could be differences in health that are not observed by the econometrician. To assess this possibility, we will analyze the experience of Medicare beneficiaries who change PCPs due to a move. The mover design allows us to control for patient fixed effects and exploit a plausibly exogenous change in patient referral concentration using an instrumental variable strategy. We describe the mover specifications in more detail and report results in Section 6C.

Together, these three strategies aim to identify the impact of PCP referral concentration on the costs of care, accounting for other differences in PCP practice style and the possibility of endogenous sorting of patients to PCPs.

C. Main results

Baseline results are in Table 2. Columns 1-3 run regressions in which the dependent variable is care utilization measured at standardized prices, while columns 4-6 use total spending as the dependent variable, combining both price and utilization effects. Column 7 parallels column 1, but is run on the 20% sample of Medicare beneficiaries. We discuss its interpretation in the next section, noting here that the Medicare magnitude is not directly comparable to the Massachusetts APCD results due to measurement error in the Medicare measure of team referral concentration.

In Table 2, columns 1 and 4, we report results with only the baseline controls for patient and insurer characteristics described above. The findings confirm the strong relationship between within team referral concentration and spending visually seen in Figure 3.

The estimated magnitude of team referral concentration's effect on utilization and spending is economically significant. To interpret magnitudes, it is helpful to remember that the above and below median average measures of team referral concentration differ by 0.1. Thus the coefficients in columns 1 and 4 mean that compared to similar patients seen by PCP's with average below median team referral concentration, patients seen by PCPs with average above median team referral concentration have 6.3% (=-0.626*0.1) lower medical care utilization and 11.6% (=-1.16*0.1) lower total spending. Alternatively, a 1 standard deviation increase in team referral concentration leads to 4.0% lower utilization and 7.4% lower spending.

These results indicate that patients of PCPs with higher team referral concentration use fewer services and also see lower priced providers. While we are not able to identify the precise mechanisms causing the negative association between team referral concentration and provider pricing, it is helpful to consider possible explanations. Our findings control for both insurer (e.g. Anthem, United, etc.) and plan type (e.g. HMO, PPO, etc.), so pricing variation due to differences in insurance plan breadth and quality are unlikely to be the primary driver of this result. Instead, the pricing effect suggests that PCPs with higher referral concentration tend to either have lower prices themselves or send patients to lower-priced specialists and hospitals.

Our model suggests that PCPs who put greater weight on containing the total costs of care will be inclined to both concentrate their referrals and recommend lower priced providers.[26] An alternative explanation for the price effect we observe is that locations with high demand for specialists relative to supply may have higher prices and also larger fixed costs of establishing specialist-PCP relationships as specialists may not have an incentive to establish new relationships with PCPs, making the setup costs higher for PCPs. These higher fixed costs in areas with short supply of specialists will lead to a negative correlation between team referral concentration and prices. If this alternative explanation is correct, then some of the association between team referral concentration and the price of services delivered may not be causal.

[26] A related explanation is that PCPs with higher team referral concentration are more likely to be in smaller (or less vertically integrated) practices with less market power and hence lower prices. PCPs in these low market power practices may be under more pressure to contain both prices and utilization to maintain their position in insurance networks. Notably, in the Massachusetts market, Partners HealthCare, which commands substantial market power due to its large scale and prestigious reputation, is not an outlier when we rank large healthcare systems by average PCP referral concentration. For details of this comparison, consult Appendix Table A1.

As a robustness check, we augment the regression with a control for patient care continuity HHI*;* results are in columns (2) and (5). The relationship between team referral concentration and spending attenuates somewhat in these specifications: compared to similar patients and holding fixed patient care continuity HHI, patients seen by PCPs with average above median team referral concentration have 3.7% lower utilization and 9.0% lower total spending than patients seen by PCPs with average below median team referral concentration.

If patient care continuity HHI simply captures exogenous and unobserved patient heterogeneity, the results in columns (2) and (5) ought to be preferred over those in columns (1) and (4). But enhanced care continuity (i.e. concentrating patient visits among a smaller set of providers) might also be part an endogenous response to the PCP practice choices that determine team referral concentration. This possibility complicates interpretation.

Suppose, for example, that PCPs with high team referral concentration are also generally more reluctant to refer to a specialist. In this case controlling for patient care continuity HHI biases our estimate of the cost effects of team referral concentration towards zero. In this scenario, a patient with a given care continuity HHI seeing a PCP with high team referral concentration would tend to be sicker than his counterpart seeing a PCP with a low team referral concentration. An alternative possibility is that care continuity may be the direct result of improved team coordination itself. This might happen if a PCP's investments in improving coordination with one specialist reduces the need to refer the patient to a different type of specialist (within the same specialty) to collect additional information, or improves clarity and agreement on which patients do not require referrals. To the extent that patient care continuity HHI is an endogenous response to PCP practice styles of the sort we modeled above, then the estimates in columns (2) and (5) are conservative tests of the hypothesis that team coordination as measured by team referral concentration is an important determinant of costs.

Finally, in columns (3) and (6), we add new controls for other dimensions of PCP heterogeneity, including the average HCC risk score of a PCP's commercially insured working age patients, a 5-knot spline in the number of working-age patients the PCP treats, and an indicator for whether the PCP's specialty is in Internal Medicine or Family Medicine.. The additional controls in this specification do not substantively change the magnitude of our results compared to the prior specification, providing reassuring evidence that variation in referral concentration is not reflecting major differences in the size of the physician's patient panel, training, or case mix.[27]

[27] Results are similar when we also include a fixed effect for PCP ZIP code (see Appendix Table A2). When we included PCP ZIP code, we identify the impact of referral concentration using variation within a physician's practice location. The fact that the results are similar suggests that the relationship between team referral concentration and spending is not driven solely by practice differences across organizations or practice sites.

D. Within-PCP Variation in Team Referral Concentration

A limitation of our approach so far is that we cannot distinguish the effects of team referral concentration from other, unobserved, dimensions of PCP practice style. To address this concern, we perform an additional analysis that includes PCP fixed effects and exploits differences in team referral concentration across specialties. For example, if a PCP is highly concentrated in her cardiology referrals but not in her endocrinology referrals, then we would expect superior coordination (and lower costs) in the former than the latter.

This approach to identifying the effects of care coordination is conservative, as it is identified only across specialties within PCP and so removes from consideration any efforts a PCP may have made in improving coordination with all specialists to whom she refers. The relationship between team referral concentration and spending may also be understated if team referral concentration in any one specialty effects a PCP's "bandwidth" for forming relationships with physicians in different specialties. Finally, estimates may be attenuated to the extent that within-PCP variance is driven by measurement error.

To estimate spending and utilization based on within PCP variation in team referral concentration, we restrict the sample to patients who saw at least one specialist in exactly one of our 5 specialty categories. (Unlike the base sample, this subsample excludes patients who saw no specialists or those who consulted more than one type of specialist.) Instead of using a PCP's average team referral concentration across all the specialties in our set, a patient is assigned the (jackknifed) team referral concentration of their PCP for the specialty in which they saw a specialist. The key independent variable of interest, team referral concentration, is now matched to the specific specialty the patient consulted.

Table 3 displays the regression results. Columns (1) and (4) include similar controls to those used in Table 2 column (1) to estimate the relationship between team referral concentration and utilization/spending in this specific subsample. The controls include patient characteristics (but not patient care continuity HHI) and insurer variables, but exclude PCP characteristics. Table 3 also introduces a new set of controls in all reported specifications: a set of indicator variable for the specialty consulted (e.g. cardiology, endocrinology, etc.). Note that the coefficient in Table 3 column (1) is similar in magnitude to the utilization results in Table 2.

In Table 3, columns (2) and (5), we add PCP fixed effects, exploiting within-PCP variation in team referral concentration across specialties. The results remain negative and statistically significant, but the magnitude is about one-half (utilization) to one-third (spending) the size of the effect reported in

columns (1) and (4).[28] This attenuation is not surprising because, as discussed above, within PCP estimates remove important sources of variation in team referral concentration and a larger component of the remaining variation may be due to measurement error. The persistence of a negative, statistically and economically significant effect after controlling for PCP fixed effects is suggestive that our main findings in Table 2 are not entirely the result of some unobserved, fixed PCP characteristic that is correlated with team referral concentration. These results suggest that unobserved differences in PCP quality or practice style (uniform across conditions the PCP treats) do not drive the estimated effect of team referral concentration.

E. Within Specialist Variation in Team Referral Concentration

Another potential concern with the results reported so far is that physicians with differing team referral concentration refer to specialists of differing quality. Perhaps PCPs with less concentrated referrals have higher costs because they are referring to "better" and therefore more expensive specialists. Alternatively, perhaps PCPs with less concentrated referrals have higher costs because they are still learning to identify the low cost specialists and will continue to experiment until they identify the best possible concentrated set of specialists. We address these concerns by estimating our set of regressions using fixed effects for the specific specialist the patient sees.

These regressions echo the sample used in the PCP fixed effect analysis. In particular, we restrict the sample to patients referred to at least one specialist in exactly one specialty. Similar to our plurality rule for PCP assignment, we then assign each patient to their plurality specialist within the specialty consulted. The regressions then include a fixed effect for the identity of the patient's plurality specialist; the independent variable of interest is the PCP's team referral concentration for the relevant specialty. These regressions effectively compare patients who share the same specialist, but who are referred by different PCPs with different levels of team referral concentration for that specialty.

