Learning on the Job: Evidence from Physicians in Training*

David C. Chan†

May 7, 2018

Abstract

Learning on the job presents a tradeoff in making team decisions: Workers with less knowledge have less to contribute to team decisions, yet in order to learn, they may need to have an experiential stake in decision-making. This paper studies learning and influence in team decision-making among physician trainees. Exploiting a discontinuity in relative experience, I find reduced-form evidence of influence due to seniority between trainees. I specify a simple structural model of Bayesian information aggregation and define a benchmark of static efficiency that allocates influence to make the best decision using knowledge at hand. The vast majority of learning occurs only after trainees are senior and can influence decisions. Influence is approximately efficient between trainees, but trainees exert much more influence than is statically efficient relative to their supervisors, possibly because such influence contributes to experiential learning.

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, 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.

For the things we have to learn before we can do them, we learn by doing them.

Aristotle, *The Nichomachean Ethics* (Ross and Brown Translation, 2009, p. 24)

1 Introduction

In a broad range of professions, workers acquire important knowledge on the job. In these settings, on-the-job training often involves working in teams with more experienced co-workers. Allowing inexperienced workers to take part in decisions has real consequences, but also offers them an opportunity to learn. Therefore, the nature of learning on the job has implications for how information should be used in team decision-making. In particular, if learning is *experiential*, such that knowledge takes root only when learners have a stake in their decisions and can explore the consequences of their actions, then there exists a tradeoff between making the correct decision at hand and training the next generation of professionals.

This paper studies the process of training new physicians as a particularly salient setting for the general issue of learning and decision-making in teams. Residency training is designed as an intensive program to impart knowledge to physicians beyond facts, "developing habits, behaviors, attitudes, and values that will last a professional lifetime" (Ludmerer, 2014). Yet, despite the intensity of selection and training in the medical profession, there exists substantial variation in practices and beliefs across physicians (Currie and MacLeod, 2017; Cutler et al., 2018), with large welfare implications (Finkelstein et al., 2016; Chandra et al., 2016). The source of such practice variation, in health care and in other professions, remains a question with little evidence to guide policy.

Medical residency provides a tractable setting for extracting and studying the general effects of workers on team decisions as they progress in training. Physician trainees almost universally begin their first jobs as physicians at the same point in their medical careers: immediately following medical school. For every patient, decisions are explicitly made in a team comprising a junior trainee in the first year of training, a senior trainee past the first year of training, and a supervising physician who has completed training. Patient cases are quasi-randomly assigned to trainee teams, and teams are reshuffled weekly so that each physician works with many co-workers throughout training. Finally, trainees are assigned a large number of patients over the course of residency and take part in dozens of medical decisions per patient-day.

Specifically, I follow a diverse group of 802 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 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 at various points in the trainees' tenure.[1] I define *practice variation* as the standard deviation of the distribution of these trainee effects across trainees in a given tenure period.

It is important to recognize that a trainee's effect reflects two conceptual objects related to knowledge—*judgment* (what the provider would have decided on her own) and *influence* (the extent to which her judgment sways the team decision). Thus, practice variation may reflect variation in judgments across providers, which should decrease as a group of providers gains more complete knowledge. But at the same time, holding judgments constant, practice variation increases with influence in team decisions, which accrues with knowledge. In reduced form, I am able to assess the role of influence by exploiting a discontinuity in the relative experience at the end of the trainees' first year of training: Trainees have relatively less experience than their teammate in their first year, and relatively more experience than their teammate immediately after their first year. Under the assumption that trainee judgment (and other characteristics) are continuous across the one-year mark, a discontinuous increase in practice variation across one year implies the effect of influence on practice variation via a change in relative experience.

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. Substantial practice variation exists across a whole range of decision types but is larger—both at baseline among junior trainees and even more so after the discontinuity for senior trainees—in domains with fewer clear rules and more discretion, such as diagnostic testing as opposed to medications. Interestingly, profiles of practice variation differ across inpatient services of general medicine, cardiology, and oncology hosting the *same* group of

[1] The strategy I employ is similar to that used in a number of papers starting with Abowd et al. (1999), which have studied effects using switching between 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 the rotations across teams.

trainees, suggesting that the development of practice variation is specific to the learning environment, even for the same group of physicians. Finally, I find little evidence supporting intrinsic heterogeneity (e.g., time-invariant preferences or skills) or learning by copying others (i.e., "learning by osmosis") as alternative sources of practice variation.[2]

In order to more directly study the process of learning during residency training, I adopt a simple model of Bayesian information aggregation, along the lines of DeGroot (2005). In this model, the optimal influence of a trainee is equal to the precision of her information in proportion to the total information used by the team to make decisions. In other words, the optimal team decision is a weighted average of physician judgments, where the weights are proportional to the precision of each team member's knowledge. Against this benchmark of *static efficiency*, I also allow for departures in which junior trainees may exert less than optimal influence (e.g., by herding around senior judgments, as in Prendergast, 1993) or, conversely, more influence than is justified by their current knowledge. The latter possibility allows for a "supervised learning" strategy (Lizzeri and Siniscalchi, 2008) which grants trainees a stake in decisions so that they may gain experiential knowledge.

The structural model maps primitives of information and influence onto previously estimated moments of practice variation at each tenure level in residency. The primitives of this model include trainee knowledge at start of residency, learning rates, deviations from static efficiency in allocating influence between trainees, and the supervisory information outside of trainee knowledge. Separate identification of learning and influence can be understood with the following intuition: While influence in team decisions always increases with learning, on net this pathway widens practice variation only for trainees who have relatively low influence; for trainees with higher influence, increasing agreement in judgments will eventually outweigh the effect of increasing influence. Given that the estimates of practice variation are nonparametrically identified and estimated in a first step, this approach to recovering model primitives can be viewed as semiparametric two-step estimation (see Ackerberg et al., 2014).

Structural estimation reveals that trainees begin residency with essentially no useful knowledge. Yet compared to their first year of training, in the second year—when they have influence as a senior trainee—they

[2]I investigate intrinsic heterogeneity in two ways: First, I exploit detailed characteristics about the physician trainees (e.g., prior degrees and honors, test scores, position on the residency rank list, and future career paths) and show that these characteristics predict an exceedingly a small fraction of the overall practice variation, even though they predict future incomes very well. Second, if practice variation is solely driven by unchanging heterogeneity, then there should be a high within-trainee correlation between effects across all time periods. Instead, I find that trainee effects are highly correlated only in adjacent tenure periods, but are very weakly correlated between more distant periods. I investigate learning from others ("learning by osmosis") by exploiting random assignment of trainees to supervisory teammates.

learn about 30 times more quickly. Learning appears to cease (i.e., "full knowledge" is acquired) around the beginning of the third year. The allocation of influence between junior and senior trainees is approximately efficient. However, the entire supervisory hospital structure contributes decision-making information that is less than 40% of the knowledge of a fresh residency graduate. The result is consistent with supervising physicians (and the rest of the hospital) granting much more autonomy to trainees than is statically efficient so that they may learn.[3] In counterfactual analyses, I quantify the tradeoff between improving decisions with supervisory information and encouraging trainees to learn: Increasing supervisory information significantly increases the time needed for trainees to acquire full knowledge, reduces the knowledge trainees contribute to decision-making, and as a result, diminishes the gain in total information by almost half.

This paper contributes to several literatures. First, it sheds new empirical light on the nature of learning on the job. Philosophers, psychologists, and educational reformers have long articulated the idea that learning, from childhood to professional development, may be most effective when it involves active exploration, participation, and experience.[4] This concept is embodied in the widespread training arrangements that exist beyond formal education in medicine, law, engineering, business, and academia. In medical training, the model has long been summarized as "see one, do one, teach one."[5] My results suggest that, on its own, *seeing* prompts relatively little learning, but that *doing*—specifically applying one's own decisions—and possibly *teaching*, are the crucial stages of training that establish knowledge.

Second, this paper 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."

[3] This last result is even more striking when considering that supervising physicians work with only one senior trainee and therefore have the same span of control in terms of patients to attend to. The "supervisory hospital structure" includes not only the supervising physician, but also nurses, pharmacists, consultants, the computer order entry system, and any information gathered that is orthogonal to trainee knowledge.

[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; 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). Economists have long been interested in job training (see, e.g., Mincer, 1962; Becker, 1965; Heckman et al., 1997; Barron et al., 1989). However, the process of knowledge acquisition has mostly remained a black box. Further, with the notable exception of Lizzeri and Siniscalchi (2008), which develops a theory of parental sheltering, experiential learning has largely been overlooked.

[5] This dictum is attributed to William Halsted, the first Chief of Surgery at Johns Hopkins Hospital, where modern residency training was first established in the US (Rodriguez-Paz et al., 2009).

I show how knowledge is aggregated across agents in a team via influence. Unlike predictions in the canonical team-theoretic models, notably Garicano (2000), I find that influence is non-monotonic in knowledge. The least experienced agents contribute little to decision-making, in proportion to their knowledge. However, agents in training drive more decision-making than is statically efficient, plausibly because experiential learning is an organizational priority.

Third, these results relate to a large literature documenting practice variation in health care.[6] Academic and policy discussions often refer to 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 with survey evidence revealing that experts often and widely disagree (Cutler et al., 2018). This paper highlights informational mechanisms behind wide practice variation in a training environment *designed* to create homogeneity. The experiential nature of learning, and the lack of intrinsic heterogeneity or "learning by osmosis," may explain why heterogeneous "practice styles" persist despite guidelines, are difficult to predict, and remain even between physicians who work together.

The organization of this paper is as follows. Section 2 describes the institutional setting and data. Section 3 presents reduced-form results on practice variation as a function of trainee tenure and on mechanisms behind practice variation. Section 4 introduces a model of learning, influence, and practice variation. Section 5 discusses structural estimates and counterfactual results. And finally, Section 6 discusses policy implications for practice variation and offers concluding comments.

2 Institutional Setting

2.1 The History and Philosophy of Residency Training

From the 1860s leading to the Flexner Report in 1910, the proliferation of medical knowledge led to the realization that physicians must be intensively trained following graduation from medical school. Between the two world wars, medicine added great improvements in surgical technique; new fields such as radiology,

[6] 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); Molitor (2017). Much of this 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. While this is mostly outside the focus of this paper, 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.

pediatrics, and psychiatry; and advances in anesthesiology, blood transfusion, and intravenous fluid replacement. Residency programs came to dominate the training of physicians to incorporate this new knowledge.

Since this time, the principles of residency training have remained relatively consistent despite continuing changes in medical technology.[7] First, trainees are to be given independence and responsibility to make clinical decisions on their own, with increased responsibility added in steps consistent with increased knowledge. Second, training emphasizes the deep exploration of relatively few cases, with close and immersive observation of each patient. Rather than memorizing facts in an increasingly rich field of information, trainees are encouraged to relate their cases to underlying principles, think critically about clinical observations and evidence, and engage in debate about diagnosis and treatment. Finally, much of training is informal, with trainees emulating the unspoken behaviors and values of senior physicians.

2.2 The Residency Program

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").

The team structure always assigns two interns to each resident, so that interns are assigned half the number of patients as residents. This allows interns todevote more attention to each patient, and they are usually the first to examine a patient and make judgments. Aside from differences in experience and span of control, the intern and resident roles have no formal distinctions in decision rights or regulatory considerations. Each trainee team is supervised by an attending physician, who has completed residency, and operates 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 number of interns plan only to spend one year in the internal medicine residency ("preliminary" versus "categorical" interns), subsequently

[7] See Ludmerer, 2014, for numerous historical details. For a current sense of how physician trainees in residency are evaluated, see `http://www.acgme.org/Portals/0/PDFs/Milestones/InternalMedicineMilestones.pdf`. Consistent with these basic principles, many of the guidelines for evaluation emphasize the acquisition of general concepts, skills, and professional norms.

proceeding to another residency program such as anesthesiology, radiology, or dermatology. Trainee schedules are arranged a year in advance.

This study focuses on inpatient ward rotations, which comprise cardiology, oncology, and general medicine services. Per residency administration, trainee rotation preferences are not collected and assignment does not consider trainee characteristics. Scheduling does not consider the teams of intern, resident, and attending physicians that will be formed as a result. Attending schedules are created independently, with neither trainee nor attending aware of one another's schedule in advance.

Patients arriving at the hospital are assigned to interns and residents by a simple algorithm that distributes patients in a rotation among on-call trainees that have not reached their patient capacity.[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.

2.3 Data

This study uses data collected from several sources. First, I observe the identities of each physician on the clinical team—intern, resident, and attending physician—for each patient on an internal medicine ward service and for each day 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]

Detailed residency application information for each trainee includes demographics, medical school, US Medical Licensing Examination (USMLE) test scores, membership in the Alpha Omega Alpha (AOA) medical honors society, other degrees, and position on the residency rank list. I also observe pre-committed career tracks for each trainee physician, including special tracks (e.g., primary care, genetics), the categorical internal medicine track, and tracks into other residencies, such as anesthesiology, dermatology, psychiatry, or

[8] Depending on the reason for admission, patients may be matched to categories of attending physicians according to the admitting service. Conditional on the service, patient types are not matched to trainees.

[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.

radiology after a preliminary internship year. In addition to trainee characteristics determined prior to residency, I observe physician specialty after 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 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.[11]

I use scheduling data and past matches between trainees and with supervising attending physicians. As described in Section 2, trainees do not choose most of their learning experiences, at least in terms of their clinical rotations and the peers, supervising physicians, and patients seen on the wards. 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 29 Elixhauser comorbidity dummies and Charlson comorbidity indices (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.

For each patient-day, I observe total cost information aggregated within 30 billing departments, which I further group into categories of diagnostic (laboratory and radiology), medication, blood bank, and nursing spending. Because costs I observe are based on the hospital's accounting of resource utilization due to physician *actions*, and not the measures of Medicare reimbursement used in recent studies (Doyle et al., 2015; Skinner and Staiger, 2015; Chandra et al., 2016), they provide new insight into welfare-relevant resource use.[12] 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

[11] 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.

[12] 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.

medication orders. Table 2 shows distributional statistics of daily spending in each category and in the services of cardiology, oncology, and general medicine.

3 Reduced-form Results

3.1 Practice Variation over Trainee Tenure

As a baseline analysis, I examine trainee effects on team decisions as a function of trainee tenure. Variation in trainee effects, or *practice variation*, may reflect three conceptual objects: (i) disagreement in trainee *judgments*, if they were allowed to make decisions on their own; (ii) *influence* on team decisions; and (iii) intrinsic heterogeneity in ability or preferences that may result in different actions using the same information.

