Influence and Information in Team Decisions: Evidence from Medical Residency*

David C. Chan†

February 19, 2020

Abstract

I study team decisions among physician trainees. Exploiting a discontinuity in team roles across trainee tenure, I find evidence that teams alter decision-making, concentrating influence in the hands of senior trainees. I also demonstrate little convergence in variation of trainee effects despite intensive training. This general pattern of trainee effects on team decision-making exists in all types of decisions and settings that I examine. In analyses evaluating mechanisms behind this pattern, I find support for the idea that significant experiential learning occurs during training and that teams place more weight on judgments of senior trainees in order to aggregate information.

JEL Codes: D83, L23, M53

*I am grateful to David Cutler, Joe Doyle, Bob Gibbons, and Jon Gruber for early guidance on this project. I also thank Achyuta Adhvaryu, Daron Acemoglu, Leila Agha, David Autor, Daniel Barron, David Bates, Amitabh Chandra, Wes Cohen, Michael Dickstein, Amy Finkelstein, Matt Gentzkow, Mitch Hoffman, Peter Hull, Emir Kamenica, Pat Kline, Jon Kolstad, Eddie Lazear, Frank Levy, David Molitor, Maria Polyakova, Maya Rossin-Slater, Jon Skinner, Doug Staiger, Chris Stanton, Caio Waisman, Chris Walters, and many seminar audiences for helpful comments. Joel Katz and Amy Miller provided invaluable context to the data. Samuel Arenberg, Atul Gupta, Uyseok Lee, and Natalie Nguyen provided excellent research assistance. I acknowledge support from the NBER Health and Aging Fellowship, under the National Institute of Aging Grant Number T32-AG000186; the Charles A. King Trust Postdoctoral Fellowship, the Medical Foundation; and the Agency for Healthcare Research and Quality Ruth L. Kirschstein Individual Postdoctoral Fellowship 1-F32-HS021044-01.

†Address: 117 Encina Commons, Room 215; Stanford, CA 94306. Phone: 650-725-9582. Fax: 650-723-1919. Email: david.c.chan@stanford.edu.

1 Introduction

Scholars have long conceptualized how teams may restructure decision-making. Teams may combine information across individuals, in order to address more complex problems (Hayek, 1945). Teams may also handle time-sensitive problems arriving at uncertain times by routing decision-making through an organizational structure (Marschak and Radner, 1972). Despite the important implications of teams on economic activity, empirical analysis of how teams and organizations may impact decision-making remains scarce.

In health care, a large body of evidence documents wide variation across organizational boundaries in decisions driven by physicians (McCarthy and Blumenthal, 2006; Institute of Medicine, 2013). Without an understanding of how teams alter decision-making, the scale of practice variation across organizations seems difficult to reconcile with a similar magnitude of variation across individual providers (e.g., Skinner, 2012; Van Parys and Skinner, 2016). Large numbers of providers practicing independently at each institution should mute any systematic variation across organizations.[1] But if decision-making is concentrated in the hands of fewer providers, then they may drive surprisingly large variation across organizations. Policy implications would then depend on the organizational and informational frictions leading to such concentration.

I address this empirical question in the setting of medical residency, which is well-suited for studying team decision-making and the roots of practice variation for several reasons. Each patient is assigned to a well-defined physician trainee team comprising a junior trainee in the first year of training and a senior trainee past the first year of training. Teams are reshuffled weekly, so that each physician trainee works with many co-workers throughout training. A large number of patient cases are quasi-randomly assigned to trainees over the course of residency, and trainees take part in dozens of medical decisions per patient-day that are captured in the electronic medical record. Finally, as a separate point of interest, residency training provides a unique window into the evolution of practice variation among providers in health care. By design, residency is an intensive program to impart knowledge to physicians beyond facts, "developing habits, behaviors, attitudes, and values that will last a professional lifetime" (Ludmerer, 2014).

[1] For example, (Molitor, 2017) finds a lack of any systematic sorting of individual providers to locations by their practices. Perhaps more intriguingly, he also finds that, upon changing locations, providers immediately change their decision-making to match a local practice style, which suggests that physician decision-making is not independent of the local environment.

Specifically, I follow a group of 799 physician trainees in a large academic hospital and exploit detailed administrative data of physician trainees to teams caring for patients. Team decisions are measured over a five-year period as detailed orders for 3.4 million medications, 3.1 million laboratory tests, and 268,065 radiology tests. I aggregate dozens of physician orders by their costs to form spending summary statistics of team decisions for each of 220,117 patient-days, in categories of laboratory testing, radiology testing, medication, blood transfusion, and nursing.

Using random assignment of patients to physician teams and frequent rotation of trainees across teams, I identify the causal trainee effects on team decisions measured by spending at various points in the trainees' tenure. Specifically, I employ a strategy similar to that used in a number of papers starting with Abowd et al. (1999), which have decomposed joint outcomes into contributions due to workers and firms (Card et al., 2013), workers and managers (Lazear et al., 2015), patients and geographic locations (Finkelstein et al., 2016), and physicians and locations (Molitor, 2017), among others. A key difference is that I estimate separate trainee effects at different points in their residency training, which is possible because of the frequency of the patient observations and of the rotations across teams. As a central object of interest, I define tenure-specific *practice variation* as the standard deviation of the distribution of these trainee effects across trainees in a given tenure period. Given the finite number of observations per trainee in a tenure period, I develop an estimation approach based on random effects in a hierarchical model (Searle et al., 1992; Gelman and Hill, 2007).

Next, I use the team structure in medical residency to decompose trainee effects into two components relevant for team decision-making: a trainee's *judgment* (what she would have decided on her own as a single agent) and her *influence* (the extent to which her judgment sways the team decision). I isolate the effect of influence by assessing trainee effects across a discontinuity at the one-year tenure mark: Before one year, trainees have relatively less experience than their teammate, while they have relatively more experience than their teammate immediately after their first year. Teams may also induce greater influence due to roles and responsibilities within the team. If trainee judgments (and other individual characteristics) are plausibly continuous across the one-year mark, then a discontinuous change in practice variation across one year reflects the contribution of influence due to any of these team-induced mechanisms.

I find a significant and discontinuous increase in practice variation across the one-year mark of training. Junior trainees before this mark show variation in total spending effects with a standard

deviation of 5%, while senior trainees beginning their second year show variation in total spending effects with a standard deviation of 24%. Subsequent practice variation remains large to the end of training, and there is remarkably little convergence in trainee effects. When I consider two-agent teams (i.e., one junior trainee and one senior trainee is responsible for each patient-day), the senior trainee is responsible for $\frac{0.24^2}{0.05^2+0.24^2} \approx 96\%$ of the variance in team-level decisions.[2]

The discontinuous change in practice variation at the one-year mark provides strong evidence that teams matter for decision-making. I consider how such a change in influence across roles might reflect three types of "team concerns." First, in classical team theory, teams may address an issue of bandwidth limits among agents by distributing problems to agents. Since senior trainees split their time working with two junior trainees, they have time to attend to fewer problems per patient and should on average have *less* influence on any given case, an idea known as "management by exception" (Marschak and Radner, 1972; Garicano, 2000). Second, teams may introduce a principal-agent problems by their hierarchical nature, inducing herding around senior trainees' beliefs, effectively increasing the influence of senior trainees (Prendergast, 1993). Third, teams may aggregate information (DeGroot, 2005), by allowing agents to confer with each other before making joint decisions. If senior trainees have more knowledge than junior trainees, they should have more influence. Additionally, unlike other team concerns, information aggregation may blunt convergence in trainee effects, since influence continuously increases with knowledge (unlike discrete hierarchical roles and titles).

In the later half of the paper, I shed some empirical light on potential mechanisms that might explain the combination of (i) a discontinuous increase in practice variation at the one-year tenure mark, and (ii) a lack of significant convergence in practice variation even as trainees near completion of residency. First, senior trainees may have greater influence independent of their knowledge, for example due to institutionalized differences in their job duties or by herding around their beleifs. Relatedly, they may simply hold decision rights over important decisions. Second, influence may reflect systematic differences in knowledge across tenure, acquired by learning during residency. Under the first mechanism, asymmetric influence arises from frictions in the efficient *application* of knowledge. Under the second mechanism, asymmetric influence may represent an efficient application of greater knowledge held by senior trainees; in contrast, this greater knowledge arises from frictions in the

[2]Other members of the care team outside of trainees include supervising physicians, nurses, pharmacists, and specialty consultants. I focus on trainee teams because they are the most clearly quasi-randomly assigned.

acquisition of knowledge.

While these mechanisms may coexist, I consider two scenarios representing extreme versions of the two mechanisms. Isolating the first mechanism, I consider the possibility of fixed trainee judgments (i.e., knowledge) across residency. In this scenario, physicians possess all their knowledge from the beginning, and no learning occurs in residency. The increase in influence is thus due only to titles and decision rights unrelated to knowledge. Contradicting the decision-rights hypothesis—based on a team-theoretic Garicano (2000) model that routes decisions to different team members, with more important decisions going to senior trainees—I find that senior trainees have much greater influence over *all* types of decisions, both great and small, and particularly over diagnostic decisions that may be more uncertain (though not particularly expensive). I further rule out the extreme hypothesis of no-learning by showing that trainee practice styles vary over time. Detailed and seemingly important time-invariant trainee characteristics predict only a small portion of practice variation. In addition, the serial correlation between trainee judgments grows weaker time. In contrast to prior literature that seemingly suggests relatively stable practice styles, this evidence suggests strong learning in the sense that physician practice styles are highly mutable, at least during physicians' early careers.[3]

To isolate the second mechanism, I assume an alternatively extreme scenario that influence is optimally allocated according to knowledge, as specified by a simple structural model of Bayesian information aggregation in decision-making, but that knowledge must accrue with training. In this model, as trainees learn, their increasing influence on team decisions may dampen or reverse any convergence in their practice styles in teams. This stands in contrast with independent decision-making, in which increasing knowledge will necessarily lead to convergence. Results from this structural model imply substantial learning in the first year of training, relative to any pre-residency knowledge. Interestingly, the results also suggest much greater learning when trainees become senior and have a larger stake in decision-making, which is consistent with large literatures on *experiential learning*, positing that learning requires active participation and experience.[4] Between trainees, I find that de-

[3]For example, Epstein and Nicholson (2009) examines practice styles of obstetricians and projects changes of other obstetricians practicing in the same hospital on the practice style of each index obstetrician. Molitor (2017) examines practice styles of cardiologist movers and similarly projects changes in the local practice style induced by the move onto the average practice style of moving cardiologists. In both studies these projections are remarkably stable over time, but they may mask significant evolution of practice styles unrelated to these projections. Doyle et al. (2010) study physician trainees from two different residency programs and find systematic differences between trainees of the two programs. However, they abstract from any variation within program or changes within trainee.

[4]Notable contributions in this area include John Dewey's (1938) thoughts on progressive education in *Experience and Education*; Maria Montessori's (1948) method of teaching children; Jean Piaget's (1971) constructivist theory of knowing;

viations from optimal influence are small. However, I also find that, relative to their supervisors, the trainee team receives much more influence than justified by the trainees' knowledge.

This paper contributes to several literatures. First, it contributes to a general literature on decision-making in organizations (e.g., Marschak and Radner, 1972; Van Zandt, 1998; Garicano, 2000). As noted by Hayek (1945, p. 519),

> "The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess."

Despite seminal theoretical contributions in this literature, empirical evidence remains scarce. The evidence in this paper highlights the important team function of aggregating information for a given decision, because the optimal decision may not be known perfectly by any single agent. This stands in contrast to canonical team-theoretic models, notably Garicano (2000), that view the function of organizations as routing problems with known solutions.

Second, this paper sheds new empirical light on the nature of learning, as defined by forming judgments to make decisions. In economics, a large empirical literature on "learning on the job" (Mincer, 1962) has mostly relied on wages as a marker of learning, while another empirical literature studying "learning by doing" (Arrow, 1962) has measured task performance (e.g., speed or accuracy) attributable to agents or firms gaining experience.[5] Neither approach seems appropriate in this setting: Physician trainees are paid fixed salaries, and complicated and potentially high-stakes decisions are made in teams with layers of experience.[6] This paper makes some progress on this problem by developing a notion of learning with empirical implications for team decisions.

Third, as noted above, these results relate to a large literature documenting practice variation in health care (Fisher et al., 2003a,b).[7] Academic and policy discussions on this topic often refer to

and Kolb and Fry's (1975) experiential learning. Similar concepts also include problem-based learning (e.g., Wood, 2003), and "learning by teaching" (Gartner et al., 1971).

[5]Examples in the empirical literature of on-the-job training that focuses on wages include Topel (1991) and Kahn and Lange (2014). Examples in the empirical literature on learning by doing include Benkard (2000), Levitt et al. (2013), and Hendel and Spiegel (2014).

[6]Indeed, team-based decision-making may be a key reason why teaching hospitals show no significant decline in patient outcomes in July, when a sudden (but scheduled) influx of fresh physician trainees arrives (Young et al., 2011; Song and Huckman, 2018).

[7]In addition to the literature reviewed by Skinner (2012), recent contributions in the economics literature include Doyle et al. (2015), Cooper et al. (2015), Chandra et al. (2016), Finkelstein et al. (2016), and Molitor (2017). Much of this

features of the health care marketplace that insulate providers from competition, but this reasoning assumes that, absent incentives, providers mostly agree on the diagnosis and treatment for any given patient (Cutler, 2010; Skinner, 2012). This view is incompatible both with survey evidence revealing that experts often widely disagree (Cutler et al., 2018). It is also inconsistent with growing evidence of simultaneous errors of commission and omission among providers in both diagnostic and treatment decisions (Abaluck et al., 2016; Chan et al., 2019). This paper highlights informational mechanisms behind wide practice variation in an intense and highly selective training environment *designed* to create homogeneity. Interestingly, team decision-making and knowledge frictions may concentrate influence behind practice variation into the hands of fewer providers who nonetheless disagree with each other. If so, appropriate policy responses to practice variation should focus on organizational and informational levers.

The organization of this paper is as follows. Section 2 describes the institutional setting and data. Section 3 introduces the empirical approach. Section 4 presents main results and discusses how they relate to team concerns. Section 5 investigates mechanisms in greater detail. Section 6 discusses policy implications for practice variation and concludes.

2 Setting and Data

2.1 The Structure of Residency

I study trainees associated with the internal medicine residency program of a large teaching hospital. The program is highly selective, and the hospital is a source of numerous clinical trials and guidelines. As is standard across internal medicine programs, training takes place over three years in teams organized by experience: Each patient is cared for by a first-year junior trainee ("intern") and a second- or third-year senior trainee ("resident").

While each patient is assigned to a team of one intern and one resident, residents split their time between two interns. Thus, interns are assigned half the number of patients as residents. This allows interns to devote more attention to each patient, and they are usually the first to examine a patient

literature focuses on differences among regions or hospitals. See Epstein and Nicholson (2009) as an example of physician-level variation that has generally been difficult to explain. Similar informational frictions can underlie differences across organizations (e.g., Bloom and Van Reenen, 2010). Particularly relevant to the setting of residency training is work by Doyle et al. (2010) comparing mean practices between two groups of trainees from different programs randomly assigned patients in the same hospital.

and make judgments. Each senior trainee (along with the two junior trainees working with her) is supervised by an "attending" physician, who has completed residency. Teams also operate within a broad practice environment that influences decision-making, including institutional rules, information systems, and other health care workers such as consulting physicians, pharmacists, and nurses. Trainees on the same teams may come from different predetermined career tracks, other programs (e.g., obstetrics-gynecology, emergency medicine), or another hospital. A sizable minority of interns plan only to spend one year in the internal medicine residency ("preliminary" versus "categorical" interns), subsequently proceeding to another residency program such as anesthesiology, radiology, or dermatology.

This study focuses on inpatient ward rotations, which comprise cardiology, oncology, and general medicine services. According to interviews with residency administration, trainee rotation preferences are not collected and assignment does not consider trainee characteristics. Scheduling is done a year in advance and does not consider matches among intern, resident, and attending physicians that will be formed as a result. Supervising physician schedules are created independently, with neither trainee nor supervising physician aware of one another's schedule in advance. Therefore, trainees and supervising physicians are as good as randomly assigned to each other.

Patients admitted to ward services are assigned to interns and residents by a simple algorithm that distributes patients in a rotation among on-call trainees.[8] Patients who remain admitted for more than one day may be mechanically transferred to other trainees as they change rotations. When one trainee replaces another, she assumes the entire patient list of the previous trainee. Because trainee blocks are generally two weeks long and are staggered for interns and residents, patients frequently experience a change in either the intern or the resident on the team.

2.2 Team Decisions

As in other small-team settings, formal decision rights are rarely invoked in patient care teams. While senior teammates may influence decisions by their general knowledge or status, junior teammates may acquire more patient-specific knowledge and are usually charged with implementing decisions. A variety of protocols and customs common in residency encourage trainees to function independently

[8]Depending on the reason for admission, patients may be matched to categories of attending physicians according to the admitting service. Trainees who have reached their capacity may also be taken out of the algorithm for accepting new patients during the remainder of a call day. Conditional on these constraints, patient types are not matched to trainees.

and to take responsibility for clinical decisions. For example, junior trainees are listed as the first point of contact, so that information from patients, nurses, and consultants generally flows through junior trainees before reaching senior trainees or supervising physicians. Similarly, junior trainees are expected to write orders and for discussing the care plan with patients and other staff, so that they are abreast of all decisions made for their patients. While trainees may consult with supervising physicians in real time, they often make and communicate decisions without prior consultation. As a practice, supervising physicians will often delay discussing new patients or new developments until after trainees have evaluated the patient and formulated a treatment plan. Thus, supervisors will often learn about decisions after they are made.

2.3 Data

I collect data from several sources. First, I observe the identities of each physician on the clinical team—intern, resident, and attending physician—for each patient on a ward service on each day that the patient is in the hospital. Over five years, I observe data for 46,091 admissions, equivalent to 220,074 patient-day observations. Corresponding to these admissions are 799 unique trainees and 531 unique attendings. of the trainees, 516 are from the same internal medicine residency, with the remainder visiting from another residency program within the same hospital or from another hospital.[9] There is no unplanned attrition across years of residency.[10]

I collect detailed information for each trainee, including demographics, medical school, US Medical Licensing Examination (USMLE) Step 1 test scores, membership in the Alpha Omega Alpha (AOA) medical honor society, other degrees, and position on the residency rank list. Summary statistics of trainees characteristics are given in Appendix Tables A-2 and A-3 and are consistent with an elite group of trainees.[11] I also observe pre-committed residency tracks for each trainee physician. In addition to trainee characteristics determined prior to residency, I observe each trainees realized specialty after her training to impute expected yearly future income in the five years immediately following this training based on industry-standard survey data from the Medical Group Management

[9] Of the 799 unique trainees, 649 are observed as interns and 407 are observed as residents. Of the 516 trainees from the same-hospital internal medicine residency, 401 are observed as interns, and 338 are observed as residents.

[10] In two cases, interns with hardship or illness in the family were allowed to redo intern year.

[11] For example, trainees in the data are almost three times more likely to be AOA inductees than the national average, a trait that predicts a 6-10 greater odds of matching to a first-choice residency program (Rinard and Mahabir, 2010). The mean USMLE Step 1 score is 244, or approximately the 76th percentile of the national distribution.

Association. The average above- and below-median future incomes for junior trainees are $424,000 and $269,000, respectively; the respective numbers for senior trainees are $409,000 and $249,000.[12]

I use scheduling data and past matches between trainees and with supervising attending physicians. Consistent with Section 2.1, Table 1 shows that interns and residents with high or low spending effects are exposed to similar types of patients and are equally likely to be assigned to high- or low-spending coworkers and attendings. Appendix A-1 presents more formal analyses on conditional random assignment of trainee physicians, including F-tests showing joint insignificance.

Patient demographic information includes age, sex, race, and language. Clinical information derives primarily from billing data, in which I observe International Classification of Diseases, Ninth Revision, (ICD-9) codes and Diagnostic-related Group (DRG) weights. I use these codes to construct Charlson comorbidity indices and 29 Elixhauser comorbidity dummies (Charlson et al., 1987; Elixhauser et al., 1998). I also observe the identity of the admitting service (e.g., "Heart Failure Team 1"), which categorizes patients admitted for similar reasons. Patients are *not* randomly assigned to supervising physicians, since supervising physicians within the same service may belong to different practice groups (e.g., HMO, private practice, hospitalist) that I do not explicitly capture and condition on.

I observe cost information for each patient-day aggregated within 30 cost departments used by the hospital for accounting purposes. I further group these departments into four categories: diagnostic (laboratory and radiology) testing, medication, blood bank, and nursing. Because costs are based on the hospital's accounting of resource utilization due to physician *actions*, not the measures of Medicare reimbursement used in recent studies (Doyle et al., 2015; Skinner and Staiger, 2015; Chandra et al., 2016), they provide more direct insight into welfare-relevant resource use.[13] Consistent with prior literature on practice variation, I consider spending as a summary statistic of the many actions involved in patient care. Laboratory costs are based on 3.1 million physician laboratory orders; radiology costs on 268,065 tests ordered in CT, MRI, nuclear medicine, and ultrasound; and medication costs on 3.4 million medication orders. Table 2 shows distributional statistics of daily spending in

[12]The difference in future incomes between junior and senior trainees reflects that the career paths for preliminary interns (e.g., future anesthesiologists, dermatologists, and radiologists) are often more lucrative.

[13]In this prior research, a difficulty in connecting practice variation in health care to the productivity literature is that "spending" input measures are actually government-set reimbursement rates that reflect hospital *revenues* rather than input costs. In large part, the Medicare reimburses inpatient care prospectively based on *diagnoses* rather than social cost of actual utilization.

each category and in the services of cardiology, oncology, and general medicine.

3 Analysis of Team Decisions

3.1 Potential Decisions

I observe a large set of decisions and the identities of agents on the team responsible for each decision. However, I do not observe an agent's contribution to the team decision, which is a key object of interest. The goal of the empirical approach is thus to decompose a team decision into such contributions made by each agent on the decision, and to allow this decomposition to depend on circumstances that may shed light on organizational considerations in team decision-making.

To characterize decision-making on a tractable and continuous scale, I reduce the dimensionality of decisions by aggregating the direct costs of the decisions, observed via the hospital's accounting system, for a given patient-day.[14] Thus, a patient-day, or the combination (i,t) for patient admission i and day t, constitutes a "case" for which a team decision is observed. I denote potential team decisions for patient-day (i,t) assigned to a two-agent team composed of agents $j \in \mathcal{J}_{it}$ and $k \in \mathcal{K}_{it}$ as $Y_{it}(j,k)$. The realized decision is

$$Y_{it} = \sum_{j \in \mathcal{J}_{it}} \sum_{k \in \mathcal{K}_{it}} D_{ijt} D_{ikt} Y_{it}(j,k). \tag{1}$$

$D_{ijt} \in \{0,1\}$ and $D_{ikt} \in \{0,1\}$ are indicator variables for assignment. Equivalently, since each case is assigned to one pair (j,k), define an assignment function $j(i,t)$ and $k(i,t)$ such that $D_{ijt} = \mathbf{1}(j = j(i,t))$ and $D_{ikt} = \mathbf{1}(k = k(i,t))$. In this setting, $\mathcal{J}_{it}$ and $\mathcal{K}_{it}$ are disjoint sets for any (i,t) since j is a junior trainee, and k is a senior trainee.

3.2 Trainee Effects

Given the potential outcome notation in Equation (1), I define trainee effects on team decision-making. For example, the effect of assignment to trainee j instead of j', holding k fixed, is $Y_{it}(j,k) - Y_{it}(j',k)$. Similarly, the effect of assignment to trainee k instead of k', holding j fixed, is $Y_{it}(j,k) -$

[14] In principle, given these micro-data, I could also study variation at the order level. However, the set of potential orders is large, and many orders are very specific to certain clinical scenarios that may not be observed frequently. Restricting study to certain types of clinical decisions, such as C-sections vs. vaginal deliveries (e.g., Currie and Gruber, 1996) or interventional treatment of heart attacks vs. medical management (e.g., Chandra and Staiger, 2007), is an approach used by many influential studies in the literature but does not capture the breadth or complementarity of physician decisions made on a daily basis. Section 5.1 provides some interesting evidence of such complementarity.

$Y_{it}(j,k')$. Because I only observe $Y_{it} = Y_{it}(j(i,t),k(i,t))$, effects for a particular case (i,t) are unobservable.

My goal is to recover expectations of trainee effects, by making use of quasi-random assignment of cases to trainees and of trainees to each other, as described in Section 2. Specifically, I consider the following conditional independence assumption:

Assumption 1 (Quasi-Random Team Assignment). *Potential team decisions are independent of team assignments, conditional on clinical service $s(i,t)$ and indicators of time $\mathbf{T}_t$ (e.g., day of the week, month-year combinations):*

$$\{Y_{it}(j,k)\}_{(j,k)\in\mathcal{J}_{it}\times\mathcal{K}_{it}} \perp\!\!\!\perp \left(D_{ijt}, D_{ikt}\right) \Big| s(i,t), \mathbf{T}_t.$$

If case potential outcomes are conditionally independent of team assignments, then trainee treatment effects are also conditionally independent of the team assignments. Appendix A-1.1 presents evidence of quasi-random assignment of patients to trainees, and Appendix A-1.2 presents evidence of quasi-random assignment of trainees to each other.

In the main analysis, I wish to capture a trainee's average treatment effect on team decisions, depending on her tenure and on the tenure of teammates she could be working with. The timing of residency implies a mechanical relationship between the tenures of the junior and senior trainees. Since trainees all begin residency at the same time of the year, a junior trainee with tenure τ_j will work with senior trainees with one or two more years of tenure, or $\tau_k \in \left\{\tau_j + 1, \tau_j + 2\right\}$. I consider a population of cases defined by a feasible combination of junior and senior trainee tenure periods, or $\mathscr{C} = \left\{(i,t) : \tau(j(i,t),t) = \tau_j, \tau(k(i,t),t) = \tau_k\right\}$, where $\tau(h,t)$ is a function that maps trainee h at time t to a tenure period.