Results of these regressions with specialist fixed effects are in Table 3, columns (3) and (6). We continue to find that PCPs with higher team referral concentration have significantly lower levels of utilization, even when their patients are referred to identical specialists. The magnitude suggests that moving from a below to above median team referral concentration PCP (a change of 0.1) is associated with a 1.9% reduction in utilization, significant at the 1% level. This effect is similar in magnitude (slightly larger) than the estimates that included PCP fixed effects, and is about two-thirds the size of the effect estimated in this sample without any physician fixed effects (cf. Table 3, column [1]). These findings provider further support for the notion that team relationships between PCPs and specialists promote lower cost care.

[28] To interpret the magnitude of the PCP fixed effect specification, compare a specialty in which a PCP refers equally to 5 versus 10 specialists (a difference in team referral concentration of 0.10); this increase in team referral concentration is associated with 1.4% lower care utilization.

6. Team Referral Concentration in Medicare

Our results so far have focused on the Massachusetts All Payer Claims Data. This data offers remarkable breadth for measuring referral networks precisely at the physician level but is also limited in two important ways. First, the APCD is limited to a single state, whose healthcare institutions may not be nationally representative. Secondly, our extract of the APCD data is essentially a cross-sectional data set; even adding more recent years would make for a very short panel. As a result, we cannot use the Massachusetts data to estimate a model with patient fixed effects. We address both these deficiencies by analyzing team referral concentration in a national sample of Medicare beneficiaries.

A. Measurement error in Medicare

A natural way to begin our analysis of Medicare beneficiaries would be to replicate the analysis we ran using the APCD data. Unfortunately, the Medicare data is only a 20% sample of Medicare fee for service enrollees over the age of 65. These restrictions are such that we only observe a small fraction of the total patients each doctor sees, which creates acute measurement error problems for network measures like our team referral concentration variable.

Consider a PCP who has 5 patients, each referred to a different specialist within a single specialty. The PCP should have a team referral concentration of 0.2. However, if we only observe 1 out of the 5 patients in the data, we will measure a team referral concentration of 1 for that PCP. Similarly, a PCP who refers all her patients to the same specialist will also have a referral concentration of 1. Measurement error in this setting differs from classical measurement error; our noisy estimates of team referral concentration are biased upwards and the size of the error is correlated with the underlying true referral concentration. In this section, we demonstrate the empirical impact of measurement error on the baseline specifications. In Section 6C, we discuss the theoretical derivation of the measurement error in our instrumental variable estimates .

To illustrate the impact of measurement error on our results, we run a series of simulations in the Massachusetts APCD. In these simulations, we draw subsamples of patients from the APCD and then estimate team referral concentration from just these patient subsamples. For example, we compare results from the full 100% Massachusetts APCD sample to results from a randomly drawn 20% Massachusetts APCD sample.

We draw a series of subsamples of patients in the APCD, using percentage samples that range from # = {10%, 20%, …, 100%} of the full sample of patients. For each percentage subsample, we repeat 50

random draws to account for sampling error. Within each subsample, we construct $ReferralCon_i^{\#\%\ sample}$ and estimate a regression of the form:[29]

$$\log y_i = \alpha_{\#} ReferralCon_i^{\#\%\ sample} + \beta X_i + \gamma Z_i + \varepsilon_i$$

We then calculate the multiplier $\lambda_{\#} = \alpha_{100}/\alpha_{\#}$ that tell us how to scale estimated coefficients on team referral concentration for each #% subsample. The scaling factor $\lambda_{\#}$ will depend on the set of controls used.

Because Medicare is different from the APCD in many ways, a range of multipliers need to be considered; a 20% sample of the APCD data may not have an identical multiplier as a 20% sample of the Medicare data. In Figure 4, we plot the results of how the multiplier $\lambda_{\#}$ depends on the size of the subsample used to measure team referral concentration for regression specification from Table 2, Column 1. Figure 4 illustrates that the magnitude of the attenuation bias falls with the size of the patient sample. Further, the multipliers for specification 1 are modest: if we observed 20% of the APCD sample, we would want to multiply our estimate by about 2 to gauge impact of team referral concentration as measured from the 100% sample.

However, adding additional controls to the regression greatly exacerbates the measurement error problem. The control variables are correlated with the "signal" in our estimated referral concentration, and remaining variation in referral concentration has a proportionally larger "noise" component. Appendix Figure A1 shows that attenuation bias is much more severe for regression specification 2, which adds a control for patient care continuity HHI. For a 20% sample, specification 2 has a multiplier of about 8, with a 95% CI from 4 to 16. At a 10% sample, the problem is severe: the estimated mean multiplier is 48 with a 95% CI that includes -49 to 225, implying we may not even estimate the correct direction of the effect. We also find large attenuation bias in specification 3 (which adds PCP controls in addition to the patient care continuity HHI), though not as severe as specification 2.

We therefore conclude that specifications 2 and 3 are uninformative in the Medicare data, but that specification 1 may shed useful light on the generalizability of our Massachusetts results. In recognition of these uncertainties, we report unadjusted coefficients, and then assess how imposing the APCD scaling factor to the national 20% Medicare sample would influence the interpretation of these results.

[29] Precisely, we estimate the regression using the value of $ReferralCon_i^{\#\%\ sample}$ for each doctor constructed from the subsample of patients. We then run the regression using all the patients, but with the noisily measured $ReferralCon_i^{\#\%\ sample}$. Using the full sample of patients in the regression should not affect the expected multiplier, but it does reduce sampling variation in measuring the multiplier.

B. Replicating Results in Medicare

Summary statistics on the Medicare sample are reported in Appendix Table A3. Similar to the Massachusetts findings, summary statistics demonstrate that patient age, sex, and disease burden are similar across patients seeing PCPs with above and below median team referral concentration.

Table 2, column 7 shows the regression results for specification 1 in Medicare.[30] Note that we only estimate utilization equations, not spending. This is because prices are administratively set in Medicare and primarily adjusted only for geographic location. The geographic variation will be largely eliminated by the patient ZIP code fixed effects we include as controls.

The estimated effect is quite substantial. An increase in measured team referral concentration of 0.1 is associated with a 2.9% decline in utilization. This analogous coefficient estimated in the parallel specification on Massachusetts data (Table 2, column 1) was about twice as large, consistent with the magnitude of attenuation predicted in our measurement error simulations.

Another advantage of the Medicare sample is that bills are easily decomposed into three categories: provider submitted claims from the Carrier files, Inpatient claims submitted by hospitals, and Outpatient claims for hospital-based outpatient care. Results from this decomposition exercise are reported in Appendix Table A5. Higher team referral concentration is associated with statistically significant reductions in all three types of billings in the least noisy specification 1 (without controls for patient care continuity HHI or PCP characteristics). Patients of PCPs with higher referral concentration are slightly less likely to have an inpatient stay.

C. Medicare movers and reversion to the mean identification strategy

We have documented a positive relationship between team referral concentration and costs in two very different patient populations: chronically ill, commercially insured working age patients in the Massachusetts APCD and elderly patients in Medicare. In both cases, our results relied on cross-sectional regressions and so there was still scope for selection on unobservable patient characteristics to bias the findings.

To address this possibility, in this section we identify the effect of team referral concentration from the experience of Medicare patients who change their PCP as a result of a move. This approach builds on work by Laird and Nielsen (2016); Agha, Frandsen, and Rebitzer (2017); and Finkelstein, Gentzkow, and Williams (2016). The technique allows the inclusion of a patient fixed effect to control for differences in

[30] Appendix Table A4 reports regression results for specifications 2 and 3 as well, though our measurement error exercise suggests these specifications are not reliable. We find statistically significant negative coefficients on team referral concentration in both specifications 2 and 3, though the magnitude of the effect is small in specification 3.

patient demand for care that are stable over time. Moreover, we also use an instrument for the change in referral concentration to address the possibility that the choice of new PCP is endogenous to a change in health status.

For the movers analysis, we expand the Medicare sample to include data from 2007-2012.[31] We calculate PCP team referral concentration on an annual basis using the full 20% Medicare sample (not restricted to physicians who treat movers). We restrict the analysis sample to enrollees who move to a new hospital referral region over this period.[32]

The first way to measure the response of utilization to change in team referral concentration is with a difference-in-difference strategy: we add a patient fixed effect, β_i, to our estimation equation, and include $\Delta ReferralCon_{-i,d}$, which measures the change in the jackknifed referral concentration of the patient's post-move PCP compared to the patient's pre-move PCP.[33] We interact $\Delta ReferralCon_{-i}$ with $Post_{it}$, an indicator variable for being in the post-move period. We include a vector of fixed effects $\rho_{R_{i,t}}$ for the event year relative to the move (denoted $R_{i,t}$), allowing movers' annual demand for care to depend on the timing of their move. (For example, year $R_{i,t}$= -1 corresponds to the year before the move, year 0 for the year of the move, etc., and we include indicator variables for each year in event time.) Further, we include characteristics of the patient's plurality PCP and year fixed effects in the control vector Z_{it}. We then get the following difference-in-differences equation:

$$\log y_{it} = \alpha \Delta ReferralCon_{-i} Post_{it} + \beta_i + \gamma Z_{it} + \rho_{R_{i,t}} + \varepsilon_{it}$$

Unlike the earlier specifications, we omit controls for patient comorbidities because the patient fixed effect should account for fixed differences in patient health over time. Patient comorbidities can change over time, and in principle could be tracked using this data. We do not include such controls in this regression because evidence suggests that there are regional differences in comorbidity coding (Song et al. 2010, Finkelstein et al. 2017), which could be endogenously related to changes in team referral concentration.

[31] We limit this movers analysis to the Medicare sample, since the Massachusetts APCD is not ideal for a number of reasons: there are not many regions in Massachusetts, our panel is short, and the APCD data do not let us accurately track a patient who changes insurers or employers.