Before specifying a model of (i) and (ii) in Section 4, and considering (iii) in Section 3.3, I will assess evidence of (ii) in reduced form by focusing on practice variation across the one-year tenure mark, when trainees experience a discontinuity in relative experience: Immediately *before* the one-year mark, trainees are junior, with at least a year less experience relative to their trainee teammate; immediately *following* one year, they are senior relative to their trainee teammate. Assuming that judgments (i.e., knowledge) and intrinsic heterogeneity are continuous across the one-year tenure mark, any discontinuity in practice variation reflects the impact of influence via relative experience.

To summarize a large number of decision types recorded as orders, I aggregate the direct costs of the decisions in each patient-day, observed via the hospital's accounting system. I model trainee effects on log total costs at the patient-day level as

$$Y_{it} = \mathbf{X}_i\beta + \mathbf{T}_t\eta + \xi_{j(i,t)}^{\tau(j(i,t),t)} + \xi_{k(i,t)}^{\tau(k(i,t),t)} + \zeta_{\ell(i,t)} + \nu_i + \varepsilon_{it}, \tag{1}$$

where i indexes the patient, and t indexes the day. $j(i,t)$, $k(i,t)$, and $\ell(i,t)$ refer to the junior trainee, senior trainee, and attending (supervising) physician, respectively, assigned to the patient i on day t. Trainee effects—$\xi_{j(\cdot)}^{\tau(\cdot)}$ and $\xi_{k(\cdot)}^{\tau(\cdot)}$ for the junior and senior trainees, respectively—depend on both the identity of the trainee and the tenure period $\tau(h,t)$ that day t falls in for trainee $h \in \{j(i,t), k(i,t)\}$. Equation (1) also includes patient and admission characteristics $\mathbf{X}_i$; time categories $\mathbf{T}_t$ (i.e., month-year combination, day of the week, and day of service relative to the admission day); and attending fixed effects $\zeta_{\ell(i,t)}$. In some

specifications, I also allow for daily costs to be correlated within patient via a patient variance component ν_i.

Conditional on service and time categories, the causal effects of trainee teams on patient outcomes are identified by quasi-random assignment of patients to trainees (see Appendix A-1.1). Junior and senior trainee effects are further separately identified by frequent and quasi-random reshuffling of trainees across teams (see Appendix A-1.2). For example, I can identify the effect of a junior trainee of a given tenure relative to another trainee of the same tenure by differences in outcomes for the two trainees while working with the *same* senior trainee.[13] Since patients are *not* randomly assigned to attending physicians, I consider attending physician "effects" as fixed and treat these as nuisance parameters capturing both true effects and unobserved patient selection to attending physicians.[14]

Because observations per trainee are finite, OLS estimates of trainee effects will include random noise, and the variance of such estimates would be biased upward relative to the true variance of trainee effects. I thus consider trainee effects as random and directly estimate practice variation as the standard deviation of the tenure-specific distribution of trainee effects.[15] In Appendix A-2, I detail a method akin to restricted maximum likelihood (REML) that allows for a large number of fixed covariates, potentially correlated with the random effects, outside of the maximum likelihood estimation. As is standard in hierarchical modeling (Gelman and Hill, 2007), I assume that the random trainee and patient effects are normally distributed and uncorrelated with one another. The former assumption is an approximation; practice variation estimates should still have an interpretation of minimizing prediction mean squared error, even if effects are not normally distributed. The latter assumption that trainee and patient effects are uncorrelated is supported by evidence of random assignment of trainees to each other and to patients (Table 1 and Appendix A-1). I estimate Equation (1) separately within bins of observations according to trainee tenure. I impose no assumption on the structure of correlation between effects of the same trainee in different periods, although I directly estimate this in Appendix A-3.2.

[13]This "mover-based" approach is similar to that in Abowd et al. (1999). As detailed in that paper and in Card et al. (2013), identification of trainee effects may allow for systematic, time-invariant selection of trainees to other team members and only require that *changes* in team composition are unrelated to time-varying characteristics of the trainees. Appendix A-1.2 demonstrates a stronger condition—that trainees are quasi-randomly assigned to teams (i.e., no systematic or time-varying selection of trainees to teams).

[14]Physician practice patterns have been found to be quite stable in the existing literature, which motivates fixed effects that are time-invariant (Epstein and Nicholson, 2009; Molitor, 2017).

[15]This approach is very much linked to Bayesian shrinkage (Morris, 1983; Chandra et al., 2016). Bayesian shrinkage requires knowledge of the variance of the effects and is primarily concerned with the "shrunken" estimates of individual effects, known as "best linear unbiased predictions" (BLUPs). In this paper, I primarily focus on estimates of the variance of the effects itself (i.e., "practice variation"), although in some analyses (e.g., in Appendix A-3.3), I will use these estimates to calculate BLUPs.

Figure 1 presents results for the estimated standard deviations of the trainee effect distributions within each tenure interval τ. In my baseline specification, I consider non-overlapping tenure intervals that are 60 days long for the first two years of residency and 120 days long for the third year, since third-year trainees have fewer inpatient days.[16] A standard-deviation increase in the effect of junior and senior trainees increases daily total spending by about 5% and 24%, respectively. After the first year of training, any convergence in practice variation is minor: The standard deviation of the trainee effect distribution remains above 20% throughout. Including or omitting admission-level random effects for the patient does not significantly alter results.

The discontinuity at the one-year mark demonstrates the large role of senior influence on practice variation. A natural potential source of senior influence is the greater knowledge that senior trainees possess relative to junior trainees. Agents with more knowledge should naturally receive more weight in making optimal team decisions. However, senior physicians may also have more influence than is warranted by knowledge because of titles, hierarchy, or prestige. The discreteness of decision-making, or herding around senior judgments due to strategic concerns (Scharfstein and Stein, 1990; Prendergast, 1993), may also inflate senior influence. On the other hand, trainees may be granted more influence by their supervisors than is warranted by their current knowledge to encourage experiential learning. In Section 4, I structurally account for influence that is either greater or less than the level justified by knowledge.

3.2 Decision Types and Ward Services

I next exploit the richness of the data and institutional setting to show how practice variation profiles, as a function of trainee tenure, may vary across decision types and ward services. In Figure 2, I show the tenure profile of practice variation for spending in different categories of decisions: diagnostic (radiology and laboratory), medication, blood transfusion, and nursing. Table 2 also shows summary statistics of spending in these categories. In all decision categories, there is an increase in practice variation at the one-year mark.

Nonetheless, practice variation and its discontinuous increase at the one-year mark does depend on the type of decision. Diagnostic spending shows a large increase in practice variation, with a standard deviation of 16% to 74% before and after the one-year tenure mark. In contrast, medication spending shows relatively small practice variation, both overall and in the increase at the relative experience discontinuity. Although

[16] 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.

I am unaware of prior evidence against which to benchmark these results, they are consistent with the idea that different decisions use knowledge differently. For example, 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.[17]

In Figure 3 I show that the practice variation profile is largely similar for patients with high and low predicted mortality, and for days earlier or later in a patient's stay. There appears to be greater convergence in practice variation among senior trainees for decisions earlier in a patient's stay. Finally, in Figure 4, I assess the path of practice variation across ward services representing different subspecialties of medicine: cardiology, oncology, and general medicine. The same trainees rotate across these services, each with different groups of supervising physicians, patients, and processes of delivering care. Cardiology appears to have more convergence relative to the other two services. Interestingly, the difference in the cardiology practice variation profile is unrelated to formal diagnostic codes.[18] Practice variation profiles do not depend on whether a patient's diagnostic code is common (Figure A-5) or linked to an official guideline (Figure A-6). Reweighting general medicine diagnostic codes to reflect a "pseudo-cardiology" service has no impact on the practice variation profile (Figure A-7). These results suggest that mechanisms underlying the difference in practice variation across services is unrelated to the formal coding of diagnoses.

3.3 Mechanisms: Intrinsic Heterogeneity and Learning by Osmosis

Finally, I directly evaluate two alternative mechanisms that may drive practice variation, both with clear policy implications. In the first mechanism, intrinsic physician heterogeneity in skills or preferences is at the root of practice variation. In the second, physicians learn primarily from others rather than from their own experience. Under these respective mechanisms, practice variation could be eliminated by selecting physicians with the correct characteristics to provide care, or by selecting the correct peers, supervisors, or training programs to instill and disseminate best practices. Although these mechanisms may seem intuitive, the evidence supporting them has been weak, and they seem inconsistent with wide practice variation among

[17] An alternative view is that diagnostic decisions could be less costly to experiment with, since unlike treatment decisions, they may not directly lead to differences in patient care. However, this interpretation is inconsistent with diagnostic practice variation being significantly greater for senior trainees than for junior trainees and persistent throughout the end of training.

[18] A cursory review of diagnostic (ICD-9) codes reveals significant overlap across services in formal diagnoses that are often insufficiently informative. For example, the most common formal diagnosis in both cardiology and general medicine is "Chest pain, not otherwise specified." Table A-5 illustrates this further by listing the 15 most common diagnoses in each service, as well as whether there exists a guideline for each of the listed ICD-9 codes.

a highly selected group of trainees in an intensive residency program.[19]

I evaluate intrinsic heterogeneity in two ways. First, I assess whether trainee effects can be predicted by detailed trainee characteristics, including demographics, prior formal degrees, place of medical school, standardized examination scores, position on the rank list, and future income. I find that these characteristics largely do not predict trainee effects. In Figure 5, I show the distribution of trainee effects in each tenure period throughout residency for high- versus low-ranked trainees, and for trainees with high versus low future income. I describe these analyses further in Appendix A-3.1 and present more exhaustive results in Table A-2. Second, I test whether trainee effects are persistently correlated throughout residency. If practice variation is due to time-invariant intrinsic heterogeneity, then trainee effects should be correlated across distant periods. Instead, I find that trainee effects are strongly correlated between adjacent tenure periods but are only weakly correlated between distant tenure periods (Appendix A-3.2 provides further details).

I evaluate the mechanism of learning from others ("learning by osmosis") by exploiting the random assignment of trainees to teams and supervising physicians. On a patient level, this assignment may substantially influence the practice patterns a trainee is exposed to.[20] I calculate several measures of prior training experience and find that they are broadly unrelated to differences in current practice across trainees. As an example, Figure 6 shows that, throughout the course of training, prior experience with supervising physicians—defined as the average supervising physician spending effect for all prior patients and for patients in the prior two, four, and six months—does not predict current spending decisions. Appendix A-3.3 describes this analysis and presents others in further detail.

4 Model of Learning and Influence

In this section, I specify a simple structural model of learning and influence in team decisions in order to interpret practice variation patterns throughout the course of training. As in the team-theoretic literature

[19] Learning from others is related to an old idea that practice variation reflects "schools of thought" transmitted during training (Phelps and Mooney, 1993). While physician selection and training may play still play a role for specific types of decisions, as shown by Schnell and Currie (2017) for opioid prescribing, many papers (e.g., Epstein and Nicholson, 2009) have attempted to explain practice variation based on physician characteristics and training history and have been unable to provide much support for this idea. In a seminal case study, Doyle et al. (2010) show differences in patient care due to random assignment of patients to trainees in two residency programs in the same hospital; in Appendix A-3.3, I describe a similar analysis and find that trainee effects predicted by training program are only moderate relative to overall practice variation.

[20] From the perspective of a junior trainee, practice patterns are driven by both the senior trainee and the supervising physician. From the perspective of a senior trainee, practice patterns are still driven by the identity of the senior trainee. A standard-deviation increase in the best linear unbiased prediction (BLUP) for senior trainees is 16.6% in overall spending. A standard-deviation increase in the BLUP for supervising physicians is 7.3%.

(e.g., Marschak and Radner, 1972; Radner, 1993; Garicano, 2000), I begin with the organizational problem of using information dispersed across agents to make decisions. In this case, I consider the team as being comprised of (i) a junior trainee j, (ii) a senior trainee k, and (iii) information from a single supervisory "agent" that in practice includes the attending physician and other actors or rules in the hospital. 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 supervisory or public knowledge 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 all of this 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 1. *The optimal action for decision d assigned to trainees j and k is*

$$a_d^* = \frac{\rho_{d,j}\mu_{d,j} + \rho_{d,k}\mu_{d,k}}{\rho_{d,j} + \rho_{d,k} + P_d^*}. \tag{2}$$

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 optimal weights on judgments in Equation (2),

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

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 or the supervisory structure contributed perfect information (i.e., $\rho_{d,-h} = \infty$ or $P^*_d = \infty$), a trainee would have no influence. Conversely, at the one-year tenure mark, influence discontinuously increases because the precision of a trainee's teammate $\rho_{d,-h}$ discontinuously decreases.

However, influence may not always be optimally allocated. For example, junior trainees may exercise less than optimal influence if they are overly deferential to seniors, either for strategic reasons (Scharfstein and Stein, 1990; Prendergast, 1993; Ottaviani and Sorensen, 2001) or due to the "prestige" of senior titles. Alternatively, trainees may be allowed more influence than is justified by their knowledge in a policy of "supervised learning" (Lizzeri and Siniscalchi, 2008), to encourage experiential learning. Because the goal of supervised learning is to improve the quality of *future* decisions, I consider only influence defined by $g^*_{d,h;-h}$ as *statically efficient*, in that it optimizes the decision at hand. In estimation, I allow for deviations from static efficiency as

$$\hat{a}_d = \frac{\tilde{\rho}_{d,i}\mu_{d,i} + \tilde{\rho}_{d,j}\mu_{d,j}}{\tilde{\rho}_{d,i} + \tilde{\rho}_{d,j} + P_d}. \tag{3}$$

$\tilde{\rho}_{d,h} = \rho_{d,h} + \delta(\tau_h)$ 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. P_d is similarly an effective "precision" for supervisory information. That is, 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.

4.1 Practice Variation Moments

To map empirical moments of practice variation to a model with learning, I introduce learning in residency as a function of the precision of trainee knowledge with respect to tenure; i.e., $\rho_{d,h} = \rho(\tau(h, t(d)))$. The goal is to consider how knowledge and decision-making for a single decision translates to empirical objects corresponding to a set of decisions randomly assigned to trainees.