I then define

$$\begin{aligned}
\text{ATE}(j|\mathscr{C}) &\equiv E_{(i,t)\in\mathscr{C}}\left[E_{k\in\mathcal{K}_{it}}\left[Y_{it}(j,k)\right] - E_{(j,k)\in\mathcal{J}_{it}\times\mathcal{K}_{it}}\left[Y_{it}(j,k)\right]\right]; \\
\text{ATE}(k|\mathscr{C}) &\equiv E_{(i,t)\in\mathscr{C}}\left[E_{j\in\mathcal{J}_{it}}\left[Y_{it}(j,k)\right] - E_{(j,k)\in\mathcal{J}_{it}\times\mathcal{K}_{it}}\left[Y_{it}(j,k)\right]\right].
\end{aligned}$$

$\text{ATE}(j|\mathscr{C})$ is junior trainee j's average effect on team decisions, working with in an "average" senior trainee, relative to "average" counterfactual teams of junior and senior trainees. $\text{ATE}(k|\mathscr{C})$ considers

a similar object for senior trainee k. In both cases, the "average" teammate and the "average" team is defined by the set of cases, $\mathscr{C}$, that specifies the tenures of the junior and senior trainees.

In these definitions, I exploit the fact that trainee assignment is independent of potential team decisions; this implies that expectations of $Y_{it}(j,k)$, holding j or k fixed, are the same regardless of whether we condition on actual assignment to j or k. Note that $\text{ATE}(j|\mathscr{C})$ and $\text{ATE}(k|\mathscr{C})$ depend not only on the indentity of the trainee j or k, but also on the set of potential teammates implied by $\mathscr{C}$. The same trainee may have different effects on team decisions in different environments, particularly depending on whether they are more or less senior to their teammate.

Assumption 1 implies that I can recover consistent estimates of $\text{ATE}(j|\mathscr{C})$ and $\text{ATE}(k|\mathscr{C})$ by the following regression, performed over a sample of observations (i,t) drawn from the population set $\mathscr{C}$:

$$Y_{it} = \xi^{\mathscr{C}}_{j(i,t)} + \xi^{\mathscr{C}}_{k(i,t)} + \gamma_{s(i,t)} + \mathbf{T}_t\eta + \varepsilon_{it}, \tag{2}$$

where $\xi^{\mathscr{C}}_j$ and $\xi^{\mathscr{C}}_k$ are trainee effects for the junior and senior trainees, $\gamma_{s(i,t)}$ is a fixed effect for the clinical service $s(i,t)$, and ε_{it} is an error term. By construction, $E[\varepsilon_{it}|s(i,t),\mathbf{T}_t,\mathbf{D}_{it}] = 0$, where $\mathbf{D}_{it}$ is a design vector indicating junior trainee and senior trainee identities. Under Assumption 1, we also have $E[\varepsilon_{it}|s(i,t),\mathbf{T}_t,\mathbf{D}_{it}] = E[\varepsilon_{it}|s(i,t),\mathbf{T}_t] = 0$. Thus, regression estimates of $\xi^{\mathscr{C}}_j$ and $\xi^{\mathscr{C}}_k$ are consistent estimators of $\text{ATE}(j|\mathscr{C})$ and $\text{ATE}(k|\mathscr{C})$, respectively.

Because $\mathscr{C}$ is defined by trainee periods τ_j and τ_k for the junior and senior trainees, respectively, I rewrite these effects as $\xi^{\tau_j;\tau_k}_j$ and $\xi^{\tau_k;\tau_j}_k$ to be more explicit about the tenure-dependence of the estimated trainee effects. In practice, I perform the following regression with additional controls:

$$Y_{it} = \mathbf{X}_{it}\beta + \xi^{\tau_j;\tau_k}_j + \xi^{\tau_k;\tau_j}_k + \mathbf{T}_t\eta + \gamma_{s(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{3}$$

The objects of interest in Equation (3) are tenure-specific trainee effects—$\xi^{\tau_j;\tau_k}_{j(\cdot)}$ and $\xi^{\tau_k;\tau_j}_{k(\cdot)}$ for the junior and senior trainees, respectively—that depend on the identity of the trainee and the tenure periods of both trainees. To improve efficiency, I also include fixed effects for the supervising physician, $\ell(i,t)$, and a rich set of patient and admission characteristics, $\mathbf{X}_{it}$.[15] These controls are unnecessary for identification under Assumption 1, and in Section 4.2, I show robustness of results to including

[15] Specifically, I control for patient race dummies, male gender, linear and quadratic age, the Charlson comorbidity score (Charlson et al., 1987), 29 Elixhauser comorbidity dummies (Elixhauser et al., 1998), Diagnostic Related Group (DRG) weights, and day of the patient's length of stay dummies.

none of these controls.

3.3 Random-Effects vs. Fixed-Effects Estimation

As described in Abowd et al. (2008), two approaches to estimating Equation (3) are to treat the trainee effects of interest as "fixed" or as "random." In this subsection, I adopt the random-effects approach for three reasons. First, I am interested in measures of practice variation, which are moments of a *distribution* of trainee effects, specifically the standard deviation of trainee effects across trainees in a given tenure period. Random effect estimation directly focuses on this measure, while fixed effect estimation focuses on individual trainee effects.

Second, relatedly, I observe a finite number of observations for each trainee and in each tenure period. Importantly, this number of observations may vary for different trainees and in different tenure periods. Random-effects estimation directly accounts for this by estimating a "prior distribution" of trainee effects by maximum likelihood. Empirical Bayes posteriors may then be obtained for each tenure-specific trainee effect, using the estimated prior and the data for each trainee and in each tenure period. This procedure will "shrink" information from the data toward the prior mean, in a way that minimizes prediction errors of trainee effects (Morris, 1983; Searle et al., 1992). In contrast, OLS estimates of a given trainee (fixed) effect make no use of information on other trainee effects, and naive (unshrunken) fixed-effect estimates of trainee effects will overstate practice variation relative to the truth.[16]

Finally, under Assumption 1, the random-effects approach is free of any notion of "connected sets" that is required under the fixed-effects setup of Abowd et al. (1999). In fixed-effects estimation, one junior trainee effect and one senior trainee effect must be dropped within each connected set in order to satisfy a rank condition. Trainees belonging to different connected sets thus cannot be compared. In finite samples, when I consider trainee effects that are tenure-specific, connections between trainees will become increasingly sparse.[17] In Appendix A-2, I compare Assumption 1 with

[16]For example, consider the simple model $Y_i = \xi_{j(i)} + \varepsilon_i$, where $\xi_j \sim N(0,1)$, $\varepsilon_i \sim N(0,1)$, and $E\left[\varepsilon_i \middle| \xi_{j(i)}\right] = 0$. Consider n observations for each agent. The estimated fixed effect for j will be $\hat{\xi}_j = n^{-1}\sum_{i:j(i)=j} Y_i$, which will be measured with error. The standard deviation of estimated agent fixed effects will be $\sqrt{1+1/n}$, which is an overestimate of the true practice variation of 1. Similarly, the difference between the fixed effect for any two agents is on average $E_{j,j'}\left[\hat{\xi}_j - \hat{\xi}_{j'}\right] = \sqrt{4\pi^{-1}(1+1/n)}$, while the difference between true effects should be $\sqrt{4\pi^{-1}}$ (Geary, 1935). Card et al. (2013) acknowledge this point in Section V.B but abstract away from this finite-sample bias using the argument that n is roughly fixed across variances they wish to compare; this is not the case in our empirical setting.

[17]Trainees switch services every week. So, in the limit, if I were to estimate a fixed effects model using only a week of

a related fixed-effects assumption in Abowd et al. (1999).

3.4 Baseline Implementation

In the random-effects approach, I estimate by maximum likelihood underlying *population* moments of trainee effects that would be consistent with the observed data. In the baseline estimation, I focus on the standard deviation of trainee effects, conditional on the trainee's tenure τ and on the teammates tenure τ^{-}, or $\sigma\left(\tau;\tau^{-}\right)$. I specify discrete tenure periods of 60 days for trainees in their first or second year of residency, or periods of 120 days for trainees in their third year of residency, since training in the third year involves fewer days spent on clinical activities.[18]

To improve the robustness of the maximum likelihood estimation, I first form a risk-adjusted measure of log spending, $\tilde{Y}_{it} = Y_{it} - \left(\mathbf{X}_{it}\hat{\beta} + \mathbf{T}_{t}\hat{\eta} + \hat{\gamma}_{s(i,t)} + \hat{\zeta}_{\ell(i,t)}\right)$, where the vector of parameters $\left(\hat{\beta},\hat{\eta},\hat{\gamma}_{s},\hat{\zeta}_{\ell}\right)$ is estimated by OLS. This approach is a version of restricted maximum likelihood (REML), which avoids the incidental parameters problem in the later maximum-likelihood stage (Patterson and Thompson, 1971). Importantly, as in Chetty et al. (2014), I estimate these OLS parameters using variation *within* interactions of trainee pairs and discrete tenure periods, which allows the remaining trainee effects in $\tilde{Y}_{it}$ to be correlated with the predicted portion of log spending due to $\mathbf{X}_{it}$, $\mathbf{T}_{t}$, $s\left(i,t\right)$, and $\ell\left(i,t\right)$.

I then specify a crossed random effects model,

$$\tilde{Y}_{it} = \xi^{\tau_j;\tau_k}_{j(i,t)} + \xi^{\tau_k;\tau_j}_{k(i,t)} + \varepsilon_{it}. \tag{4}$$

Fixing τ_j and τ_k, I aim to simultaneously estimate $\sigma\left(\tau_j;\tau_k\right)$ and $\sigma\left(\tau_k;\tau_j\right)$ by restricting estimation of Equation (4) to the set of observations $\mathscr{C}\left(\tau_j,\tau_k\right) = \left\{(i,t) : \tau\left(j\left(i,t\right),t\right) = \tau_j, \tau\left(k\left(i,t\right),t\right) = \tau_k\right\}$. I can recover the full set of possible standard deviations, $\left\{\sigma\left(\tau;\tau^{-}\right)\right\}_{(\tau,\tau^{-})}$, by considering different sets of observations corresponding to combinations of τ_j and τ_k. In this way, I impose no functional form on the shape of practice variation over time, since practice variation in each pair of junior-senior tenure periods is estimated on a separate sample of observations.

data in which no trainees switch, then in fact *no* week-specific trainee effects would be identifiable.

[18] I observe approximately half as many patient-days for trainees in the third year, because third-year trainees spend more time in research and electives than in the first two years of training.

Equation (4) can be stated in matrix form:

$$\tilde{\mathbf{Y}} = \mathbf{D}\mathbf{u} + \varepsilon, \tag{5}$$

where $\tilde{\mathbf{Y}}$ is the vector of differenced outcomes, $\mathbf{D}$ is a selection matrix, and $\mathbf{u}$ is a stacked vector of trainee random effects. Let N be the number of observations, N_J be the number of junior trainees, and N_K be the number of senior trainees in the sample $\mathscr{C}\left(\tau_j, \tau_k\right)$. Then the selection matrix $\mathbf{D}$ is $N \times (N_J + N_K)$ and assigns each observation (i,t) to a junior trainees with tenure τ_j and a senior trainee of tenure τ_k. The vector $\mathbf{u}$ is $(N_J + N_K) \times 1$ and contains the stacked effects of the N_J junior trainees and N_K senior trainees.

Assumption 1 implies that junior and senior trainee effects are independent of each other. So the variance-covariance matrix of $\mathbf{u}$ is diagonal:

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \sigma^2\left(\tau_j;\tau_k\right)\mathbf{I}_{N_J} & \mathbf{0} \\ \mathbf{0} & \sigma^2\left(\tau_k;\tau_j\right)\mathbf{I}_{N_K} \end{bmatrix}.$$

Further assuming that trainee random effects and the error term are normally distributed, the log likelihood function is

$$\mathscr{L} = -\frac{1}{2}\left\{N\log(2\pi) + \log|\mathbf{V}| + \tilde{\mathbf{Y}}'\mathbf{V}^{-1}\tilde{\mathbf{Y}}\right\}, \tag{6}$$

where $\mathbf{V} = \mathbf{D}\mathbf{G}\mathbf{D}' + \sigma_\varepsilon^2\mathbf{I}_N$. In each sample of data $\mathscr{C}\left(\tau_j, \tau_k\right)$, I estimate $\sigma\left(\tau_j;\tau_k\right)$ and $\sigma\left(\tau_k;\tau_j\right)$ by maximizing Equation (6).

The estimated variance components can be treated as empirical Bayes prior distributions. Treating $\tilde{\mathbf{Y}}$ as data, I can obtain empirical Bayes posterior means as

$$\tilde{\mathbf{u}} = \tilde{\mathbf{G}}\mathbf{D}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Y}},$$

where $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{V}}$ are $\mathbf{G}$ and $\mathbf{V}$ with random-effects estimates of $\sigma^2\left(\tau_j;\tau_k\right)$, $\sigma^2\left(\tau_k;\tau_j\right)$, and σ_ε^2 plugged in. These posterior means are also known as "best linear unbiased predictions" or BLUPs (Searle et al., 1992).

In Appendix A-3, I detail two extensions of the baseline model. First, I allow for patient admission random effects, since most patients are admitted for multiple days and may be cared for by multiple

trainees. Results are qualitatively unchanged when including patient random effects. Second, I allow for estimation of the correlation between trainee effects of the same trainee in different tenure periods, which I employ in Section 5.2.

4 Results

4.1 Baseline Results

Figure 1 presents results for practice variation from the baseline implementation described in Section 3.4. For each tenure interval τ_h, the figure displays an estimate of practice variation, or the estimated standard deviation of trainee effects among trainees with in the given tenure period.[19] A standard-deviation increase in the effect of junior and senior trainees increases daily total spending by about 5% and 24%, respectively. The difference in practice variation between junior and senior trainees occurs entirely and discontinuously at the one-year tenure mark. Changes in practice variation are otherwise muted. In particular, there exists little convergence in practice styles within either the junior role or the senior role. After the one-year discontinuity, the standard deviation of the trainee effect distribution remains above 20% throughout. Including or omitting admission-level random effects for the patient does not qualitatively alter results.

These results suggest that team decision-making is highly concentrated among much fewer agents than would be the case with independent physician practice. One way to quantify this concentration is to consider two-agent teams of one junior trainee and one senior trainee for each case. In this construction, the senior trainee is responsible for $\frac{0.24^2}{0.05^2+0.24^2} \approx 96\%$ of the variance in team-level decisions across cases. This degree of concentration is even higher when accounting for the fact that a single senior trainee works with two junior trainees: In this case, the practice variation due to each of the two junior trainees are orthogonal to each other, but the common single senior trainee will drive practice variation for all patients under her span of control. Senior agents explain 99% of decision-making variance given this construction.

[19] As described in Section 3.2, senior trainees of tenure τ_k only work with junior trainees of a given tenure $\tau_j = \tau_k - \lfloor \tau_k \rfloor$, where τ_j and τ_k are stated in continuous years. Junior trainees of tenure τ_j may work with senior trainees of tenure $\tau_j + 1$ or $\tau_j + 2$. Differences between $\sigma\left(\tau_j; \tau_j + 1\right)$ and $\sigma\left(\tau_j; \tau_j + 2\right)$ are small and statistically insignificant. I therefore average these two estimates to plot practice variation for tenure τ_j.

4.2 Robustness

I perform two robustness exercises to evaluate the validity of the baseline results. In the first robustness exercise, I address the institutional fact that a group of junior trainees known as "preliminary interns" who are not scheduled to continue in the same internal medicine residency but instead will switch to other specialties (e.g., anesthesiology, dermatology, anesthesiology) after their first year of training. While these trainees make up a minority of the overall sample of trainees, if they as a group have lower practice variation than the remaining group, then their inclusion in the analysis could bias downward an estimate of the practice-variation discontinuity at the one-year tenure mark for a *fixed* group of trainees. Since the identities of preliminary interns are known in advance, I exclude preliminary interns and re-estimate the practice variation profile. Results of this robustness exercise, shown in Panel A of Figure 2, are qualitatively unchanged.

Second, I consider the possibility that patients may not be quasi-randomly assigned to trainee teams. In particular, although Appendix A-1 supports Assumption 1 in terms of observable patient characteristics, patients may differ along unobservable characteristics across different trainee teams. Since there is only one senior trainee on each team, compared to two junior trainees, systematic sorting of patients across teams would not only bias estimates of trainee effects but would also spuriously induce greater practice variation among senior trainees. To assess this possibility, I re-estimate the practice variation profile with no patient controls. As shown in Panel B of Figure 2, the practice variation profile from this exercise also remains qualitatively unchanged from the baseline implementation. This shows that including or removing rich patient controls has no qualitative effect on the key moments of practice variation and is consistent with the causal interpretation of trainee effects implied by Assumption 1.

4.3 Team Concerns

Variation in trainee effects may reflect two conceptual objects: (i) differences in trainee *judgments*, if they were allowed to make decisions on their own, and (ii) *influence* on team decisions, or the extent to which trainees may sway team decisions. Judgments may reflect prior knowledge, beliefs, or preferences, and exist outside of a team setting. In contrast, if agents make decisions independently, then they will have full and invariant influence. In other words, for influence to matter for practice

variation, teams must alter the process of decision-making.

I consider three types of "team concerns." First, in team theory, organizations allocate decisions to individuals under bandwidth constraints (Marschak and Radner, 1972; Garicano, 2000). Organizations are naturally hierarchical, and higher levels in this hierarchy, where agents have a greater "span of control," handle fewer decisions. Second, in principal-agent models, teams may induce "herding" of decisions around the beliefs of senior agents, simply based on the prestige, rank, or power of senior agents (Scharfstein and Stein, 1990; Prendergast, 1993). Junior agents may act as "yes men" to further their careers, at the cost of worse team decisions. Third, teams may aggregate information, as agents may confer with each other before making team decisions. Joint decisions cannot be fully separated and distributed to individual agents but instead pool input across agents.

What do team concerns imply for the empirical pattern of practice variation with respect to tenure in medical residency? I consider two features of medical residency to answer this question and summarize informal implications in Table 3. The first feature concerns the structure of residency teams relative to the one-year tenure mark. When trainees pass the one-year mark, their span of control, rank, and relative experience all discontinuously increase. Assuming that judgments and preferences are continuous across the one-year tenure mark, any discontinuity in practice variation reflects the impact of influence via team concerns. Limited bandwidth would imply a *decrease* in practice variation at the one-year discontinuity, since senior trainees have greater span of control. However, either career concerns or information aggregation would imply an *increase* in practice variation at this discontinuity. If no team concerns are at play (i.e., physicians practice independently), then there should be no discontinuity in practice variation at the one-year mark.

The second feature concerns implications for practice convergence under learning, since an immense amount of learning occurs in medical residency, at least according to qualitative reports (Ludmerer, 2014). If physicians practice independently or if decisions are separable across trainees, learning would imply convergence in practices over time (i.e., practice variation should decrease with tenure).[20] But if teams aggregate information across agents in joint decisions, then influence may grow endogenously as judgments become more precise, and there may be no practice convergence despite dramatic learning in residency.

[20] In classical team theory, an agent knows how to solve a problem completely or not at all, problems are fully separable across agents, and the organization is structured so that problems are distributed efficiently to the proper agents.

5 Mechanisms

In this section, I delve further into mechanisms that may underlie the basic empirical results that (i) influence jumps discontinuously when trainees assume the senior role, and that (ii) convergence in practice variation is generally muted even as trainees progress in residency. I consider two types of mechanisms introduced in Section 4.3. First, senior trainees may arbitrarily exercise greater influence, regardless of their knowledge. For example, they may hold prestige, rank, or power that is unrelated to knowledge, or they may simply have decision rights in their jobs for "important" decisions. Second, influence in teams may depend on systematic differences in knowledge between teammates with different tenures. This mechanism might allow for differences in knowledge that are simply correlated across tenure groups, as opposed to within tenure groups. Thus, this mechanism may arise even in the case that knowledge in individual cases is not directly observable, if general relationships between tenure and knowledge are known.

The analyses in this section proceed in three parts under the following reasoning. First, if junior and senior trainees simply have different jobs with different decisions rights, then in the Garicano (2000) model, junior trainees could have greater control over some types of decisions. On the other hand, if all decisions require knowledge gained with experience, then we should find the same practice variation profile over all types of decisions. I will thus first explore heterogeneity across different types of decisions. Second, in order for the first category of mechanisms to *fully* explain the practice variation pattern in Figure 1, there must be close to no learning, since there is almost no convergence in practice variation. I will therefore examine the extreme proposition of no learning. Third, I will examine the opposite proposition that influence is optimally allocated according to knowledge. While the two mechanisms are not mutually exclusive, these analyses may shed light on the relative importance of each mechanism.

5.1 Decision Types

I re-estimate practice variation profiles, using the same approach described in Section 3, by subsetting decisions in several ways. First, I consider decisions in the four main clinical cost departments of diagnosis (radiology and laboratory), medication, blood transfusion, and nursing. Rather than aggregating the direct costs of all orders for a given patient-day case (i,t), I only aggregate the direct costs

of the subset of orders in a given clinical category. Second, I consider how practice variation profiles may differ by patient severity or whether decisions are early vs. late in a patient's stay. Finally, I subset cases $(i,t) \in \mathscr{C}$ according to formal diagnostic codes, grouped by the frequency of the diagnostic code or by whether there exists a formal guideline for the diagnostic code in `guidelines.gov`.

In all of these cases, the practice variation profile is qualitatively similar: Variation increases discontinuously at the one-year tenure mark and remains stable to the end of training. Figure 3 shows practice variation profiles across different clinical cost categories. Figures 4 and 5 show virtually identical practice variation profiles across patient severity, patient-days that earlier or later in a patient's stay, and patients with different formal diagnoses.[21]

Despite qualitative similarities in Figure 3, the magnitudes of practice variation and its discontinuous increase at the one-year mark do vary meaningfully across clinical cost categories. Diagnostic spending shows the largest increase in practice variation, with a standard deviation of 16% to 74% before and after the one-year tenure mark. In contrast, medication and nursing spending shows relatively small practice variation, both overall and in the increase at the relative experience discontinuity. These differences may be consistent with greater uncertainty and greater control by trainees of diagnostic and transfusion decisions.[22] On the other hand, decisions types with greater proportional increases in influence at the one-year tenure mark do not account for a larger share of total spending (see Table 2 for summary statistics by clinical cost department). Given the magnitudes of trainee effects on overall spending (Figure 1), this suggests spillovers across clinical cost categories, driven by interconnected decisions.

[21] Interestingly, practice variation is remarkably similar between formal diagnoses with and without formal guidelines. This possibly reflects the coarseness of formal diagnoses and formal guidelines. For example, "Chest pain, not otherwise specified" is the most common formal diagnostic code both for patients admitted to general medicine and for patients admitted to the subspecialty cardiology service. The coarseness of formal diagnostic codes and a review of the guidelines strongly suggest that very little meaningful clinical information can be formally encoded (Shaneyfelt et al., 1999). Another explanation is that, while guidelines may decrease practice uncertainty, diagnoses with more uncertainty may warrant guidelines.

[22] Medication decisions are better described in publicly accessible sources of knowledge, while diagnostic decisions draw more on clinical reasoning that would be difficult to pre-specify and reference for trainees who have never before encountered a patient presentation. Similarly, blood transfusion reflects an important decision with large variation across providers and surprisingly little guidance for how to tailor the transfusion decision to individual cases (Carson et al., 2016). On the other hand, nursing decisions are intuitively outside the scope of most physician decision-making, and it seems intuitive that physician trainees will have little influence on these decisions.

5.2 No-Learning Scenario

I next evaluate the extreme case in which the influence differential between senior and junior trainees is unrelated to any differential in knowledge. This case is tantamount to no learning in residency. Instead, physicians may differ in their judgments, due for example to heterogeneous preferences or beliefs, in ways that predate residency and are time-invariant during residency. Given the intensity of residency training, this scenario seems unlikely on its face. However, a general version of intrinsic heterogeneity that is relatively stable over time has been invoked in many settings, several of them in health care (e.g., Doyle et al., 2010; Fox and Smeets, 2011; Bartel et al., 2014; Currie and MacLeod, 2017). I therefore evaluate the relative importance of time-invariant heterogeneity in explaining practice variation using two complementary approaches.

In the first approach, I exploit detailed trainee characteristics that should be highly correlated with preferences and ability, including demographics, prior formal degrees, place of medical school, standardized examination scores, position on the rank list, and future income. Indeed, these characteristics are the key summary statistics considered by residency programs in accepting future physicians and may represent important differences in ability and future careers. Empirically, I show that *ex ante* trainee characteristics strongly predict position on the rank list (i.e., desirability to the residency program) and the probability of higher-than-median future income, which is at least 50% greater than the future income below median.[23] However, despite these important relationships between trainee characteristics and career-changing outcomes, I strikingly find that these trainee characteristics are broadly uncorrelated with trainee effects on clinical decisions. In Figure 6, I show the *distribution* of trainee effects in each tenure period throughout residency is also unchanged regardless of conditioning on trainees rank or future income. I describe these analyses further in Appendix A-4 and present more exhaustive results in Table 4.[24]

In the second approach, I measure the serial correlation between random trainee effects in two different tenure periods and provide further details of the statistical approach in Appendix A-3.2. The

[23] Trainees with a predictive score one standard-deviation above mean are two to three times more likely to be ranked in the upper half of the rank list than those with a predictive score one standard-deviation below mean. Trainees with a predictive score one standard-deviation above mean are more than three times as likely to obtain above-median future income than those with a predictive score one-standard deviation below mean.

[24] In Appendix A-5, I also explore whether trainee practice styles can be predicted by supervising physicians and senior trainees whom they have worked with in the past. Interestingly, practice variation is orthogonal to the practice styles of these past teammates. This suggestive evidence is consistent with tacit knowledge and experiential learning.

conceptual reason for examining serial correlation is as follows: If practice variation reflects intrinsic heterogeneity and no learned beliefs, then effects in different time periods within the same trainee should be constantly and highly correlated, regardless of the time between the time periods. However, if trainees are learning, then adjacent time periods should exhibit higher correlation in trainee effects than do distant time periods. Figure 7 presents averages of serial correlation estimates between trainee effects as a function of the distance between the tenure periods.[25] Serial correlation in trainee effects across two adjacent periods is moderately positive, while the correlation quickly decreases to zero with more distance between the two periods. Interestingly, correlation eventually becomes negative, though statistically indistinguishable from 0, between trainee effects in distant periods. These results strongly suggest that judgments during residency are quite dynamic. In other words, consistent with numerous qualitative accounts, trainees are engaging in active learning during residency.