[32] The sample restricts to patients with exactly one move over this period, and requires that at least 75% of a patient's claims are in the hospital referral region that corresponds to their listed address zip code in each year (excluding the year of the move).

[33] Note that patients are assigned to PCPs using our plurality assignment rule on an annual basis, allowing for patients to switch PCPs across years, even in the absence of a move. Pre-move team referral concentration is calculated as the average level of PCP referral concentration over the year(s) prior to the move. Similarly, post-move team referral concentration is calculated as the average level of PCP referral concentration over the year(s) after the move. (Note the year of the move is excluded from both calculations.) The change in PCP team referral concentration is the difference of post- and pre-move average team referral concentrations.

Results from the difference-in-differences specification are in Panel A of Table 4. The role of measurement error in this specification is derived in Appendix D. The baseline specification estimates an increase in team referral concentration of 0.1 is associated with a 4.3% decrease in care utilization, significant at the 1% level. However, the difference-in-differences framework faces an identification threat: $\Delta ReferralCon_{-i}$ may be endogenous to real changes in patients' clinical conditions. If movers with deteriorating health selected new PCPs (after moving) who relied on more diverse specialists, our estimate of parameter α would be biased.

Analogous to the approach of Laird and Nielsen (2016), we apply an instrumental variables strategy that exploits mean reversion in PCP team referral concentration to identify exogenous variation in PCP team referral concentration. The patient's PCP team referral concentration in the pre-move period is used as an instrumental variable for the *change* in the patient's PCP team referral concentration, $\Delta ReferralCon_{-i}$.

The first stage equation proceeds as follows:

$$\Delta ReferralCon_{-i}Post_{it} = \tilde{\alpha} PreMoveReferralCon_{-i}Post_{it} + \tilde{\beta}_i + \tilde{\gamma} Z_{it} + \tilde{\rho}_{R_{it}} + \widetilde{\varepsilon_{it}}$$

And the reduced form is given as:

$$\log y_{it} = \alpha PreMoveReferralCon_{-i}Post_{it} + \beta_i + \gamma Z_{it} + \rho_{R_{it}} + \varepsilon_{it}$$

These regressions measure the causal effect of team referral concentration on utilization under two key assumptions. First, the patient's initial PCP team referral concentration must be uncorrelated with future *changes* in the patient's demand for care after the move. For example, if poor underlying health status is correlated with the initial PCP's low level of referral concentration, and health status mean reverts, then this identification approach will overstate the relationship between team referral concentration and spending. Conversely, if patients who initially sorted to PCPs with low team referral concentration are on deteriorating health trends relative to other movers, then this instrumental variables approach will understate the relationship between team referral concentration and spending. Second, as with the baseline set of regression results, the PCP's referral concentration must be independent of other dimensions of PCP practice style that might influence care utilization.

Note that the instrumental variable approach may still suffer bias from measurement error due to the correlation in the error of the instrument, $PreMoveReferralCon_{-i}Post_{it}$, and the error in the endogenous variable $\Delta ReferralCon_{-i}Post_{it}$.

Table 4 reports first stage (Panel B) and two-stage least squares (Panel C) results of these instrumental variable Medicare utilization regressions on the movers sample. The specification in column 1 includes baseline controls for patient fixed effects, year fixed effects, and a series of indicators for event year relative to the move. In column 2, we add in a time varying control for the patient care continuity HHI. Finally, in column 3, we add additional controls for the PCP's specialty, sex, and patient volume.

The first stage estimates confirm the predicted mean reversion pattern. Patients treated by PCPs with high team referral concentration prior to their move experience relative reductions in PCP team referral concentration after their move, compared to patients initially treated by PCPs with low team referral concentration.

The two stage least squares estimates of the impact of team referral concentration on utilization are consistently larger than the cross-sectional estimates in Medicare data. This pattern is consistent with a reduced degree of attenuation bias due to measurement error of PCP referral concentration using the instrumental variables strategy; it may also reflect a larger true effect size due to the movers being an older sample of patients with a higher prevalence of heart conditions than the 2012 Medicare 20% sample (Appendix Table A3). In Table 4, column 1, an increase in team referral concentration of 0.1 is associated with a 5.8% reduction in utilization. Adding control variables for patient care continuity HHI and PCP characteristics reduces the estimated effect size only slightly, so that a 0.1 increase in team referral concentration is associated with a 4.7% reduction in utilization. These findings provide further evidence that unobserved patient characteristics associated with the PCP's team referral concentration are not driving our main results.

7. Conclusion

Teams are pervasive in economic organizations so the performance of teams matters a great deal for organizational efficiency. Researchers, however, know very little about how the structure of teams influences economic performance. We address this issue with an application to healthcare by examining the teams that primary care physicians (PCPs) assemble when they refer patients to specialists. Our theoretical model analyzes how PCPs trade off costly coordination against beneficial specialization and finds that team coordination improves when PCPs concentrate their referrals within a small set of specialists. Empirically, we find that patients of PCPs who concentrate their referrals have lower healthcare costs. This effect exists for both commercially insured and Medicare populations; is statistically and economically significant; and holds under various identification strategies that account for unobserved patient and physician characteristics.

More specifically, for commercially insured working-age patients in Massachusetts, those treated by PCPs with below median team referral concentration have 6.3% lower utilization, compared to those treated by above median PCPs after controlling for detailed patient and insurer characteristics. Smaller effect sizes are estimated using PCP fixed effects estimates that rely only on within-PCP variation in team referral concentration across different specialties. For Medicare beneficiaries, we study those who switch doctors as the result of a move across regions. Using both a difference-in-difference analysis with patient fixed effects and a regression to the mean instrumental variables strategy, we find that an increase in team referral concentration is also associated with lowered utilization.

Our analysis has a number of limitations that may inspire future research. One important limitation is that we do not investigate the link between team referral concentration and quality of care due to the challenges of measuring care quality in insurance claims data. Physician fixed effect specifications suggest, however that differences in unobserved PCP or specialist quality are not a major confounder. Nevertheless, the cost effects we estimate suggest a bound on the value of any associated quality change needed to offset the additional care utilization associated with lower team referral concentration.

Our empirical research also gives only limited insights into how concentrated team referrals lead to reduced utilization. Our theoretical model points towards heightened relationship specific investments that improve care coordination, but we do not directly observe the coordination effort or investments of PCPs.

Another unanswered question is whether the gains from concentrated team referrals are best realized within firms. If so, integration of PCP practices with specialty physicians may lead to more concentrated referral patterns or lower coordination costs for referrals within the integrated firm.[34] If present, these coordination benefits may partially offset the higher prices of integrated practices.[35]

Taken literally, our results suggest that insurance companies, providers or policymakers seeking to reduce costs and improve care coordination may want to encourage concentrated referral networks. Narrow insurance networks, for example, may promote team referral concentration by limiting the set of in-network specialists, although they could also disrupt existing investments in PCP-specialist relationships. Certain cost containment incentives, such as those found in Accountable Care Organizations, patient centered medical homes, or physician payment adjustments by the Medicare Merit-based Incentive Payment System may potentially induce formation of more concentrated referral networks. The potential for policy instruments or management strategies to shape referral networks

[34] Baker et al. (2016) find that hospital ownership of physician practices leads physicians to increase patient admissions to the owning hospital, possibly increasing concentration of admissions.
[35] See Baker et al. (2014) for estimates of the relationship between vertical integration and pricing.

and the downstream consequences of any such changes for costs and quality is a promising avenue for future work.

We conclude by observing that our measure of team referral concentration may be applied to examine team relationships in settings beyond healthcare. Teams in many other sectors can be designed in ways that enhance or reduce repeat interactions between team members with specialized knowledge. Design choices that facilitate repeat interactions likely enhance coordination at the cost of reducing the match value from drawing from a larger pool. The magnitude of this tradeoff and its consequences for optimal team design and performance, however, likely vary. There is much still to be learned about how the structure of teams influences economic performance.

8. References

Agha, Leila, Brigham Frandsen and James B. Rebitzer (2017). "Causes and Consequences of Fragmented Care Delivery: Theory, Evidence, and Public Policy." *NBER Working Paper 23078.*

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek. 2008. "When Does Coordination Require Centralization?" *American Economic Review* 98 (1): 145–79. doi:10.1257/aer.98.1.145.

Baker, Laurence C., M. Kate Bundorf, and Daniel P. Kessler. 2014. "Vertical Integration: Hospital Ownership Of Physician Practices Is Associated With Higher Prices And Spending." *Health Affairs* 33(5): 756-763.

Baker, Laurence C., M. Kate Bundorf, and Daniel P. Kessler. 2016. "The effect of hospital/physician integration on hospital choice." *Journal of Labor Economics* 50: 1-8.

Barnett, Michael L., Nicholas A. Christakis, A. James O'Malley, Jukka-Pekka Onnela, Nancy L. Keating, and Bruce E. Landon. 2012a. "Physician Patient-Sharing Networks and the Cost and Intensity of Care in US Hospitals." *Medical Care* 50 (2):152–60. https://doi.org/10.1097/MLR.0b013e31822dcef7.

Barnett, Michael L., Nancy L. Keating, Nicholas A. Christakis, A. James O'Malley, and Bruce E. Landon. 2012a. "Reasons for Choice of Referral Physician Among Primary Care and Specialist Physicians." *Journal of General Internal Medicine* 27 (5):506–12. https://doi.org/10.1007/s11606-011-1861-z.