Since I estimate Equation (1) using observations in a tenure block, I first assume that the precision of knowledge is constant across trainees in the same tenure.[21] Further, if $P^*_d = P^*$, then we can state influence between a pair of trainees in terms of the tenure-dependent precision of knowledge.[22] For all d by any pair

[21] This assumes that trainees learn at the same pace, which to some extent is justified by evidence in Section 3.3 that practice styles do not seem to reflect intrinsic heterogeneity or variation in experience within residency. Further, while I make this assumption to make precise statements about statistical moments, similar statements can be made qualitatively, viewing $\rho(\tau)$ as *average* knowledge across trainees at τ.

[22] In Appendix Figure A-4, 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.

of trainees with tenures τ_h and τ_{-h}, statically efficient influence would be implied by the weights

$$g^*(\tau_h;\tau_{-h}) = \frac{\rho(\tau_h)}{\rho(\tau_h)+\rho(\tau_{-h})+P^*}.$$

Similarly, if $P_d = P$, effective influence would be implied by the weights

$$g(\tau_h;\tau_{-h}) = \frac{\tilde{\rho}(\tau_h)}{\tilde{\rho}(\tau_h)+\tilde{\rho}(\tau_{-h})+P}. \tag{4}$$

Given tenure-specific influence weights, we can state trainee effects in terms of influence and expected judgments averaged over decisions. Denote the set of decisions $\mathcal{D}_h^\tau$ involving trainee h in tenure block τ. For intern j and resident k in respective tenure blocks τ_j and τ_k, the expected action in Equation (3) for these decisions can be decomposed into two parts corresponding to their effects:

$$\begin{aligned} E_d\left[\hat{a}_d \middle| d \in \mathcal{D}_j^{\tau_j} \cap \mathcal{D}_k^{\tau_k}\right] &= E_d\left[\hat{a}_d \middle| \tau_j, \tau_k\right] \\ &= g\left(\tau_j;\tau_k\right) E_d\left[\mu_{d,j}\middle|\tau_j\right] + g\left(\tau_k;\tau_j\right) E_d\left[\mu_{d,k}\middle|\tau_k\right]. \end{aligned}$$

The first equality holds because patients (and decisions) are randomly assigned; the second because influence is constant conditional on the tenures of the junior and senior trainees, regardless of their identities. In the second line, the first term corresponds to the intern trainee effect, or $\xi_j^{\tau_j}$, and the second term corresponds to the resident trainee effect, or $\xi_k^{\tau_k}$, on an average decision as given in Equation (1).

The next step is to state practice variation in terms of the precision of trainee knowledge. Note that the precision of knowledge for decision d, a function of tenure, implies variation in judgments across the population of h for a given d: $Var_h\left(\mu_{d,h}\middle|\tau_h=\tau\right) = 1/\rho(\tau)$.[23] Since trainee effects only capture the component of a trainee's judgment that is common across decisions, the variance of trainee effects is naturally lower than $Var_h\left(\mu_{d,h}\middle|\tau_h=\tau\right)$. Based on a simple variance decomposition, we have $Var_h\left(\xi_h^\tau\right) = Var_h\left(E_d\left[\mu_{d,h}\middle|\tau_h=\tau\right]\right) = \kappa/\rho(\tau)$, for some $\kappa \in (0,1)$.[24] κ represents within-provider agreement across

[23] To conceptualize this, consider a pool of experience that trainees may draw from to make decision d. Each "draw of experience" x provides information about θ_d, and $x \sim N\left(\theta_d, \sigma_d^2\right)$. Now consider a pool of trainees $h \in \mathcal{H}$, each with the same number of draws N: $x_{1,h}, x_{2,h}, \ldots, x_{N,h}$. The knowledge of any trainee in this pool should be $\rho_d = N/\sigma_d^2$. The judgment of trainee h is $\mu_{h,d} = \sum_{i=1}^N x_{i,h}/N$. The variance in judgments across trainees should be $Var_h\left(\mu_{h,d}\right) = \sigma_d^2/N$.

[24] For example, consider $\mu_{d,h} = \eta_d + \overline{\mu}_h + \tilde{\mu}_{d,h}$ in a given tenure period τ, where $E_d\left[\tilde{\mu}_{d,h}\right] = 0$ for all h. By the law of total variance, $Var_h\left(\overline{\mu}_h\middle|\tau_h=\tau\right) + E\left[Var_h\left(\tilde{\mu}_{d,h}\middle|\tau_h=\tau\right)\right] = Var_h\left(\mu_{d,h}\middle|\tau_h=\tau\right) = 1/\rho(\tau)$. The first term is the same as $Var_h\left(\xi_h^\tau\right)$, and thus $\kappa = Var_h\left(\overline{\mu}_h\middle|\tau_h=\tau\right)/Var_h\left(\mu_{d,h}\middle|\tau_h=\tau\right) \in (0,1)$. If learning occurs at the same rate for the trainee mean judgment (the trainee effect) and deviations from the trainee mean judgment, then κ is a constant.

decisions, which could reflect the degree to which a provider extrapolates knowledge across related decisions.[25] The observed standard deviation of trainee effects for trainees with tenure τ_h, working with teammates with tenure τ_{-h}, is then

$$\sigma(\tau_h, \tau_{-h}) = g(\tau_h; \tau_{-h}) \sqrt{\kappa / \rho(\tau_h)}. \tag{5}$$

It is easy to see that the practice variation profile with respect to tenure is scaled by $\sqrt{\kappa}$. In the limit, if there is no common component across decisions for a given trainee, then even if judgments differ across trainees for a given decision, we will not be able to observe such variation from trainee effects in reduced form. Similarly, the implied knowledge and rates of learning, measured in precision, will be scaled by $1/\kappa$. However, regardless of κ, ratios of knowledge, learning, and practice variation at different points in training will remain identical.

4.2 Interpretation and Identification

In the efficient benchmark, as the precision of knowledge increases for trainee h, holding fixed everything else, the influence of the trainee, $g^*(\tau_h; \tau_{-h})$, will increase. On the other hand, increasing the precision of knowledge also implies that the dispersion in average judgments, $\sqrt{\kappa / \rho(\tau_h)}$, will decrease. In order to understand how learning can be identified from a profile of practice variation over trainee tenure, one must consider the relative importance of changes in influence versus changes in dispersion in judgments across trainees, as knowledge increases through learning.

Influence depends on the *relative* size of $\rho(\tau_h)$ compared to the total information used in the decision, or $\rho(\tau_h) + \rho(\tau_{-h}) + P$. Increases in influence will be large when $\rho(\tau_h)$ is relatively small compared to $\rho(\tau_h) + \rho(\tau_{-h}) + P$. In contrast, holding fixed κ, dispersion in judgments depends on the *absolute* size of $\rho(\tau_h)$. Thus, increases in influence will outweigh decreases in judgment dispersion when trainees have little influence, or when the total information is large relative to their knowledge. In this case, learning will cause practice variation to increase. Conversely, when trainees have much influence, decreases in judgment dispersion will outweigh increases in influence, and practice variation will decrease with learning. In the limit, of course, when a single agent is responsible for making decisions, learning unambiguously decreases practice variation. Appendix A-4 shows this formally in Proposition A-2 and provides numerical examples.

[25] Another interpretation of κ is that it could reflect intrinsic heterogeneity (i.e., heterogeneous skills or preferences). However, as discussed in Section 3.3, I find little evidence of intrinsic heterogeneity as a major factor in practice variation.

Given that trainee teams have a clear structure (i.e., senior trainees with tenure τ_k work with junior trainees having tenure $\tau_j = \tau_k - \lfloor \tau_k \rfloor$), the relative importance of pre-training knowledge, learning on the job, and outside information is identified (up to a scale) by the shape of the practice variation profile with respect to tenure. This shape includes (i) the discontinuity in practice variation at the one-year tenure mark, (ii) the smaller potential discontinuity at the two-year tenure mark (Proposition A-1 in Appendix A-4), and (iii) the change in practice variation in continuous portions with respect to tenure. In each of the continuous segments, the curvature (second derivative) of practice variation with respect to tenure should be negative. That is, practice variation may be increasing, decreasing, or increasing and then decreasing with respect to tenure. But according to the model, it should never be decreasing then increasing within a continuous segment (Proposition A-2 in Appendix A-4).

Greater precision of information in both $\rho(\tau)$ and P will result in smaller practice variation, holding constant the shape of the practice variation profile. However, as noted above, smaller commonality in judgments across decisions for the same trainee, κ, would also reduce the scale of practice variation. Because I cannot separately identify κ and precision in $\rho(\tau)$ and P, I normalize $\kappa = 1$ and attribute smaller scale to larger $\rho(\tau)$ and P.[26]

Finally, I allow for deviations of influence from the efficient benchmark. To do so, I assume that knowledge is continuous, while deviations from efficient influence come from a step function with respect to years of training. That is, the "effective" trainee precision relevant for influence is

$$\begin{aligned} \tilde{\rho}(\tau) &= \rho(\tau) + \delta(\tau) \\ &= \rho(\tau) + \delta_1 \mathbf{1}(\tau \geq 1) + \delta_2 \mathbf{1}(\tau \geq 2). \end{aligned} \tag{6}$$

δ_1 represents the deviation in effective precision based on the title of "senior trainee" alone, and δ_2 represents a potentially additional deviation in effective precision when trainees enter the third year of training. As illustrated in Appendix A-4 (Figure A-10), under the efficient benchmark, a continuous $\rho(\tau)$ implies a practice variation along the entire profile of training, both in the continuous portions and at the year discontinuities. Thus, for a given shape in the continuous (within-year) portions, deviations in the practice-variation changes at the year discontinuities imply δ_1 and δ_2, which represent deviations from efficient allocation of

[26] In principle, one could separately identify κ by estimating a joint model of different types of decisions. Because this is not the focus of my analysis, I instead consider different types of decisions in separate estimations, so that I might allow more flexibility in learning and influence parameters in each type of decision.

influence *between* trainees.

Any deviation from efficient influence between the trainee team and the supervisory hospital structure is implicitly captured by P. It is also worth noting that trainee effects only capture the *prior knowledge* trainees bring to the decision, independent of any information they may gather in the process. Any outside information gathered by trainees is instead included in P. Thus, P is the precision from any information outside of trainee prior knowledge, including supervising physician knowledge, informational inputs from outside staff (e.g., nursing, consultants), or any information gathered by the trainees themselves.[27] Thus, under the very weak assumption that the precision of outside information must be at least as large as that of trainee knowledge at the end of training, or $P^* \geq \rho(\tau = 3)$, I identify a lower bound on this deviation between trainees and supervisors of $\rho(3) - P$, if $P < \rho(3)$.

5 Structural Parameters and Counterfactual Results

5.1 Estimation Approach

I approach estimation of 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 (1) and were previously discussed in Sections 3.1 and 3.2. The second step takes these moments of practice variation and, from the model in Section 4, recovers underlying primitives of knowledge and influence using minimum distance estimation.

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

$$\rho(\tau) = \begin{cases} \rho_0 + \rho_1 \tau, & \tau \in [0,1]; \\ \rho_0 + \rho_1 + \rho_2 (\tau - 1), & \tau \in [1,2]; \\ \rho_0 + \rho_1 + \rho_2 + \rho_3 (\tau - 2), & \tau \in [2,3], \end{cases} \tag{7}$$

where ρ_0 represents the precision of knowledge before starting residency, ρ_1 is the yearly rate of learning in the first year of residency as a junior trainee, and ρ_2 and ρ_3 are analogous rates of learning as a senior

[27] 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 (1), including the error term.

trainee in the second and third years, respectively.

The model primitives $\boldsymbol{\theta} = (\rho_0, \rho_1, \rho_2, \rho_3, \delta_1, \delta_2, P)$ can estimated by minimum distance:

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

where $\hat{\boldsymbol{\sigma}}$ is the vector of empirical estimates of practice variation from the first step, with elements corresponding to $(\tau_h, \tau_{-h}) \in \mathcal{T}$; $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is the corresponding vector of model-implied practice variation from Equation (5) given $\boldsymbol{\theta}$; and $\mathbf{W}$ is a weighting matrix.

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 $\sqrt{n}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0}, \Omega)$, then the asymtotic variance of $\hat{\boldsymbol{\theta}}$ is given by

$$\text{Asy. Var}\, \hat{\boldsymbol{\theta}} = \frac{1}{n} (\Gamma(\boldsymbol{\theta}_0)' \mathbf{W} \Gamma(\boldsymbol{\theta}_0))^{-1} (\Gamma(\boldsymbol{\theta}_0)' \mathbf{W} \Omega \mathbf{W} \Gamma(\boldsymbol{\theta}_0)) (\Gamma(\boldsymbol{\theta}_0)' \mathbf{W} \Gamma(\boldsymbol{\theta}_0))^{-1},$$

where $\boldsymbol{\theta}_0$ is the true parameter vector, and $\Gamma(\boldsymbol{\theta}_0) = \text{plim}\, \partial \boldsymbol{\sigma}\left(\hat{\boldsymbol{\theta}}\right) / \partial \hat{\boldsymbol{\theta}}$ is an 18×7 matrix of analytical derivatives of Equation (5) with respect to $\boldsymbol{\theta}$, evaluated at $\hat{\boldsymbol{\theta}}$. The optimal weighting matrix is $\mathbf{W} = \hat{\Omega}^{-1}$, which I obtain from the first-step estimation of practice variation. This yields for inference

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

I also calculate likelihood ratio tests for the joint-significance of learning and influence parameters against a restricted model with no learning and completely inefficient senior influence (i.e., only ρ_0, δ_1, and P are non-zero).

5.2 Parameter Estimates

In Column 1 of Table 3, I show baseline parameter estimates based on practice variation in overall spending. In Figure 7, I show the implied path of practice variation according to the model and estimated parameters, overlaid on reduced-form estimates of practice variation shown earlier in Figure 1. Several of the parameter

estimates can be understood by the shape and scale of the figure.

First, the large discontinuous increase in practice variation at the one-year tenure mark implies that trainees with no experience must have relatively small knowledge compared to trainees with one year of experience. The model estimates trainees at the beginning of residency with only $\rho_0 = 0.04$ units of precision compared to $\rho_1 = 0.20$ units of precision gained in the first year. In the second year of training, practice variation increases, then begins to decrease around halfway through the year. This implies a pace of learning that is rapid enough so that the senior trainee starts the year with relatively little influence but has enough influence by the middle of the year to enable practice variation to decrease with learning. I estimate that $\rho_2 = 7.5$, or 30 times the rate of learning in the second year than in the first year. As a related implication, supervisory information P must be small enough to allow senior trainees to overtake the majority of influence; the size of P also determines the size of trainee practice variation. I estimate that $P = 3.7$.