5.3 Optimal Influence

In the other extreme, I consider a simple model of optimal influence by Bayesian information aggregation. In order to optimize the decision at hand, teams allocate influence in proportion to the knowledge of each team member (DeGroot, 2005). The more precise the signal from her prior knowledge relative to other sources of information, the greater her influence will be. At the one-year tenure mark, influence discontinuously increases because the knowledge of a trainee's teammate discontinuously decreases.

In this model, as trainees gain knowledge, their judgments will converge, but their influence will increase. These two factors have opposing implications for practice variation, so that practice variation may not always decrease with learning. In contrast, agents who practice independently have constant (full) influence and should always exhibit convergence in their practice styles as they learn. In this way, I use tenure-specific knowledge and Bayesian aggregation in teams to imply tenure-specific practice variation. An assumption of continuous knowledge places restrictions on patterns of influence—and therefore practice variation—over trainees' tenure. A further assumption that supervising physicians possess at least as much knowledge as a senior trainee imposes another restriction on the scale of practice variation. The actual pattern and scale of practice variation therefore allow identification of deviations from optimal influence.

[25] Appendix Table A-4 and Appendix Figure A-4 show results for individual pairs of tenure periods.

In Appendix A-6, I provide details of the model setup, identification, estimation procedure, and results. In brief, the model uses tenure-specific moments of practice variation from the random effects model in Equation (3) to recover underlying primitives of learning (i.e., the rate at which knowledge increases with tenure), as well as potential deviations from optimal influence capturing other team concerns in Section 4.3. I specify influence as divided between the junior trainee, the senior trainee, and "external information," which may be drawn from the supervising physician, any other personnel, or guidelines and protocols.

In results, I find very little knowledge at the beginning of residency compared to learning in the first year. Learning in the second year occurs at a much faster than in the first year but appears to cease by the third year.[26] Between junior and senior trainees, influence approximates the Bayesian benchmark. In likelihood ratio tests, I cannot reject a model with learning and optimal influence between trainees, compared to a less-restrictive model that allows for deviations from optimal influence. However, I find that external information (including the supervising physician) influences decisions by less than half of the influence of a graduating trainee. This suggests that trainees, as a group, are given much more influence than warranted under the Bayesian benchmark. Many of these patterns persist when re-estimating the model with practice variation profiles in specific spending categories and types of cases.

Although the model is highly stylized and is based on relatively few empirical moments, the idea that learning increases when trainees become senior and have a greater stake in decision-making is consistent with experiential learning (Dewey, 1938). Experiential learning implies a tradeoff in the use of information to make team decisions. While supervisory information improves the quality of decision-making at hand, it may constrain experiential learning by trainees. Perhaps for this reason, external information receives much less weight than it should in a Bayesian framework that optimizes only decisions at hand. Appendix A-6 undertakes counterfactual analyses to quantify the welfare consequences of this tradeoff.

[26] There exists a large theoretical literatures on why learning may stop, related to learning costs or knowledge constraints (e.g., Rogerson et al., 2005; Caplin and Dean, 2015), uncertainty in the mapping between beliefs and data (Acemoglu et al., 2006), and social learning (Ellison and Fudenberg, 1993).

6 Discussion and Conclusion

In this paper, I study decision-making in teams, in the setting of physician trainees in medical residency. As in other settings involving teams, I observe decisions attributable to teams and the team members at the time of each decision, but I do not directly observe the agents' contributions to the decisions. Building on a "movers literature" starting with Abowd et al. (1999), I develop and apply a method to extract each team member's average contribution to decisions over time, using quasi-experimental variation in the assignment of cases and physician trainees to teams, as well as frequent switches of trainees across teams. By tracking the effects of trainees on team decisions over their tenure, I also shed light on how teams may alter decision-making relative to agents who make decisions "on their own."

In my primary finding, I show that senior trainees explain the vast majority of practice variation across teams. This suggests that differences across organizations in health care and in other settings may be driven by a few individuals. Furthermore, by exploiting a discontinuity in team roles at the one-year tenure mark, I show that team dynamics are responsible for this outsize influence of senior trainees. From multiple analytical lenses, I find evidence suggestive of an intriguing interplay between experiential learning during residency training and the allocation of influence in teams. While this evidence on its own may be suggestive, it is consistent with a large body of work, much of it outside of economics, suggesting "tacit knowledge" that is difficult to pass on to others (Polanyi, 1966) and "experiential learning" that accrues only through experience (Dewey, 1938).

At a minimum, these results suggest the importance of team concerns in analyzing decisions made within organizations. In health care, a large and influential literature has focused primarily on variation across regions or institutions. If decision-making is concentrated in the hands of a few individuals, then understanding micro-level foundations of decision-making will be essential for characterizing variation that has long been noted at more aggregate levels.

Moreover, if an important task for teams is to aggregate information, particularly for complex and consequential decisions, then policy-makers may need to focus more on the informational frictions that underlie the skewed concentration of knowledge and the remaining practice variation even among experts with the most knowledge. Such informational levers could be more effective at reducing practice variation (Institute of Medicine, 2013), compared to previously proposed policy levers of

financial incentives, simple reporting of variation, and patient cost-sharing (see Skinner, 2012, for a summary). This idea is consistent with a growing literature that suggests that skill, or productivity, plays an important role in practice variation in both diagnostic and treatment decisions.[27] If providers simultaneously under- and over-treat patients, then instituting policy levers to encourage providers to treat either uniformly more or uniformly less will be ineffective. Similarly, imposing a uniform treatment rate can be counterproductive if under-treatment is costlier and if providers who treat more do so because they are less skilled at targeting (Chan et al., 2019). Instead, policy levers need to accomplish better targeting of resources by improving the use of information in decision-making.

Further, if learning requires experience and feedback, then the usual forms of spreading information, such as clinical guidelines, formal instruction (e.g., "continuing medical education"), or formal testing (e.g., board recertification), may do little to change practice or generate consensus (Shaneyfelt et al., 1999). While effective policies are beyond the scope of this paper, such policies will likely need to improve the use of existing information within an organization and to encourage its spread across team members and organizations. For example, process innovations might invite feedback from peers, or they might explicitly use consensus-building to specify nuanced "clinical pathways." These approaches aim to promote learning among "experts"—well beyond residency training—and by organizations themselves (Institute of Medicine, 2012; Bohmer et al., 2013).

References

ABALUCK, J., L. AGHA, C. KABRHEL, A. RAJA, AND A. VENKATESH (2016): "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, 106, 3730–3764.

ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333.

[27] A recent literature in economics has begun to directly consider skill in diagnosis, decision-making, and treatment. Abaluck et al. (2016) investigate whether providers decide to test for pulmonary embolisms and find that misallocation of resources has much larger welfare consequences than systematic overuse; Mullainathan and Obermeyer (2019) similarly find over- and under-testing of heart attack. Currie and MacLeod (2017) show variation in allocation of cesarean sections to patients according to their characteristics ("diagnostic skill") that could be as important as variation in procedural skill. Gowrisankaran et al. (2017) investigate diagnosis and treatment of specific potential conditions in the emergency department. Chandra and Staiger (2017) apply a framework to study variation in spending across hospitals and examine to what extent this variation reflects allocative inefficiency versus comparative advantage.

ABOWD, J. M., F. KRAMARZ, AND S. WOODCOCK (2008): "Econometric Analyses of Linked Employer-Employee Data," in *The Econometrics of Panel Data*, ed. by L. Matyas and P. Sevestre, Springer Berlin Heidelberg, no. 46 in Advanced Studies in Theoretical and Applied Econometrics, 727–760.

ACEMOGLU, D., V. CHERNOZHUKOV, AND M. YILDIZ (2006): "Learning and Disagreement in an Uncertain World," Working Paper 12648, National Bureau of Economic Research.

ARROW, K. J. (1962): "The Economic Implications of Learning by Doing," *The Review of Economic Studies*, 29, 155–173.

BARTEL, A. P., N. BEAULIEU, C. PHIBBS, AND P. W. STONE (2014): "Human Capital and Productivity in a Team Environment: Evidence from the Healthcare Sector," *American Economic Journal: Applied Economics*, 6, 231–259.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

BENKARD, C. L. (2000): "Learning and Forgetting: The Dynamics of Aircraft Production," *American Economic Review*, 90, 1034–1054.

BLOOM, N. AND J. VAN REENEN (2010): "Why Do Management Practices Differ across Firms and Countries?" *Journal of Economic Perspectives*, 24, 203–224.

BOHMER, R. M. J., A. C. EDMONDSON, AND L. FELDMAN (2013): "Intermountain Health Care," Harvard Business School Case 603-066.

CAPLIN, A. AND M. DEAN (2015): "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 105, 2183–2203.

CARD, D., J. HEINING, AND P. KLINE (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, 128, 967–1015.

CARSON, J. L., G. GUYATT, N. M. HEDDLE, B. J. GROSSMAN, C. S. COHN, M. K. FUNG, T. GERNSHEIMER, J. B. HOLCOMB, L. J. KAPLAN, L. M. KATZ, N. PETERSON, G. RAMSEY, S. V. RAO, J. D. ROBACK, A. SHANDER, AND A. A. R. TOBIAN (2016): "Clinical Practice

Guidelines From the AABB: Red Blood Cell Transfusion Thresholds and Storage," *The Journal of the American Medical Association*, 316, 2025.

CHAN, D. C., M. GENTZKOW, AND C. YU (2019): "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," Working Paper 26467, National Bureau of Economic Research.

CHANDRA, A., A. FINKELSTEIN, A. SACARNY, AND C. SYVERSON (2016): "Healthcare Exceptionalism? Productivity and Allocation in the U.S. Healthcare Sector," *American Economic Review*, 106, 2110–2144.

CHANDRA, A. AND D. STAIGER (2017): "Identifying Sources of Inefficiency in Health Care," Tech. Rep. w24035, National Bureau of Economic Research, Cambridge, MA.

CHANDRA, A. AND D. O. STAIGER (2007): "Productivity spillovers in healthcare: evidence from the treatment of heart attacks," *The Journal of Political Economy*, 115, 103–140.

CHARLSON, M. E., P. POMPEI, K. L. ALES, AND C. R. MACKENZIE (1987): "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, 40, 373–383.

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

COOPER, Z., S. CRAIG, M. GAYNOR, AND J. VAN REENEN (2015): "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured," Tech. Rep. w21815, National Bureau of Economic Research, Cambridge, MA.

CURRIE, J. AND J. GRUBER (1996): "Saving Babies: The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women," *Journal of Political Economy*, 104, 1263–1296.

CURRIE, J. AND W. B. MACLEOD (2017): "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians," *Journal of Labor Economics*, 35, 1–43.

CUTLER, D. (2010): "Where Are the Health Care Entrepreneurs?" *Issues in Science and Technology*, 27, 49–56.

CUTLER, D., J. SKINNER, A. D. STERN, AND D. WENNBERG (2018): "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," *American Economic Journal: Economic Policy*, Forthcoming.

DEGROOT, M. H. (2005): *Optimal Statistical Decisions*, John Wiley & Sons, google-Books-ID: dtVieJ245z0C.

DEWEY, J. (1938): *Experience and Education*, New York: Kappa Delta Pi.

DOYLE, J. J., S. M. EWER, AND T. H. WAGNER (2010): "Returns to physician human capital: Evidence from patients randomized to physician teams," *Journal of Health Economics*, 29, 866–882.

DOYLE, J. J., J. A. GRAVES, J. GRUBER, AND S. KLEINER (2015): "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, 123, 170–214.

ELIXHAUSER, A., C. STEINER, D. R. HARRIS, AND R. M. COFFEY (1998): "Comorbidity Measures for Use with Administrative Data," *Medical Care*, 36, 8–27.

ELLISON, G. AND D. FUDENBERG (1993): "Rules of thumb for social learning," *Journal of Political Economy*, 101, 612–643.

EPSTEIN, A. J. AND S. NICHOLSON (2009): "The formation and evolution of physician treatment styles: an application to cesarean sections," *Journal of Health Economics*, 28, 1126–1140.

FINKELSTEIN, A., M. GENTZKOW, AND H. WILLIAMS (2016): "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *Quarterly Journal of Economics*, 131, 1681–1726.

FISHER, E. S., D. E. WENNBERG, T. A. STUKEL, D. J. GOTTLIEB, F. L. LUCAS, AND E. L. PINDER (2003a): "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care," *Annals of Internal Medicine*, 138, 273–287.

——— (2003b): "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine*, 138, 288–298.

FOX, J. T. AND V. SMEETS (2011): "Does Input Quality Drive Measured Differences in Firm Productivity?" *International Economic Review*, 52, 961–989.

GARICANO, L. (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108, 874–904.

GARTNER, A., M. C. KOHLER, AND F. RIESSMAN (1971): *Children teach children: learning by teaching*, Harper & Row.

GEARY, R. C. (1935): "The Ratio of the Mean Deviation to the Standard Deviation as a Test of Normality," *Biometrika*, 27, 310–332.

GELMAN, A. AND J. HILL (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

GOWRISANKARAN, G., K. JOINER, AND P.-T. LEGER (2017): "Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments," Tech. Rep. w24155, National Bureau of Economic Research, Cambridge, MA.

HAYEK, F. A. (1945): "The Use of Knowledge in Society," *The American Economic Review*, 35, 519–530.

HENDEL, I. AND Y. SPIEGEL (2014): "Small Steps for Workers, a Giant Leap for Productivity," *American Economic Journal: Applied Economics*, 6, 73–90.

INSTITUTE OF MEDICINE (2012): "Best Care at Lower Cost: The Path to Continuously Learning Health Care in America," Tech. rep., National Academies Press, Washington, D.C.

——— (2013): *Variation in Health Care Spending: Target Decision Making, Not Geography*, National Academies Press.

KAHN, L. B. AND F. LANGE (2014): "Employer Learning, Productivity, and the Earnings Distribution: Evidence from Performance Measures," *The Review of Economic Studies*, 81, 1575–1613.

KOLB, D. A. AND R. FRY (1975): "Toward an applied theory of experiential learning," in *Theories of Group Process*, ed. by C. Cooper, London: Wiley.

LAZEAR, E. P., K. L. SHAW, AND C. STANTON (2015): "The Value of Bosses," *Journal of Labor Economics*, 33.

LEVITT, S. D., J. A. LIST, AND C. SYVERSON (2013): "Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant," *Journal of Political Economy*, 121, 643–681.

LIZZERI, A. AND M. SINISCALCHI (2008): "Parental Guidance and Supervised Learning," *Quarterly Journal of Economics*, 123, 1161–1195.

LUDMERER, K. M. (2014): *Let Me Heal: The Opportunity to Preserve Excellence in American Medicine*, New York: Oxford University Press.

MANSKI, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60, 531–542.

MARSCHAK, J. AND R. RADNER (1972): *Economic Theory of Teams*, New Haven, CT: Yale University Press.

MAS, A. AND E. MORETTI (2009): "Peers at Work," *The American Economic Review*, 99, 112–145.

MCCARTHY, D. AND D. BLUMENTHAL (2006): "Stories from the sharp end: case studies in safety improvement," *Milbank Quarterly*, 84, 165–200.

MINCER, J. (1962): "On-the-Job Training: Costs, Returns, and Some Implications," *Journal of Political Economy*, 70, 50–79.

MOLITOR, D. (2017): "The evolution of physician practice styles: Evidence from cardiologist migration," *American Economic Journal: Economic Policy*, 10, 326–356.

MONTESSORI, M. (1948): *The Discovery of the Child*, Madras: Kalkshetra Publications Press., google-Books-ID: G3EvGGUKS14C.

MORRIS, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55.

MULLAINATHAN, S. AND Z. OBERMEYER (2019): "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions," Working Paper 26168, National Bureau of Economic Research.

OTTAVIANI, M. AND P. SORENSEN (2001): "Information aggregation in debate: who should speak first?" *Journal of Public Economics*, 81, 393–421.

PATTERSON, H. D. AND R. THOMPSON (1971): "Recovery of inter-block information when block sizes are unequal," *Biometrika*, 58, 545–554.

PHELPS, C. E. AND C. MOONEY (1993): "Variations in medical practice use: causes and consequences," *Competitive Approaches to Health Care Reform*, 139–175.

PIAGET, J. (1971): *Psychology and Epistemology: Towards a Theory of Knowledge*, New York: Grossman.

POLANYI, M. (1966): *The Tacit Dimension*, New York: Doubleday Press.

PRENDERGAST, C. (1993): "A Theory of Yes Men," *The American Economic Review*, 83, 757–770.

RINARD, J. R. AND R. C. MAHABIR (2010): "Successfully Matching Into Surgical Specialties: An Analysis of National Resident Matching Program Data," *Journal of Graduate Medical Education*, 2, 316–321.

ROGERSON, R., R. SHIMER, AND R. WRIGHT (2005): "Search-Theoretic Models of the Labor Market: A Survey," *Journal of Economic Literature*, 43, 959–988.

SCHARFSTEIN, D. S. AND J. C. STEIN (1990): "Herd Behavior and Investment," *The American Economic Review*, 80, 465–479.

SEARLE, S. R., G. CASELLA, AND C. E. MCCULLOCH (1992): *Variance Components*, Wiley New York.

SHANEYFELT, T. M., M. F. MAYO-SMITH, AND J. ROTHWANGL (1999): "Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature," *The Journal of the American Medical Association*, 281, 1900–1905.

SKINNER, J. (2012): "Causes and Consequences of Regional Variations in Healthcare," in *Handbook of Health Economics*, ed. by M. V. Pauly, T. G. McGuire, and P. Barros, San Francisco: Elsevier, vol. 2, 49–93.

SKINNER, J. AND D. STAIGER (2015): "Technology Diffusion and Productivity Growth in Health Care," *Review of Economics and Statistics*, 97, 951–964.

SONG, H. AND R. S. HUCKMAN (2018): "Cohort Turnover and Operational Performance: The July Phenomenon in Teaching Hospitals," SSRN Scholarly Paper ID 3037753, Social Science Research Network, Rochester, NY.

TOPEL, R. (1991): "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority," *Journal of Political Economy*, 99, 145–176.

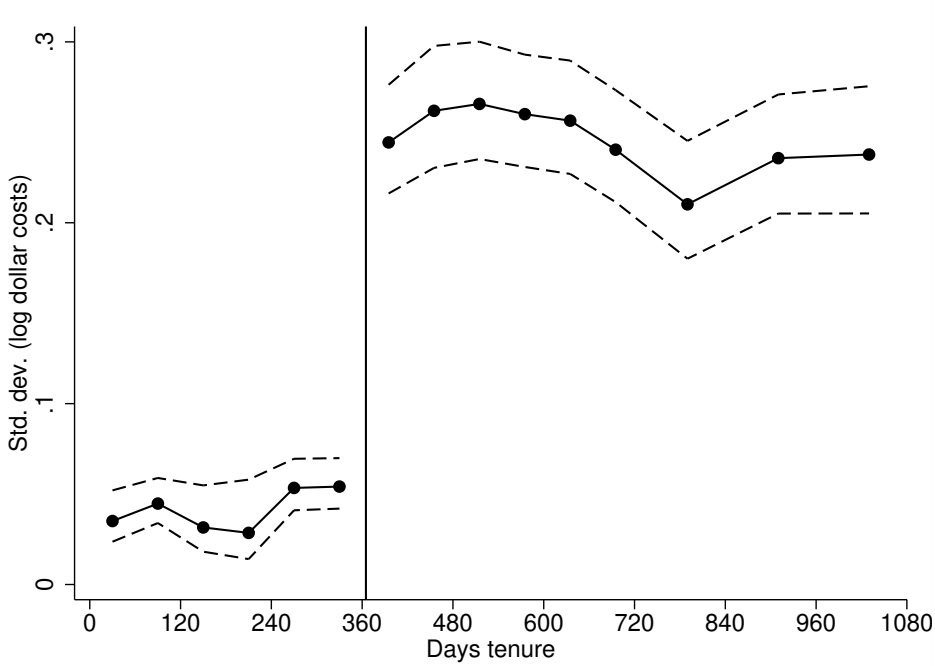
VAN PARYS, J. AND J. SKINNER (2016): "Physician Practice Style Variation: Implications for Policy," *JAMA Internal Medicine*, 176, 1549.

VAN ZANDT, T. (1998): "Organizations with an Endogenous Number of Information Processing Agents," in *Organizations with Incomplete Information: Essays in Economic Analysis*, ed. by M. Majumdar, Cambridge, UK: Cambridge University Press.

WOOD, D. F. (2003): "ABC of learning and teaching in medicine: Problem based learning," *BMJ*, 326, 328–330.

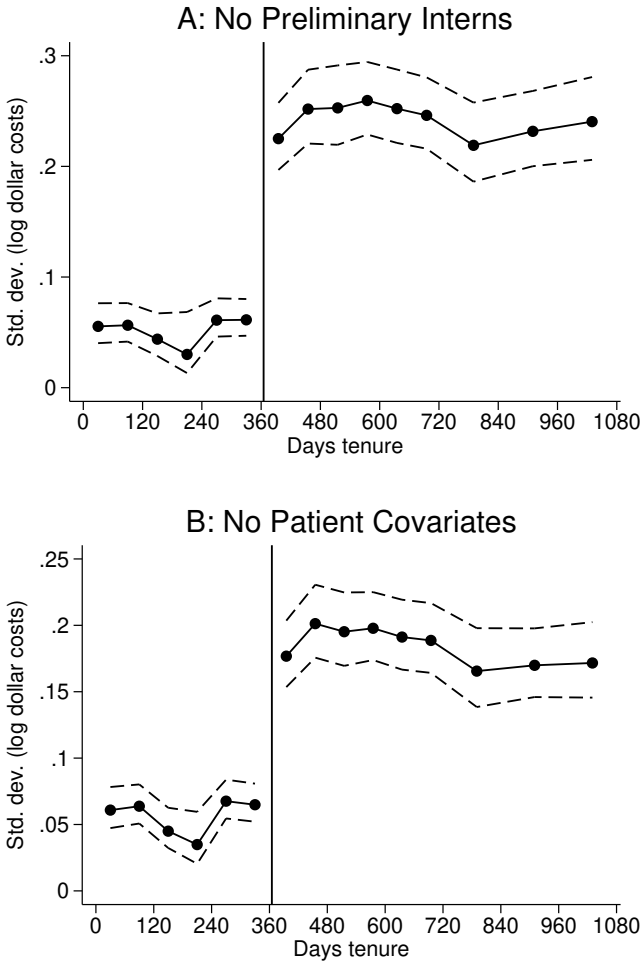
YOUNG, J. Q., S. R. RANJI, R. M. WACHTER, C. M. LEE, B. NIEHAUS, AND A. D. AUERBACH (2011): ""July effect": impact of the academic year-end changeover on patient outcomes: a systematic review," *Annals of Internal Medicine*, 155, 309–315.

Figure 1: Profile of Practice Variation by Tenure



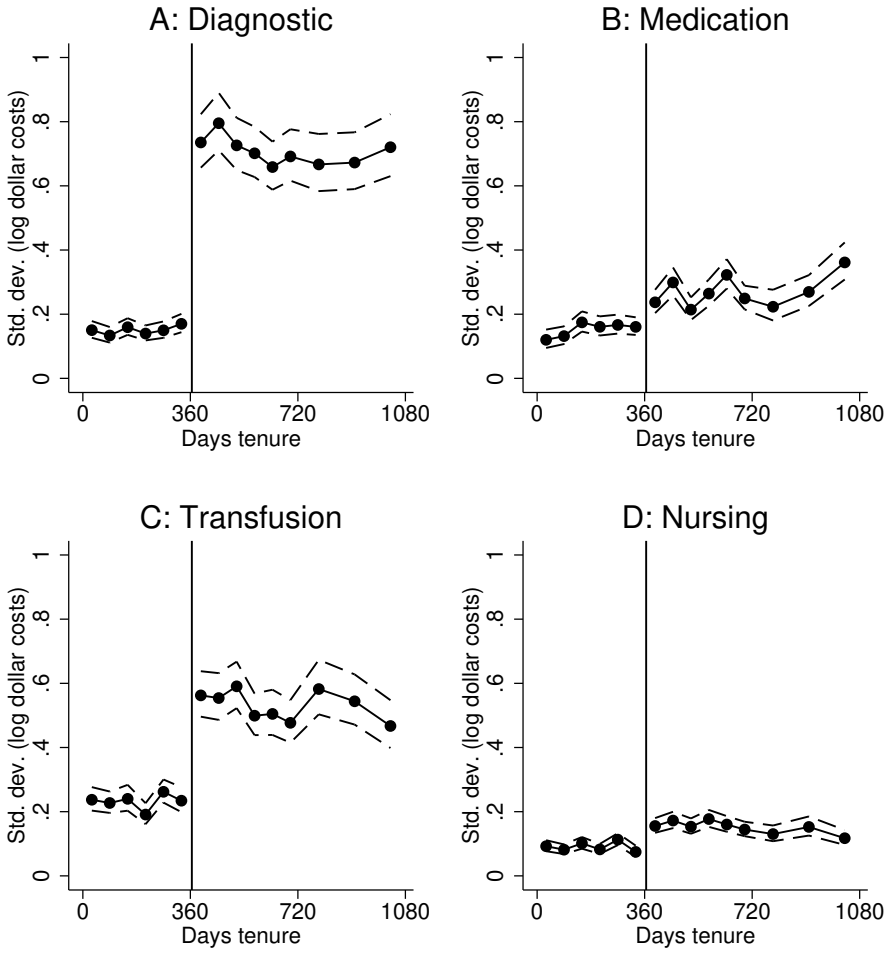
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily total costs at each non-overlapping tenure period. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark. The model controls for patient and admission observable characteristics, time dummies (month-year interactions, day of the week), and attending identities (as fixed effects). Patient characteristics include demographics, Elixhauser indices, Charlson comorbidity scores, and DRG weights. Admission characteristics include the admitting service (e.g., "Heart Failure Team 1"). Estimates for junior trainees are done separately for second-year senior trainees and for third-year senior trainees, then subsequently averaged for purposes of presentation. An alternative approach estimating junior-trainee practice variation by pooling observations by junior-trainee tenure yields qualitatively similar results.