Barton H. Hamilton, Jack A. Nickerson, and Hideo Owan. 2003. "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy* 111(3): 465-497.

Baumgardner, James R. 1988. "Physicians' Services and the Division of Labor Across Local Markets." *Journal of Political Economy* 96 (5): 948–82. doi:10.2307/1837242.

Becker, Gary S. and Kevin M. Murphy. 1992. "The Division of Labor, Coordination Costs, and Knowledge" *Quarterly Journal of Economics*, 107(4): 1137-1160.

Bloom, Nicholas, and John Van Reenen. 2011. "Chapter 19 - Human Resource Management and Productivity." In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 4:1697–1767. Elsevier. http://www.sciencedirect.com/science/article/pii/S0169721811024178.

Bodenheimer, Thomas, Bernard Lo, Lawrence Casalino. 1999. “Primary Care Physicians Should Be Coordinators, Not Gatekeepers.” *Journal of the American Medical Association.* 281(21): 2045–2049. doi:10.1001/jama.281.21.2045

Cebul, Randall D., James B. Rebitzer, Lowell J. Taylor, and Mark E. Votruba. 2008. "Organizational Fragmentation and Care Quality in the U.S. Healthcare System" *Journal of Economic Perspectives*, 22(4): 93-114.

Chillemi, Ottorino, and Benedetto Gui. "Team Human Capital and Worker Mobility." *Journal of Labor Economics* 15, no. 4 (1997): 567-85. doi:10.1086/209838.

Crawford, Vincent P. 1990. “Relationship-Specific Investment.” *Quarterly Journal of Economics.* 105(2): 561-574.

Ericson, Keith and Amanda Starc. 2015. " "Measuring Consumer Valuation of Limited Provider Networks." *American Economic Review*, 105(5): 115-19.

Feri, Francesco, Bernd Irlenbusch, and Matthias Sutter. 2010. "Efficiency Gains from Team-Based Coordination—Large-Scale Experimental Evidence." *American Economic Review*, 100(4): 1892-1912.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. “Sources of Geographic Variation in Health Care: Evidence From Patient Migration.” *Quarterly Journal of Economics,* 131 (4): 1681–1726. doi:10.1093/qje/qjw023.

Finkelstein, Amy, Matthew Gentzkow, Peter Hull, and Heidi Williams. 2017. “Adjusting Risk Adjustment – Accounting for Variation in Diagnostic Intensity” *New England Journal of Medicine,* 376 (6): 608-610.

Frandsen, Brigham R., Karen E. Joynt, James B. Rebitzer, and Ashish K. Jha. 2015. "Care Fragmentation, Quality, and Costs among Chronically Ill Patients" *American Journal of Managed Care*, 21(5): 355-326.

Frandsen, Brigham and James B. Rebitzer. 2014. "Structuring Incentives within Accountable Care Organizations" *Journal of Law, Economics, and Organization,* 31(S1): i77-i103.

Garicano, Luis, and Thomas N. Hubbard. 2009. “Specialization, Firms, and Markets: The Division of Labor within and between Law Firms.” *Journal of Law, Economics, and Organization,* 25 (2): 339–71. doi:10.1093/jleo/ewn003.

Geissler, Kimberley, Ben Lubin, and Keith M Ericson. (2017). “The role of organizational affiliations in physician referral patterns.” Working Paper.

Grosse, Stefan, Louis Putterman, and Bettina Rockenbach. 2011. “Monitoring in Teams: Using Laboratory Experiments to Study a Theory of the Firm.” *Journal of the European Economic Association* 9 (4):785–816. https://doi.org/10.1111/j.1542-4774.2011.01026.x.

Hart, Oliver. 2017. "Incomplete Contracts and Control." *American Economic Review*, 107(7): 1731-52.

Holmström, Bengt. 1982. “Moral hazard in teams.” *Bell Journal of Economics*, 13(2): 324-340.

Huckman, Robert S., Bradley R. Staats, David M. Upton. 2009. “Team Familiarity, Role Experience, and Performance: Evidence from Indian Software Services.” *Management Science* 55(1): 85-100.

Hussey, Peter S., Eric C. Schneider, Robert S. Rudin, Steven Fox, Julie Lai, and Craig E. Pollack. 2014. "Continuity and the Costs of Care for Chronic Disease" *JAMA Internal Medicine*, 174(5): 742-748.

Jackson, Matthew O., Brian W. Rogers, and Yves Zenou. 2017. “The Economic Consequences of Social-Network Structure.” *Journal of Economic Literature* 55 (1):49–95. https://doi.org/10.1257/jel.20150694.

Kautter, John et al. 2014. “The HHS-HCC Risk Adjustment Model for Individual and Small Group Markets under the Affordable Care Act.” *Medicare & Medicaid Research & Review*. DOI: http://dx.doi.org/10.5600/mmrr.004.03.a03.

Kolber, Morey J. 2006. “Stark Regulation: A Historical and Current Review of the Self-Referral Laws.” *HEC Forum: An Interdisciplinary Journal on Hospitals’ Ethical and Legal Issues* 18 (1):61–84.

Kozlowski SWJ, Ilgen DR. 2006. “Enhancing the effectiveness of work groups and teams.” *Psychological Science in the Public Interest*, 7(3):77–124. http://dx.doi.org/10.1111/j.1529-1006.2006.00030.x.

Laird, Jessica and Torben Nielsen. 2016. “The Effects of Physician Prescribing Behaviors on Prescription Drug Use and Labor Supply: Evidence from Movers in Denmark.” Harvard University Working Paper.

Mailath, George J., Andrew Postlewaite. 1990. “Workers Versus Firms: Bargaining Over a Firm's Value.” *The Review of Economic Studies*, 57 (3):369-380.

Marschak, J and R. Radner. 1972. *Economic Theory of Teams* Yale University Press, New Haven, CT.

Massachusetts Health Quality Partners. 2016. “MHQP 2016 Massachusetts Provider Database (MPD)”. Online: http://www.mhqp.org/products_and_tools/?content_item_id=226

McWilliams, J. Michael. 2016. “Cost Containment and the Tale of Care Coordination.” *New England Journal of Medicine,* 375 (23): 2218–20. doi:10.1056/NEJMp1610821.

McWilliams, J. Michael, Michael E. Chernew, and Bruce E. Landon. 2017. “Medicare ACO Program Savings Not Tied To Preventable Hospitalizations Or Concentrated Among High-Risk Patients.” Health Affairs, 36 (12): 2085-2093. doi.org/10.1377/hlthaff.2017.0814.

Meltzer, David. 2001. "Hospitalists and the Doctor Patient Relationships" *Journal of Legal Studies*, 30: 589.

Meltzer, David O. and Jeanette W. Chung. 2010. "Coordination, Switching Costs and the Division of Labor in General Medicine: An Economic Explanation for the Emergence of Hospitalists in the United States" NBER Working Paper #16040.

Milstein, Arnold and Elizabeth Gibertson. 2009. "American Medical Home Runs" *Health Affairs*, 28(5): 1317-1326.

Pham HH, O'Malley AS, Bach PB, Saiontz-Martinez C, Schrag D. 2009. “Primary Care Physicians’ Links to Other Physicians Through Medicare Patients: The Scope of Care Coordination.” *Annals of Internal Medicine*. 17;150(4):236-42. PMID: 19221375.

Pollack, Craig E., Peter S. Hussey, Robert S. Rudin, Steven D. Fox, Julie Lai, and Eric C. Schneider. 2016. "Measuring Care Continuity: A Comparison of Claims-Based Methods" *Medical Care*. 54(4):e30-4.

Press, Matthew J. (2014) “Instant Replay—A Quarterback’s View of Care Coordination.” *New England Journal of Medicine.* 371: 489-491.

Reagans, Ray, Linda Argote, and Daria Brooks. 2005. “Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together.” *Management Science* 51 (6):869–81. https://doi.org/10.1287/mnsc.1050.0366.

Rebitzer, James B. and Mark E. Votruba. 2011. "Organizational Economics and Physician Practices." NBER Working Paper #17535.

Romano, M. J., J. B. Segal, and C. E. Pollack. 2015. "The Association between Continuity of Care and the Overuse of Medical Procedures." *JAMA Internal Medicine*. 175(7): 1148-54.

Samer Faraj, Lee Sproull. 2000. "Coordinating Expertise in Software Development Teams." *Management Science* 46(12):1554-1568. https://doi.org/10.1287/mnsc.46.12.1554.12072

Seltz, David, David Auerbach, Kate Mills, Marian Wrobel, and Aaron Pervin. (2016) "Addressing Price Variation on Massachusetts." *Health Affairs Blog*. Online: http://healthaffairs.org/blog/2016/05/12/addressing-price-variation-in-massachusetts/

Small, K. A. and Rosen, H. S. (1981) "Applied Welfare Economics with Discrete Choice Models." *Econometrica*. 49, pp. 105-130.

Song, Yunjie, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E. Wennberg, and Elliott S. Fisher. 2010. "Regional Variations in Diagnostic Practices". *New England Journal of Medicine,* 363:45-53. DOI: 10.1056/NEJMsa0910881

Stille, Christopher J., Anthony Jerant, Douglas Bell, David Meltzer, Joann G. Elmore. 2005. "Coordinating Care across Diseases, Settings, and Clinicians: A Key Role for the Generalist in Practice." *Annals of Internal Medicine*, 142:700–708. doi: 10.7326/0003-4819-142-8-200504190-00038

Weber, Roberto A. 2006. "Managing Growth to Achieve Efficient Coordination in Large Groups." American Economic Review, 96(1): 114–26

Williamson, O. 1985. *The Economic Institutions of Capitalism*, New York: Free Press.