Finally, practice variation is relatively flat in the third year despite having begun to converge in the second year. This implies essentially no learning in the third year ($\rho_3 = 0$), which I interpret as trainees having reached "full knowledge" prior to the start of the third year. Relevant for counterfactual analyses below, in Appendix A-5 I estimate a more flexible version of Equation (7) for trainee learning:

$$\rho(\tau) = \begin{cases} \rho_0 + \rho_1 \tau, & \tau \in [0,1]; \\ \rho_0 + \rho_1 + \rho_2 (\tau - 1), & \tau \in [1, \tau_c]; \\ \rho_0 + \rho_1 + \rho_2 (\tau_c - 1) + \rho_3 (\tau - \tau_c), & \tau \in [\tau_c, 3]. \end{cases} \tag{8}$$

where $\tau_c \in (1,3)$ is a kink point around which $\rho(\tau)$ changes slope. I estimate that $\tau_c = 1.87$, or less than two months short of the two-year tenure mark. Other parameter estimates remain similar to the baseline model with Equation (7), particularly with $\rho_2 = 8.0$ and $\rho_3 = 0$.

For trainee influence relative to static efficiency, I first estimate that $\delta_1 = 0.23$. Although this deviation from static efficiency 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). Thus, a reasonable view is that influence is close to statically efficient between junior and senior trainees. I also estimate that $\delta_2 = -1.4$, which implies that third-year trainees have *less* influence than is statically efficient, although this parameter is imprecisely estimated and small relative to ρ_2. Finally, to consider the static efficiency of trainees relative to supervisors, I consider a

lower bound for statically efficient supervisory information to be the full knowledge as a graduating trainee (e.g., $\underline{P} \equiv \rho(3) \approx 7.74$ under Equation (7)). The rationale for this lower bound is that, at a minimum, the hospital supervisory structure includes the supervising physician, who has completed residency training. Thus, $P = 3.7$ represents a strikingly low contribution of information—less than half of the lower bound $\underline{P}$.

I also estimate model parameters based on practice variation in spending for subgroups of decisions along ward services (the remaining columns of Table 3), decision types (Table 4), and types of patient-days (Table 5). Learning is often greatest in the second year of training, regardless of the set of decisions. One exception is the set of decisions on the cardiology ward service, which shows continued convergence in the third year, implying significant learning in the third year. 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.

5.3 Counterfactual Results

In the results above, I find that the rate of learning increases dramatically when trainees have influence in team decision-making. This implies a tradeoff in the use of information to make team decisions: While supervisory information improves decision-making in the static sense, it may constrain experiential learning by trainees. Presumably for this reason, I find that trainees are supervised with much less supervisory information than would be statically efficient.

To quantify the implications of this tradeoff, I perform two sets of counterfactual analyses. First, I alter the level of supervisory information used in decision-making, which amounts to altering the influence of trainees relative to their supervisors. Second, I alter the relative influence between junior and senior trainees while holding fixed supervisory information. Under both sets of counterfactual scenarios, I characterize (i) the time required for trainees to attain "full knowledge," (ii) the amount of information contributed by the junior and senior trainee team, averaged across patients, and (iii) the total amount of information used in decision-making, which includes both (ii) and the supervisory information. Details of the counterfactual analyses are given in Appendix A-5.

In Panel A of Figure 8, I illustrate results under counterfactual scenarios of supervisory information. Increasing supervisory information to the lower bound of static efficiency, $\underline{P}$, would increase the time for trainees to attain full knowledge from 1.87 years to 2.36 years. Increasing supervisory information to a very plausible $1.5\underline{P}$ would increase this time to 3.59 years, greater than the current three years of residency

training in internal medicine. I also show that the average information from trainee knowledge decreases as supervisory information increases. A gain of 10 precision units in supervisory information reduces the average information from trainee knowledge by about 4 precision units, which implies a net gain of only 6 precision units in total average information. In Panel B of Figure 8, I show results under counterfactual scenarios of allocating influence between junior and senior trainees. While $\delta_1 = 0.23$ is large relative to $\rho(1) = \rho_0 + \rho_1$, it is small relative to $\rho(3)$. The range of counterfactual values of δ_1 is thus relatively small, and implications for counterfactual learning and decision-making information are similarly limited.[28]

6 Discussion and Conclusion

I follow physicians in residency training as they acquire professional knowledge and make decisions in teams. I find wide spending variation attributable to providers aggregated from a large number of randomly assigned decisions. There is little evidence to suggest that the variation either reflects intrinsic characteristics of the providers or is driven by practices of others. This suggests that decisions are driven by "tacit" knowledge not easily transferred across providers, and that the knowledge, while incomplete, is used to extrapolate practices to patients who are unique but who share features with previously experienced patients (Polanyi, 1966).

Exploiting a discontinuity in relative experience, I find reduced-form evidence that physicians with more experience exert more influence in team decisions. I then specify a simple model of Bayesian information aggregation to study more directly learning and influence in teams. I find that the allocation of influence between senior and junior trainees is approximately efficient, in that decisions are weighted by relative knowledge, but that trainees are allowed more autonomy than is statically efficient, presumably to allow "supervised learning" (Lizzeri and Siniscalchi, 2008). I find support for this hypothesis in the sense that learning is 30 times faster when trainees are beginning their role as senior trainee on the team as compared to when they are beginning residency. The structural evidence is consistent with a large body of work, mostly outside of economics, that learning is *experiential*; specifically, that agents must have a stake in their decisions in order to learn.

At the same time, I find that practice variation persists through the end of training, even in this elite

[28] It is noteworthy, however, that increasing the influence of senior trainees relative to junior trainees increases learning and information. Appendix A-5 discusses the intuition for this result, which is briefly that the rate of learning is convex with respect to influence. In this appendix, I also consider counterfactual values of δ_2 and find similarly limited effects, with results shown in Appendix Figure A-11.

and intensive residency program, and structural estimates largely suggest that learning is minimal in the final year of training. While this suggests that training could be shorter without consequences for practice variation or learning, perhaps the more important questions are (i) why does learning stop, and (ii) what are the policy implications for widely documented persistent practice variation (Syverson, 2011; Gibbons and Henderson, 2012).[29]

Although these questions are mostly outside the scope of this study, the possibility of experiential learning has implications for policy, particularly in health care, where practice variation has received a large amount of policy attention but remains poorly understood. If the knowledge required for decision-making is complex, then previously proposed policy levers of financial incentives, reporting, and patient cost-sharing (see Skinner, 2012, for a summary) will have little impact. Targeting decision-making rather than aggregate measures is likely to be more important and more effective (Institute of Medicine, 2013).[30] Similarly, if learning requires experience, then this may explain why the usual forms of spreading information, such as clinical guidelines or formal instruction in "continuing medical education," have done little to change practice (Shaneyfelt et al., 1999).

Instead, effective policies may identify and disseminate organizational practices associated with both productivity and consistent decision-making. These process innovations—alternatively termed "continuous quality improvement," "lean management," and even "learning health care"—appear to elicit, codify, and disseminate information to workers at the local level while crucially engaging them in the process (Institute of Medicine, 2012; Bohmer et al., 2013). Consistent with learning via experience, experimentation and tinkering are often explicitly encouraged, even when the answer may be publicly known from another setting. There are predictable frictions in instilling these practices across organizations (Nelson and Winter, 1982), but the returns may be great enough to justify policy efforts in this direction (Bloom et al., 2013, 2014).

[29]There exists a large theoretical literature relevant to the first question. For example, when learning is costly, then it may be efficient to stop learning. This intuition is related to a large literature on search theory and learning by doing (see e.g., Rogerson et al. 2005, for a review). See Caplin and Dean (2015) for a broader discussion of rational decision-making under knowledge constraints and information cost functions. An alternative formulation by Acemoglu et al. (2006) allows for a lack of asymptotic agreement if there is sufficient uncertainty in the subjective distributions that map signals onto underlying parameters. Also, Ellison and Fudenberg (1993) show that, under social learning, there will be less convergence if agents observe greater diversity in choices made. However, empirical evidence has been scarce.

[30]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. 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.

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.

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.

ACKERBERG, D., X. CHEN, J. HAHN, AND Z. LIAO (2014): "Asymptotic Efficiency of Semiparametric Two-step GMM," *The Review of Economic Studies*, 81, 919–943.

BARRON, J. M., D. A. BLACK, AND M. A. LOEWENSTEIN (1989): "Job Matching and On-the-Job Training," *Journal of Labor Economics*, 7, 1–19.

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.

BECKER, G. S. (1965): *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, New York, NY: National Bureau of Economic Research, distributed by Columbia University Press.

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.

BLOOM, N., B. EIFERT, A. MAHAJAN, D. MCKENZIE, AND J. ROBERTS (2013): "Does Management Matter? Evidence from India," *The Quarterly Journal of Economics*, 128, 1–51.

BLOOM, N., R. LEMOS, R. SADUN, D. SCUR, AND J. VAN REENEN (2014): "The New Empirical Economics of Management," *Journal of the European Economic Association*, 12, 835–876.

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.

CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. Intriligator, Amsterdam: North Holland, vol. Chapter 22, 1248–1318.

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.

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 II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–79.

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 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 & education*, New York: Kappa Delta Pi, oCLC: 972898376.

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.

FLEXNER, A. (1910): *Medical education in the United States and Canada: a report to the Carnegie Foundation for the Advancement of Teaching*, Carnegie Foundation for the Advancement of Teaching.

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, google-Books-ID: SbGcAAAAMAAJ.

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

GIBBONS, R. AND R. HENDERSON (2012): "What do managers do? Exploring persistent performance differences among seemingly similar enterprises," in *The Handbook of Organizational Economics*, ed. by R. Gibbons and J. Roberts, Princeton, NJ: Princeton University Press, 680–732.

GOWRISANKARAN, G., K. JOINER, AND P.-T. LÃ©GER (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.

HECKMAN, J. J., H. ICHIMURA, AND P. E. TODD (1997): "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, 605–654.

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*, Washington, D.C.: National Academies Press, google-Books-ID: vF2fAwAAQBAJ.

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.

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.

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.

NELSON, R. R. AND S. G. WINTER (1982): *An Evolutionary Theory of Economic Change*, Harvard University Press.

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.

RADNER, R. (1993): "The Organization of Decentralized Information Processing," *Econometrica*, 61, 1109–46.

RODRIGUEZ-PAZ, J. M., M. KENNEDY, E. SALAS, A. W. WU, J. B. SEXTON, E. A. HUNT, AND P. J. PRONOVOST (2009): "Beyond "see one, do one, teach one": toward a different training paradigm," *BMJ Quality & Safety*, 18, 63–68.

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.

SCHNELL, M. AND J. CURRIE (2017): "Addressing the opioid epidemic: Is there a role for physician education?" *American Journal of Health Economics*, 1–37.

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.

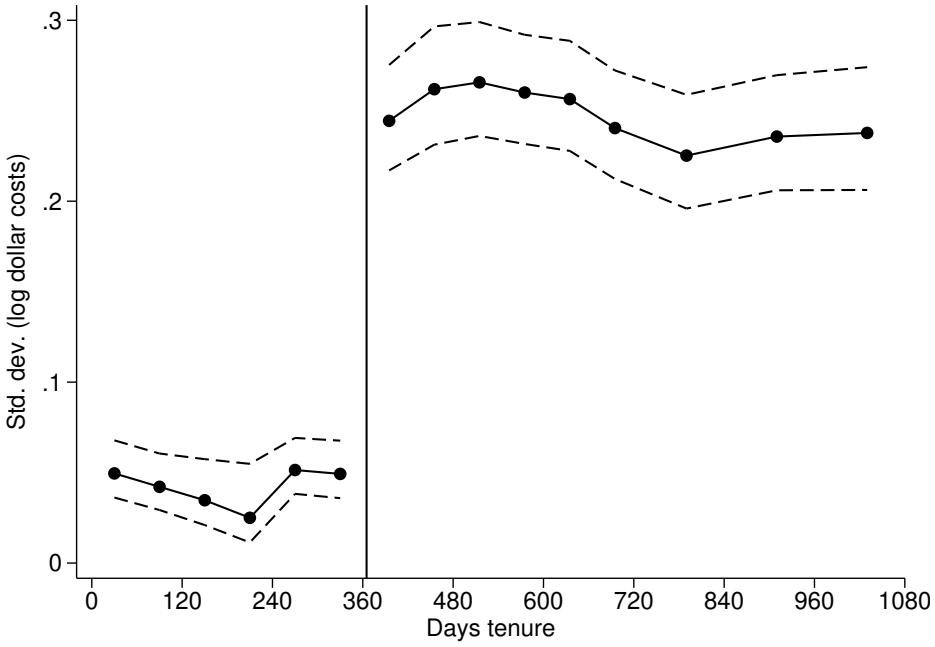
SKINNER, J. AND D. STAIGER (2015): "Technology Diffusion and Productivity Growth in Health Care," *Review of Economics and Statistics*, 97, 951–964.

SYVERSON, C. (2011): "What Determines Productivity?" *Journal of Economic Literature*, 49, 326–365.

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.