Figure 2: Robustness of Baseline Results



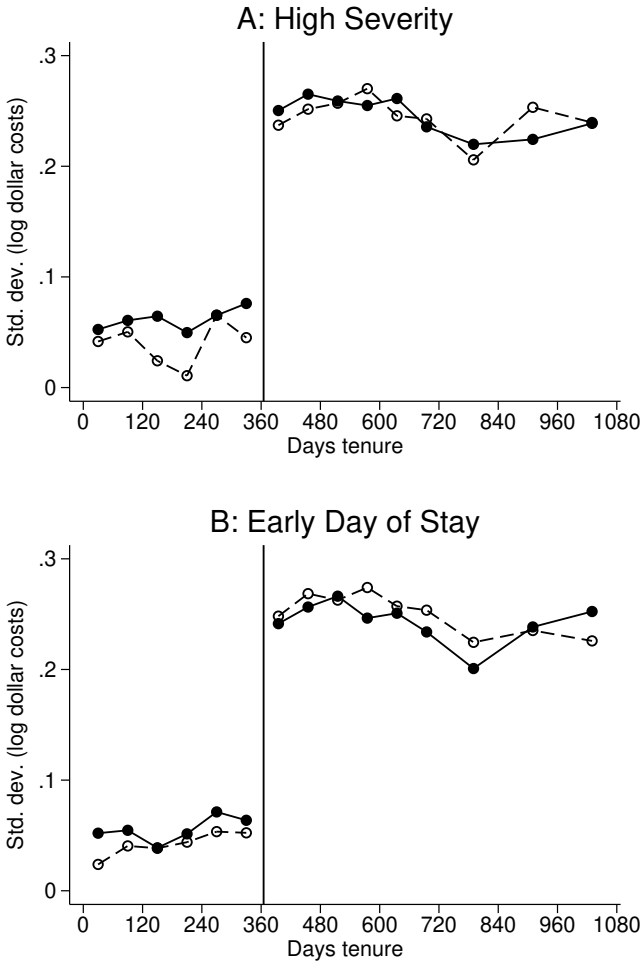
Note: This figure shows robustness of baseline results shown in Figure 1 along two dimensions. In Panel A, I drop "preliminary interns," or junior trainees who are not scheduled to continue as senior trainees in internal medicine. This leaves only trainees that will continue on as senior trainees in internal medicine. In Panel B, I estimate the model with no patient characteristics as covariates. The estimation approach is otherwise the same as for Figure 1. The model estimates practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily total costs at each non-overlapping tenure period. The model controls are as stated for Figure 1. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 3: Practice Variation Profile by Spending Category



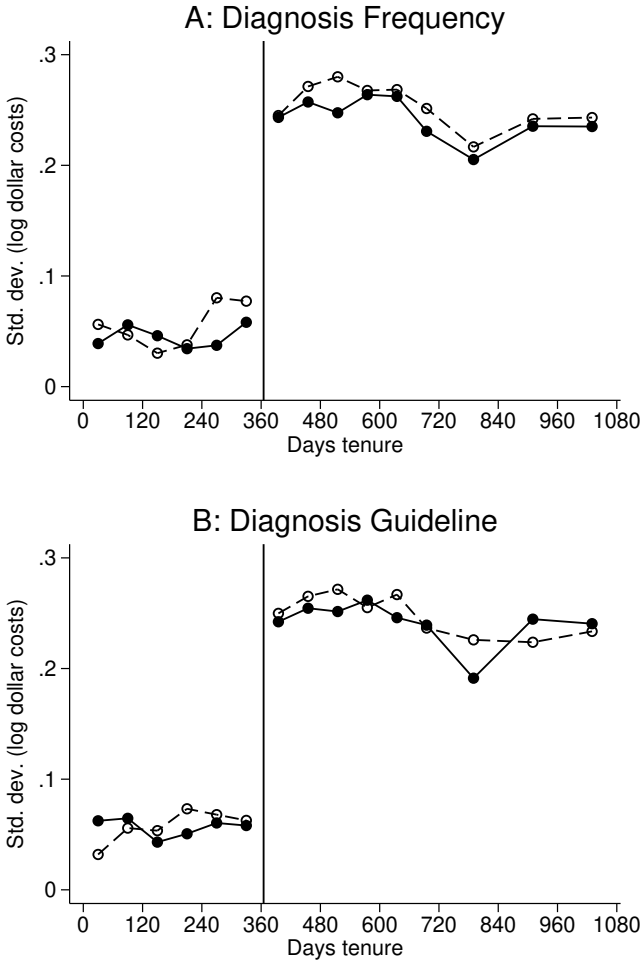
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily costs at each non-overlapping tenure period. Each panel shows a different spending category. The model controls are as stated for Figure 1. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 4: Practice Variation Profile by Patient Severity and Day of Stay



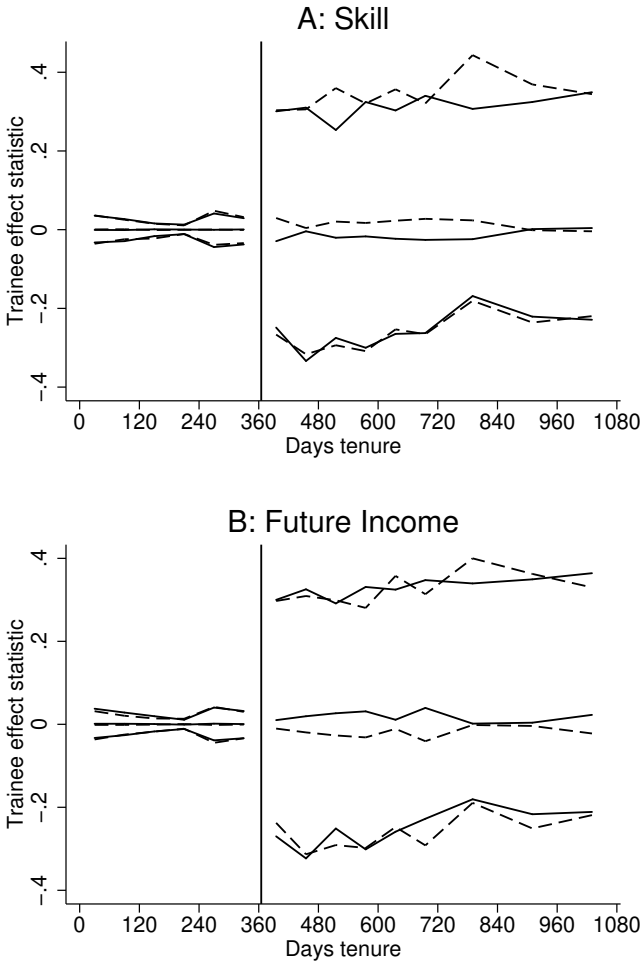
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily total costs at each non-overlapping tenure period. Panel A estimates the model separately in two samples of patients with above- (solid dots) versus below-median (hollow dots) expected 30-day mortality. Panel B estimates the model separately in two samples of days before (solid dots) versus after (hollow dots) the middle of each patient's stay (with the middle day, if it exists, randomized between the two groups). The model controls are as stated for Figure 1. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 5: Practice Variation Profile by Diagnosis Type



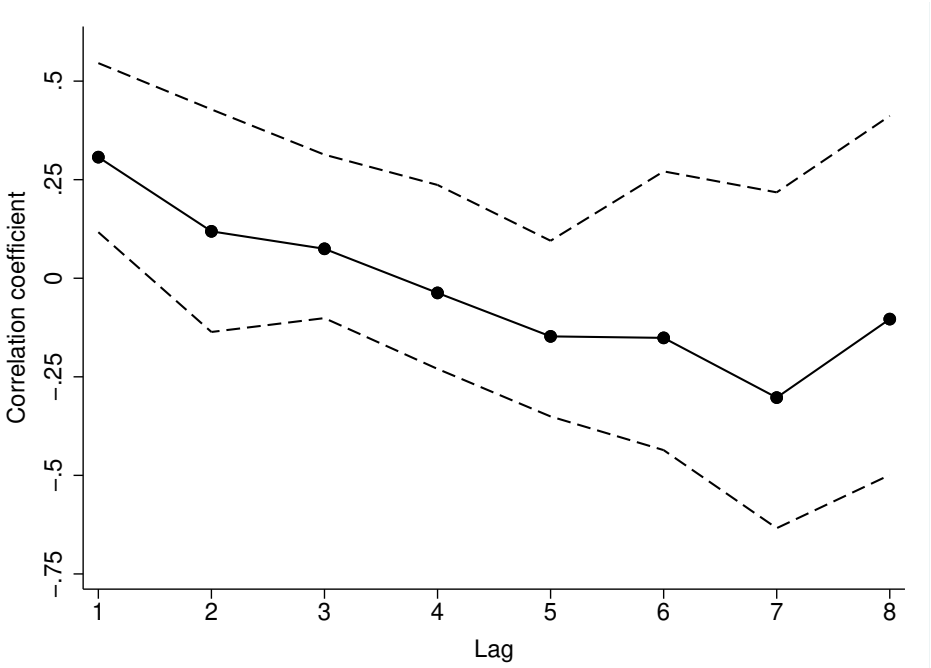
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily total costs at each non-overlapping tenure period. Panel A estimates the model separately in two samples of patients with diagnosis (ICD-9) codes with above- (solid dots) versus below-median (hollow dots) frequency in the data. Panel B estimates the model separately in two samples of patients with diagnosis codes with (solid dots) and those without (hollow dots) published guidelines cataloged by the US Agency for Healthcare Research and Quality (`guidelines.gov`). The model controls are as stated for Figure 1. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 6: Practice Style Distribution by Trainee Type



Note: This figure shows the patient-day-weighted 90th percentile, mean, and 10th percentile of the practice style (trainee effect) distribution according to trainee type. The unconditional distribution in each tenure period is normalized to have mean 0. Panel A shows the distribution for high-skill trainees (solid lines) relative to low-skill trainees (dashed lines), where "skill" is defined as position on the rank list more favorable than median when defined, and above-median USMLE test score when position on the rank list is missing. Panel B shows the distribution for trainees with above-median expected future income relative (solid lines) to those with below-median future income (dashed lines), where future income is based on known subsequent subspecialty training (if any) and imputed with national average yearly income in the first five years of practice after training. The average yearly future incomes of above- and below-median junior trainees are 424,000 and $268,000, respectively; the respective yearly future incomes for senior trainees are $409,000 and $249,000 (junior trainees include "preliminary interns," described in Section 2, who generally move on to more lucrative specialties). Practice styles are calculated as empirical Bayes posterior means from the random-effects model specified in Equation (3), where estimated variance components of the random-effects model are treated as prior distributions. The model controls are as stated for Figure 1. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark. Results for other trainee characteristics are shown in Appendix Tables A-2 and A-3.

Figure 7: Average Serial Correlation by Tenure Period Lag



Note: This figure shows the average serial correlation in trainee effects between 120-day tenure periods as a function of the lag between the tenure periods. Serial correlation parameters are estimated for each pair of tenure periods by a maximum likelihood method described further in Appendix A-3.2. There are a total of 9 non-overlapping tenure periods across the three years of training. The x-axis corresponds to the lag between the tenure periods, such that when the lag is 1, the y-axis displays the average of the serial correlations across the pairs of tenure periods $(1,2),(2,3),\ldots,(8,9)$. In general, for lag L, the y-axis displays an average of the serial correlations computed for $9-L$ tenure periods $(1,1+L),\ldots,(9-L,9)$. Thus, for the lag of 1, the average is across 8 serial correlation cells, while for the lag of 8, the "average" simply contains the serial correlation between tenure periods $(1,9)$. Underlying results for each pair of tenure periods are given in Appendix Table A-4 and are also shown graphically in Appendix Figure A-4. Confidence intervals are calculated by bootstrap. The model controls are as stated for Figure 1.

Table 1: Quasi-Random Assignment for Trainees with Above or Below Average Spending

	Interns		Residents	
	Below-median spending	Above-median spending	Below-median spending	Above-median spending
Patient characteristics				
Age	62.04 (16.91)	62.14 (16.85)	62.03 (16.92)	62.14 (16.83)
Male	0.483 (0.500)	0.482 (0.500)	0.484 (0.500)	0.482 (0.500)
White race	0.707 (0.455)	0.705 (0.456)	0.703 (0.457)	0.709 (0.454)
Black race	0.161 (0.367)	0.156 (0.363)	0.156 (0.363)	0.161 (0.368)
Predicted log total costs	8.477 (0.142)	8.478 (0.139)	8.498 (0.140)	8.477 (0.140)
Physician teammates				
Above-median-spending residents	0.504 (0.500)	0.495 (0.500)	N/A	N/A
Above-median-spending attendings	0.486 (0.500)	0.509 (0.500)	0.484 (0.500)	0.510 (0.500)

Note: This table shows evidence of quasi-random assignment for trainees with below-median or above-median averaged spending effects. Trainee spending effects, not conditioning by tenure, are estimated as fixed effects by a regression of log daily spending on patient characteristics and physician (intern, resident, and attending) identities. Lower- and higher-spending interns are identified by their fixed effect, relative to the median fixed effect, in a regression of admission spending that controls for patient characteristics (race, age, and gender), admission service dummies, and month-year interaction dummies. For each of these groups of interns, this table shows average patient characteristics and spending effects for supervising physicians. Averages are shown with standard deviations in parentheses.

Table 2: Summary Statistics of Spending in Categories and Services

	Log daily total costs				
	(1) Radiology	(2) Laboratory	(3) Medication	(4) Transfusion	(5) Nursing
Cardiology					
5th percentile	0	11	4	0	189
10th percentile	0	16	14	0	244
Median	0	34	67	16	658
Mean	54	51	113	33	662
90th percentile	125	103	233	56	1,075
95th percentile	375	145	417	87	1,212
Oncology					
5th percentile	0	3	0	0	192
10th percentile	0	13	13	0	256
Median	0	34	94	12	673
Mean	66	58	155	78	682
90th percentile	248	124	350	204	1,033
95th percentile	423	212	542	411	1,270
General Medicine					
5th percentile	0	8	2	0	160
10th percentile	0	12	10	0	205
Median	0	35	69	14	561
Mean	66	62	99	38	577
90th percentile	234	139	210	48	959
95th percentile	385	222	286	95	1,130

Note: This table reports summary statistics of patient-daily spending in categories across columns, and in ward services of cardiology, oncology, and general medicine. The statistics are calculated based on 56,780, 66,662, and 96,632 patient-day observations on the cardiology, oncology, and general medicine services, respectively.

Table 3: Team Concerns

	Discontinuity	Convergence
Independent practice	None	Yes
Span of control	Decrease	Yes
Principal-agent, rank	Increase	Yes
Information aggregation	Increase	Depends

Note: This table summarizes implications of three types of team concerns on two features of the practice variation profile with respect to trainee tenure: (i) the existence and direction of a discontinuity in practice variation as trainees move from junior to senior at the one-year tenure mark, and (ii) whether practice variation decreases (i.e., practices "converge") with tenure as trainees learn. Section 4.3 discusses team concerns in detail. "Independent practice" represents the benchmark with no team concerns. "Span of control" refers to the team theoretic idea that agents have limited bandwidth, and agents higher in the hierarchy (i.e., senior trainees) assigned to more than one agent lower in the hierarchy (i.e., junior trainees) will attend to fewer problems per case (Marschak and Radner, 1972; Garicano, 2000). "Principal-agent, rank" refers to models in which "senior" agents may have power over other "junior" agents, which may induce junior agents to suggest decisions that herd around the beliefs of senior agents (Prendergast, 1993). "Information aggregation" refers to the possibility of Bayesian information aggregation across agents to make a single decision (DeGroot, 2005). Although team concerns may coexist, each row represents one team concern in the absence of the other two team concerns. For example, in "span of control," predictions are for classical team theory, in which there are no principal-agent issues, and decisions are separable across agents and do not aggregate information.

Table 4: Effect of Trainee Characteristics on Spending

	Log daily total costs					
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	High USMLE	Highly ranked	High future income	Other hospital	Overall score
Panel A: Interns						
Effect of trainee with characteristic	-0.001	0.002	0.010*	0.007*	0.017*	0.003
	(0.004)	(0.005)	(0.006)	(0.004)	(0.010)	(0.002)
Observations	186,398	185,201	131,247	215,678	219,727	190,331
Adjusted R^2	0.089	0.089	0.090	0.088	0.087	0.090
Sample characteristic mean	0.596	0.258	0.234	0.415	0.055	N/A
Panel B: Residents						
Effect of trainee with characteristic	-0.013***	0.010**	-0.004	-0.001	0.013	0.004*
	(0.004)	(0.005)	(0.007)	(0.004)	(0.011)	(0.002)
Observations	206,455	199,371	129,281	218,376	219,727	206,455
Adjusted R^2	0.089	0.089	0.083	0.087	0.088	0.090
Sample characteristic mean	0.564	0.235	0.214	0.332	0.060	N/A

Note: This table reports results for some regressions of the effect of indicators of some trainee characteristics, including other hospital status, and a normalized predictive score (with standard deviation 1) based on *all* observed trainee characteristics. Panel A shows results for interns; Panel B shows results for residents. Columns (1) to (5) are regressions of the form in Equation (A-9), where the coefficient of interest is on an indicator for a group of trainees identified by either pre-residency characteristics, whether the trainee is from the other academic hospital, or whether the trainee is expected to have above-median future income based on known subspecialty training following residency The effect of many other characteristics of interest (or groups) were estimated as insignificant and omitted from this table for brevity. Column 6 reports results where the covariate of interest is a normalized predictive score based on predetermined characteristics of age, sex, minority status, track, rank on matching rank list, USMLE score, medical school rank in *US News & World Report*, indicators for whether the medical school is foreign or "rare," AOA medical honor society membership, and additional degrees at time of residency matriculation. By comparison, a predictive score for being highly ranked (in the top 50 rank positions) based on the same characteristics (except rank) changes the probability of being highly ranked by about 20% for both interns and residents. All models control for patient and admission characteristics, time dummies, and fixed effects for attending and the other trainees on the team (e.g., the resident is controlled for if the group is specific to the intern). Standard errors are clustered by admission.

Appendix

A-1 Random Assignment

This appendix presents two sets of randomization tests for quasi-random assignment, complementing evidence in Table 1. Section A-1.1 presents results regarding the assignment of patients to trainees. Section A-1.2 presents the assignment of trainees to supervising physicians.

A-1.1 Assignment of Patients to Trainees

First, I test for the joint significance of trainee identities in regressions of this form:

$$X_a = \mathbf{T}_{t(a)}\eta + \mu_{s(a)} + \zeta^{\tau<T}_{j(a)} + \zeta^{\tau>T}_{k(a)} + \zeta_{\ell(a)} + \varepsilon_a, \tag{A-1}$$

where a is a patient admission and X_a is some patient characteristic or linear combination of patient characteristics for the patient in admission a, described in Section 2.3. $t(a)$ refers to the day of admission, $s(a)$ is the service of admission, $j(a)$ is the junior trainee, $k(a)$ is the senior trainee, and $\ell(a)$ is the supervising physician. $\mathbf{T}_{t(a)}$ is a set of time categories for the admission day, including the day of the week and the month-year interaction; μ_s is a fixed effect that corresponds to the admitting service s (e.g., "heart failure service" or "oncology service"). $\zeta^{\tau<T}_i$, $\zeta^{\tau>T}_j$, and ζ_k are fixed effects for the intern i, resident j, and attending k, respectively. I do not impose any relationship between the fixed effect of a trainee as an intern and the fixed effect of the same trainee as a resident. I then test for the joint significance of the fixed effects $\left(\zeta^{\tau<T}_j, \zeta^{\tau>T}_k\right)_{j\in\mathcal{J},k\in\mathcal{K}}$.

In Column 1 of Table A-1, I show F-statistics and the corresponding p-values for the null hypothesis that $\left(\zeta^{\tau<T}_j, \zeta^{\tau>T}_k\right)_{j\in\mathcal{J},k\in\mathcal{K}} = \mathbf{0}$. I perform the regression (A-1) separately each of the following patient characteristics X_a as a dependent variable: patient age, a dummy for male gender, and a dummy for white race.[28] I also perform (A-1) using as dependent variables the linear prediction of log admission total spending based on patient age, race, and gender. I fail to find joint statistical significance for any of these tests.

Second, I test for the significance of trainee characteristics in regressions of this form:

$$X_a = \mathbf{T}_{t(a)}\eta + \mu_{s(a)} + \gamma_1 Z_{j(a)} + \gamma_2 Z_{k(a)} + \zeta_{\ell(a)} + \varepsilon_a. \tag{A-2}$$

Equation (A-2) is similar to Equation (A-1), except for the use of a vector of trainee characteristics $Z_{j(a)}$ and $Z_{k(a)}$ for the junior and senior trainee, respectively, on day of admission to test whether certain types of residents are more likely to be assigned certain types of patients. Trainee characteristics include the following: position on the rank list; USMLE Step 1 score; sex; age at the start of training; and dummies for foreign medical school, rare medical school, AOA honor society membership, PhD or another graduate degree, and racial minority.

[28] I do not test for balance in patient diagnoses, because these are discovered and coded by physicians potentially endogenous. Including or excluding them in the baseline specification of Equation (3) does not qualitatively affect results.

Columns 2 and 3 of Table A-1 show F-statistics and the corresponding p-values for the null hypothesis that $(\gamma_1, \gamma_2) = \mathbf{0}$. Column 2 includes all trainee characteristics in Z_h; column 3 excludes position on the rank list, since this information is missing for a sizable proportion of trainees. Patient characteristics for dependent variables in (A-2) are the same as in (A-1). Again, I fail to find joint significance for any of these tests.

Third, I compare the distributions of patient age and of predicted total costs across patients admitted to interns and residents with high or low spending. I consider trainee spending effects that are fixed within junior or senior role using this regression:

$$Y_a = \mathbf{X}_a\beta + \mathbf{T}_{t(a)}\eta + \zeta_{j(a)}^{\tau<T} + \zeta_{k(a)}^{\tau>T} + \zeta_{\ell(a)} + \varepsilon_a, \tag{A-3}$$

where Y_a is log total spending for admission a, and other variables are defined similarly as in Equation (A-1). Figure A-1 shows kernel density plots of the age distributions for patients assigned to interns and residents, respectively, each of which compare trainees with practice styles above and below the mean. Figure A-2 plots the distribution of predicted spending for patients assigned to trainees with above- or below-mean spending practice styles. There is essentially no difference across the distribution of age or predicted spending for patients assigned to trainees with high or low spending practice styles. Kolmogorov-Smirnov statistics cannot reject the null that the underlying distributions are different.

A-1.2 Assignment of Trainees to Other Providers

To test whether certain types of trainees are more likely to be assigned to certain types of other trainees and attending physicians, I perform the following regression to examine the correlation between two trainees and between a trainee and the supervising physician assigned to the same patient:

$$\hat{\zeta}_{h(a)}^{r} = \gamma_h\hat{\zeta}_{-h(a)}^{1-r} + \gamma_\ell\hat{\zeta}_{\ell(a)} + \varepsilon_a, \tag{A-4}$$

where $r \equiv \mathbf{1}(\tau > T)$ is an indicator for whether the fixed effect for trainee h was calculated while h was a junior trainee ($r = 0$) or a senior trainee ($r = 1$). As in Equation (A-1), I assume no relationship between $\hat{\zeta}_h^{\tau<T}$ and $\hat{\zeta}_h^{\tau>T}$. Each observation in Equation (A-4) corresponds to an admission a, but where error terms are clustered at the level of the intern-resident-attending team, since there are multiple observations for a given team. $\hat{\zeta}_\ell$ is the estimated fixed effect for attending k.[29] Estimates for γ_h and γ_ℓ are small, insignificant, and even slightly negative.

Second, I perform a similar exercise as in the previous subsection, in which I plot the distribution of estimated attending fixed effects working with trainees with above- or below-mean spending practice styles. In Figure A-3, the practice-style distribution for attendings is similar for those assigned to

[29] I use two approaches to get around the reflection problem due to the first-stage joint estimation of ζ_j^0, ζ_k^1, and ζ_ℓ (Manski, 1993). First, I perform (A-4) using "jack-knife" estimates of fixed effects, in which I exclude observations with $-h$ and ℓ to compute the $\hat{\zeta}_h^{r}$ estimate that I use with $\hat{\zeta}_{-h}^{1-r}$ and $\hat{\zeta}_k$. Second, I use the approach by Mas and Moretti (2009), in which I include nuisance parameters in the first stage to absorb team fixed effects for (j, k, ℓ).

high- versus low-spending trainees. As for distributions of patient characteristics in Appendix A-1.1, differences in the distributions are not qualitatively significant, and Kolmogorov-Smirnov statistics cannot reject the null that these distributions are different, at least when clustering at the level of the intern-resident-attending team.

A-2 Random-Effects vs. Fixed-Effects Identification

The fixed-effects estimation approach (e.g., Abowd et al., 1999; Card et al., 2013) relies on a version of Assumption 1 that is only slightly weaker:

Assumption 2 (Quasi-Random Team Assignment within Connected Sets (Abowd et al., 1999)). *Potential team decisions are independent of team assignments, conditional on clinical service $s(i,t)$, indicators of time t, and connected sets $g(i,t)$:*

$$\{Y_{it}(j,k)\}_{(j,k)\in\mathcal{J}_{it}\times\mathcal{K}_{it}} \perp\!\!\!\perp \left(D_{ijt}, D_{ikt}\right) \Big| s(i,t), t, g(i,t).$$

As discussed in Abowd et al. (2008), a "connected set" g comprises cases (i,t) such that $j(i,t) \in \mathcal{J}^g$ or $k(i,t) \in \mathcal{K}^g$. $\mathcal{J}^g$ includes any junior trainee who has worked with a senior trainee in $\mathcal{K}^g$, and $\mathcal{K}^g$ includes any senior trainee who has worked with a junior trainee in $\mathcal{J}^g$. Any pair of trainees $(j,k) \in \mathcal{J}^g \times \mathcal{K}^g$, whose observations are in the same connected set, can be "connected" via a chain of trainees that have worked together.

Assumption 1 implies Assumption 2, and if $\mathcal{J}^{g(i,t)} \supseteq \mathcal{J}_{it}$ and $\mathcal{K}^{g(i,t)} \supseteq \mathcal{K}_{it}$, then Assumption 1 is equivalent to Assumption 2. Fixed-effects estimation, under Assumption 2, comes with the cost that the effects of trainees in different connected sets are not comparable: For each g, one junior-trainee effect and one senior-trainee effect need to be dropped from estimation to satisfy the rank condition. Stated differently, to identify *any* trainee effects, the fixed-effects framework requires trainee "movers," who work with more than one teammate. While our setting involves and exploits such movers, this requirement is not strictly necessary in the random-effects approach, under Assumption 1. The sense in which Assumption 2 is weaker than Assumption 1 mostly results from the rank condition and not a necessarily substantive difference in the quasi-experimental design. In finite samples, if we observed fewer cases for the same set of trainees, the sets $\mathcal{J}^{g(i,t)}$ and $\mathcal{K}^{g(i,t)}$ could contain fewer elements, even though $\mathcal{J}_{it}$ and $\mathcal{K}_{it}$ would be unchanged.