Figures and Tables

Figure 1. Patient-Level Measures of Care Coordination versus Team Referral Concentration.

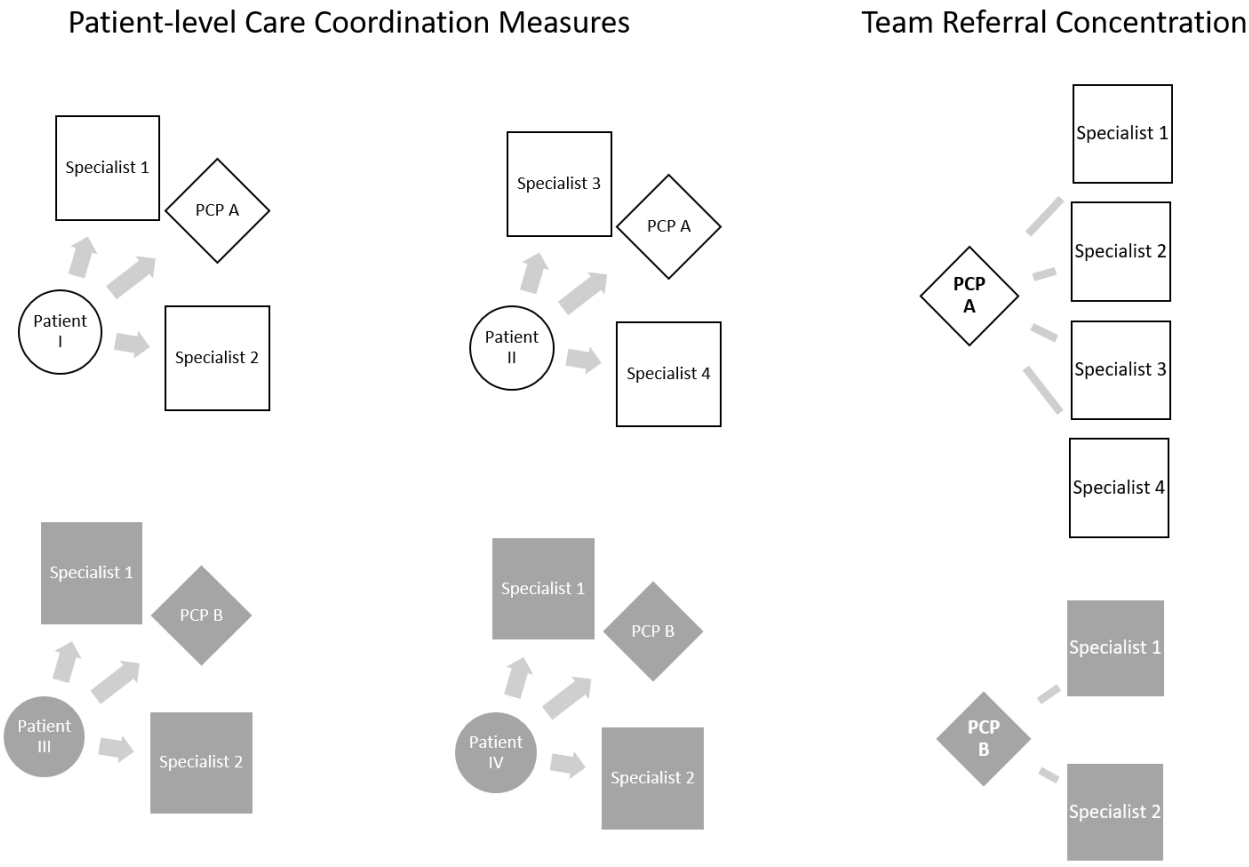
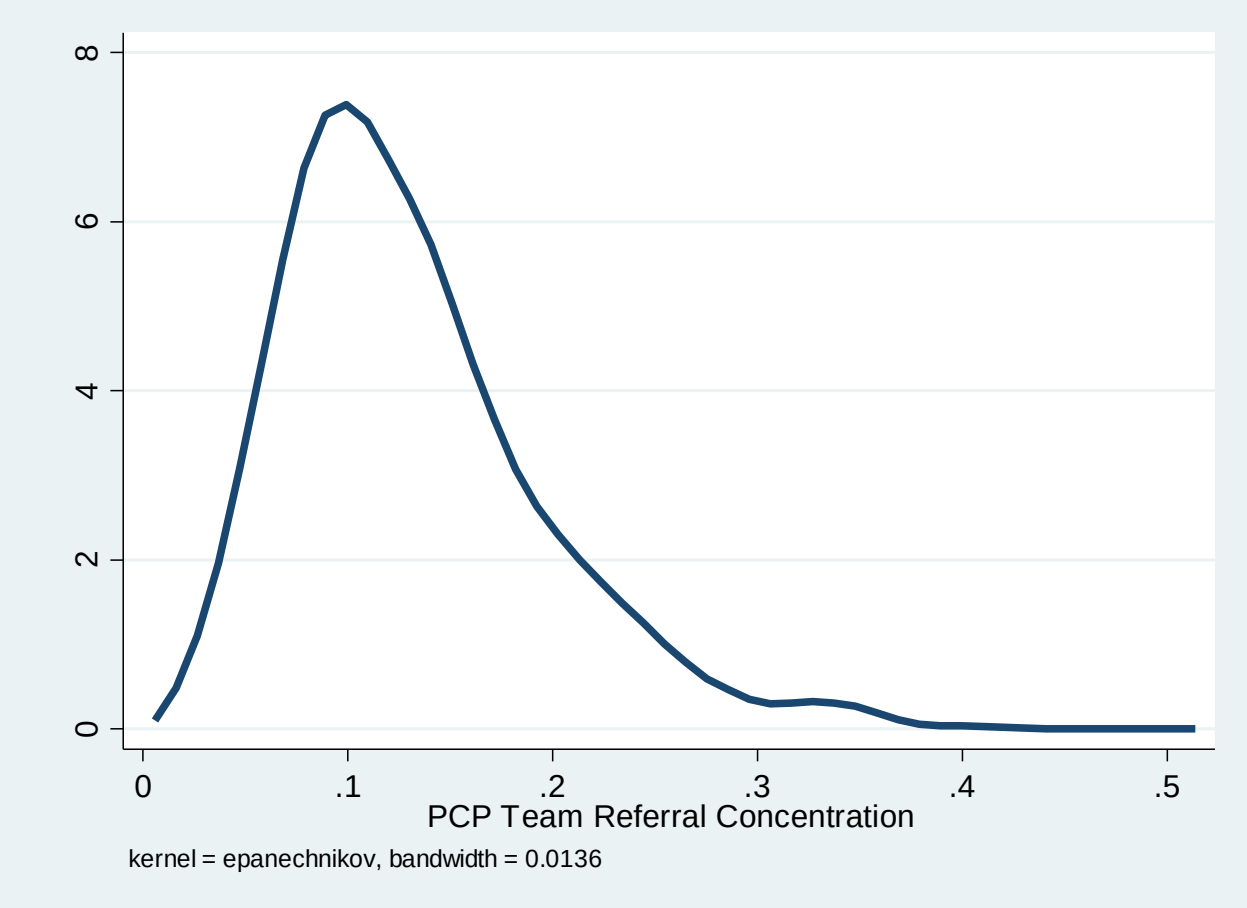
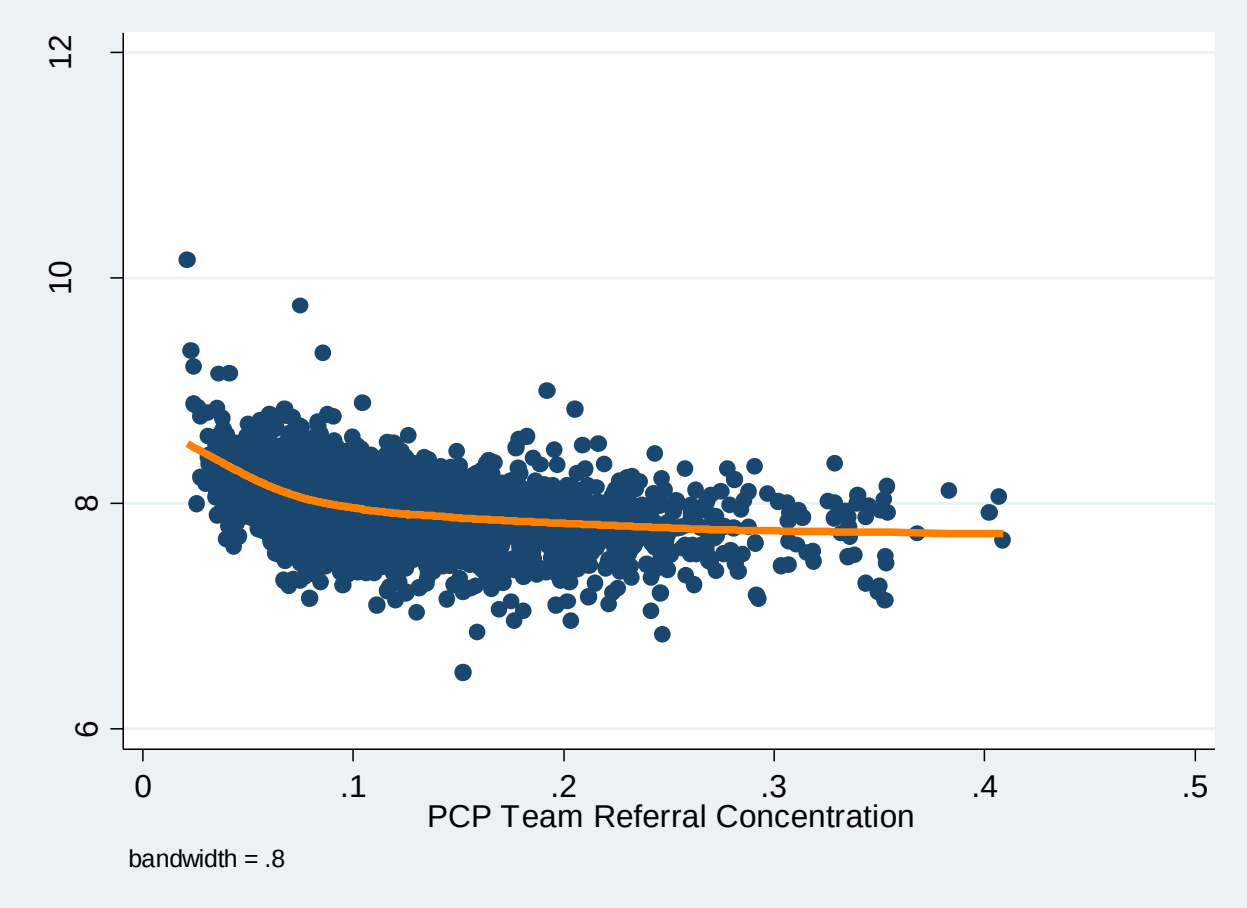