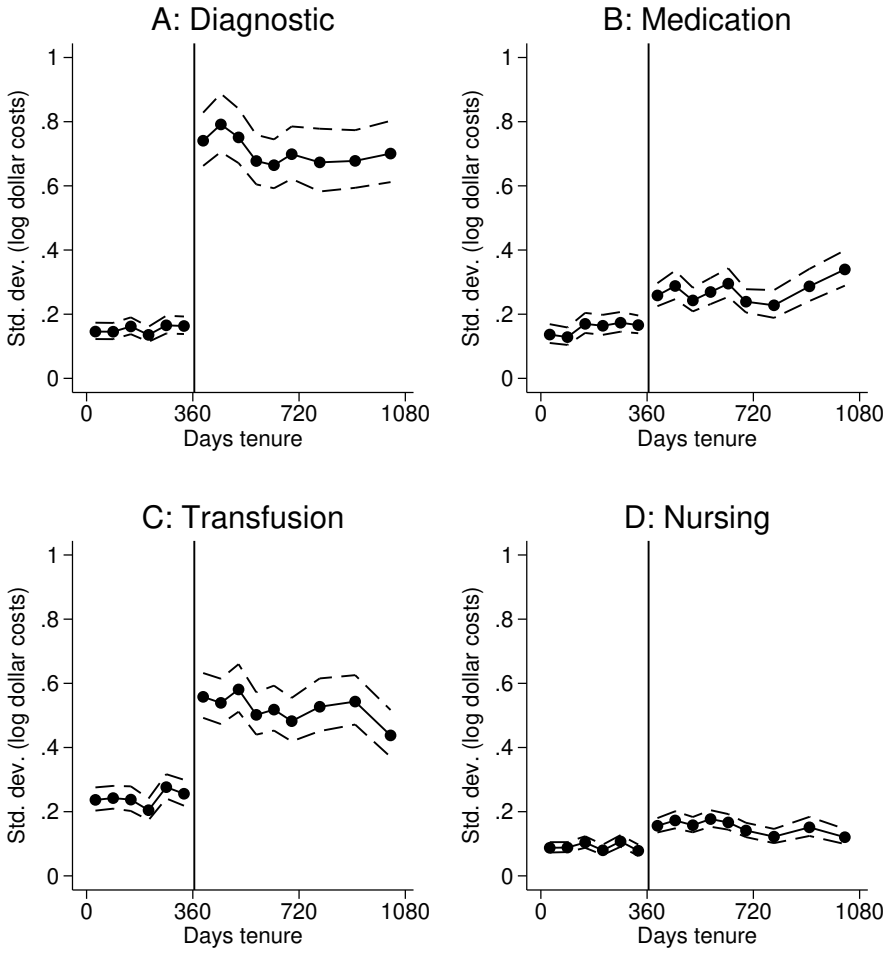
WOOD, D. F. (2003): "ABC of learning and teaching in medicine: Problem based learning," *BMJ*, 326, 328–330.

Figure 1: Profile of Practice Variation by Tenure



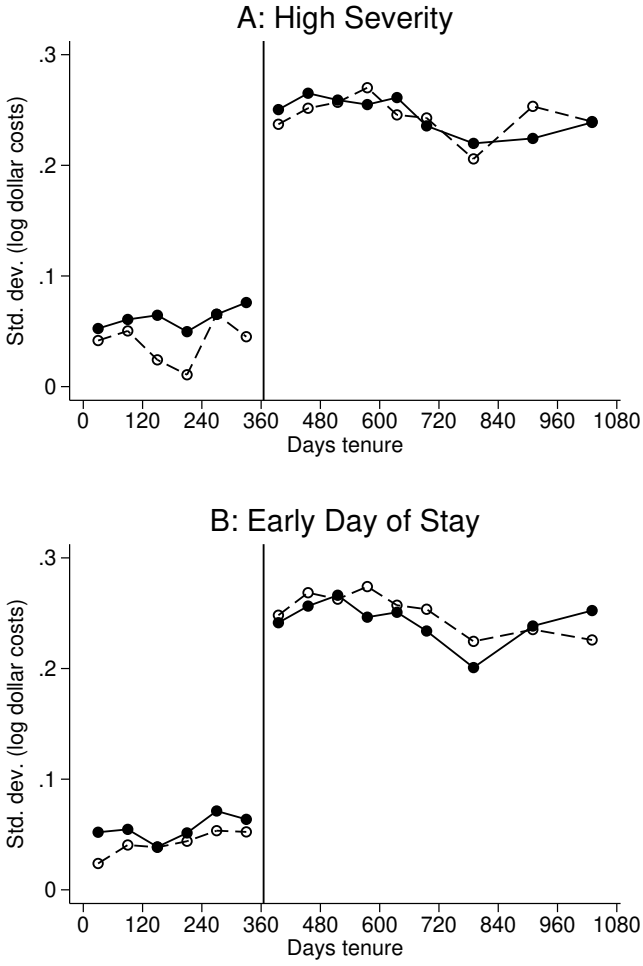
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (1), in log daily total costs at each non-overlapping tenure period. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. The model 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"). 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 2: Practice Variation Profile by Spending Category



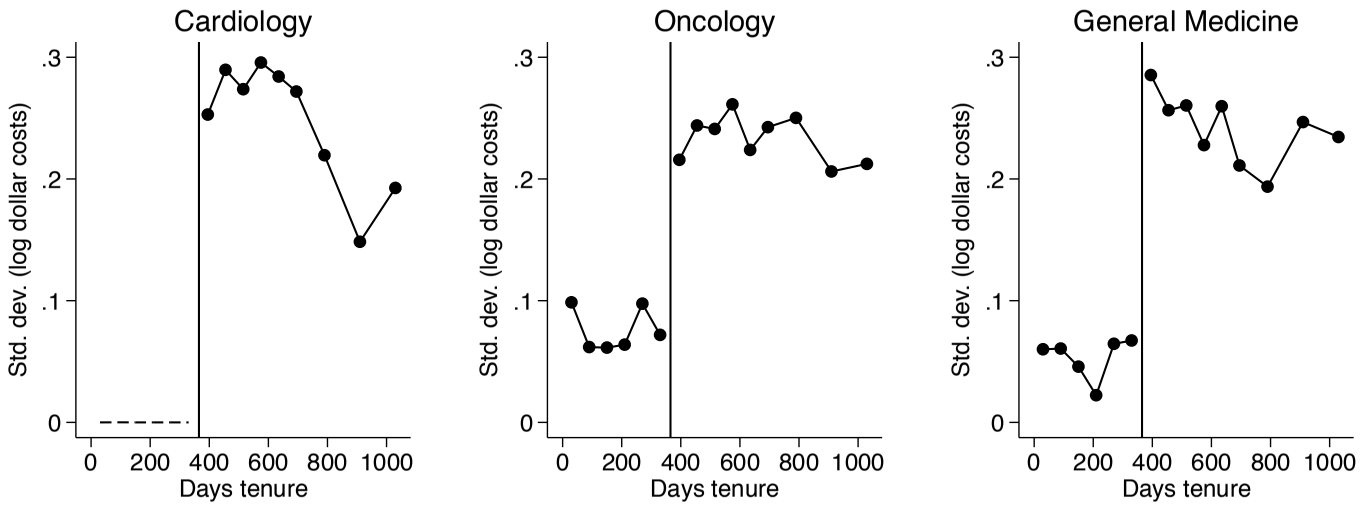
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (1), in log daily costs at each non-overlapping tenure period. Each panel shows a different spending category. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. 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 3: Practice Variation Profile by Patient Severity and Day of Stay



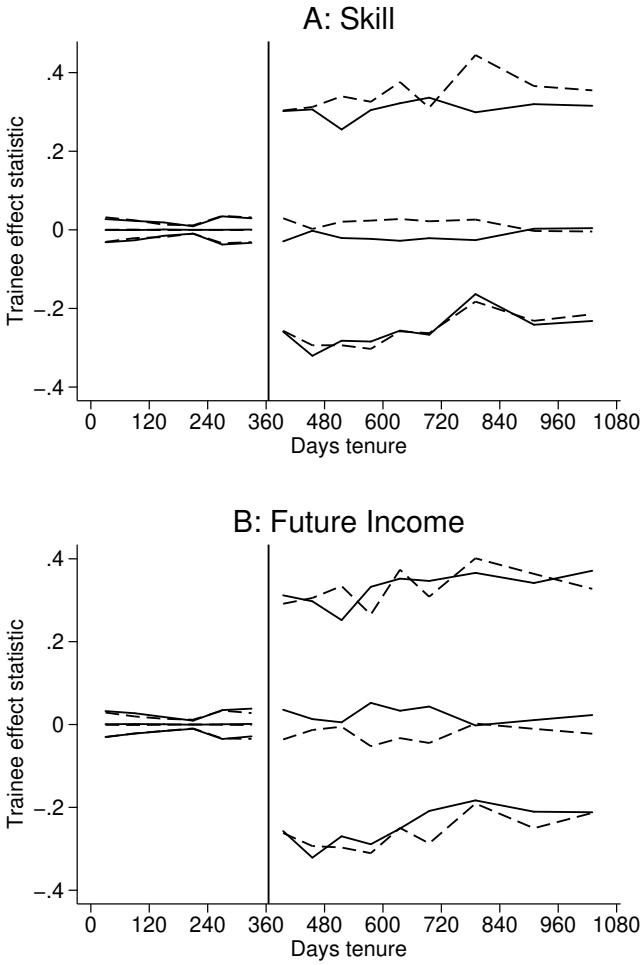
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (1), in log daily total costs at each non-overlapping tenure period. Panel A estimates the model separately in two samples of patients with above- (solid dots) versus below-median (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). Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. 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 4: Practice Variation Profile by Service



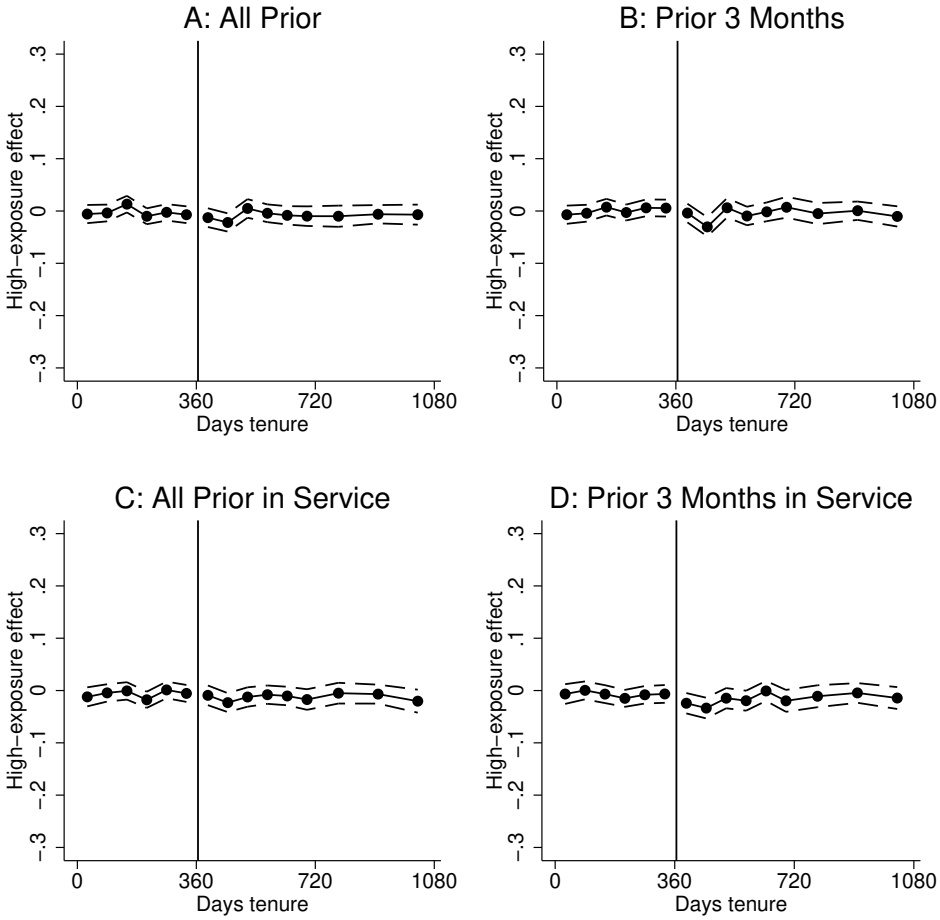
Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (1), in log daily total costs at each non-overlapping tenure period. Each panel shows results estimated from within a service of cardiology, oncology, or general medicine. The dashed line prior to 365 days in the cardiology service indicates that no significant positive standard deviation was estimated for observations corresponding to junior trainees on the cardiology service. Controls are the same as those listed in the caption for Figure 1. Trainee prior to one year in tenure are junior trainees and become senior trainees after one year in tenure; vertical lines denote the one-year tenure mark.

Figure 5: Practice Style Distribution by Trainee Type



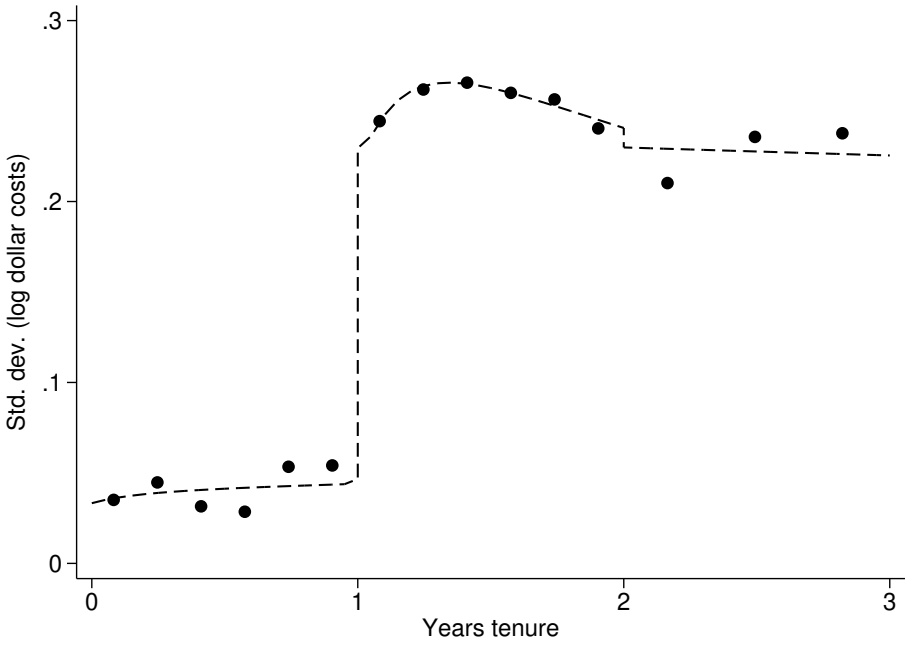
Note: This figure shows the patient-day-weighted 90th percentile, mean, and 10th percentile of the practice style (trainee effect) distribution according to trainee type. The unconditional distribution in each tenure period is normalized to have mean 0. Panel A shows the distribution for high-skill trainees (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 the Best Linear Unbiased Predictor (BLUP) posterior mean from the random effects model specified in Equation (1), of log daily total costs at each non-overlapping tenure period. The parameter of this regression is the standard deviation of trainee effects in each tenure period and is shown in Figure 1. 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: Effect of High Prior Exposure to Spending