A-3 Statistical Model of Trainee Effects

A-3.1 Patient Admission Random Effects

We may augment Equation (4) to allow for patient admission random effects, since the same patient may stay for more than one day and be exposed to different trainees:

$$\tilde{Y}_{it} = \xi^{\tau_j;\tau_k}_{j(i,t)} + \xi^{\tau_k;\tau_j}_{k(i,t)} + \nu_i + \varepsilon_{it}, \tag{A-5}$$

where ν_i is a random effect for the patient admission.[30] Under Assumption 1, $\xi_j^{\tau_j;\tau_k}$, $\xi_k^{\tau_k;\tau_j}$, and ν_i are uncorrelated with one another.

Let N_I be the number of patient admissions in sample $\mathscr{C}\left(\tau_j, \tau_k\right)$. Then in Equation (5), $\mathbf{D}$ is an $N \times (N_J + N_K + N_I)$ selection matrix for junior trainees, senior trainees, and patient admissions. $\mathbf{u}$ is an $(N_J + N_K + N_I) \times 1$ stacked vector of junior trainee, senior trainee, and patient admission random effects. We can then restate the variance-covariance matrix of $\mathbf{u}$ as

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \sigma^2\left(\tau_j;\tau_k\right)\mathbf{I}_{N_J} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2\left(\tau_k;\tau_j\right)\mathbf{I}_{N_K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\nu^2\mathbf{I}_{N_I} \end{bmatrix}.$$

The log likelihood function in Equation (6) remains the same, with $\mathbf{V} = \mathbf{DGD}' + \sigma_\varepsilon^2\mathbf{I}_N$. I maximize this log likelihood with respect to $\sigma^2\left(\tau_j;\tau_k\right)$, $\sigma^2\left(\tau_k;\tau_j\right)$, σ_ν^2, and σ_ε^2. Estimates of $\sigma^2\left(\tau_j;\tau_k\right)$ and $\sigma^2\left(\tau_k;\tau_j\right)$ in this augmented model are qualitatively unchanged relative to the baseline implementation in Section 3.4.

A-3.2 Correlation of Trainee Effects

I augment models in (4) and (A-5) to estimate the correlation between trainee effects in two separate tenure periods, τ_1 and τ_2, which I denote by $\rho(\tau_1, \tau_2)$. Although I observe each trainee across her entire training, I only observe a subset of these trainees in each period. The number of trainees observed in both tenure periods in the pair (τ_1, τ_2) is even smaller. Because trainees that I do not observe in both τ_1 and τ_2 do not contribute to the estimate of $\rho(\tau_1, \tau_2)$, I include in the estimation sample only observations associated with a trainee observed in both tenure periods. I also redefine tenure periods to be 120 days in order to enlarge the sample of trainees whom I observe in both periods in a tenure-period pair.

Specifically, in place of Equation (4), I consider

$$\tilde{Y}_{it} = \xi_{h(i,t)}^{\tau} + \xi_{-h(i,t)} + \varepsilon_{it}, \tag{A-6}$$

where $\tau \in \{\tau_1, \tau_2\}$ may be one of two tenure periods in a pair.. This specifies that effects of trainees in the tenure periods of interest (τ_1 and τ_2) may be drawn from two separate distributions depending on the tenure period τ_1 or τ_2 corresponding to observation t; I pool the effects of the teammates into a single distribution that does not depend on tenure. Because I focus on the correlation between trainee effects, I am unconcerned about the scale of practice variation and I thus do not specify the tenure of the teammate. The analog for Equation (A-5) is

$$\tilde{Y}_{it} = \xi_{h(i,t)}^{\tau} + \xi_{-h(i,t)} + \nu_i + \varepsilon_{it}. \tag{A-7}$$

[30] This specification requires the use of sparse matrices for estimation. In specifications without the use of sparse matrices, I nest this effect within interns, i.e., I include ν_{ai} as an intern-admission effect. While it is easier to estimate a specification with ν_{ai}, I will describe this specification for ease of explication. In practice, results are materially unaffected by whether I use ν_a or ν_{ai}, or in fact whether I include an admission-related effect at all.

I estimate (A-6) or (A-7) in a sample of observations, which I define as follows: $\mathscr{C}(\tau_1,\tau_2) = \{(i,t,h) : h \in \{j(i,t),k(i,t)\}, \tau(h,t) \in \{\tau_1,\tau_2\}\}$. I require that, for every trainee h in $\mathscr{C}(\tau_1,\tau_2)$, there are observations in the sample in which she has tenure τ_1 and other observations in the sample in which she has tenure τ_2. Otherwise, we cannot use trainee h to estimate the correlation in trainee effects between these two periods.

As above, I can represent both Equation (A-6) and Equation (A-7) in matrix form, as Equation (5). Denote the number of trainees h in $\mathscr{C}(\tau_1,\tau_2)$ as N_H. Denote the number of teammates trainees interacted with their tenure periods as N_H^-. The selection matrix $\mathbf{Z}$ is of size $N \times \left(2N_H + N_H^-\right)$, since it now maps observations onto one of two random effects, depending on whether $\tau = \tau_1$ or $\tau = \tau_2$, for each trainee h observed in both τ_1 and τ_2 tenure periods. The stacked vector of random effects $\mathbf{u}$ is similarly of size $(2N_\tau + N_\tau^-) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \mathbf{G}_H & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\xi^-}\mathbf{I}_{N_H^-} \end{bmatrix},$$

where $\mathbf{G}_H$ is a $2N_H \times 2N_H$ block-diagonal matrix of the form

$$\mathbf{G}_H = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A} \end{bmatrix}, \tag{A-8}$$

with each block being the 2×2 variance-covariance matrix $\mathbf{A}$ of random effects within trainee and across tenure periods:

$$\begin{aligned} \operatorname{Var}\begin{bmatrix} \xi_h^{\tau_1} \\ \xi_h^{\tau_2} \end{bmatrix} &= \mathbf{A}, \text{ for all } h, \text{ where} \\ \mathbf{A} &\equiv \begin{bmatrix} \sigma^2(\tau_1) & \rho(\tau_1,\tau_2)\sigma(\tau_1)\sigma(\tau_2) \\ \rho(\tau_1,\tau_2)\sigma(\tau_1)\sigma(\tau_2) & \sigma^2(\tau_2) \end{bmatrix}. \end{aligned}$$

Representing (A-7) as (5) is a similar exercise. The selection matrix $\mathbf{Z}$ is of size $N \times \left(2N_H + N_H^- + N_I\right)$, and the vector of random effects $\mathbf{u}$ is of size $\left(2N_H + N_H^- + N_I\right) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \mathbf{G}_H & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\xi^-}\mathbf{I}_{N_H^-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma^2_\nu\mathbf{I}_{N_I} \end{bmatrix},$$

where $\mathbf{G}_H$ is the same as in Equation (A-8). The log likelihood is the same as in Equation (6), but using revised definitions of $\mathbf{G}$ that allow for covariance between random effects of the same trainees across tenure periods. The correlation parameter of interest $\rho(\tau_1,\tau_2)$ is constrained to be between -1 and 1.

A-4 Intrinsic Heterogeneity: Trainee Characteristics

The key alternative explanation for persistent variation that I explore in this section is that physicians may intrinsically differ for reasons unrelated to knowledge and learning, such as preferences or ability (e.g., Doyle et al., 2010; Fox and Smeets, 2011; Bartel et al., 2014). To assess the possibility of intrinsic heterogeneity, I first exploit detailed trainee characteristics that should be highly correlated with preferences and ability. For example, USMLE scores measure medical knowledge as a medical student; position on the residency rank lists reflects overall desirability; and specialty tracks, mostly predetermined relative to the beginning of residency, reflect important career decisions and lifestyle preferences, such as a decision to become a radiologist rather than a primary care physician. To capture the variety of future career paths across internal medicine trainees, I impute future yearly incomes after specialty training based on the final specialty choices of trainees. As cited in Section 2.3, trainees with above-median future incomes will earn substantially more than their peers with below-median future incomes.

I assess the relationship between each of these characteristics and daily spending totals for either the junior or senior trainee:

$$Y_{it} = \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{aijkt}, \tag{A-9}$$

where $\text{Characteristic}^m_h$ is an indicator for whether the junior (or senior) trainee h has the characteristic m, ζ_{-h} is a fixed effect for the other senior (or junior) trainee $-h$, and ζ_ℓ is a fixed effect for attending ℓ.[31] The coefficient of interest, α_m, quantifies the predictive effect of a trainee with characteristic m on patient spending decisions. I also evaluate the combined predictive effect of trainee characteristics in two steps. First, I regress outcomes on all direct trainee characteristics, with continuous characteristics like position on rank list entered linearly, along with the other admission and time regressors in Equation (A-9):

$$Y_{it} = \sum_m \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{A-10}$$

This yields a predicted score Z_h for each trainee h, $Z_h = \sum_m \hat{\alpha}_m \text{Characteristic}^m_h$, which I normalize to $\tilde{Z}_h = Z_h / \sqrt{\text{Var}(Z_h)}$ with standard deviation 1. Second, I regress daily total spending on this normalized score:

$$Y_{it} = \alpha \tilde{Z}_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{A-11}$$

In addition, I evaluate the predictive power of trainee characteristics more flexibly by allowing splines of continuous characteristics and two-way interactions between characteristics, while assuming an "approximately sparse" model and using LASSO to select for significant characteristics (e.g., Belloni et al., 2014). This approach guards against overfitting in finite data when the number of po-

[31] In principle, I could include trainee characteristics as mean shifters in the baseline random effects model in Equation (3). However, since characteristics are generally insignificant predictors of variation, results of (residual) variation attributable to trainees are unchanged.

tential characteristics becomes large. In total, excluding collinear characteristics, I consider 36 and 32 direct characteristics for interns and residents, respectively, and 285 and 308 two-way interactions, as potential regressors in Equation (A-9).

Table 4 shows results for Equation (A-11) and a subset of results for Equation (A-9). Considering characteristics individually in Equation (A-9), only two characteristics (gender and high USMLE test score) are statistically significant at the 5% level, and no characteristic approaches the one-standard deviation benchmark effect in the trainee effect distribution. Likewise, a standard-deviation change in the overall predictive score has no economically significant effect on spending for either interns or residents. LASSO selected no intern characteristic as significant and selected only resident gender as significant. Although it is possible that there are other unmeasured and orthogonal characteristics that are more relevant for practice variation, this seems *a priori* unlikely given that these are the characteristics on which the residency program bases acceptance decisions,[32] and that they are also highly predictive of future career paths and incomes.

Finally, I investigate the *distribution* of trainee effects as a function of tenure for trainees in different groups. As shown in Figure 6, the distributions of trainee effects throughout training are not meaningfully different between groups of trainees separated by their test scores, rank list positions, or future earnings. This finding implies that trainees who differ significantly along meaningful dimensions still practice similarly not only on average, but also in terms of variation over time. That is, trainees evaluated with higher test scores, more desirable rankings, or higher future earnings do not exhibit lower variation or higher convergence over training.

A-5 Learning by Osmosis: Predictable Learning

Finally, I assess whether trainee practice styles can be predicted by the sequence of observable learning experiences. This evaluation tests two concepts. First, practice styles may predictably change if they reflect acquired skill that may grow with greater experience. Second, trainees may absorb spending patterns from supervising physicians or from a broader practice environment.[33]

To explore the potential effect of learning from others in greater detail, I estimate supervising physician "effects" by shrinking their observed fixed effects, and I similarly calculate best linear unbiased predictions (BLUPs) of senior trainee effects. The standard deviation of shrunken supervising physician effects is 7.3%, and the standard deviation of the senior trainee BLUPs is 16.6% in terms of overall spending. I then form measures of prior exposure to spending due to supervising physicians by averaging spending effects of supervising physicians who have previously worked with a given trainee, weighted by patient-days, at a given point in time. This exposure measure may or may not be

[32] Using the same characteristics to predict whether a trainee was ranked in the upper half on the residency program's rank list (excluding rank as a characteristic) yields a predictive score that with one standard deviation changes the probability of being highly ranked by about 20%.

[33] The related concept of "schools of thought," in which physicians may have systematically different training experiences, has been proposed as a mechanism for geographic variation (e.g., Phelps and Mooney, 1993). This hypothesis is not inconsistent with tacit knowledge and in fact relies partly on it, but it does not by itself explain large variation within the same training program.

restricted to patient-days on the same ward service (e.g., cardiology, oncology, or general medicine). Similarly, the measure may be calculated for all prior patient-days or only for patient-days in the last three months. I also calculate similar measures of exposure to senior trainees for trainees based on their previous team matches when they were junior.

For a given prior exposure measure, I define trainees with above-median measures in a given tenure period as having "high exposure" to spending and trainees with below-median measures as having "low exposure" to spending. Compared to other trainees with the same tenure, these trainees have worked with attending physicians or residents trainees (while they were interns) with higher average spending effects. Table A-5 shows the difference between high-exposure and low-exposure trainees for various spending-exposure measures at different trainee tenure periods. Differences between high and low exposure to supervising-physician spending range from 1.9% to 6.7%. Differences between high and low exposure to senior-trainee spending range from 17.5% to 23.4%.

I then estimate the effect of high exposure to spending over each tenure period of training with a regression of the form

$$\begin{aligned} Y_{it} \quad = \quad & \sum_{\tau:\tau<1} \alpha_\tau \mathbf{1}\left(\tau\left(j\left(i,t\right),t\right)=\tau\right) \cdot \text{HighSpendingExposure}^m_{j(i,t),t} + \\ & \sum_{\tau:\tau\geq 1} \alpha_\tau \mathbf{1}\left(\tau\left(k\left(i,t\right),t\right)=\tau\right) \cdot \text{HighSpendingExposure}^m_{k(i,t),t} + \\ & \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{\ell(i,t)} + \varepsilon_{it}, \end{aligned} \tag{A-12}$$

where, as in Equation (3), $j\left(i,t\right)$ is the junior trainee, $k\left(i,t\right)$ is the senior trainee, and $\tau\left(j\left(i,t\right),t\right)$ and $\tau\left(k\left(i,t\right),t\right)$ are the relevant tenure periods of the junior and senior trainees at t. The variables $\text{HighSpendingExposure}^m_{j,t}$ and $\text{HighSpendingExposure}^m_{k,t}$ are indicators for high exposure to spending under measure m for the junior and senior trainee, respectively. The effect of this exposure can vary by τ. Figure A-6 shows results for exposure to spending by supervising physicians, and Figure A-7 shows similar results for exposure to spending by senior trainees. Results among the wide range of exposure measures are broadly insignificant.

More broadly, I also consider several measures of prior experience—including days on ward service, patients seen, and supervising physicians for a given trainee prior to a patient encounter—for either the junior or senior trainee. For each of these experience measures, I estimate a regression of the form

$$Y_{it} = \alpha_m \text{Experience}^m_{h(i,t),t} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}, \tag{A-13}$$

where $\text{Experience}^m_{h,t}$ is an indicator for whether trainee h at time t has experienced a measure (e.g., number of days on service, average supervising physician spending effect) above median for the relevant tenure period, where both the measure and the median are calculated using observations prior to the relevant tenure period. In my baseline specification, I control for the other trainee and supervising physician identities, although this does not qualitatively affect results. Results are shown in Table A-6 and are broadly insignificant. A LASSO implementation that jointly considers a larger number of summary experience measures in early or more recent months relative to the patient encounter, as

well as two-way interactions between these measures (112 and 288 variables for interns and residents, respectively), also fails to select any measure as significant.

In addition to trainees in the main residency program, I observe visiting trainees based in a hospital with 20% lower Medicare spending according to the Dartmouth Atlas. I evaluate the effect of these trainees on teams, as interns and as residents, using Equation (A-9). This effect includes both differences in selection (i.e., intrinsic heterogeneity) into the different program and in training experiences across the programs. Table 4 shows that visiting trainees do not have significantly different spending effects, either as interns or as residents.[34]

Overall, these results indicate that summary measures of trainee experience are poor predictors of practice and outcomes, especially relative to the large variation across trainees. The results fail to support "learning by osmosis" as a major source of practice variation, at least within an organization with *ex ante* uniform training experiences but nonetheless large practice variation.

A-6 Model of Information Aggregation and Experiential Learning

A-6.1 Setup

Each decision d can be summarized perfectly by an unknown parameter θ_d. If θ_d were known, then the optimal action would be $a_d = \theta_d$. Each agent has only *partial* knowledge about the correct action, in the form of a Bayesian prior about θ_d. A team decision is made as follows:

1. Each agent $h \in \{j, k\}$ has prior knowledge bearing on the decision; specifically, a Bayesian prior distribution, $\theta_{d,h}$. $\theta_{d,h}$ is a normal distribution and can be summarized by mean $\mu_{d,h}$ and precision $\rho_{d,h}$. One may describe $\mu_{d,h}$ as the *judgment* (due to prior knowledge) that agent h has about d.

2. There may also be external information about d. Some of this knowledge is held by the attending physician, but other sources derive from hospital nurses, consultants, and protocols. Each agent may also collect information about the decision, which I assume to be independent of prior knowledge. I consider external information as a public judgment with mean 0 and precision P_d^*.

3. The team takes an action and derives utility $u = -(\theta_d - a_d)^2$. As in the standard team-theoretic environment, there is no conflict of interest between agents.

Proposition A-1. *The optimal (Bayesian) action for decision d assigned to trainees j and k is*

$$a_d^* = \frac{\rho_{d,j}\mu_{d,j} + \rho_{d,k}\mu_{d,k}}{\rho_{d,j} + \rho_{d,k} + P_d^*}. \tag{A-14}$$

[34] This result of course does not rule out that training programs can matter. Doyle et al. (2010) studies the effect of trainee teams from two different programs and find that trainees from the higher-prestige program spend less. However, this result does suggest that even when trainees come from significantly different hospitals, differences in their mean practice styles can be dwarfed by variation within training program.

This expression aggregates information as a weighted average of judgments in proportion to the precisions of the respective judgments (DeGroot, 2005). Supervisory information, measured by precision P_d^*, reduces the effect of either trainee's judgment on a_d^*.

The weights on judgments in the Bayesian action in Equation (A-14),

$$g_{d,h;-h}^* \equiv \frac{\rho_{d,h}}{\rho_{d,h} + \rho_{d,-h} + P_d^*},$$

have a natural interpretation as the *influence* of trainee h on the action a_d^*. The more precise the signal from her prior knowledge relative to her teammate and any supervisory information, the greater her influence will be. In the limit, if either her teammate's knowledge or external information is perfect (i.e., $\rho_{d,-h} = \infty$ or $P_d^* = \infty$), a trainee would have no influence. On the other hand, if a trainee has perfect knowledge, then she would have full influence. At the one-year tenure mark, influence discontinuously increases because the precision of a trainee's teammate $\rho_{d,-h}$ discontinuously decreases.

Influence may deviate from the Bayesian benchmark due to other team concerns. Career concerns or the "prestige" of senior titles may underweight the knowledge of junior trainees (Scharfstein and Stein, 1990; Prendergast, 1993; Ottaviani and Sorensen, 2001), or trainees may be given more influence than justified by their knowledge if supervisors wish to encourage experiential learning that requires a stake in decision-making (Lizzeri and Siniscalchi, 2008; Ludmerer, 2014). In estimation, I allow for actions that deviate from the Bayesian benchark:

$$\hat{a}_d = \frac{\tilde{\rho}_{d,j}\mu_{d,j} + \tilde{\rho}_{d,k}\mu_{d,k}}{\tilde{\rho}_{d,j} + \tilde{\rho}_{d,k} + P_d}. \tag{A-15}$$

$\tilde{\rho}_{d,h} = \rho_{d,h} + \delta(\tau_h)$ as an effective "precision" that equals the true precision of h's knowledge adjusted by $\delta(\tau_h)$, depending on the tenure of h, τ_h. The influence of trainees with tenure τ_h relative to their peers may receive less influence than the Bayesian benchmark if $\delta(\tau_h) < 0$ or more influence if $\delta(\tau_h) > 0$. Similarly, for external and supervisory information, P_d is an effective "precision": Even though supervising physicians and the broader supervisory structure may have access to information relevant for d with precision P_d^*, this information may be underweighted ($P_d < P_d^*$) or overweighted ($P_d > P_d^*$) in decision-making.

I consider the precision of knowledge as a function of tenure for given class of decisions, c: $\rho_{d,h} = \rho_{c(d)}(\tau_h(t(d)))$. I similarly specify external information as depending on the class of decisions: $P_d = P_{c(d)}$.[35] Effective influence of a trainee with tenure τ_h working with teammate with tenure τ_{-h} is

$$g_c(\tau_h; \tau_{-h}) = \frac{\tilde{\rho}_c(\tau_h)}{\tilde{\rho}_c(\tau_h) + \tilde{\rho}_c(\tau_{-h}) + P_c}. \tag{A-16}$$

Model-predicted practice variation (i.e., standard deviation of trainee effects) for trainees with

[35] In Appendix Figure A-5, I support for this assumption by showing that both the trainee-related variation and the residual variation in spending are relatively constant across July, when old interns transition to residents and new interns begin training.

tenure τ_h, working with teammates with tenure τ_{-h}, is then

$$\sigma_c(\tau_h, \tau_{-h}) = g_c(\tau_h; \tau_{-h}) \sqrt{\kappa_c / \rho_c(\tau_h)}, \tag{A-17}$$

where $\kappa_c \in [0, 1]$ reflects the similarity of judgments across different decisions in class c within the same provider. Systematic practice variation across trainees, requires that $\kappa_c > 0$, or that trainees practice similarly across different decisions. While levels of knowledge, learning, and practice variation are scaled by κ_c, ratios comparing different points in training will be unaffected by κ_c.

A-6.2 Identification

As trainees learn, the precision of their knowledge, or $\rho_c(\tau_h)$, increases with tenure. Greater knowledge increases influence, or $g_c(\tau_h; \tau_{-h})$, holding teammates and external information fixed, while it reduces dispersion in judgments, or $\sqrt{1/\rho_c(\tau_h)}$. Thus practice variation may not always decrease even as trainees learn. In general, the effect of increasing influence on practice variation will tend to dominate when a trainee's influence is relatively low, while when a trainee has relatively high influence, the effect of reducing dispersion in judgments will tend to dominate. In the extreme, agents who practice independently (i.e., they have full influence over their decisions) will show convergence in their decisions as they learn.

A-6.2.1 Analytical Evaluation

Consider practice variation—or the standard deviation of trainee effects—under Bayesian-benchmark influence:

$$\begin{aligned} \sigma(\tau_h, \tau_{-h}) &= \frac{g^*(\tau_h; \tau_{-h})}{\sqrt{\rho(\tau_h)}} \\ &= \frac{\sqrt{\rho(\tau_h)}}{\rho(\tau_h) + \rho(\tau_{-h}) + P}, \end{aligned} \tag{A-18}$$

where I assume that $\kappa = 1$ in (A-17) without loss of generality.

As a first observation, note that the discontinuity in practice variation is greater across the one-year tenure mark than it is across the two-year tenure mark.

Proposition A-2. *Define* $\sigma(1^-) \equiv \lim_{\tau \to 1^-} E_{-h}[\sigma(\tau_h, \tau_{-h}) | \tau_h]$, *and* $\sigma(1^+) \equiv \lim_{\tau \to 1^+} E_{-h}[\sigma(\tau_h, \tau_{-h}) | \tau_h]$; *similarly define* $\sigma(2^-) \equiv \lim_{\tau \to 2^-} E_{-h}[\sigma(\tau_h, \tau_{-h}) | \tau_h]$, *and* $\sigma(T^+) \equiv \lim_{\tau \to 2^+} E_{-h}[\sigma(\tau_h, \tau_{-h}) | \tau_h]$. *Then*

$$\frac{\sigma(1^+)}{\sigma(1^-)} > \frac{\sigma(2^+)}{\sigma(2^-)} > 1.$$

Proof. Assume that interns work with second-year residents in λ proportion of the time and work with third-year residents in the remaining $1 - \lambda$ proportion of the time. At the first-year discontinuity,

$$\frac{\sigma(1^+)}{\sigma(1^-)} = \frac{\rho(1) + \lambda\rho(2) + (1 - \lambda)\rho(3) + P}{\rho(1) + \rho(0) + P}.$$

At the second-year discontinuity,

$$\frac{\sigma(2^+)}{\sigma(2^-)} = \frac{\rho(2)+\rho(1)+P}{\rho(2)+\rho(0)+P}.$$

Since $\rho(\cdot)$ is increasing in τ, $\rho(0) \leq \rho(1) \leq \rho(2) \leq \rho(3)$, which yields our result. □

Because there is a change in the tenure of the other trainees as new interns arrive at the beginning of each academic year, there is in principle a discontinuous increase in influence (and therefore practice variation) at the beginning of each year. However, the increase at $\tau_h = 1$ is always larger than the increase at $\tau_h = 2$ for two reasons, both related to the monotonic increase in precision with tenure: First, trainees at $\tau_h = 1$ have less precise subjective priors than those at $\tau_h = 2$, so any decrease in the relative tenure of their peer trainee increases their influence by more. Second, the decrease in the relative tenure of the peer is greater at $\tau_h = 1$ (from $\tau_{-h} = 2$ to $\tau_{-h} = 0$) than at $\tau_h = 2$ (from $\tau_{-h} = 1$ to $\tau_{-h} = 0$). I show below in the numerical examples that, within this framework, this difference in the discontinuous increases at $\tau_h = 1$ and at $\tau_h = 2$ can be quite large, and that the discontinuity at $\tau_h = 2$ can be quite trivial.