Figure 2. Distribution of PCP Team Referral Concentration.



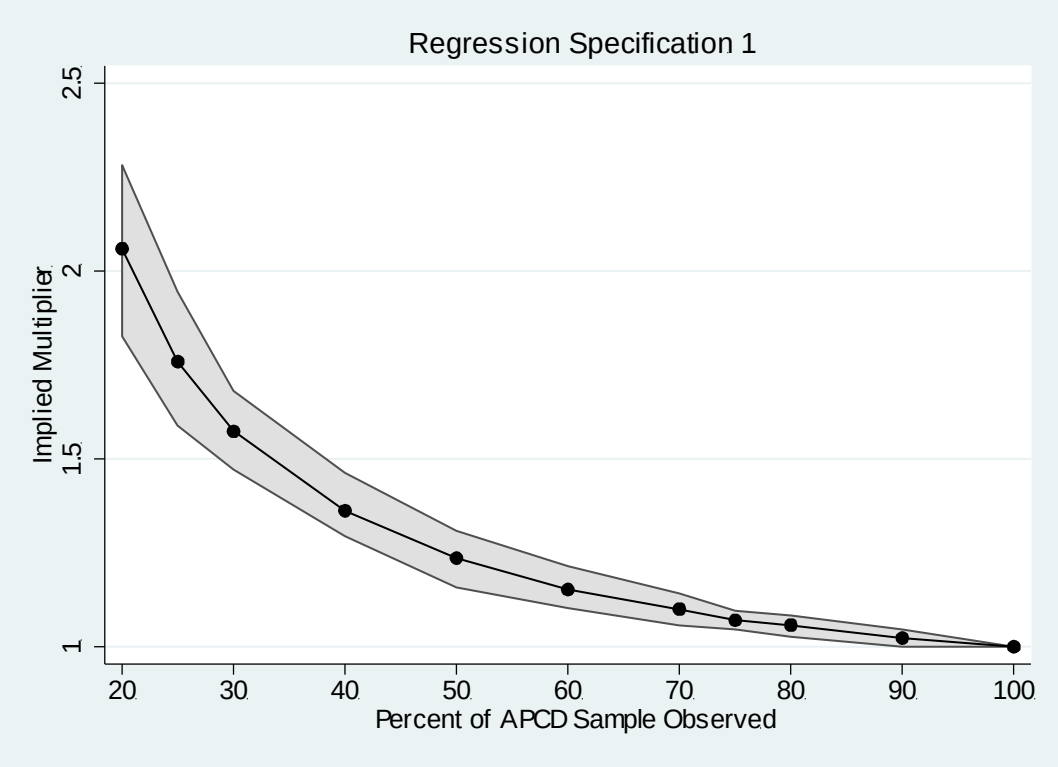
Notes: Data from Massachusetts All Payer Claims Data chronic illness analysis sample. One observation per patient.

Figure 3: Scatterplot of PCP Team Referral Concentration and Patient Spending



Notes: Unit of observation is PCP. Scatterplot with lowess smoothed line. Sample: MA APCD chronic illness analysis sample. Sample restricted to PCPs with at least 50 chronically ill patients in analysis sample.

Figure 4. Measurement Error Multiplier Simulations Using Subsamples of APCD Data



Note: Plots mean multiplier. Shaded area shows 5th, and 95th percentiles bootstrapped from 50 random samples per percent subsample. Follows the regression specification for Column 1 of Table 2.

Table 1. Descriptive Statistics of the Massachusetts APCD Analysis Sample

	Below Median Team Referral Concentration	Above Median Team Referral Concentration
Patient characteristics:		
Mean Spending ($)	8155	6407
Median Spending ($)	2952	2450
Age	48.9	49.1
Male	0.49	0.49
Pr(Any Inpatient Admission)	0.08	0.06
Patient care continuity HHI	0.44	0.47
Pr(Diabetes)	0.11	0.12
Pr(Heart Condition)	0.14	0.12
Pr(Bipolar and Major Depressive)	0.07	0.07
Pr(Asthma)	0.10	0.10
PCP characteristics:		
PCP's Team Referral Concentration	0.08	0.18
Pr(Internal Medicine)	0.75	0.71
Fraction Capitated Encounters	0.04	0.08
Fraction HMO Patients	0.60	0.62
PCP is Male	0.61	0.60
N Patients	157,360	157,318

Notes: Data from Massachusetts All Payer Claims Data, chronically ill analysis sample. Conditions are defined by HCC codes. Diabetes: 15-20. Heart Condition: 79-88, 92-93. Bipolar and Major Depressive: 55. Asthma: 110. The columns represent mean values for patients whose PCP has levels of team referral concentration that are respectively below or above the median. PCP characteristics are weighted by number of assigned patients.

Table 2. Referral Concentration and Healthcare Utilization and Spending

	Massachusetts All Payer Claims Data						Medicare 20% Sample
	Dependent variable: ln(Utilization)			Dependent variable: ln(Spending)			Dependent variable: ln(Utilization)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PCP Team Referral Concentration	-0.626***	-0.369***	-0.307***	-1.160***	-0.900***	-0.869***	-0.286***
	(0.056)	(0.055)	(0.056)	(0.069)	(0.069)	(0.070)	(0.011)
Patient Care Continuity HHI	No	Yes	Yes	No	Yes	Yes	No
PCP controls	No	No	Yes	No	No	Yes	No
Patient controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Insurer controls	Yes	Yes	Yes	Yes	Yes	Yes	NA
N patients	314,678	314,678	314,678	314,678	314,678	314,678	1,848,071

Notes: PCP's team referral concentration is jackknifed (a patient's own visits are excluded from the calculation for their PCP). Standard errors are clustered at the PCP level. Utilization is measured using standardized prices as described in the text. Insurer controls are not applicable for Medicare; for patients in Massachusetts they are a fixed effect for each payer and a fixed effect for each of 13 types of insurance plan (i.e. HMO, PPO, EPO, indemnity, etc.). Patient controls are patient ZIP code fixed effects, sex (male/female), age (included as a 5-knot spline), and comorbidity controls. The patient comorbidity control in the Massachusetts data is the HCC risk score included as a 5-knot spline; in the Medicare data, it is a vector of comorbidity fixed effects for each of the 27 conditions recorded in the Chronic Condition Warehouse. Massachusetts PCP controls include the average HCC risk score of the PCP's commercial patients (included as a 5-knot spline), total number of patients (included as 5-knot spline), indicator for PCP sex, and indicator for whether the PCP's specialty is Internal Medicine or Family Medicine. Medicare PCP controls include PCP specialty (i.e., family medicine or internal medicine), gender, a 5-knot spline in number of patients, and the PCP-level mean value of each of the 27 Chronic Condition Warehouse comorbidities.

Data: Columns 1-6, 2012 APCD analysis sample: commercially insured Massachusetts residents with chronic illness.

Data for Column 7: 2012 Medicare beneficiaries in 20% sample. See discussion of measurement error in text for interpretation.

Table 3. Robustness of Relationship between Referral Concentration and Healthcare Utilization and Spending to PCP & Specialist Fixed Effects

	Dependent variable: ln(utilization)			Dependent variable: ln(spending)		
	(1)	(2)	(3)	(4)	(5)	(6)
PCP Team Referral Concentration in relevant specialty	-0.297***	-0.143***	-0.192***	-0.403***	-0.115***	-0.268***
	(0.032)	(0.038)	(0.040)	(0.036)	(0.039)	(0.041)
PCP fixed effect	No	Yes	No	No	Yes	No
Specialist fixed effect	No	No	Yes	No	No	Yes
Patient and insurer controls	Yes	Yes	Yes	Yes	Yes	Yes

Note: PCP's team referral concentration is jackknifed (a patient's own visits are excluded from the calculation for their PCP). Standard errors are clustered at the patient level. Sample: N = 108,442. Sample is restricted to patients who saw at least one specialist in exactly one specialty: cardiology, orthopedics, endocrinology, dermatology, or OB/GYN. All specifications include fixed effects for the specialty consulted by the patient. Column 1 specification replicates the results on the restricted sample without the new physician fixed effects. Column 2 includes PCP fixed effects. Column 3 includes specialist fixed effects. Patient and insurer controls are the same as in Table 2, column 1.