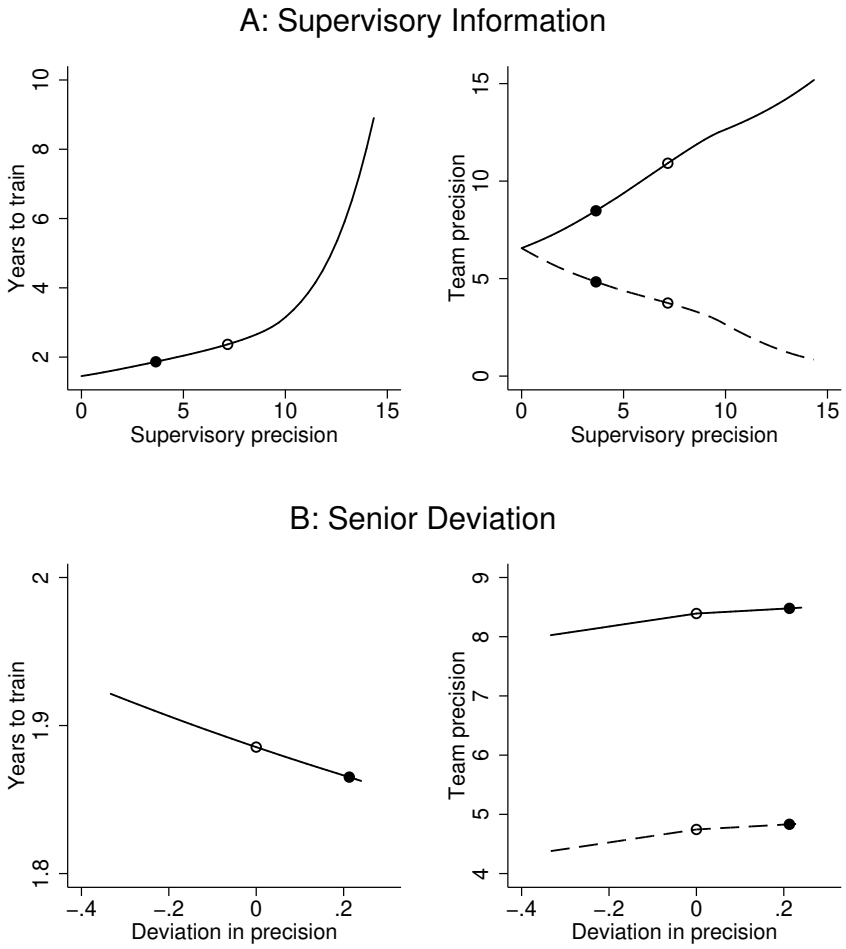
Note: This figure shows the effect of high prior exposure to supervising-physician spending. This exposure measure is discussed in further detail in Appendix A-3.3 and in Table A-3 and reflects the average spending effects of supervising physicians that a given trainee was matched to in the past. The tenure-specific effect of having high prior exposure to spending is estimated as in Equation (A-14). Panel A uses an exposure measure that includes all prior matches, regardless of service (corresponding to Column 1, Panel A of Table A-3). 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-3). 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-3). 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-9.

Figure 7: Model Fit to Practice Variation Profile



Note: This figure shows practice variation, defined as the standard deviation of random trainee effects specified in Equation (1), 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. Estimates for senior trainees are the same as shown in Figure 1; estimates for junior trainees are done separately for second-year senior trainees and for third-year senior trainees, but averaged in this figure for ease of presentation. Practice variation implied by the model of learning and influence, specifically Equation (5), is shown as a dashed line. Estimation of parameters of this model is described in Section 5. 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 8: 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-11 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 static efficiency; static efficiency in Panel A is a lower bound for supervisory information 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-5.

Table 1: Exogenous 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 exogenous 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	32	661
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	77	681
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: Model Parameter Estimates by Service

	Service			
	(1)	(2)	(3)	(4)
	All	General Medicine	Cardiology	Oncology
Knowledge parameters				
Prior to training (ρ_0)	0.039	0.291	0.000	0.093
	(0.030)	(0.206)	(0.000)	(0.484)
First year (ρ_1)	0.204	0.233		0.410
	(0.129)	(0.308)		(1.189)
Second year (ρ_2)	7.542	7.674	6.289	4.595
	(2.627)	(2.927)	(3.439)	(3.575)
Third year (ρ_3)	0.000	0.000	18.272	0.000
	(0.000)	(0.000)	(23.060)	(0.000)
Influence parameters				
Deviation after first year (δ_1)	0.231	1.013	0.187	0.162
	(0.238)	(0.809)	(0.140)	(1.855)
Deviation after second year (δ_2)	−1.366	−0.185		−1.410
	(1.171)	(1.174)		(1.591)
Supervisory information (P)	3.678	4.629	3.162	3.905
	(0.503)	(1.121)	(0.511)	(1.698)
Likelihood ratio test p-value	0.003	0.022	0.000	0.793

Note: This table shows parameter estimates of the model of learning and influence described in Section 4 and specified in Section 5.1. Columns correspond to models estimated on observations by ward service. The first column includes all observations. Parameters are estimated from reduced-form practice variation moments, as shown in Figures 1 and Figure 4, for each ward service. Knowledge parameters represent units of precision as function of tenure, as in Equation (7): 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 statically efficient benchmark in terms of effective precision as a function of completed years of training, as given in Equation (6). 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). Standard errors are displayed in parentheses.

Table 4: Model Parameter Estimates by Spending Category

	Service			
	(1)	(2)	(3)	(4)
	Diagnostic	Transfusion	Medication	Nursing
Knowledge parameters				
Prior to training (ρ_0)	0.936	0.225	1	0.000
	(0.277)	(0.200)		(0.000)
First year (ρ_1)	0.296	0.361		4.172
	(0.272)	(0.278)		(1.173)
Second year (ρ_2)	0.263	0.245		15.357
	(0.172)	(0.236)		(4.654)
Third year (ρ_3)	0.000			4.501
	(0.000)			(3.099)
Influence parameters				
Deviation after first year (δ_1)	4.388	0.349	0.730	−2.784
	(1.123)	(0.346)	(0.099)	(2.207)
Deviation after second year (δ_2)	−0.682			−10.284
	(0.563)			(3.182)
Supervisory information (P)	0.000	0.941	3.784	4.326
	(0.000)	(0.229)	(0.190)	(1.356)
Likelihood ratio test p-value	0.009	0.001	N/A	0.022

Note: This table shows parameter estimates of the model of learning and influence described in Section 4 and specified in Section 5.1. Columns correspond to models estimated on observations by spending category. Parameters are as described in the note for Table 3 and are estimated from reduced-form practice variation moments, as shown in Figure 2 for each spending category. 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 transfusion model, as the estimated model is in fact a no-learning model. In this model, $\rho_0 = 1$ as a normalization. Standard errors are displayed in parentheses.

Table 5: Model Parameter Estimates by Patient Severity and Day of Stay

	Service			
	(1)	(2)	(3)	(4)
	Early	Late	High Severity	Low Severity
Knowledge parameters				
Prior to training (ρ_0)	0.000	0.371	0.019	0.021
	(0.000)	(0.154)	(0.007)	(0.014)
First year (ρ_1)	0.513	0.126	0.091	0.123
	(0.210)	(0.137)	(0.209)	(0.193)
Second year (ρ_2)	8.854	7.215	6.672	6.359
	(3.221)	(2.600)	(2.498)	(3.382)
Third year (ρ_3)	0.000	0.000	0.038	0.000
	(0.000)	(0.000)	(0.041)	(2.000)
Influence parameters				
Deviation after first year (δ_1)	0.030	0.177	0.304	0.417
	(0.575)	(0.220)	(0.428)	(0.243)
Deviation after second year (δ_2)	−0.549	−0.100	−0.795	−0.839
	(0.859)	(2.189)	(1.176)	(2.294)
Supervisory information (P)	3.237	2.348	3.863	4.320
	(0.995)	(0.517)	(0.850)	(0.508)
Likelihood ratio test p-value	0.012	0.201	0.021	0.182

Note: This table shows parameter estimates of the model of learning and influence described in Section 4 and specified in Section 5.1. 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 3 and are estimated from reduced-form practice variation moments, as shown in Figure 3 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.

Appendix (for Online Publication per Referees / Editor)

A-1 Random Assignment

This appendix presents two sets of randomization tests for exogenous 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:

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

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"). $\zeta_i^{\tau<T}$, $\zeta_j^{\tau>T}$, 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 $\left(\zeta_j^{\tau<T}, \zeta_k^{\tau>T}\right)_{j\in\mathcal{J},k\in\mathcal{K}}$.

In Column 1 of Table A-1, I show F-statistics and the corresponding p-values for the null hypothesis that $\left(\zeta_j^{\tau<T}, \zeta_k^{\tau>T}\right)_{j\in\mathcal{J},k\in\mathcal{K}} = \mathbf{0}$. 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.[31] 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.

Columns 2 and 3 of Table A-1 show F-statistics and the corresponding p-values for the null hypothesis

[31] 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 (1) does not qualitatively affect results.

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.[32] 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 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

[32] 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).

these distributions are different, at least when clustering at the level of the intern-resident-attending team.

A-2 Statistical Model of Trainee Effects

In this appendix I introduce a statistical model to estimate the standard deviation $\sigma(\tau)$ of trainee effects ξ_h^τ in discrete tenure period τ and the correlation $\rho(\tau_1, \tau_2)$ between trainee effects $\xi_h^{\tau_1}$ and $\xi_h^{\tau_2}$ in two discrete periods, τ_1 and τ_2. Random assignment of patients to trainee, conditional on time categories, allows me to estimate trainee effects.[33] Finite observations per trainee-period means that effects will be estimated with error, which implies that standard deviations of unshrunken effects will overstate the true $\sigma(\tau)$. Further, correlations of fixed effect estimates of $\xi_h^{\tau_1}$ and $\xi_h^{\tau_2}$ will generally understate true correlations, and comparing the relative magnitudes of correlations between two pairs of periods will be invalid.

I adopt a random effects approach in which I simultaneously estimate both distributions of intern and resident effects by maximum likelihood. First, similar in spirit to Chetty et al. (2014) and closely related to the idea of restricted maximum likelihood (REML) (Patterson and Thompson, 1971), I create the differenced outcome $\tilde{Y}_{it} = Y_{it} - \left(\mathbf{X}_i\hat{\beta} + \mathbf{T}_t\hat{\eta} + \hat{\zeta}_{\ell(i,t)}\right)$. Importantly, $\hat{\beta}$, $\hat{\eta}$, and $\hat{\zeta}_\ell$ are estimated using variation *within* trainee pairs and discrete tenure periods, so that $\tilde{Y}_{it}$ can be thought of as including the trainee effects of interest. This differencing procedure allows the trainee effects to be correlated with $\mathbf{X}_i$, $\mathbf{T}_t$, and ζ_ℓ.[34] Note that $E\left[\tilde{Y}_{it}\right] = 0$. In practice, given random assignment of attending physicians and patients to trainees conditional on schedules, I am concerned only with correlations between trainee effects and $\mathbf{T}_t$. However, differencing out projections due to $\mathbf{X}_i$ and ζ_ℓ simplifies computation and avoids the incidental parameters problem in the later maximum-likelihood stage. In the next two subsections I will describe how I calculate $\sigma(\tau)$ and $\rho(\tau_1, \tau_2)$.

A-2.1 Standard Deviation of Trainee Effects

To estimate $\sigma(\tau)$, I specify a crossed random effects model for each set of days comprising a trainee tenure period τ,

$$\tilde{Y}_{it} = \xi_{j(i,t)}^{\tau(j(i,t),t)} + \xi_{k(i,t)}^{\tau(k(i,t),t)} + \varepsilon_{it}, \tag{A-5}$$

using observations for which $\tau(h,t) = \tau$. In other specifications, I consider a random effect model that allows for unobserved heterogeneity in patients:

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

[33] I do not strictly require conditional random assignment of patients to trainees if I use patients that are shared by multiple interns or residents due to lengths of stay spanning scheduling shifts. However, I do not rely on this in my baseline specification, in order to use more of the data.

[34] An alternative albeit slightly more involved approach involves estimating "correlated random effects," as described by Chamberlain (1984) and Abowd et al. (2008).

where ν_i is a random effect for the patient admission.[35] Because trainees are assigned conditionally randomly to each other and to patients, $\xi_{j(i,t)}^{\tau(j(i,t),t)}$, $\xi_{k(i,t)}^{\tau(k(i,t),t)}$, and ν_i are uncorrelated with one another. Assuming ξ_j^{τ}, $\xi_k^{\tau'}$, and ν_i are normally distributed, their standard deviations $\sigma_{\xi,\tau}$, $\sigma_{\xi,\tau'}$, and σ_ν are the parameters of interest in the following maximum-likelihood estimation, done in separate samples selected on τ.

Equations (A-5) and (A-6) can be stated in vector form:

$$\tilde{\mathbf{Y}} = \mathbf{Z}\mathbf{u} + \varepsilon, \tag{A-7}$$

where $\tilde{\mathbf{Y}}$ is the $n \times 1$ vector of differenced outcomes, $\mathbf{Z}$ is a selection matrix, and $\mathbf{u}$ is a stacked vector of random effects.