Second, I consider whether practice variation is likely to increase or decrease with tenure. Since trainees and their teammates gain tenure together, I consider $\tau_{-h} = \tau_h + \Delta$, where Δ is fixed in a continuous portion of practice variation (i.e., not at the one- or two-year discontinuities). Applying the quotient rule to $\sigma(\tau_h, \tau_{-h}) = \sigma(\tau_h, \tau_h + \Delta)$,

$$\begin{aligned}
\sigma'(\tau_h) &\equiv \frac{\partial \sigma(\tau_h, \tau_h+\Delta)}{\delta \tau_h} \\
&= \frac{\frac{1}{2}\rho(\tau_h)^{-1/2}\rho'(\tau_h)(\rho(\tau_h)+\rho(\tau_{-h})+P) - \rho(\tau)^{1/2}(\rho'(\tau)+\rho'(\tau_{-h}))}{(\rho(\tau)+\rho(\tau_{-h})+P)^2}.
\end{aligned}$$

Focusing on the numerator to determine the sign of $\sigma'(\tau)$, I arrive at the following necessary and sufficient condition for convergence (i.e., decreasing practice variation with tenure, or $\sigma'(\tau_h) < 0$):

Proposition A-3. *Practice variation decreases if and only if*

$$\frac{\rho'(\tau_h)}{\rho'(\tau_h)+\rho'(\tau_{-h})} < 2g^*(\tau_h; \tau_{-h}). \tag{A-19}$$

Learning (i.e., $\rho'(\tau_h) > 0$) does not guarantee convergence. Instead, convergence requires that the "share of learning," defined as $\rho'(\tau_h)/(\rho'(\tau_h)+\rho'(\tau_{-h}))$, is smaller than twice the influence. Since this "share" is always less than 1, convergence is guaranteed whenever the trainee has full influence, or $g^*(\tau_h; \tau_{-h}) = 1$, as is the case in a single decision-maker. The larger the trainee's influence, the more likely convergence will occur. Since influence grows with tenure, this also implies that practice variation generally increases and then decreases. Special cases may involve practice variation only increasing or only decreasing, but not decreasing and then increasing with tenure.

A-6.2.2 Numerical Examples

Figure A-8 presents a few numerical examples of variation profiles under various learning profiles described by functions of the piecewise linear form in Equation (A-20). The three parameters of interest are ρ_0, or initial knowledge; ρ_1, or the rate of increase in the precision during the first year as a junior trainee; and $\rho_2 = \rho_3$, or the rate of increase during the subsequent two years as a senior trainee. The precision of judgments at the end of training is $\rho(3) = \rho_0 + \rho_1 + 2\rho_2$. I also normalize $P = 1$, so that whether precisions of beliefs are greater than the precision of the supervisory prior simply depends on whether they are greater or less than 1. I consider this normalization as only relevant for the scale of the variation profile, since any scale keeping the same shape over the overall variation profile $\sigma(\tau)$ can be implemented by multiplying ρ_0, ρ_1, ρ_2, and P by some constant.

I discuss each panel of Figure A-8 in turn:

- Panel A considers equal $\rho_0 = \rho_1 = \rho_2 = 0.2$, which are relatively small compared to $P = 1$. The result is broadly non-convergence, as greater experience primarily results in greater influence against a relatively strong supervisory practice environment. The discontinuity in variation is significantly larger at $\tau = 1$ than at $\tau = 2$. Variation increases in intern year and decreases but only slightly in the next two years as resident.

- Panel B imposes no resident learning ($\rho_2 = 0$) and presents the limiting case in which discontinuous increases in variation at $\tau = 1$ and $\tau = 2$ are the same. Variation is still at least as big during the two years as resident as during the year as intern, driven by influence. Variation seems relatively constant over training.

- Panel C generates a similar variation profile as in Panel B with a non-zero ρ_2 by increasing the ratios of ρ_0 and ρ_1 to ρ_2. The scale of variation is smaller than in Panel B, which reflects that precision in trainee beliefs are now larger. A rescaled version with smaller precisions (and smaller P) would reveal larger relative increases in variation at the discontinuities.

- Panel D examines increasing ρ_1 relative to ρ_0, so that more learning occurs in the first year of training compared with knowledge possessed before starting training. Influence more obviously increases in the first year, and increases in variation are sharper at the discontinuities, since intern experience matters more. Note that working with a resident is equivalent to working with an end-of-year intern, and increases in variation at $\tau = 1$ and $\tau = 2$ are the same (as in Panel B).

- Panel E asserts that most of the learning occurs during the role as resident. There is much greater variation across residents than across interns, and the discontinuous increase in variation is much larger at $\tau = 1$, while the increase is negligible at $\tau = 2$. There is significant convergence during the two years as resident.

- Panel F is similar to panel E but shows less convergence during role as resident. The ratio of learning as intern to learning as resident (ρ_1/ρ_2) is similar, but learning during training is

reduced relative to knowledge from prior to training (ρ_0) and to supervisory information (P).

A-6.3 Specification and Estimation

I specify the precision of knowledge as a piecewise-linear function of trainee tenure:

$$\rho_c(\tau) = \begin{cases} \rho_{0,c} + \rho_{1,c}\tau, & \tau \in [0,1]; \\ \rho_{0,c} + \rho_{1,c} + \rho_{2,c}(\tau - 1), & \tau \in [1,2]; \\ \rho_{0,c} + \rho_{1,c} + \rho_{2,c} + \rho_{3,c}(\tau - 2), & \tau \in [2,3], \end{cases} \tag{A-20}$$

where $\rho_{0,c}$ represents the precision of knowledge before starting residency, and $\rho_{1,c}$, $\rho_{2,c}$, and $\rho_{3,c}$ are the yearly rate of learning in the first, second, and third years of residency, respectively, for decisions in class c.

Assuming that knowledge is continuous with tenure, I also identify deviations from efficient influence that come from a step function with respect to years of training. That is, the "effective" trainee precision relevant for influence is

$$\tilde{\rho}_c(\tau) \quad = \quad \rho_c(\tau) + \delta_{1,c}\mathbf{1}(\tau \geq 1) + \delta_{2,c}\mathbf{1}(\tau \geq 2). \tag{A-21}$$

$\delta_{1,c}$ and $\delta_{2,c}$ represent deviations in influence from the efficient benchmark that may result from titles (e.g., "senior trainee") that discontinuously change at years of training, $\lfloor \tau \rfloor$.[36] Finally, I identify deviations from the Bayesian benchmark from the fact that $P_c^* \geq \rho_c(\tau = 3)$: At a minimum, external information must be greater than the knowledge held by a senior trainee, since all supervising physicians have completed training, and since supervisory information includes informational inputs from outside staff (e.g., nursing, consultants), or any information gathered by the trainees themselves.[37] $P_c < \rho_c(3)$ would strongly imply that trainees are granted *more* influence than warranted by their knowledge.

I estimate learning and influence parameters as a two-step process. The first step recovers moments of practice variation, specifically the standard deviation of the distribution of trainee effects, for trainees of tenure τ_h working with teammates of tenure τ_{-h}. These empirical moments, $\hat{\sigma}(\tau_h, \tau_{-h})$, are estimated from the random effects model in Equation (3) and were previously discussed in Section 3. The second step takes these moments of practice variation and, from the model in Section 5.3, recovers underlying primitives of knowledge and influence using minimum distance estimation.

For each class of decisions c, I estimate model primitives $\boldsymbol{\theta}_c = (\rho_{0,c}, \rho_{1,c}, \rho_{2,c}, \rho_{3,c}, \delta_{1,c}, \delta_{2,c}, P_c)$

[36] Common title conventions may refer to trainees by their year of training: PGY1, PGY2, and PGY3 use the acronym "PGY" for "post-graduate year"; R1, R2, and R3 simply use "R" for "resident."

[37] While I consider the distribution of this "supervisory" information as having mean 0 in the simple model, this assumption is inconsequential, as it is by definition orthogonal to trainee knowledge. The "judgment" of the supervisory information can be viewed as captured by all terms other than the trainee effects in the regression Equation (3), including the error term.

by minimum distance:

$$\hat{\boldsymbol{\theta}}_c = \arg\min_{\boldsymbol{\theta}_c \in \Theta} \left(\hat{\boldsymbol{\sigma}}_c - \boldsymbol{\sigma}\left(\boldsymbol{\theta}_c\right)\right)' \mathbf{W} \left(\hat{\boldsymbol{\sigma}}_c - \boldsymbol{\sigma}\left(\boldsymbol{\theta}_c\right)\right),$$

where $\hat{\boldsymbol{\sigma}}_c$ is the vector of empirical estimates of practice variation corresponding to decisions in class c from the first step, with elements corresponding to $(\tau_h, \tau_{-h}) \in \mathcal{T}$; $\boldsymbol{\sigma}\left(\boldsymbol{\theta}_c\right)$ is the corresponding vector of model-implied practice variation from Equation (A-17) given $\boldsymbol{\theta}_c$; and $\mathbf{W}$ is a weighting matrix. Primitives may also be estimated on overall practice variation moments, in which case I omit labels of c.

Consistent with previous reduced-form estimation, I fit the model on $\|\mathcal{T}\| = 18$ moments of practice variation: I divide observations with residents in the second year of training into resident tenure blocks of 60 days, resulting in 6 resident moments and 6 intern moments of practice variation; I also divide observations with residents in the third year of training into resident tenure blocks of 120 days, resulting in 3 resident moments and 3 intern moments of practice variation. If $\sqrt{n}\left(\hat{\boldsymbol{\sigma}}_c - \boldsymbol{\sigma}\left(\boldsymbol{\theta}_c\right)\right) \xrightarrow{d} N\left(\mathbf{0}, \Omega_c\right)$, then the asymtotic variance of $\hat{\boldsymbol{\theta}}_c$ is given by

$$\text{Asy. Var}\, \hat{\boldsymbol{\theta}}_c = \frac{1}{n} \left(\Gamma\left(\boldsymbol{\theta}_{0,c}\right)' \mathbf{W} \Gamma\left(\boldsymbol{\theta}_{0,c}\right)\right)^{-1} \left(\Gamma\left(\boldsymbol{\theta}_{0,c}\right)' \mathbf{W} \Omega_c \mathbf{W} \Gamma\left(\boldsymbol{\theta}_{0,c}\right)\right) \left(\Gamma\left(\boldsymbol{\theta}_{0,c}\right)' \mathbf{W} \Gamma\left(\boldsymbol{\theta}_{0,c}\right)\right)^{-1},$$

where $\boldsymbol{\theta}_{0,c}$ is the true parameter vector, and $\Gamma\left(\boldsymbol{\theta}_{0,c}\right) = \text{plim}\, \partial \boldsymbol{\sigma}\left(\hat{\boldsymbol{\theta}}_c\right) / \partial \hat{\boldsymbol{\theta}}_c$ is an 18×7 matrix of analytical derivatives of Equation (A-17) with respect to $\boldsymbol{\theta}_c$, evaluated at $\hat{\boldsymbol{\theta}}_c$. The optimal weighting matrix is $\mathbf{W} = \hat{\Omega}_c^{-1}$, which I obtain from the first-step estimation of practice variation. This yields for inference

$$\widehat{\text{Var}}\, \hat{\boldsymbol{\theta}}_c = \frac{1}{n} \left(\Gamma\left(\hat{\boldsymbol{\theta}}_c\right)' \hat{\Omega}_c^{-1} \Gamma\left(\hat{\boldsymbol{\theta}}_c\right)\right)^{-1}.$$

I also calculate likelihood ratio tests for the joint-significance of learning and influence parameters against a restricted model with no learning but potentially inefficient senior influence via "status" (i.e., only $\rho_{0,c}$, $\delta_{1,c}$, and P_c are non-zero).

A-6.4 Results

In Table A-7, Column 1, I show baseline parameter estimates based on practice variation in overall spending. In Figure A-9, I show the implied path of practice variation according to the model and estimated parameters, overlaid on reduced-form estimates from Section 3. Structural estimates imply very little knowledge at the beginning of residency ($\rho_0 = 0.04$) compared to learning in the first year ($\rho_1 = 0.20$). Learning in the second year occurs at a rate 30 times faster than in the first year ($\rho_2 = 7.5$), but appears to cease by the third year ($\rho_3 = 0$). Between junior and senior trainees, influence approximates the Bayesian benchmark.[38] However, I find that the contribution of external information

[38] I estimate that $\delta_1 = 0.23$. Although this deviation from the Bayesian benchmark for senior trainees is large relative to knowledge at the end of the first year ($\rho_0 + \rho_1 = 0.24$), it is relatively small compared to learning that occurs in the second year ($\delta_1 / \rho_2 \cdot 365$ days = 11 days worth of second-year learning). I also estimate that $\delta_2 = -1.4$, which implies that third-year trainees have *less* influence than under the Bayesian benchmark, although this parameter is imprecisely estimated and small relative to ρ_2.

($P = 3.7$) is much lower than the knowledge of a graduating trainee ($\underline{P} \equiv \rho(3) \approx 7.74$). Since external information includes knowledge of supervising physicians who have completed training, this suggests that trainees are given much more influence than under the Bayesian benchmark.

I also estimate model parameters based on practice variation in spending specific to classes of decisions (Table A-7) and by types of patient-days (Table A-8). Learning is often greatest in the second year of training, regardless of the set of decisions. Decisions broken into components of diagnostic testing, prescriptions, blood transfusions, and nursing orders show somewhat less pronounced learning in the second year, which suggests potential interactions between components that are important for learning.

Based on likelihood ratio tests comparing the baseline model and more restrictive models, I can reject a model with no learning (i.e., $\rho_1 = \rho_2 = \rho_3 = 0$) and only senior prestige (i.e., $\delta_1 > 0$) for overall spending decisions (Column 1 of Table A-7) and for the majority of other outcomes or subsets of the data (Tables A-7 and A-8). On the other hand, if I allow for learning but impose the Bayesian benchmark influence between trainees (i.e., $\delta_1 = \delta_2 = 0$), the restricted model (Panel B of Figure A-10) fits the data quite well and cannot be rejected by the likelihood ratio test. Finally, I can strongly reject a model with strictly Bayesian influence between trainees and supervisors (i.e., $\delta_1 = \delta_2 = 0$, $P \geq \rho_0 + \rho_1 + \rho_2 + \rho_3$); the graphical fit of this model (Panel C of Figure A-10) is obviously problematic.

A-6.5 Counterfactual Analyses

A-6.5.1 Model of Learning

In my baseline results, I find that learning is low as a junior trainee in the first year, high as a senior trainee in the second year, and null in the third year. I interpret the first switch in the rate of learning—from low learning in the first year to high in the second—as due to the effect of influence on learning. $\tau = 1$ serves as an intuitive kink point for this switch.

I interpret the second switch in learning—from high learning in the second year to none in the third—as an indication that trainees have reached "full knowledge," after which learning stops, due to the relative benefits and costs of learning. It is not obvious why this kink in the rate of learning should occur at $\tau = 2$. Thus, the first step in my approach for counterfactual analyses is to specify a more flexible model of trainee learning, in which this kink point occurs at any $\tau = \tau_c \in (1,3)$ during the two years of the senior trainee role. In this model, trainee knowledge takes this form:

$$\rho(\tau) = \begin{cases} \rho_0 + \rho_1 \tau, & \tau \in [0,1]; \\ \rho_0 + \rho_1 + \rho_2 (\tau - 1), & \tau \in [1, \tau_c]; \\ \rho_0 + \rho_1 + \rho_2 (\tau_c - 1) + \rho_3 (\tau - \tau_c), & \tau \in [\tau_c, 3]. \end{cases} \tag{A-22}$$

Estimation of this more flexible model yields similar results to those from the baseline model: $\hat{\rho}_0 = 0.04$, $\hat{\rho}_1 = 0.20$, $\hat{\rho}_2 = 8.01$, $\hat{\rho}_3 = 0$, $\hat{\tau}_c = 1.87$, $\hat{\delta}_1 = 0.21$, $\hat{\delta}_2 = -1.42$, and $\hat{P} = 3.65$.

In counterfactual scenarios of learning, I assume that the rate of learning depends on influence, but

that learning continues until full knowledge has been reached. Parameters in Equation (A-22) imply that full knowledge is $\overline{\rho} = \hat{\rho}_0 + \hat{\rho}_1 + \hat{\rho}_2 (\hat{\tau}_c - 1) \approx 7.17$, which I consider as fixed in counterfactual scenarios. For the key relationship that drives learning from influence, I assume that the rates of learning during training, ρ_1 and ρ_2, are piecewise linear functions of the average influence of the trainee during the respective tenure intervals, $T_1 \equiv [0, 1]$ and $T_2 \equiv [1, \tau_c]$.

In notation, first define average influence over tenures uniformly distributed in interval T as

$$\overline{g}(T; \boldsymbol{\theta}) \equiv E_{\tau_h} \left[g(\tau_h; \tau_{-h}) \,|\, \boldsymbol{\theta} \right], \tag{A-23}$$

where influence $g(\tau_h; \tau_{-h})$ is given in Equation (A-16) and depends on $\boldsymbol{\theta} = (\rho_0, \rho_1, \rho_2, \rho_3, \delta_1, \delta_2, P)$. Consider a counterfactual scenario as defined by key parameters of supervisory information or influence, and denote the corresponding set of counterfactual parameters as $\boldsymbol{\theta}^\Delta$. Then a counterfactual rate of learning takes the following form: For $t \in \{1, 2\}$,

$$\rho_t^\Delta = \begin{cases} \hat{\rho}_1 \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right), & \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) \leq \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right), \\ \hat{\rho}_1 + \frac{\hat{\rho}_2 - \hat{\rho}_1}{\overline{g}\left(T_2; \hat{\boldsymbol{\theta}}\right) - \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right)} \left(\overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) - \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right) \right), & \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) > \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right). \end{cases} \tag{A-24}$$

Under estimated parameters $\hat{\boldsymbol{\theta}}$, the implied rates of learning are similar for $\overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right)$ above and below $\overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right)$: $\hat{\rho}_1 / \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right) \approx 13.2$, and $(\hat{\rho}_2 - \hat{\rho}_1) / \left(\overline{g}\left(T_2; \hat{\boldsymbol{\theta}}\right) - \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right) \right) \approx 14.6$.

A-6.5.2 Counterfactual Scenarios and Outcomes

I consider counterfactual scenarios defined by counterfactual supervisory information (P^Δ) or influence between trainees (δ_1^Δ and δ_2^Δ). A counterfactual scenario implies varying levels of influence along the entire course of training, as given by Equations (A-16) and (A-21). Influence also depends on knowledge, as given by Equation (A-22), which in turn depends on learning via influence, as given by (A-24).

Thus, I must find an internally consistent set of parameters $\boldsymbol{\theta}^\Delta$ that contains P^Δ. In all counterfactual scenarios, I hold fixed $\rho_0^\Delta = \hat{\rho}_0$ and $\rho_3^\Delta = \hat{\rho}_3 = 0$. In counterfactual scenarios involving P^Δ, I also hold fixed $\tilde{\delta}_1^\Delta \equiv \delta_1^\Delta / \left(\rho_0^\Delta + \rho_1^\Delta \right) = \delta_1 / (\rho_0 + \rho_1)$, since it is not possible to have $\delta_1^\Delta - \left(\rho_1^\Delta + \rho_0^\Delta \right) < 0$; I similarly hold fixed $\tilde{\delta}_2^\Delta \equiv \delta_2^\Delta / \min\left(\overline{\rho}, \rho_0^\Delta + \rho_1^\Delta + \rho_2^\Delta \right) = \delta_2 / \min\left(\overline{\rho}, \rho_0 + \rho_1 + \rho_2 \right)$. Conversely, for counterfactual scenarios involving influence between trainees, I vary $\tilde{\delta}_1^\Delta$ or $\tilde{\delta}_2^\Delta$ while holding fixed $P^\Delta = P$. Given these constraints, I identify an internally consistent $\boldsymbol{\theta}^\Delta$ by solving for ρ_1^Δ and ρ_2^Δ in the nonlinear system of two equations implied by Equations (A-16), (A-21), (A-22), (A-23), and (A-24), for $t \in \{1, 2\}$.

For each of the counterfactual scenarios, I consider the following outcomes of learning and decision-making information:

1. Time for trainees to acquire full knowledge:

$$\overline{\tau}^{\Delta} = 1 + \frac{\overline{\rho} - \left(\rho_0 + \rho_1^{\Delta}\right)}{\rho_2^{\Delta}}.$$

This calculated time summarizes the counterfactual rates of learning, ρ_1^{Δ} and ρ_2^{Δ}. Since learning is always incomplete in the first year of training under all counterfactual scenarios (i.e., $\rho_1^{\Delta} < \overline{\rho}$), this time is always greater than one year.

2. Average information from trainee knowledge: A trainee can contribute no more information than her knowledge, but she can contribute less if decision-making departs from the Bayesian benchmark. In other words, when working with peers of tenure τ_{-h}, trainees of tenure τ_h contribute precision equal to

$$\underline{\rho}^{\Delta}(\tau_h; \tau_{-h}) = \min\left(1, \frac{g(\tau_h; \tau_{-h})}{g^*(\tau_h; \tau_{-h})}\right) \rho^{\Delta}(\tau_h).$$

Counterfactual knowledge, $\rho^{\Delta}(\tau_h)$, is given by Equation (A-22) using the counterfactual parameters ρ_1^{Δ} and ρ_2^{Δ}; $\tilde{\rho}^{\Delta}(\tau)$, as given by Equation (A-21), may differ from $\rho^{\Delta}(\tau)$ if $\delta_1^{\Delta} \neq 0$ or $\delta_2^{\Delta} \neq 0$. For patients uniformly distributed over the course of an academic year, the average information from trainee teams is then

$$Q^{\Delta} = \int_0^1 \left(\lambda\left(\underline{\rho}^{\Delta}(\tau; \tau+1) + \underline{\rho}^{\Delta}(\tau+1; \tau)\right) + (1-\lambda)\left(\underline{\rho}^{\Delta}(\tau; \tau+2) + \underline{\rho}^{\Delta}(\tau+2; \tau)\right)\right) d\tau,$$

where $\lambda = 0.7$ is the approximate fraction of patients seen by teams with second-year trainees, and $1 - \lambda$ is the remaining fraction of patients seen by teams with third-year trainees. The three terms inside the integral represent levels of information contributed by first-, second-, and third-year trainees, respectively.

3. Average total information in decision-making: $P^{\Delta} + Q^{\Delta}$, or the sum of supervisory information and average information from trainee knowledge.

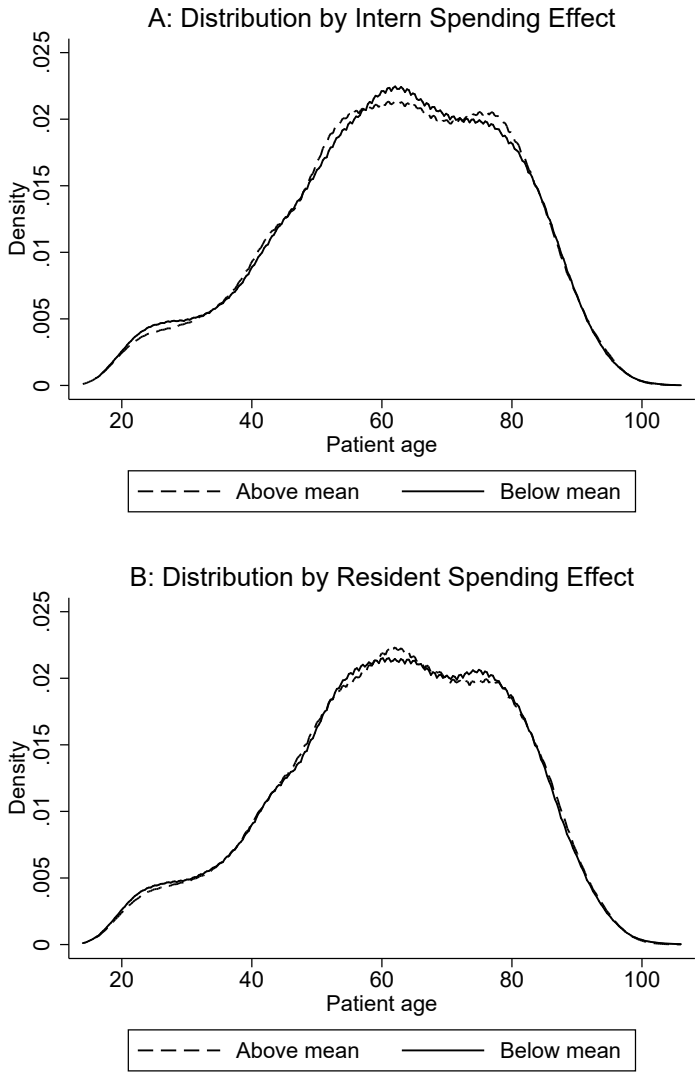
A-6.5.3 Discussion of Results

In Figure A-11, I show outcomes under counterfactual scenarios varying P^{Δ} and $\tilde{\delta}_1^{\Delta}$. As expected, increasing P^{Δ} slows the rate of learning and increases the time for trainees to acquire full knowledge. There are direct effects of P^{Δ} in decreasing trainee influence as well as indirect effects, as trainees with less influence acquire less knowledge to contribute to decision-making. Thus, increasing supervisory information decreases the information from trainee knowledge used in decision-making. The gain in total decision-making information is reduced by about 40% by this mechanism of diminishing trainee knowledge. In contrast, there is only limited impact of varying $\tilde{\delta}_1^{\Delta}$ on learning and trainee knowledge over the course of residency, at least in the range of $\tilde{\delta}_1^{\Delta} \in [-1, 1]$. By decreasing $\tilde{\delta}_1^{\Delta}$, trainees gain more knowledge when they are junior but less when they are senior. The effect of influence on learning

is slightly steeper for senior trainees, which explains why there are some slight returns to increasing $\tilde{\delta}_1^{\Delta}$ in terms of decreasing years to acquire full knowledge and increasing information from trainee knowledge in the average team decision.