Table 4. Difference-in-Differences and Instrumental Variables Results with Patient Movers in Medicare

	(1)	(2)	(3)
		A. Difference-in-Differences	
		Dependent variable:	
		ln(utilization)	
(Δ Team Referral Concentration)*Post	-0.613***	-0.434***	-0.448***
	(0.176)	(0.136)	(0.131)
		B. First Stage	
		Dependent variable:	
		(Δ Team Referral Concentration)*Post	
Pre-move Referral Concentration*Post	-0.859***	-0.858***	-0.851***
	(0.041)	(0.041)	(0.043)
F statistic of excluded instrument	441	438	393
		C. Two Stage Least Squares	
		Dependent variable:	
		ln(utilization)	
(Δ Team Referral Concentration)*Post	-0.584**	-0.445**	-0.468**
	(0.237)	(0.186)	(0.180)
Patient fixed effect	Yes	Yes	Yes
Patient care continuity HHI	No	Yes	Yes
PCP controls	No	No	Yes
Number of individual patients	1639	1639	1639
Number of observations (patient X year)	6230	6230	6230

Note: PCP's team referral concentration is jackknifed (a patient's own visits are excluded from the calculation for their PCP). Standard errors are clustered at the patient level. Data: patients in Medicare 20% sample who move across regions during 2007-2012.In panels B and C, the PCP's team referral concentration in the origin region is used as an instrumental variable for the change in team referral concentration experienced after the move, exploiting mean reversion in PCP team referral concentration at the time of a switch in PCP.

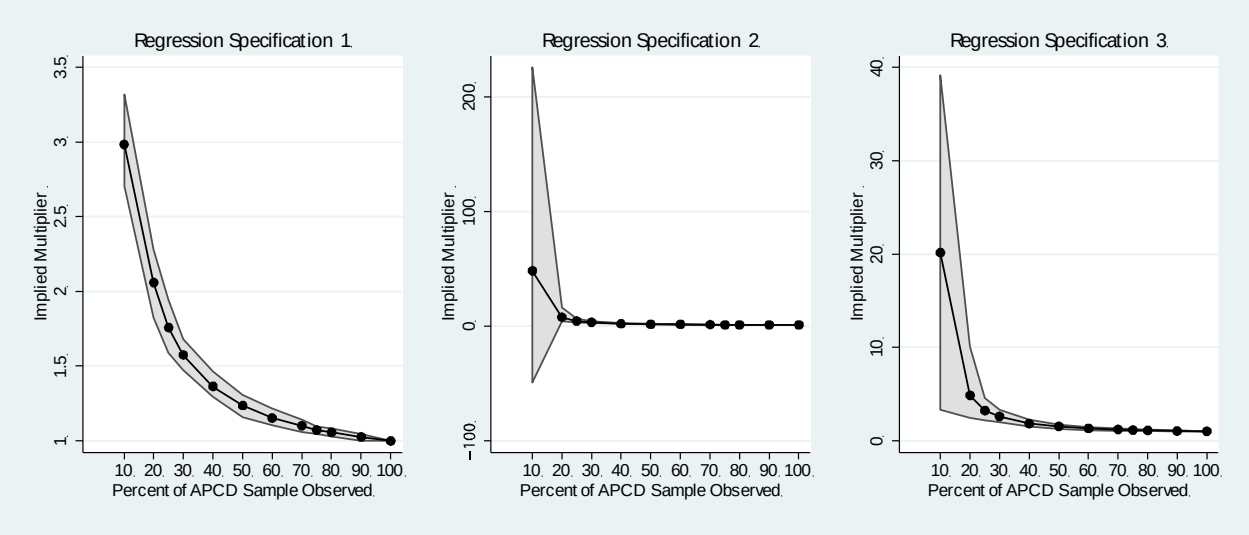
All regressions include an individual patient fixed effect, year fixed effect, and a series of indicators for event year relative to the move. The regression reported in column (2) adds a time varying measure of the patient's care continuity HHI. The regression reported in column (3) adds time varying PCP controls: specialty (internal medicine vs. family medicine), gender, and 5-knot spline of PCP patient volume. Recall that these specifications, particularly columns 2 and 3, are likely to suffer from substantial attenuation bias due to measurement error.

Online Appendix

For Agha et al., "Team Formation and Performance: Evidence from Healthcare Referral Networks"

Appendix A: Additional Empirical Results

Figure A1. Measurement Error Multiplier Simulations Using Subsamples of APCD Data



Note: Plots mean multiplier, 5th, and 95th percentiles bootstrapped from 50 random samples per percent subsample. Follows the regression specifications for Column 1, 2, and 3 respectively of Table 2.

Table A1. Average Team Referral Concentration by Physician Contracting Network, MA APCD

	Average Team Referral Concentration	Std. Dev. Of Team Referral Concentration	N PCPs
Physician Contracting Network			
Beth Israel Deaconess P.O.	0.098	0.065	296
Lahey Clinic	0.103	0.072	91
UMass Memorial Health Care	0.110	0.071	233
Partners Community Health Care	0.111	0.069	909
New England Quality Care Alliance	0.124	0.059	226
Baycare Health Partners	0.130	0.070	136
Southcoast Physicians Network	0.134	0.038	53
Atrius Health	0.136	0.052	253
Caritas Christi Network Service	0.151	0.075	203
Fallon Clinic	0.189	0.064	94
No Physician Contracting Network	0.169	0.104	1974

Notes: Physician Contracting Network information obtained from the 2010 Massachusetts Provider Database.

Table A2. Referral Concentration and Healthcare Utilization and Spending, Robustness

	Commercial Patients in Massachusetts	
	Dependent variable: ln(Utilization)	Dependent variable: ln(Spending)
	(1)	(2)
PCP Team Referral Concentration	-0.420***	-0.816***
	(0.0692)	(0.0858)
PCP ZIP code fixed effects	Yes	Yes
Patient care continuity HHI	Yes	Yes
PCP controls	Yes	Yes
Patient controls	Yes	Yes
Insurer controls	Yes	Yes
N patients	314,678	314,678

Note: Specifications are the same as Table 2, columns 3 and 6, but with the addition of PCP ZIP code fixed effects.

Table A3. Medicare Sample Descriptive Statistics

	Below Median PCP Team Referral Concentration (2012)	Above Median PCP Team Referral Concentration (2012)	Sample of moving beneficiaries (2007-2012)
Patient characteristics:			
Age	77.1	76.8	79.3
Male	0.41	0.41	0.34
Mean Spending ($)	13,127	10,786	10,917
Median Spending ($)	4543	3647	4098
Pr(Any Inpatient Admission)	0.27	0.23	0.28
Patient care continuity HHI	0.35	0.39	0.38
Pr(Diabetes)	0.30	0.30	0.28
Pr(Heart Condition)	0.45	0.42	0.50
Pr(Depression)	0.16	0.15	0.18
Pr(Asthma)	0.06	0.05	0.05
PCP characteristics:			
PCP's Team Referral Concentration	0.21	0.46	0.33
Pr(Internal Medicine)	0.69	0.56	0.66
PCP is Male	0.79	0.77	0.79
N Patients	925,754	925,645	1639

Notes: Similar to our findings in Massachusetts, we find that patient's age, sex, and disease burden are similar among patients seeing PCPs with above and below median team referral concentration. Because of measurement error, there is both more concentration and more variation in the measured PCP team referral concentration: below the median, the average referral concentration is 0.21 versus 0.46 above the median.

Table A4. PCP Team Referral Concentration and Medicare Spending, With Measurement Error

	Ln(utilization)		
	(1)	(2)	(3)
PCP Team Referral Concentration	-0.286***	-0.139***	-0.035***
	(0.011)	(0.009)	(0.008)
Patient controls	Yes	Yes	Yes
Patient care continuity HHI	No	Yes	Yes
PCP controls	No	No	Yes
N patients	1,848,071	1,848,071	1,848,071

Note: Specifications likely suffer from substantial attenuation bias due to measurement error; simulations suggest measurement error is particularly acute in columns (2) and (3) after additional controls are incorporated.

PCP's team referral concentration is jackknifed (a patient's own visits are excluded from the calculation for their PCP). Standard errors are clustered at the PCP level. Data: 2012 Medicare beneficiaries in our 20% sample. Specifications parallel Columns 1-3 of Table 2. However, there are no insurer controls, since all patients are enrolled in traditional Medicare. Further, patient HCC risk scores are replaced with a simple vector of comorbidity fixed effects for each of the 27 conditions recorded in the Chronic Condition Warehouse. Column 3 includes PCP controls: PCP specialty (family medicine or internal medicine), gender, a 5-knot spline in number of patients, and the PCP-level mean value of each of the 27 Chronic Condition Warehouse comorbidities.

Table A5. Decomposition of Medicare Spending

	Independent variable: PCP team referral concentration			
	(1)	(2)	(3)	*N*
Dependent variable:				
ln(Inpatient claims)	-0.186***	-0.144***	-0.0527***	468,611
	(0.012)	(0.011)	(0.011)	
ln(Outpatient claims)	-0.111***	0.114***	0.0852***	1,848,086
	(0.028)	(0.027)	(0.027)	
ln(Provider submitted claims)	-0.151***	-0.034***	0.030***	1,848,086
	(0.008)	(0.007)	(0.007)	
Any Inpatient spending	-0.096***	-0.077***	-0.038***	1,848,086
	(0.004)	(0.004)	(0.003)	
Patient controls	Yes	Yes	Yes	
Patient care continuity HHI	No	Yes	Yes	
PCP controls	No	No	Yes	

Note: Specifications likely suffer from substantial attenuation bias due to measurement error; simulations suggest measurement error is particularly acute in columns (2) and (3) after additional controls are incorporated.