Let N_τ be the number of trainees with some tenure interval τ (e.g., 1 to 60 days) and N_{-h}^{τ} be the corresponding teammates observed in the sample. Then, in the case that (A-7) represents (A-5), $\mathbf{Z}$ is an $n \times (N_\tau + N_\tau^-)$ selection matrix for trainees with tenure τ and their peers, and $\mathbf{u}$ is an $(N_\tau + N_\tau^-) \times 1$ stacked vector of trainees and peer random effects. The variance-covariance matrix of $\mathbf{u}$ is diagonal:

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \sigma_{\xi,\tau}^2 \mathbf{I}_{N_\tau} & \mathbf{0} \\ \mathbf{0} & \sigma_{\xi,\tau'}^2 \mathbf{I}_{N_\tau^-} \end{bmatrix}.$$

Similarly, in the case that (A-7) represents (A-6), $\mathbf{Z}$ is an $n \times (N_\tau + N_\tau^- + N_i)$ selection matrix for trainees of tenure τ, teammates, and patient admissions, and $\mathbf{u}$ is an $(N_\tau + N_\tau^- + N_i) \times 1$ stacked vector of intern, resident, and admission random effects, where N_i is additionally the number of admissions in the sample. The diagonal variance-covariance matrix of $\mathbf{u}$ is

$$\operatorname{Var}\mathbf{u} = \mathbf{G} = \begin{bmatrix} \sigma_{\xi,\tau}^2 \mathbf{I}_{N_\tau} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{\xi,\tau+\Delta}^2 \mathbf{I}_{N_\tau^-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\nu^2 \mathbf{I}_{N_i} \end{bmatrix}.$$

Using the definition $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_\varepsilon^2 \mathbf{I}_{it}$, the log likelihood function under either of the above specifications is

$$L = -\frac{1}{2}\left\{ n \log(2\pi) + \log|\mathbf{V}| + \tilde{\mathbf{Y}}'\mathbf{V}^{-1}\tilde{\mathbf{Y}} \right\}. \tag{A-8}$$

I estimate (A-5) or (A-6) by maximum likelihood, for each τ separately. Holding the tenure of h fixed at τ, the tenure of the other teammate will possibly be a mixture if τ is less than one year (i.e., corresponds to junior trainees). I thus focus on $\sigma(\tau) \equiv \sigma_{\xi,\tau}$ and not $\sigma_{\xi,\tau'}$, although results do not qualitatively depend on this.

[35] 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.

A-2.2 Correlation of Trainee Effects

To estimate $\rho(\tau_1, \tau_2)$, I augment models in (A-5) and (A-6) to account for two separate tenure periods, τ_1 and τ_2, across which trainee effects may be correlated. Although I observe each trainee across her entire training, I only observe a subset of these trainees in each 60-day or 120-day tenure period. The number of trainees observed in two different tenure periods 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.

Specifically, in place of Equation (A-5), I consider

$$\tilde{Y}_{it} = \xi^{p(i,t)}_{h(i,t)} + \xi_{-h(i,t)} + \varepsilon_{it}, \tag{A-9}$$

which features the function $p(i,t) \in \{\tau_1, \tau_2\}$. 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, while effects of the teammates are pooled into a single distribution not dependent on tenure. The analog for Equation (A-6) is

$$\tilde{Y}_{it} = \xi^{p(i,t)}_{h(i,t)} + \xi_{-h(i,t)} + \nu_i + \varepsilon_{it}. \tag{A-10}$$

As above, both (A-9) and (A-10) can be written in the vector form of (A-7). When representing (A-9) as (A-7), the selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_\tau^-)$, since it now maps observations onto one of two random effects, depending on whether $p(i,t) = \tau_1$ or $p(i,t) = \tau_2$, 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

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

where $\mathbf{G}_\tau$ is a $2N_\tau \times 2N_\tau$ block-diagonal matrix of the form

$$\mathbf{G}_\tau = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A} \end{bmatrix},$$

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

$$\operatorname{Var} \begin{bmatrix} \xi^{\tau_1}_h \\ \xi^{\tau_2}_h \end{bmatrix} = \mathbf{A}, \text{ for all } h.$$

Representing (A-10) as (A-7) is a similar exercise. The selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_\tau^- + N_i)$,

and the vector of random effects $\mathbf{u}$ is of size $(2N_\tau + N_\tau^- + N_i) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

$$\text{Var}\,\mathbf{u} = \mathbf{G} = \begin{bmatrix} \mathbf{G}_\tau & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\xi,\tau'}\mathbf{I}_{N_\tau^-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma^2_v\mathbf{I}_{N_i} \end{bmatrix},$$

where $\mathbf{G}_\tau$ is the same as before. The log likelihood is the same as in Equation (A-8), 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 estimated from $\hat{\mathbf{A}}$ and is constrained to be between -1 and 1.

A-3 Alternative Mechanisms

A-3.1 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:

$$Y_{it} = \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{aijkt}, \tag{A-11}$$

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.[36] 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

[36] In principle, I could include trainee characteristics as mean shifters in the baseline random effects model in Equation (1). However, since characteristics are generally insignificant predictors of variation, results of (residual) variation attributable to trainees are unchanged.

rank list entered linearly, along with the other admission and time regressors in Equation (A-11):

$$Y_{it} = \sum_m \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{A-12}$$

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:

$$Y_{it} = \alpha \tilde{Z}_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{A-13}$$

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 potential 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-11).

Table A-2 shows results for Equation (A-13) and a subset of results for Equation (A-11). Considering characteristics individually in Equation (A-11), 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,[37] 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 5, 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-3.2 Intrinsic Heterogeneity: Serial Correlation

As a second method to evaluate intrinsic heterogeneity, I examine serial correlation in trainee effects. In the case of unchanging heterogeneity, physician practice styles should be constantly and highly correlated across time periods, regardless of the time between the periods. However, if patients are incorporating new knowledge and evolving in their practice styles, then adjacent time periods should exhibit higher correla-

[37] 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%.

tion in trainee effects than do distant time periods. Appendix A-2.2 describes details of estimating serial correlation across tenure periods in my random-effects framework.

In Figure A-8, I show correlation estimates between each pair of tenure periods. Serial correlation in trainee effects across two adjacent periods are generally very high and above 0.9, while the correlation decreases with more distance between the two periods. Interestingly, correlation is uniformly high between any two periods within the first year of training, when trainees are junior. However, correlation diminishes at a quicker pace when trainees are senior, in the second and third years of training. This implies that practice styles change more rapidly when trainees are senior. There also appears to be a uniform drop in correlation across the one-year mark, and to a lesser extent across the two-year mark, which could be consistent with changes in practice style that are induced by changes in relative seniority or changes in the cohort of teammates.

A-3.3 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.[38]

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 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-3 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 regres-

[38]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.

sion of the form

$$
\begin{aligned}
Y_{it} \quad = \quad & \sum_{\tau:\tau<1} \alpha_\tau \mathbf{1}\left(\tau\left(j\left(i,t\right),t\right)=\tau\right) \cdot \text{HighSpendingExposure}^m_{j(i,t),t} + \\
& \sum_{\tau:\tau\geq 1} \alpha_\tau \mathbf{1}\left(\tau\left(k\left(i,t\right),t\right)=\tau\right) \cdot \text{HighSpendingExposure}^m_{k(i,t),t} + \\
& \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{\ell(i,t)} + \varepsilon_{it},
\end{aligned}
\tag{A-14}
$$

where, as in Equation (1), $j\left(i,t\right)$ 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. $\text{HighSpendingExposure}^m_{j,t}$ 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 6 shows results for exposure to spending by supervising physicians, and Figure A-9 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

$$
Y_{it} = \alpha_m \text{Experience}^m_{h(i,t),t} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it},
\tag{A-15}
$$

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-4 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-11). This effect includes both differences in selection (i.e., intrinsic heterogeneity) into the different program and in training experiences across the programs. Table A-2 shows that visiting trainees do not have significantly different spending effects, either as interns or as residents.[39]

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*

[39]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.

uniform training experiences but nonetheless large practice variation.

A-4 Identification from Practice Variation Profiles

A-4.1 Analytical Evaluation

I first make analytical observations on the shape of practice variation profiles as a function of underlying learning and influence. Consider practice variation—or the standard deviation of trainee effects—under statically efficient influence:

$$\begin{aligned}\sigma(\tau_h, \tau_{-h}) &= \frac{g^*(\tau_h; \tau_{-h})}{\sqrt{\rho(\tau_h)}} \\ &= \frac{\sqrt{\rho(\tau_h)}}{\rho(\tau_h) + \rho(\tau_{-h}) + P},\end{aligned} \tag{A-16}$$

where I assume that $\kappa = 1$ in (5) 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-1. *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*

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

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-2. *Practice variation decreases if and only if*

$$\frac{\rho'(\tau_h)}{\rho'(\tau_h) + \rho'(\tau_{-h})} < 2g^*(\tau_h; \tau_{-h}). \tag{A-17}$$

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-4.2 Numerical Examples

Figure A-10 presents a few numerical examples of variation profiles under various learning profiles described by functions of the piecewise linear form in Equation (7). 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-10 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-5 Counterfactual Analyses

A-5.1 Model of Learning

As discussed in Section 5.2, under my baseline model of knowledge, with constant rates of learning within each training year specified in Equation (7), 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, for example, as discussed in footnote 29. 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 the form of Equation (8), which I reproduce here:

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

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 (8) 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

$$\overline{g}(T; \boldsymbol{\theta}) \equiv E_{\tau_h} \left[g(\tau_h; \tau_{-h}) | \boldsymbol{\theta} \right], \tag{A-18}$$

where influence $g(\tau_h; \tau_{-h})$ is given in Equation (4) 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 $t \in \{1,2\}$,

$$\rho_t^\Delta = \begin{cases} \hat{\rho}_1 \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right), & \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) \leq \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right), \\ \hat{\rho}_1 + \frac{\hat{\rho}_2 - \hat{\rho}_1}{\overline{g}\left(T_2; \hat{\boldsymbol{\theta}}\right) - \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right)} \left(\overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) - \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right) \right), & \overline{g}\left(T_t; \boldsymbol{\theta}^\Delta\right) > \overline{g}\left(T_1; \hat{\boldsymbol{\theta}}\right). \end{cases} \tag{A-19}$$

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-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 (4) and (6). Influence also depends on knowledge, as given by Equation (8), which in turn depends on learning via influence, as given by (A-19).

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 (4), (6), (8), (A-18), and (A-19), 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 is statically inefficient between trainees. 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 (8) using the counterfactual parameters ρ_1^{Δ} and ρ_2^{Δ}; $\tilde{\rho}^{\Delta}(\tau)$, as given by Equation (6), 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.

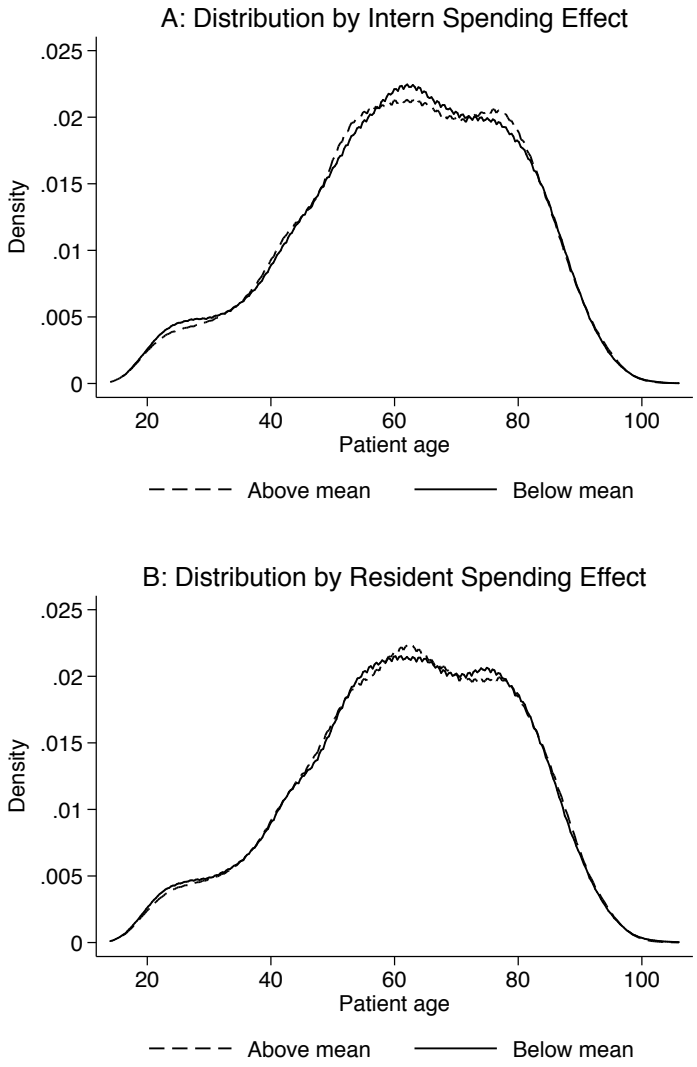
A-5.3 Discussion of Results

In Figure 8, 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.

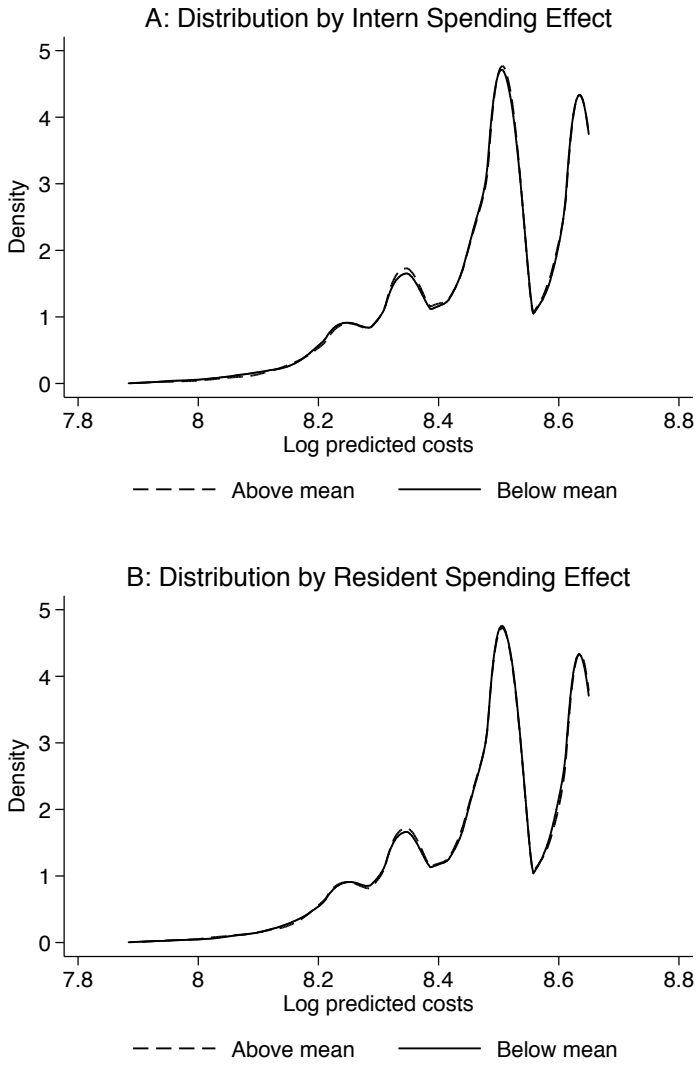
In Figure A-11, I show outcomes under counterfactual scenarios varying $\tilde{\delta}_2^{\Delta}$. The effects of increasing $\tilde{\delta}_2^{\Delta}$ on learning and decision-making information are similar to those of increasing $\tilde{\delta}_1^{\Delta}$: Increasing senior influence speeds up training 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