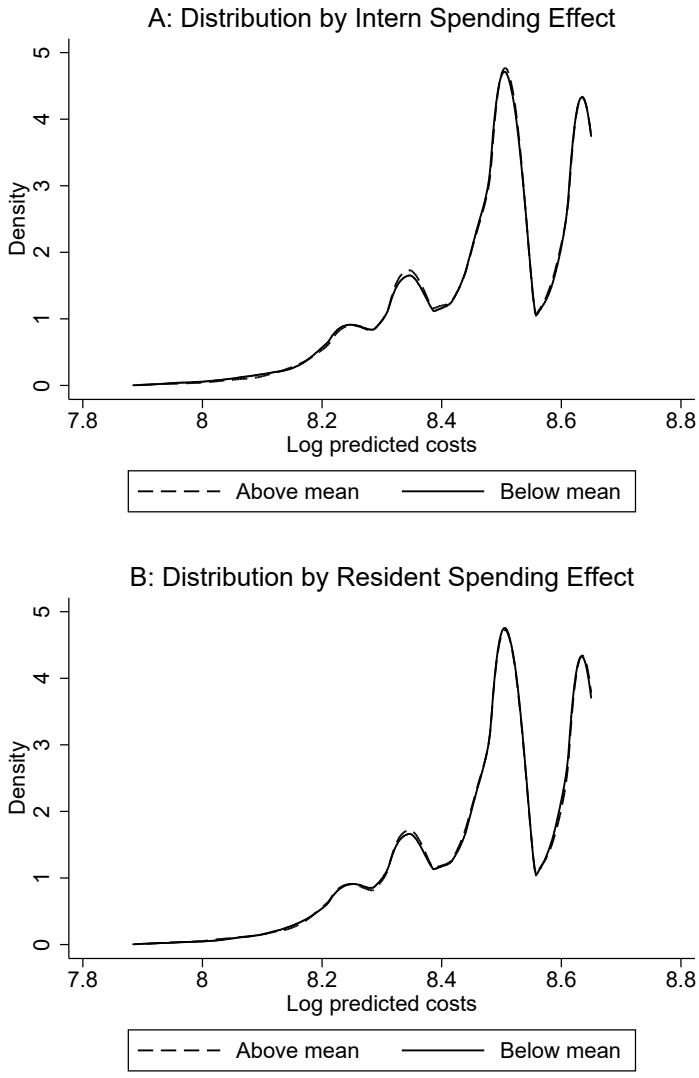
In Figure A-12, I show outcomes under counterfactual scenarios varying $\tilde{\delta}_2^{\Delta}$. The effects of increasing $\tilde{\delta}_2^{\Delta}$ on learning and decision-making information are similar to those of increasing $\tilde{\delta}_1^{\Delta}$: Increasing senior influence speeds up training and increases overall trainee knowledge. The effect range of counterfactual values of δ_2^{Δ} is larger, since the denominator in $\tilde{\delta}_2^{\Delta}$ (i.e., $\rho^{\Delta}(2)$) is larger. Interestingly, around $\tilde{\delta}_2^{\Delta} = 0$, decreasing $\tilde{\delta}_2^{\Delta}$ has a larger effect on Q^{Δ} than does increasing $\tilde{\delta}_2^{\Delta}$, due to the following intuition: Near baseline parameters, much of the third year involves no learning. Therefore, increasing the influence of third-year trainees does not aid learning for those trainees, and learning among junior trainees will suffer. However, learning indirectly increases for second-year trainees who then work with less knowledgeable junior trainees. Nonetheless, the effects on learning are generally small relative to those for varying P^{Δ}.

Figure A-1: Patients Age by Housestaff Spending Effect



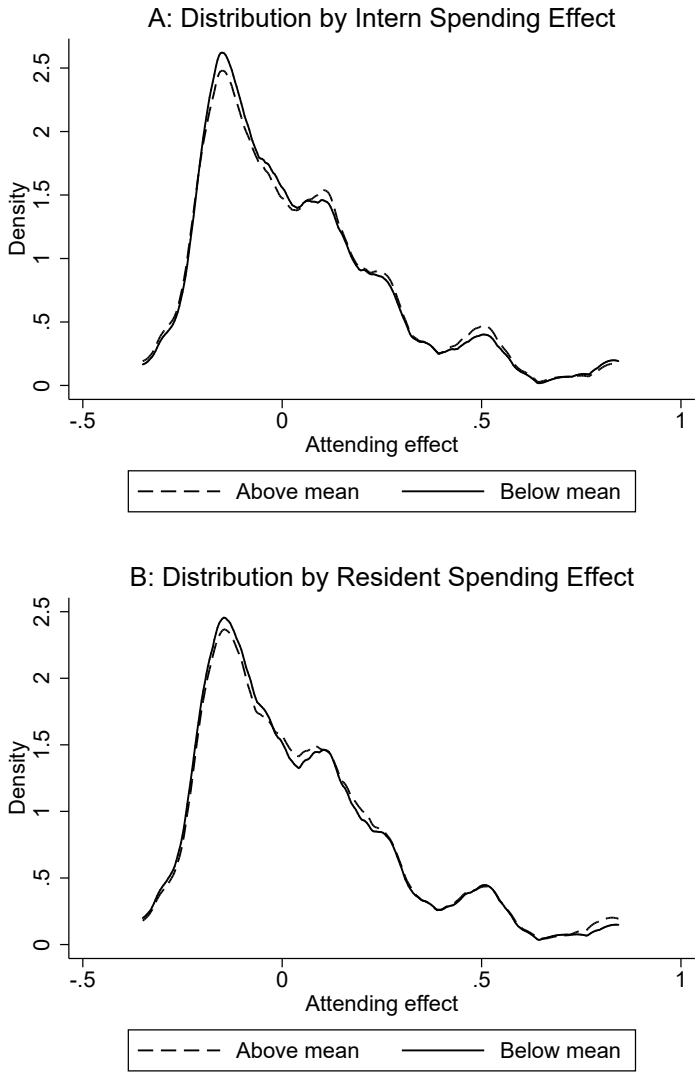
Note: This figure shows the distribution of the age of patients assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.496 and 0.875 for interns (Panel A) and residents (Panel B), respectively.

Figure A-2: Demographics-predicted Spending by Trainee Spending Effect



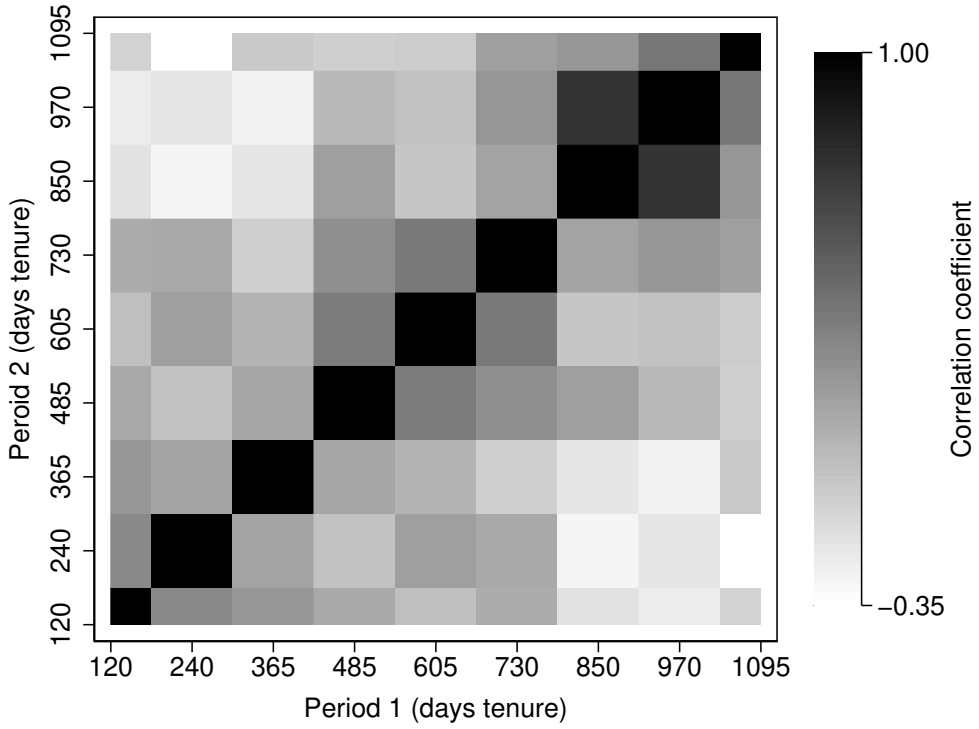
Note: This figure shows the distribution of predicted log costs (based on patient age, race, and gender) for patients assigned interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.683 and 0.745 for interns (Panel A) and residents (Panel B), respectively.

Figure A-3: Attendings Spending Effects by Trainee Spending Effect



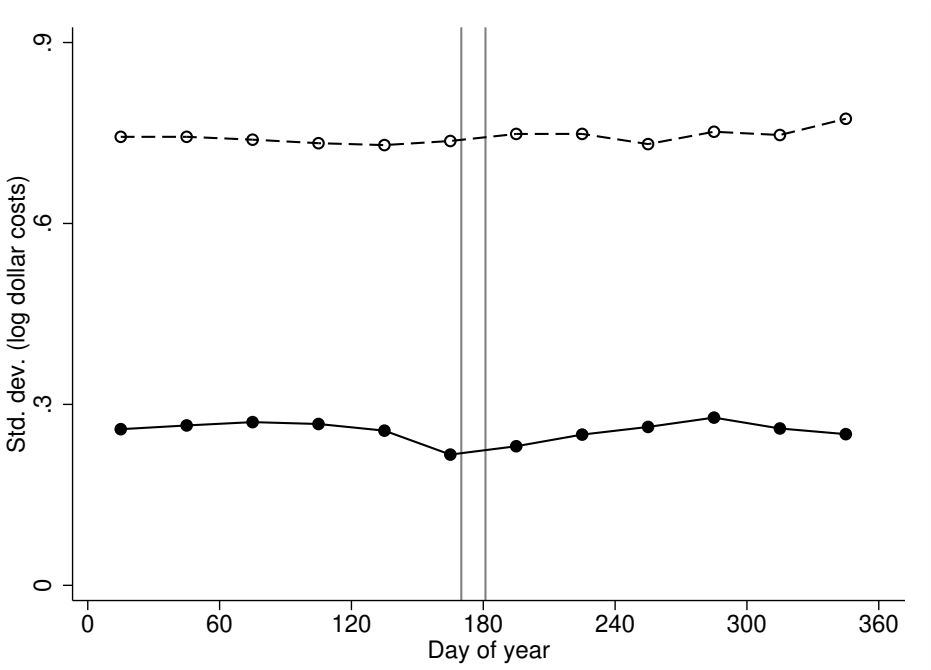
Note: This figure shows the distribution of spending fixed effects for attendings assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee and attending spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.059 and 0080 for interns (Panel A) and residents (Panel B), respectively.

Figure A-4: Serial Correlation of Trainee Random Effects



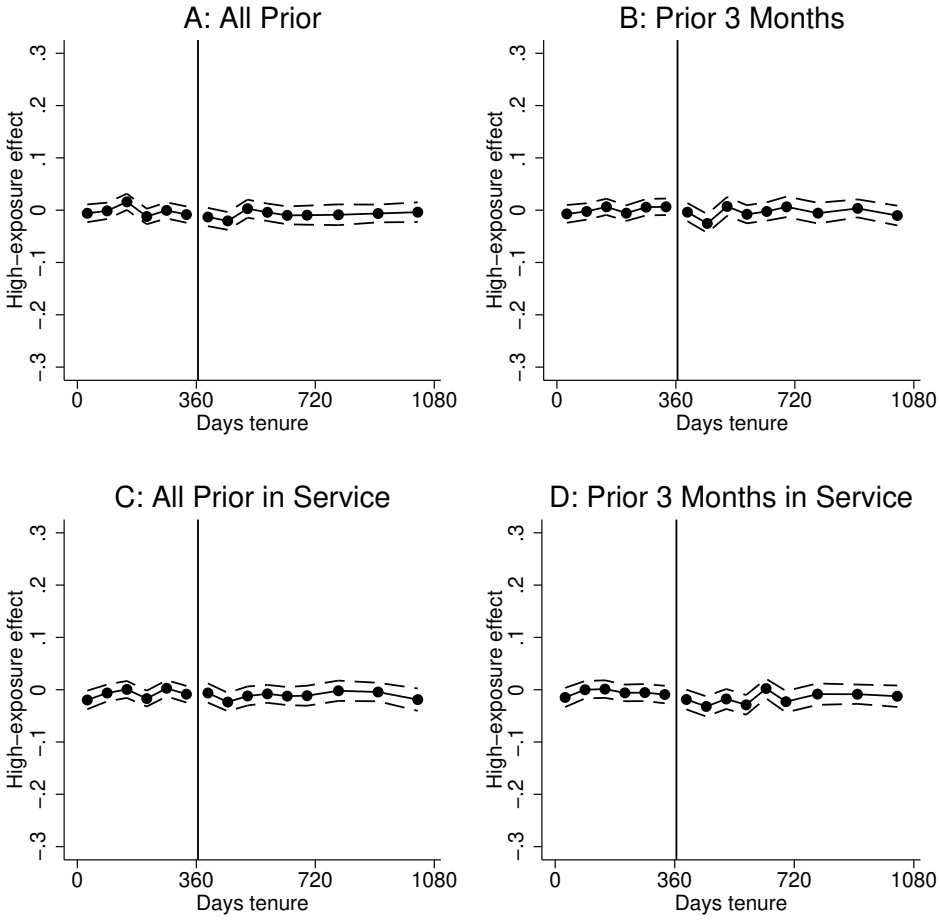
Note: This figure shows the serial correlation between random effects within trainee between two tenure periods. Details of the estimation routine are given in Appendix A-3.2. The random effect model of log daily total costs is given in Equation (3). The model controls are as stated for Figure 1. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure. Numerical values and confidence intervals are given in Table A-4.

Figure A-5: Trainee-associated and Residual Variation by Day of Year



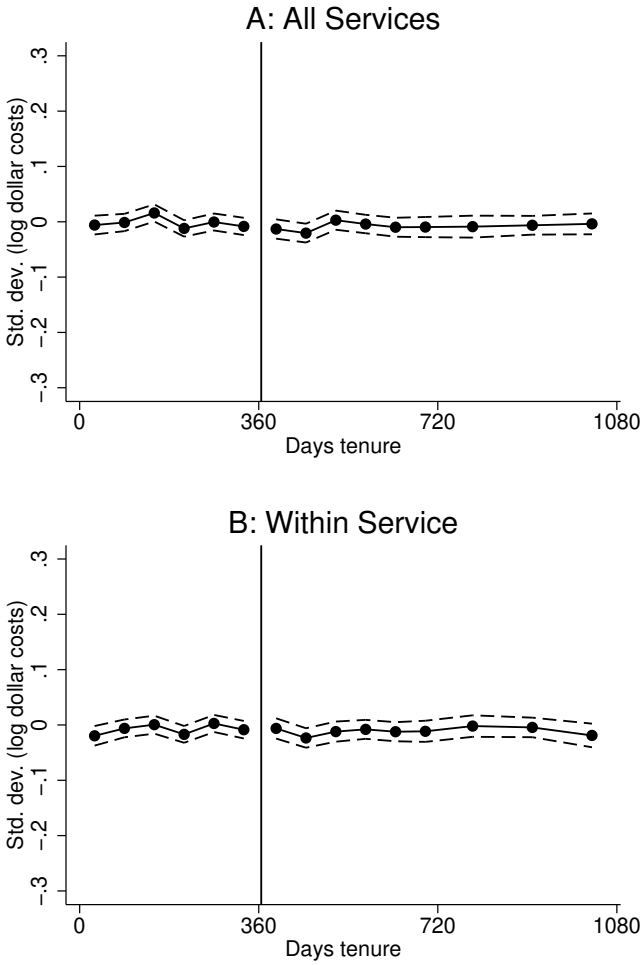
Note: This figure shows the standard deviation of random effects due to junior and senior trainee teams (solid dots) and the standard deviation of the residual (hollow dots) in 30-day periods by day of the year. Residual variation can be interpreted as variation due to independent observation. The two vertical gray lines indicate when new junior trainees begin residency on July 19 and when senior trainees advance a year on July 28 (i.e., becoming a new second-year senior trainee, becoming a third-year trainee, or completing residency). The model is similar to Equation (3), except that a single random effect is modeled for the junior and senior trainee combination, instead of two additively separable random effects for the respective trainees. Controls are given in the note for Figure 1.

Figure A-6: Effect of High Prior Exposure to Spending



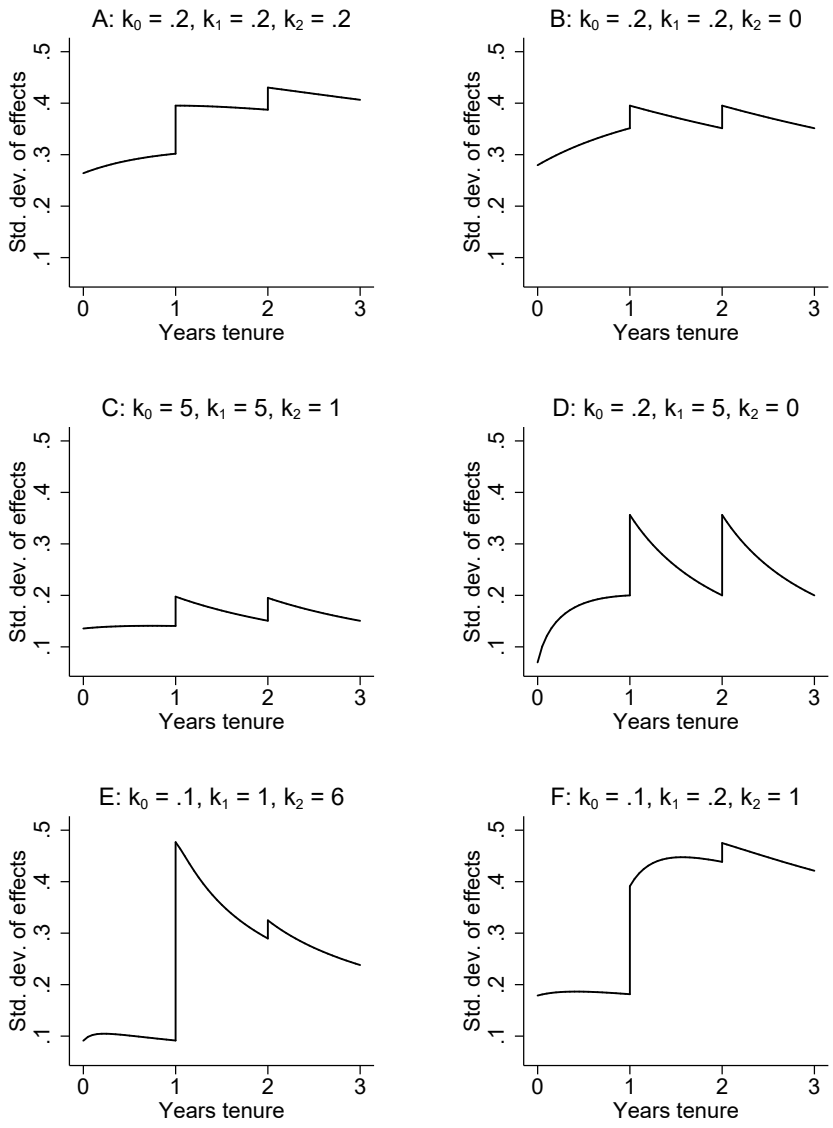
Note: This figure shows the effect of high prior exposure to supervising-physician spending. This exposure measure is discussed in Section and in Table A-5 and reflects the average spending effects of supervising physicians that a given trainee was matched to in the past. The tenure-specific effect of having high prior exposure to spending is estimated as in Equation (A-12). Panel A uses an exposure measure that includes all prior matches, regardless of service (corresponding to Column 1, Panel A of Table A-5). Panels B and D use an exposure measure that includes matches within the last three months with supervising physicians (corresponding to Columns 2 and 4, Panel A of Table A-5). Panels C and D use an exposure measure that is restricted to prior matches on the same service (corresponding to Columns 3 and 4, Panel A of Table A-5). The vertical line indicates the one-year mark of training; trainees are junior prior to this and senior after this. The model controls are as stated for Figure 1. The effect of high prior exposure to senior-trainee spending is shown in Figure A-7. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure A-7: Effect of High Exposure to Senior-trainee Spending



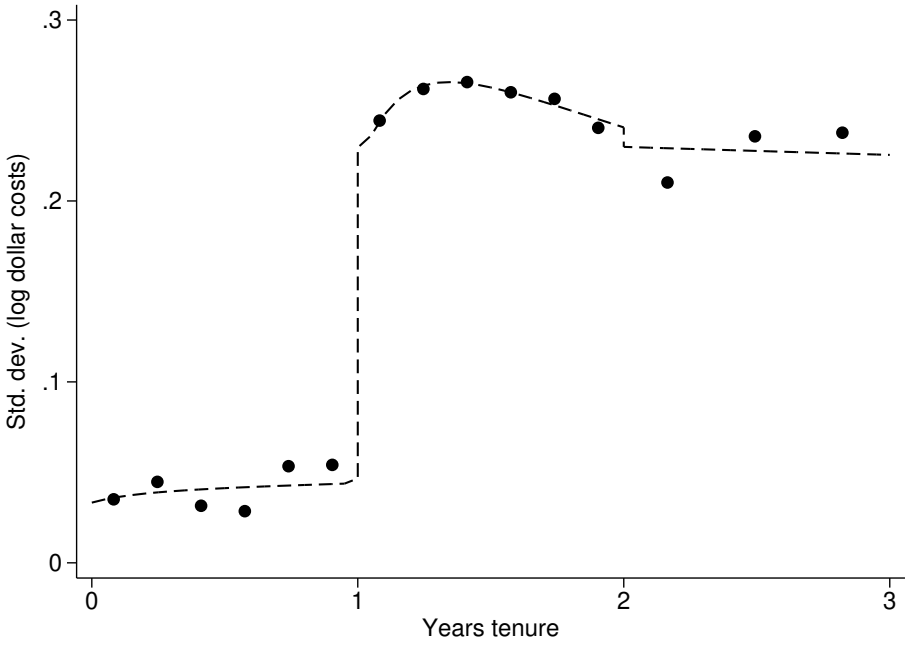
Note: This figure shows the effect of high prior exposure to senior-trainee spending. This exposure measure is discussed in further detail in Appendix A-5 and in Table A-5 and reflects the average spending effects of senior trainees that a given trainee was matched to in the past as a junior trainee. The tenure-specific effect of having high prior exposure to spending is estimated as in Equation (A-12). Panel A uses an exposure measure that includes all prior matches with senior trainees, regardless of the ward service (corresponding to Column 1, Panel B of Table A-5); Panel B uses an exposure measure that is restricted to prior matches on the same service (corresponding to Column 3, Panel B of Table A-5). For tenure periods after the one-year mark (shown as the vertical line), the trainee of interest is senior, and matches with senior trainees all date back to the trainee's first year of training as a junior trainee. The model controls are as stated for Figure 1. The effect of high prior exposure to supervising-physician spending is shown in Figure A-6.

Figure A-8: Numerical Examples of Variation Profiles



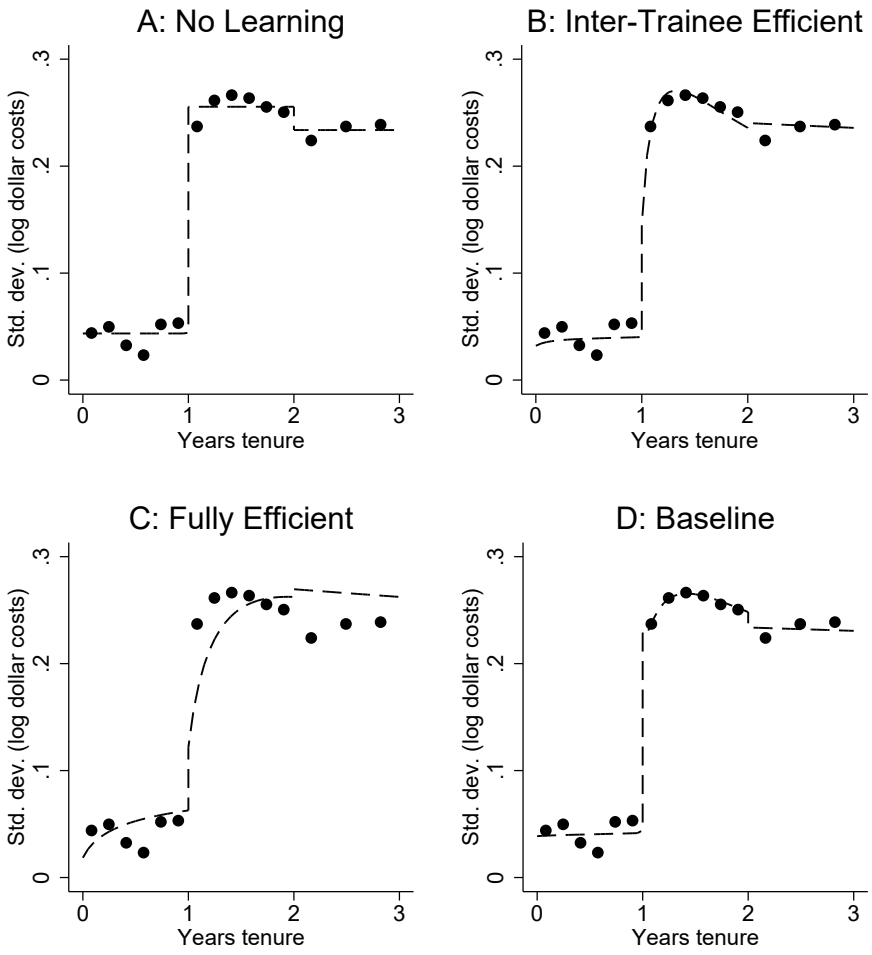
Note: This figure shows variation profiles of the expected standard deviation of trainee effects over tenure, $\sigma(\tau)$, differing by the underlying profile of learning over tenure. Learning is parameterized as a piecewise linear function $g(\tau)$ that describes how the precision of subjective priors increases over tenure. In particular, this figure considers piecewise linear functions of the form (A-20), parameterized by ρ_0, ρ_1, and $\rho_2 = \rho_3$. Each panel considers a different set of parameters of $\rho(\tau)$. Given $\rho(\tau)$, I calculate the expected standard deviation of trainee effects over tenure using Equation (A-18). I assume that interns are equally likely to work with second-year residents and third-year residents. These profiles are discussed further in Appendix A-6.2.