PCP's team referral concentration is jackknifed (a patient's own visits are excluded from the calculation for their PCP). Standard errors are clustered at the PCP level. Data: 2012 Medicare beneficiaries in our 20% sample. However, there are no insurer controls, since all patients are enrolled in traditional Medicare. Further, patient risk scores are replaced with a simple vector of comorbidity fixed effects for each of the 27 conditions recorded in the Chronic Condition Warehouse. Column 3 includes PCP controls: PCP specialty (family medicine or internal medicine), gender, a 5-knot spline in number of patients, and the PCP-level mean value of each of the 27 Chronic Condition Warehouse comorbidities.

Appendix B. Chronic Illness Definitions for Massachusetts Population

We use the definition of Chronic Illness from Frandsen et al. (2015). A patient is included for having a chronic illness if they received an ICD diagnostic code in one of the following categories:

- Coronary artery disease: 410.xx-414.xx
- Cerebrovascular disease: 433.xx-438.xx, 441.xx-442.xx
- Peripheral arterial disease: 443.xx-445.xx
- Mesenteric vascular disease: 557.xx
- Other ischemic vascular disease or conduction disorders: 391.xx, 394.xx-398.xx, 440.xx, 426.xx-427.xx
- Heart failure: 402.01, 402.11, 402.91, 401.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.xx
- Migraine and cluster headache: 346.xx, 339.xx
- Hypertension: 401.xx-405.xx
- Hyperlipidemia: 272.xx
- Diabetes mellitus: 249.xx-250.xx, 362.0x
- Asthma: 493.xx
- Chronic obstructive pulmonary disease: 491.xx-492.xx, 494.xx, 496.xx, 416.xx
- Hypercoagulability disorders: 415.xx, 451.xx-454.xx
- Osteoarthritis: 715.xx, 717.xx, 721.xx, 726.xx
- Rheumatoid arthritis: 714.xx, 720.x

Appendix C. Proofs of Results in Section 2

Result 1: The PCP's optimal team structure has a number of specialists N*. The PCP invests the same amount of coordination effort r_s^* with each specialist and refers to each specialist with the probability $\frac{1}{N*}$.

Proof: The first order condition for each choice of effort is given by $r_s = \frac{1}{\sum_{s=1}^{N} e^{V_{is}}} e^{V_{is}}(\omega + \theta)$*. It is easy to see that the symmetric solution satisfies the vector of first order conditions. However,* $ln[\sum_{s=1}^{N} e^{V_{is}}]$ *is convex in the vector of efforts* r_s*. To guarantee this symmetric solution uniquely satisfies the first order conditions, we need the Hessian matrix of second order derivatives to be negative definite everywhere. The assumption* $\omega + \theta < \sqrt{2}$ *guarantees that.*

Result 2: Coordination effort r_s^* is inversely proportional to the number of specialists in the team, *N. A PCP with a higher fixed cost φ of working with an additional specialist will work with fewer specialists, invest more coordination effort with each specialist, and have lower expected healthcare costs for their patients.**

Proof:

From the first order condition for the choice of effort, we have $r_s^* = \frac{1}{N^*}(\omega + \theta)$*. Moreover the PCP's choice of N* is chosen to* $max_{N,\{r_s\}_{s\in N}} ln[\sum_{s=1}^{N} e^{V_{is}}] - \varphi N - \sum_{s=1}^{N}[\frac{1}{2} {r_s}^2]$.

A higher value of φ leads to a weakly lower N being chosen. Note that φ only directly affects the choice of N*, and the first order condition for* r_s^* *determines the level of coordination effort. This decreases in N*. Finally, Assumption 1 gives that higher* r_s^* *leads to lower healthcare costs.*

Appendix D. Measurement error in Medicare difference-in-differences regressions

In this section, we derive the expected impact of measurement error on the difference-in-differences results in the Medicare sample.

For simplicity, we consider a first-differenced specification where we keep one observation per patient who moves across regions. The dependent variable is the patient's change in care utilization (denoted $\Delta \log y_i$) and the independent variable is the change in patient's PCP team referral concentration (denoted ΔTRC_i) after the move. (For brevity, we notate team referral concentration as *TRC* in this appendix rather than *ReferralCon* used in the text.) The regression takes the following form:

$$\Delta \log y_i = \alpha \Delta TRC_i + \beta + \varepsilon_i$$

In the absence of any measurement error, we would have a coefficient on the change in referral concentration that takes the following form:

$$\hat{\alpha} = \frac{Cov(\Delta \log y_i \, , \Delta TRC_i)}{Var(\Delta TRC_i)}$$

We do not observe ΔTRC_i directly in the Medicare data, because we only have a 20% sample of Medicare patients for each doctor. As a result, team referral concentration is measured with error. We denote these noisy signals $\widetilde{\Delta TRC}$, and suppress subscripting notation below for simplicity.

Specifically, define:

$$\widetilde{\Delta TRC} \; = \; \Delta TRC + \Delta \mu$$

We will consider two cases. First, we will assume a case with classical measurement error, so that $\Delta \mu$ is independently distributed, and therefore is not correlated with the change in team referral concentration nor with care utilization. Then we will consider the more realistic case that $\Delta \mu$ is not independently distributed.

In the classical measurement error case, the independence assumption implies that $\Delta \mu$ is uncorrelated with ΔTRC and $\Delta \log y$. When we estimate the difference-in-differences specification, we will find the following coefficient:

$$\hat{\alpha}^{classical\ m.e.} = \frac{Cov(\Delta \log y \, , \Delta TRC)}{Var(\Delta TRC) + Var(\Delta \mu)}$$

This coefficient suffers from attenuation bias, as in the classical derivations; this is seen in the addition of the term to the denominator.

Now consider the more complicated, but also more realistic, possibility that the error in the team referral concentration measure is related to the level of team referral concentration. It is easy to see why the independence assumption may be violated in our setting if you consider the behavior of measurement error near the bounds of the referral concentration measure. A doctor who is perfectly concentrated and only refers to 1 specialist of each type will have no error in his team referral

concentration measure when measured using a 20% sample. As long as we observe 1 referred patient, we would be able to perfectly calculate his TRC=1. By contrast, consider a doctor who is not at all concentrated in his referrals. Within each specialty, he refers each of his patients to a different specialist. The more patients we observe, the closer his TRC comes to 0, but in any finite subsample of his patient panel, we will overestimate his TRC. Extending this intuition, we expect measurement error in team referral concentration to be negatively correlated with the true referral concentration.

The difference-in-differences regression coefficient now becomes:

$$\hat{\alpha}^{non-independent\ m.e.} = \frac{Cov(\Delta \log y\,, \Delta TRC) + Cov(\Delta \log y\,, \Delta \mu)}{Var(\Delta TRC) + Var(\Delta \mu) + 2Cov(\Delta \mu, \Delta TRC)}$$

Unlike the classical measurement error case, the sign and size of the bias is no longer obvious, and will depend on the particular relationships in our setting. We expect that $Cov(\Delta \log y\,, \Delta TRC) < 0$, given the predictions of our model and the results in the Massachusetts data, which have minimal measurement error. By contrast, we expect that $Cov(\Delta \log y\,, \Delta \mu) > 0$, given the intuition about measurement error and its relationship to team referral concentration described in the previous paragraph. As long as $|Cov(\Delta \log y\,, \Delta TRC)| > |\ Cov(\Delta \log y\,, \Delta \mu)|$, the changes in the numerator will tend to attenuate the measured coefficient.

In the denominator, the variance terms are positive. We expect that $Cov(\Delta \mu, \Delta TRC) < 0$. This implies that the net effect of measurement error on the denominator depends on the relative size of the $Var(\Delta \mu)$ and $Cov(\Delta \mu, \Delta TRC)$ terms. If $|Var(\Delta \mu)| > |2\ Cov(\Delta \mu, \Delta TRC)|$, then the denominator will inflated, and there will be attenuation bias. On the other hand, if $|Var(\Delta \mu)| < |2\ Cov(\Delta \mu, \Delta TRC)|$, then the denominator will be smaller relative to the case without measurement error. In this case, the coefficient could be inflated (or even, in the extreme, wrong-signed).

In sum, the net effect of measurement error on the coefficient is theoretically ambiguous in the difference-in-differences setting. The coefficient could be either inflated or attenuated depending on the strength of the correlation of measurement error with the other terms. Note that the cross-sectional OLS regressions would have a very similar formulation for bias from measurement error; eliminating the Δ terms from the formulas above would yield the OLS coefficients. The simulations of measurement error we run in the Massachusetts APCD suggest that the attenuating terms dominate, at least in the OLS specification (cf. Section 6A).

A similar logic extends to the instrumental variable (IV) case, and the net effect of measurement error is theoretically ambiguous. We have non-classical measurement error, since measurement error in our instrument (pre-move team referral concentration) is mechanically correlated with measurement error in the endogenous variable (change in team referral concentration).

To compare attenuation bias in the difference-in-differences or IV specifications to attenuation bias in the OLS Medicare results (e.g. coefficients reported in Table 2, column 7), we must also draw one more distinction. When we estimate the mover results, we average the patient's PCP team referral concentration over the year(s) of the pre-move period to form the patient's PCP's pre-move team referral concentration (and similarly for the post period). This change will tend to reduce the noise in our signal of PCP team referral concentration, reducing relative to the single year from the static OLS

model. This change would lead us to predict a smaller role for bias from measurement error in the mover specifications.