Figure A-9: Model Fit to Practice Variation Profile



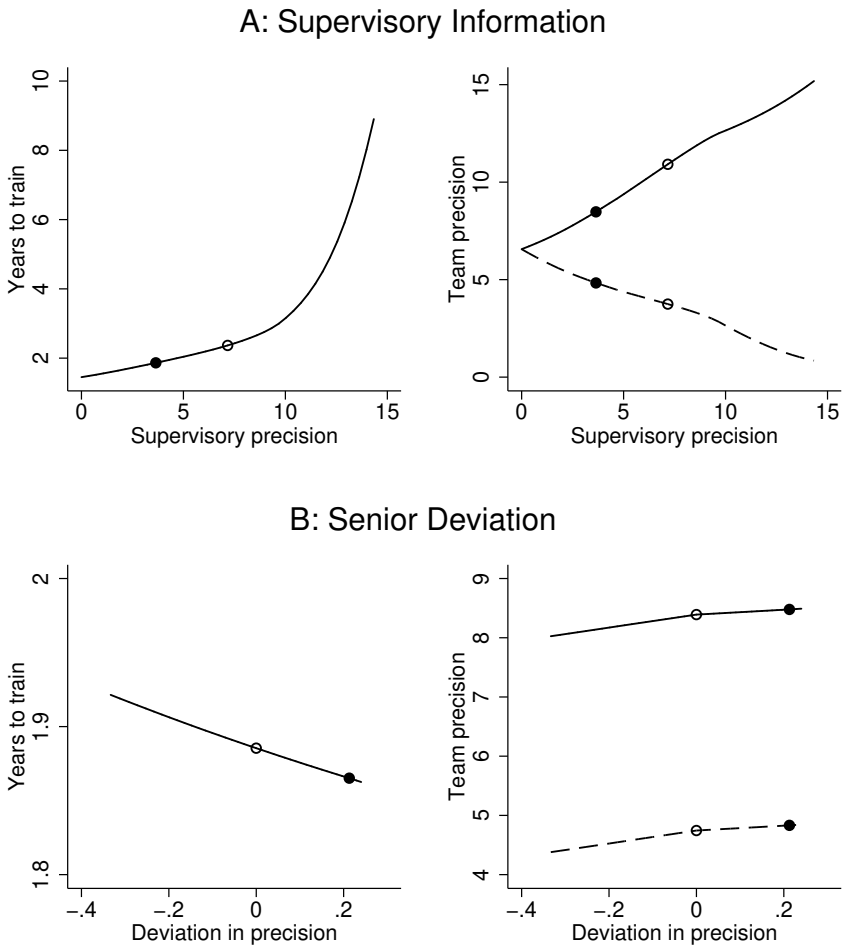
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (3), in log daily total costs at each non-overlapping tenure period. Trainee prior to one year in tenure are junior trainees and become senior trainees after one year in tenure. Reduced-form estimates of practice variation are shown in dots and are the same as shown in Figure 1. Practice variation implied by the model of learning and influence, specifically Equation (A-17), is shown as a dashed line. Estimation of parameters of this model is described in Section 5.3. The Sargan-Hansen over-identification J-test statistic of the model is $J = 8.60$, which is less than the 95th percentile value of 19.7 the χ^2_{18-7} distribution (the p-value corresponding to $J = 8.60$ is 0.67)

Figure A-10: Model Restrictions



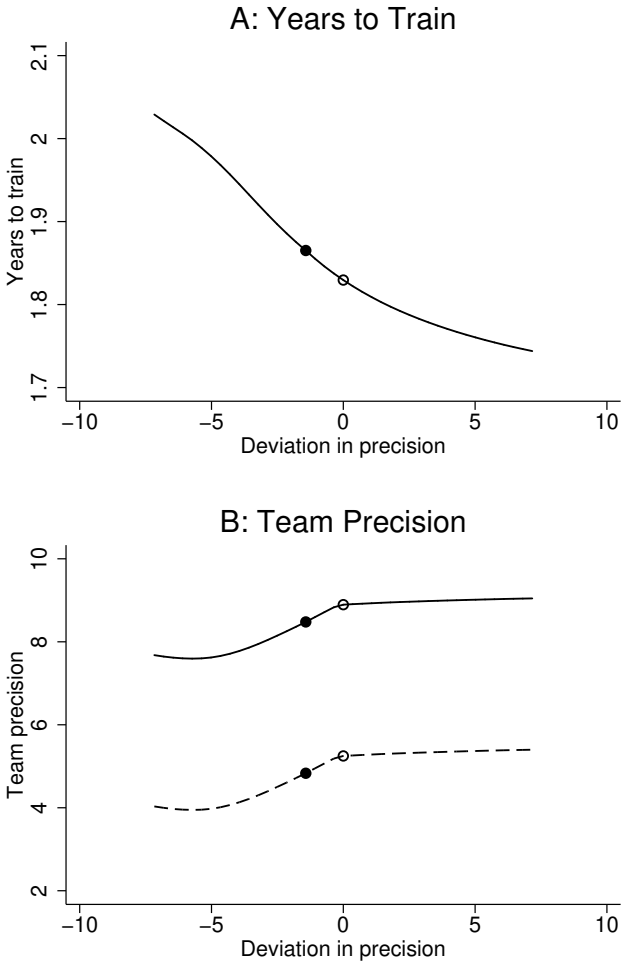
Note: This figure shows the fit of restricted models of learning and influence, with parameters described in Table A-7. Each panel shows the same reduced-form moments of practice variation for each tenure period, which are also the same as those shown in Figure A-9, reproduced in Panel D. Panel A restricts the model to no learning (i.e., $\rho_1 = \rho_2 = \rho_3 = 0$). Panel B restricts the model to the Bayesian benchmark of influence between trainees (i.e., $\delta_1 = \delta_2 = 0$). Panel C additionally restricts the model so that supervisors receive as much influence as warranted by the lower bound of their knowledge (i.e., $\delta_1 = \delta_2 = 0$, $P \geq \rho_0 + \rho_1 + \rho_2 + \rho_3$). The likelihood ratio test comparing a no-learning model (Panel A) with the baseline model (Panel D) rejects the restricted model with a p-value less than 0.01. Likelihood ratio tests for other outcomes or for subsets of the data are also given in Tables A-7 and A-8. Sargan-Hansen over-identification J-test statistics are 15.66 (p-value = 0.405 under χ^2_{18-3} distribution) for Panel A, 13.42 (p-value = 0.416 under χ^2_{18-5} distribution) for Panel B, and 65.97 (p-value < 0.01 under χ^2_{18-4} distribution).

Figure A-11: Counterfactual Training Time and Team Information



Note: This figure shows counterfactual results on time for trainees to acquire "full knowledge" and on information used in decision-making. I consider two types of counterfactual scenarios: In subpanels in Panel A, I alter on the x-axes the amount of supervisory information used in decision-making, or P in the model, while holding fixed the relative influence between junior and senior trainees. In subpanels in Panel B, I alter on the x-axes the relative influence between junior and senior trainees, or δ_1 in the model, while holding fixed the amount of supervisory information. Appendix Figure A-12 shows results for varying δ_2 in the model. The time for trainees to acquire full knowledge (or "years to train") is measured on the y-axes of the left subpanels, and the information used in decision-making is measured on the y-axes of the right subpanels. The right subpanels show both information from trainee knowledge (dashed lines) and total information (solid lines) used in decision-making. On each line, I plot a solid dot indicating actual results and a hollow dot indicating counterfactual results under Bayesian-benchmark influence; supervisory influence in Panel A is a lower bound for the Bayesian benchmark that equals full trainee knowledge, or $\underline{P} = \rho_0 + \rho_1 + \rho_2 (\tau_c - 1)$. Lines in Panel A are plotted for counterfactual $P^{\Delta} \in [0, 2\underline{P}]$; lines in Panel B are plotted for counterfactual $\delta_1^{\Delta} / \left(\rho_0^{\Delta} + \rho_1^{\Delta}\right) \in [-1, 1]$. Further details are given in Appendix A-6.5.

Figure A-12: Counterfactual Results, Varying δ_2



Note: This figure shows results for counterfactual scenarios in which I vary the additional deviation in effective precision for third-year trainees, or δ_2 in the model and shown in the x-axes of both panels. The y-axis of Panel A plots the time for trainees to acquire "full knowledge" (or "years to train"). The y-axis of Panel B plots information from trainee knowledge (dashed lines) and total information (solid lines) used in decision-making. On each line, I plot a solid dot indicating actual results and a hollow dot indicating counterfactual results under the Bayesian benchmark. Lines are plotted for counterfactual $\delta_2^{\Delta}/\rho^{\Delta}(2) \in [-1, 1]$. Further details are given in Appendix A-6.5.

Table A-1: Tests of Joint Significance of Trainee Identities and Characteristics

	Independent variables		
	Trainee identities	Trainee characteristics	
Patient characteristic	(1)	(2)	(3)
Age	$F(1055, 46364) = 0.98$ $p = 0.661$	$F(22, 14568) = 0.49$ $p = 0.978$	$F(20, 32434) = 0.55$ $p = 0.945$
Male	$F(1055, 46364) = 1.01$ $p = 0.378$	$F(22, 14568) = 1.16$ $p = 0.276$	$F(20, 32434) = 1.07$ $p = 0.376$
White	$F(1055, 46364) = 1.02$ $p = 0.356$	$F(22, 14568) = 0.72$ $p = 0.829$	$F(20, 32434) = 0.77$ $p = 0.756$
Predicted spending	$F(1055, 46364) = 0.98$ $p = 0.685$	$F(22, 14568) = 0.71$ $p = 0.836$	$F(20, 32434) = 1.12$ $p = 0.322$

Note: This table reports tests of joint significance corresponding to Equations (A-1) and (A-2). Column 1 corresponds to Equation (A-1); Columns 2 and 3 correspond to (A-2). Column 2 includes all trainee characteristics: trainee's position on the rank list; USMLE Step 1 score; sex; age at the start of training; and dummies for whether the trainee graduated from a foreign medical school, whether he graduated from a rare medical school, whether he graduated from medical school as a member of the AOA honor society, whether he has a PhD or another graduate degree, and whether he is a racial minority. Column 3 includes all trainee characteristics except for position on the rank list. Rows correspond to different patient characteristics as the dependent variable of the regression equation; the last row is predicted spending using patient demographics (age, sex, and race). F-statistics and p-values are reported for each joint test.

Table A-2: Practice Style Distribution by Junior Trainee Characteristic

Characteristic	Above median	Cases	Trainees	Characteristic		Trainee empirical Bayes posterior		
				Mean	Range	Mean	10th percentile	90th percentile
Age	Y	92,482	183	30.6	[28.1, 39.4]	0.000	−0.025	0.024
	N	92,381	184	26.8	[24.1, 28.1]	0.001	−0.025	0.025
AOA honor society	Y	63,864	124	1	[1, 1]	0.001	−0.023	0.027
	N	121,429	243	0	[0, 0]	0.000	−0.025	0.023
Future Income ($ thousands)	Y	107,080	275	424	[357, 850]	0.001	−0.022	0.026
	N	107,222	391	268	[199, 357]	0.000	−0.026	0.024
Male	Y	110,316	216	1	[1, 1]	0.000	−0.024	0.026
	N	74,977	151	0	[0, 0]	0.001	−0.027	0.023
Medical school rank	Y	87,149	175	24.7	[6, 84]	0.000	−0.022	0.024
	N	87,254	176	2.2	[1, 6]	0.000	−0.027	0.025
Minority	Y	22,564	43	1	[1, 1]	0.003	−0.022	0.023
	N	162,729	324	0	[0, 0]	0.000	−0.025	0.025
Other advanced degree	Y	62,433	119	1	[1, 1]	0.000	−0.024	0.024
	N	122,860	248	0	[0, 0]	0.001	−0.026	0.024
Rank list position	Y	65,267	132	80	[51, 255]	0.001	−0.024	0.026
	N	65,221	128	26	[1, 51]	0.001	−0.024	0.024
USMLE Step 1	Y	91,842	186	256	[246, 272]	0.001	−0.024	0.027
	N	92,254	192	233	[184, 246]	0.000	−0.025	0.023

Note: This table reports the distribution of empirical Bayes posterior means for junior trainees divided in groups by characteristics. Posterior means are calculated as in the note for Figure 6 and pooled across all tenure periods in the first year. Results for senior trainees are given in Table A-3. Cases denote the number of observations (i,t) such that the junior trainee belongs to the group. The mean, 10th percentile, and 90th percentile of each Empirical Bayes posterior distribution are all calculated over the set of relevant cases. Characteristics are discussed further in Section 2.3. Age refers to the trainee age in years at the start of residency. AOA (Alpha Omega Alpha) honor society inducts 13.5% of US medical graduates, and inclusion in AOA is associated with an odds ratio of 6-10 of successful matching into the first-choice residency (Rinard and Mahabir, 2010). Future income is imputed by future observed specialty or subspecialty entry of the trainee. Medical school rank is obtained from the *US News & World Report*. The national mean and 95th percentile of the United States Medical Licensing Examination (USMLE) Step 1 test scores are approximately 229 and 260, respectively.

Table A-3: Practice Style Distribution by Senior Trainee Characteristic

Characteristic	Above median	Cases	Trainees	Characteristic		Trainee empirical Bayes posterior		
				Mean	Range	Mean	10th percentile	90th percentile
Age	Y	98,703	167	30.3	[28, 37.9]	−0.020	−0.256	0.321
	N	100,712	169	26.8	[24.1, 28.0]	0.050	−0.234	0.359
AOA honor society	Y	70,334	112	1	[1, 1]	0.010	−0.243	0.344
	N	129,668	223	0	[0, 0]	0.020	−0.247	0.341
Future Income ($ thousands)	Y	100,571	161	411	[357, 540]	0.047	−0.241	0.351
	N	98,087	187	246	[199, 357]	−0.013	−0.256	0.330
Male	Y	112,520	188	1	[1, 1]	0.012	−0.260	0.338
	N	87,482	147	0	[0, 0]	0.022	−0.228	0.344
Medical school rank	Y	91,874	149	26.9	[7, 84]	−0.005	−0.243	0.330
	N	91,722	165	2.7	[1, 7]	0.047	−0.244	0.353
Minority	Y	23,108	40	1	[1, 1]	0.000	−0.246	0.320
	N	176,894	295	0	[0, 0]	0.019	−0.244	0.345
Other advanced degree	Y	48,765	83	1	[1, 1]	−0.009	−0.261	0.324
	N	151,237	252	0	[0, 0]	0.025	−0.242	0.349
Rank list position	Y	64,025	101	80	[54, 246]	0.089	−0.255	0.363
	N	63,967	104	28	[1, 54]	0.053	−0.253	0.342
USMLE Step 1	Y	97,224	160	255	[246, 275]	0.031	−0.247	0.350
	N	96,916	172	232	[156, 246]	0.005	−0.243	0.333

Note: This table reports the distribution of empirical Bayes posterior means for senior trainees divided in groups by characteristics. Posterior means are calculated as in the note for Figure 6 and pooled across all tenure periods in the second and third years. Results for junior trainees are given in Table A-2. Cases denote the number of observations (i,t) such that the junior trainee belongs to the group. The mean, 10th percentile, and 90th percentile of each Empirical Bayes posterior distribution are all calculated over the set of relevant cases. Characteristics are discussed further in Section 2.3. Age refers to the trainee age in years at the start of residency. AOA (Alpha Omega Alpha) honor society inducts 13.5% of US medical graduates, and inclusion in AOA is associated with an odds ratio of 6-10 of successful matching into the first-choice residency (Rinard and Mahabir, 2010). Future income is imputed by future observed specialty or subspecialty entry of the trainee. Medical school rank is obtained from the *US News & World Report*. The national mean and 95th percentile of the United States Medical Licensing Examination (USMLE) Step 1 test scores are approximately 229 and 260, respectively.

Table A-4: Serial Correlation in Trainee Effects

	Correlation in Trainee Effects							
	Period 2 (days)							
Period 1 (days)	121-240	241-365	366-485	486-605	606-730	731-850	851-970	971-1095
0-120	0.27 [−0.02, 0.58]	0.20 [−0.09, 0.58]	0.11 [−0.15, 0.40]	−0.02 [−0.29, 0.27]	0.09 [−0.22, 0.43]	−0.19 [−0.72, 0.51]	−0.25 [−0.61, 0.39]	−0.10 [−0.50, 0.39]
121-240		0.14 [−0.14, 0.54]	−0.03 [−0.32, 0.39]	0.15 [−0.15, 0.44]	0.11 [−0.25, 0.52]	−0.30 [−0.80, 0.28]	−0.21 [−0.58, 0.33]	−0.35 [−0.86, 0.28]
241-365			0.12 [−0.15, 0.43]	0.06 [−0.22, 0.35]	−0.09 [−0.45, 0.29]	−0.21 [−0.67, 0.42]	−0.27 [−0.61, 0.21]	−0.05 [−0.45, 0.40]
366-485				0.33 [0.03, 0.67]	0.24 [−0.12, 0.69]	0.16 [−0.28, 0.59]	0.02 [−0.44, 0.45]	−0.10 [−0.45, 0.27]
486-605					0.36 [0.03, 0.72]	−0.05 [−0.47, 0.42]	−0.03 [−0.31, 0.28]	−0.08 [−0.46, 0.39]
606-730						0.13 [−0.16, 0.46]	0.20 [−0.12, 0.51]	0.15 [−0.27, 0.58]
731-850							0.72 [0.37, 1.00]	0.21 [−1.00, 1.00]
851-970								0.38 [−0.02, 0.75]

Note: This table displays bootstrapped correlation parameters estimated from the model described in Section 3.4 and Appendix A-3.2. In each bootstrapped run, observations are drawn with replacement from strata defined by the junior and senior trainees, the admission service, and the tenure periods of each trainee. As described in Section 3.4, log spending is residualized by patient characteristics and fixed effects for time categories, clinical service, and supervising physician. These residuals are then used in maximum likelihood to estimate the correlation between trainee effects, as described in Appendix A-3.2. Each cell in the table corresponds to the correlation between trainee effects in two tenure periods. The mean bootstrapped correlation parameter is shown in the first line of each cell; the 95% confidence interval is shown in the second line in brackets. Mean correlation parameters are also shown in Figure (A-4).

Table A-5: Differences in Prior Exposure to Spending

	Differences Between High and Low Exposure			
	(1)	(2)	(3)	(4)
	All services		Within service	
Tenure period (days)	All prior	Prior 3 months	All prior	Prior 3 months
Panel A: Exposure to Spending by Supervising Physicians				
0-60	5.31%	5.65%	4.62%	4.84%
61-120	5.16%	5.52%	4.81%	5.03%
121-180	4.64%	5.41%	4.39%	4.87%
181-240	4.47%	5.43%	3.85%	4.41%
241-300	4.06%	5.21%	3.85%	4.41%
301-365	3.81%	4.92%	3.31%	4.28%
366-425	3.54%	5.80%	3.87%	5.41%
426-485	3.70%	6.06%	4.05%	6.04%
486-545	3.30%	5.71%	3.31%	4.83%
546-605	3.15%	5.27%	3.67%	5.47%
606-665	3.34%	6.01%	4.05%	6.26%
666-730	3.39%	5.91%	3.44%	5.24%
731-850	3.53%	4.97%	2.22%	3.78%
851-970	3.52%	5.82%	2.56%	4.05%
971-1095	3.03%	3.91%	1.80%	3.02%
Panel B: Exposure to Spending by Senior Trainees				
0-60	19.08%	19.50%	20.82%	20.92%
61-120	19.88%	20.32%	22.89%	23.02%
121-180	19.54%	21.03%	21.51%	23.23%
181-240	19.52%	20.54%	22.12%	23.53%
241-300	19.04%	20.12%	21.95%	23.61%
301-365	17.76%	17.99%	19.88%	20.28%

Note: This table presents differences in average spending effects of supervising physicians (Panel A) and of senior trainees (Panel B) who worked with trainees in the past at each tenure period for the trainees. Columns 1 and 2 include prior team pairings in all services, while Columns 3 and 4 only include prior team pairings within the same service. For example, for an observation in the cardiology service, Columns 3 and 4 only include prior team pairings for a trainee while working in the cardiology service. Columns 2 and 4 further restrict prior team pairings to those within the last three months. The spending effect of the relevant supervising physician or senior trainee is the empirical Bayes posterior mean from a random-effects model of log daily overall spending. Of the set of eligible prior team pairings, the exposure to spending measure is a weighted average (by patient-day) of the spending effects of the relevant matched physician (i.e., either the supervising physician or the senior trainee). Trainees in a given tenure period are categorized as having "high exposure" to spending if this measure is above the median measure for trainees in the same tenure period. The difference in exposure to spending between high and low exposure is simply the average measure for high-exposure trainees subtracted by the average measure for low-exposure trainees in a given tenure period.

Table A-6: Effect of Trainee Experience on Spending

	Log daily total costs				
	(1) Number of days	(2) Number of patients	(3) Number of attendings	(4) Attending spending	(5) Attending spending
Panel A: Interns					
Effect of trainee with	0.001	0.003	-0.001	-0.010**	-0.001
measure above median	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Observations	182,500	182,500	182,500	156,545	131,654
Adjusted R^2	0.088	0.088	0.088	0.089	0.089
Panel B: Residents					
Effect of trainee with	0.005	-0.005	-0.001	0.010*	0.013***
measure above median	(0.007)	(0.008)	(0.007)	(0.005)	(0.005)
Observations	200,266	200,266	200,266	182,982	176,086
Adjusted R^2	0.089	0.089	0.089	0.086	0.086
Measure and median within service	Y	Y	Y	N	Y

Note: This table reports results for some regressions of the effect of indicators of trainee experience. Panel A shows results for interns; Panel B shows results for residents. Regressions are of the form in Equation (A-9), where the coefficient of interest is on an indicator for a group of trainees identified whether their measure (e.g., number of days) is above the median within a 60-day tenure interval (across all trainees). The relevant tenure interval is the tenure interval before the one related to the day of the index admission. All columns except for (4) represent measures and medians that are calculated within service (e.g., number of days is calculated separately for a trainee within cardiology, oncology, and general medicine and compared to medians similarly calculated within service). Columns 4 and 5 feature a measure of attending spending, which is the average cumulative effect of attending physicians who worked with the trainee of interest up to the last prior tenure interval. Attending "effects" are calculated by a random effects method that adjusts for finite-sample bias; since patients are not as good as randomly assigned to attending physicians, these effects do not have a strict causal interpretation at the level of the attending physician. Other specifications (e.g., calculating all measures across services, or not conditioning on trainee identity) were similarly estimated as insignificant and omitted from this table for brevity. All models control for patient and admission characteristics, time dummies, and fixed effects for attending and the other trainees on the team (e.g., the resident is controlled for if the group is specific to the intern). Standard errors are clustered by admission.

Table A-7: Model Parameter Estimates for Overall Spending and by Spending Category

		Spending Category			
	(1)	(2)	(3)	(4)	(5)
	Overall	Diagnostic	Transfusion	Medication	Nursing
Knowledge parameters					
Prior to training (ρ_0)	0.039	0.936	0.225	1.005	0.000
	(0.032)	(0.235)	(0.341)	(0.043)	(0.000)
First year (ρ_1)	0.204	0.296	0.361		4.172
	(0.138)	(0.332)	(0.339)		(0.654)
Second year (ρ_2)	7.542	0.263	0.245		15.357
	(2.307)	(0.140)	(0.343)		(2.556)
Third year (ρ_3)	0.000	0.000			4.501
	(0.000)	(0.000)			(4.344)
Influence parameters					
Deviation after first year (δ_1)	0.231	4.388	0.349	0.725	−2.784
	(0.223)	(0.433)	(0.439)	(0.053)	(0.533)
Deviation after second year (δ_2)	−1.366	−0.682			−10.284
	(0.800)	(0.563)			(2.175)
Supervisory information (P)	3.678	0.000	0.941	3.755	4.326
	(0.503)	(0.000)	(0.150)	(0.181)	(0.427)
Likelihood ratio test p-value	0.003	0.009	0.001	N/A	0.022

Note: This table shows parameter estimates of the model of learning and influence described in Section 5.3 and specified in Section 5.3. Parameters are estimated from reduced-form practice variation moments, as shown in Figure 1 overall (Column 1) and Figure 3 for each spending category (the remaining columns). Knowledge parameters represent units of precision as function of tenure, as in Equation (A-20): ρ_0 is precision of knowledge prior to training; ρ_1, ρ_2, and ρ_3 are increases in precision (learning) in the first, second, and third years, respectively. Influence parameters δ_1 and δ_2 are deviations from the Bayesian benchmark benchmark in terms of effective precision as a function of completed years of training, as given in Equation (A-21). Specifically, a trainee who has completed one year of training receives influence that is δ_1 more (if positive) or less (if negative) units of effective precision than the efficient benchmark would imply. Similarly, a trainee who has completed two years of training receives an additional deviation of δ_2 relative to the efficient benchmark. P is the effective precision of supervisory information, including knowledge from supervisors, consultants, rules, or information produced by the trainees. Cells with missing values indicate that the model was estimated with these values constrained to 0, as less-constrained models did not converge. The likelihood ratio test p-value compares the estimated model against a restricted model of no learning (i.e., only ρ_0, δ_1, and P are non-zero). Note that this test is not relevant for the medication model, as the estimated model is in fact a no-learning model. Standard errors are displayed in parentheses.

Table A-8: Model Parameter Estimates by Day of Stay and Patient Severity

	Day of Stay		Patient Severity	
	(1)	(2)	(3)	(4)
	Early	Late	High Severity	Low Severity
Knowledge parameters				
Prior to training (ρ_0)	0.076	0.006	0.091	0.060
	(0.056)	(0.000)	(0.078)	(0.059)
First year (ρ_1)	0.346	0.294	0.371	0.207
	(0.223)	(0.087)	(0.299)	(0.253)
Second year (ρ_2)	6.681	6.655	6.242	7.644
	(2.528)	(1.414)	(2.719)	(3.572)
Third year (ρ_3)	0.000	0.845	0.000	0.000
	(0.000)	(0.007)	(0.000)	(2.000)
Influence parameters				
Deviation after first year (δ_1)	0.271	0.192	0.294	0.204
	(0.288)	(0.198)	(0.315)	(0.300)
Deviation after second year (δ_2)	−0.912	−1.554	−1.347	−0.367
	(0.719)	(0.082)	(0.780)	(1.597)
Supervisory information (P)	3.850	3.495	3.725	3.759
	(0.545)	(0.419)	(0.608)	(0.622)
Likelihood ratio test p-value	0.151	0.000	0.020	0.182

Note: This table shows parameter estimates of the model of learning and influence described in Section 5.3. Columns correspond to models estimated on observations by patient-day: Columns 1 and 2 are for days respectively before or after the middle of each patient's stay; Columns 3 and 4 are for patients with above- or below-median expected 30-day mortality, respectively. Parameters are as described in the note for Table A-7 and are estimated from reduced-form practice variation moments, as shown in Figure 4 for type of patient-day. The likelihood ratio test p-value compares the estimated model against a restricted model of no learning (i.e., only ρ_0, δ_1, and P are non-zero). Standard errors are displayed in parentheses.