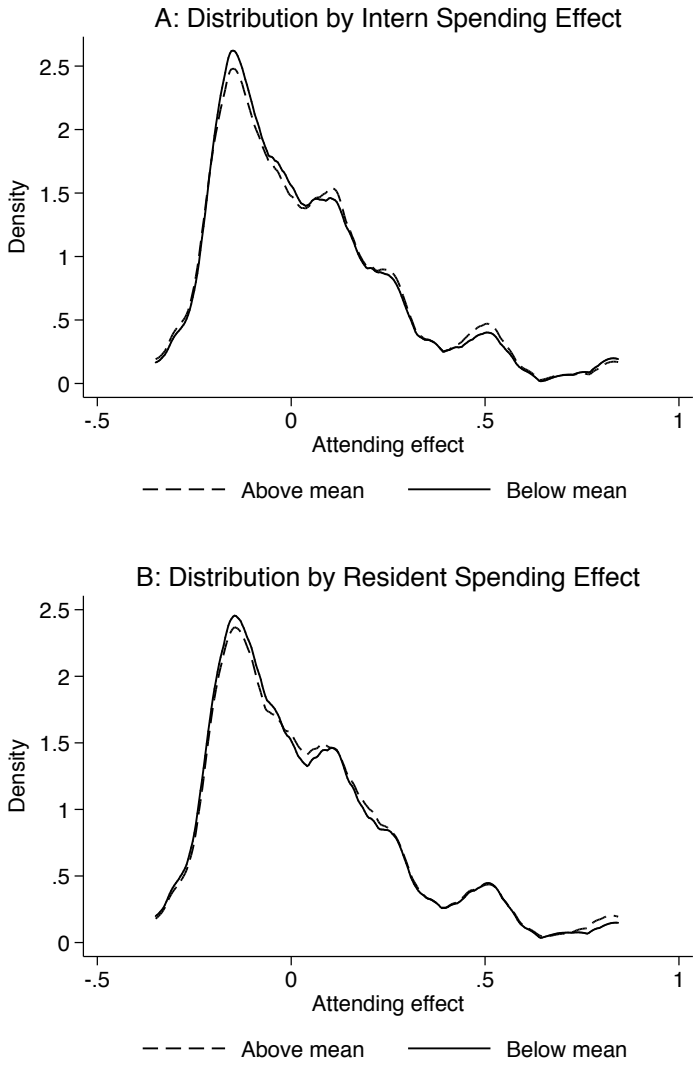
Note: This figure shows the distribution of the age of patients assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed 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



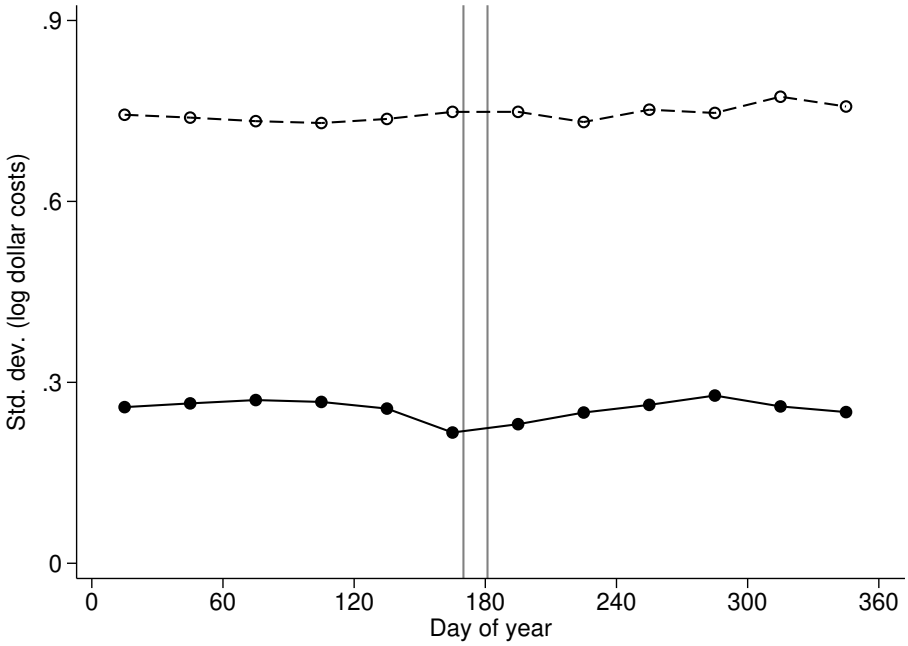
Note: This figure shows the distribution of predicted log costs (based on patient age, race, and gender) for patients assigned interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning 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



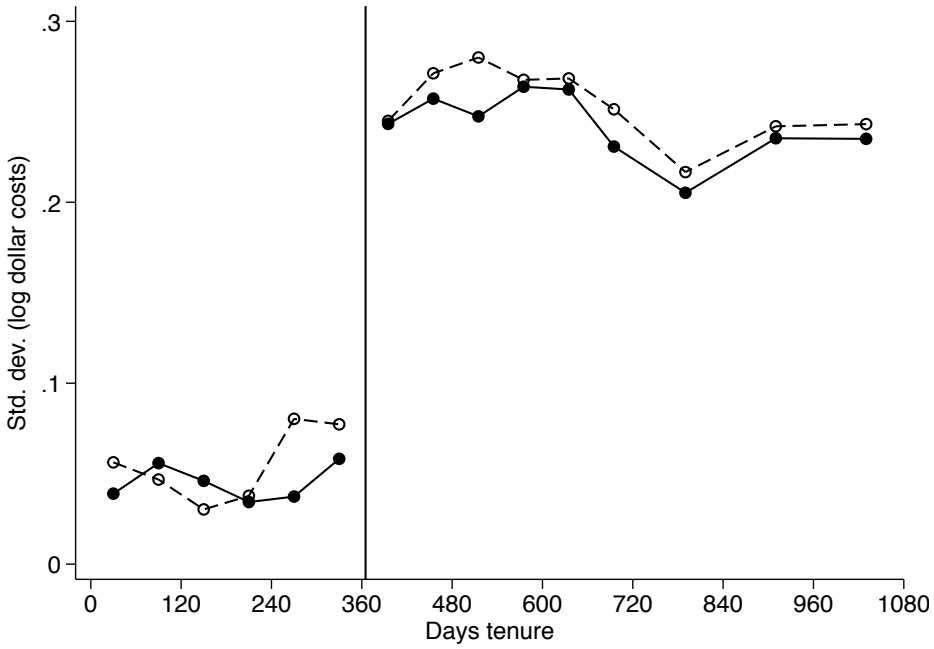
Note: This figure shows the distribution of spending fixed effects for attendings assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee and attending spending effects, not conditioning by tenure, are 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: Trainee-associated and Residual Variation by Day of Year



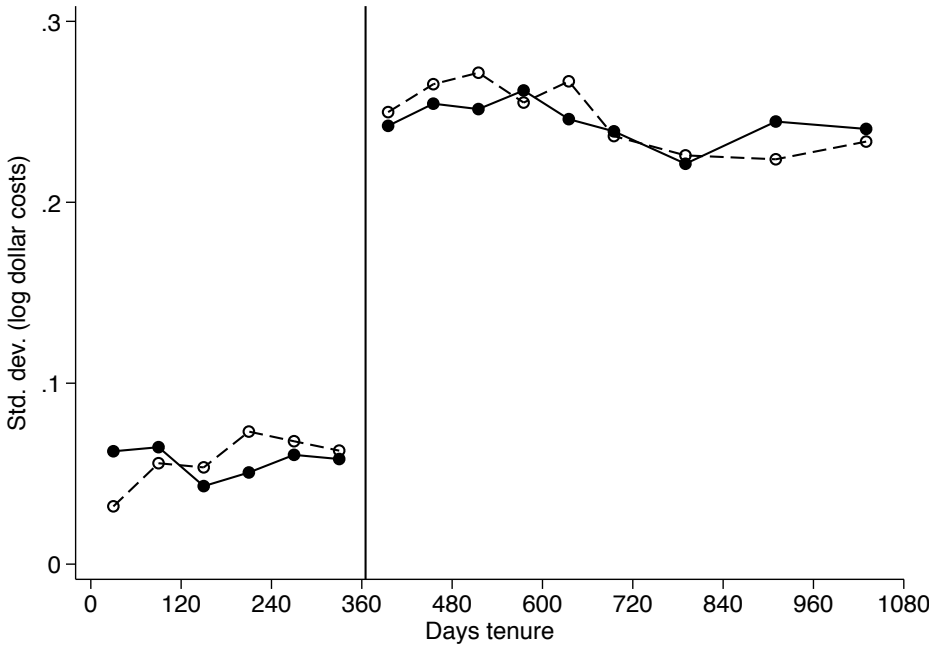
Note: This figure shows the standard deviation of random effects due to junior and senior trainee teams (solid dots) and the standard deviation of the residual (hollow dots) in 30-day periods by day of the year. Residual variation can be interpreted as variation due to independent observation. The two 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 (1), 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-5: Practice Variation Profile by ICD-9 Code Frequency



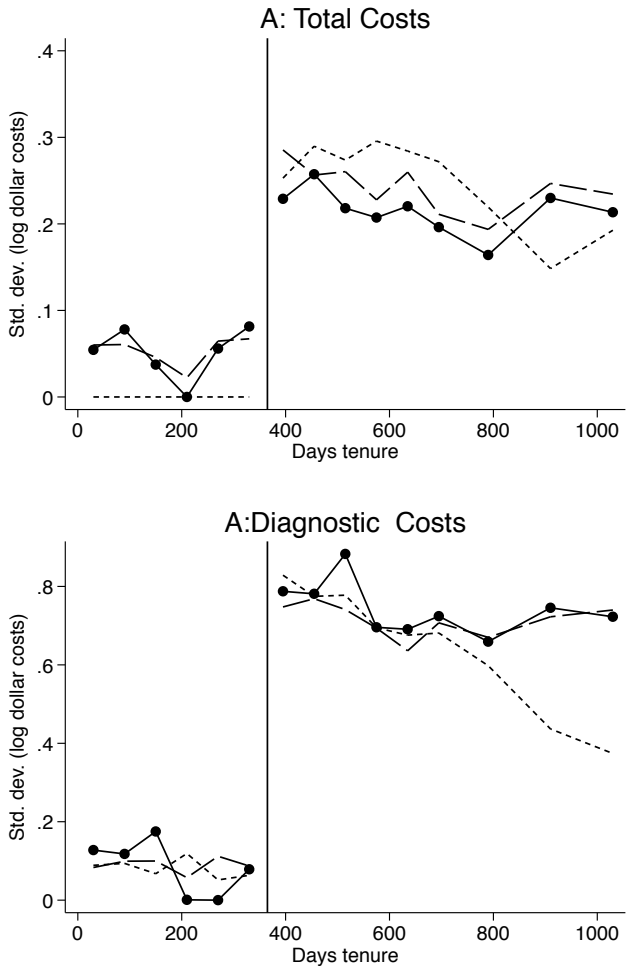
Note: This figure shows the standard deviation in a random effects model, as in Equation (1), of log daily total costs at each non-overlapping tenure interval estimated separately using observations with relatively common ICD-9 diagnostic codes (within service) (solid dots) and those with uncommon diagnoses (hollow dots). Controls are the same as those listed in the caption for Figure 1. Trainees prior to one year in tenure are interns and become residents after one year in tenure; vertical lines denote the one-year tenure mark.

Figure A-6: Practice Variation Profile by Guideline Existence



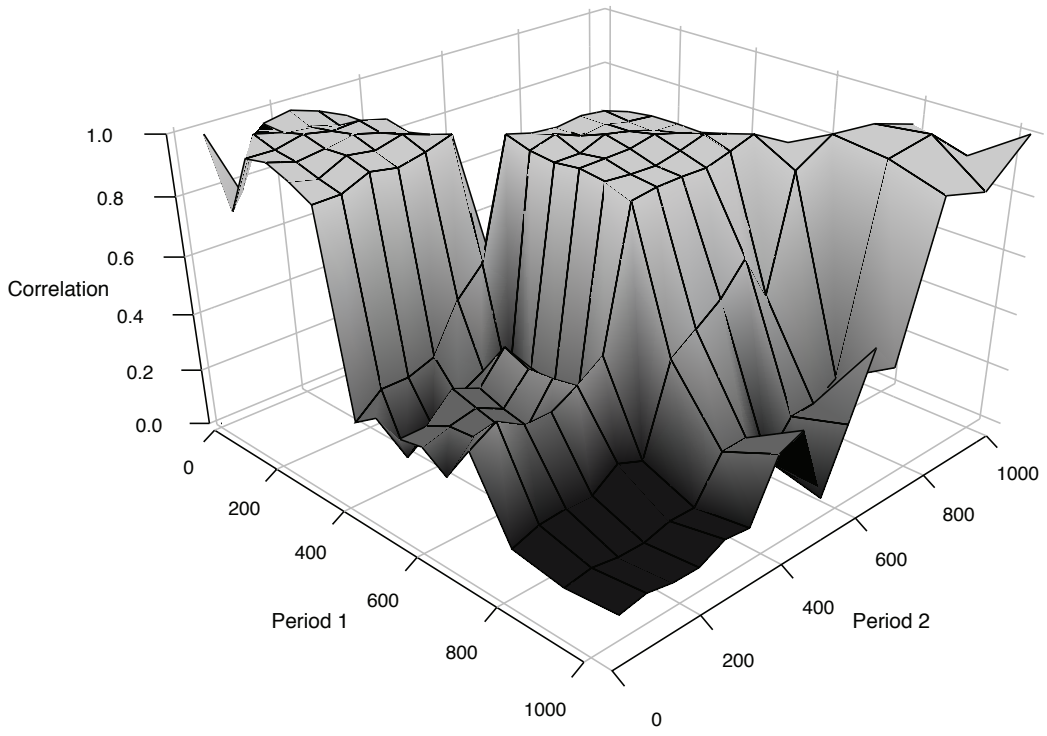
Note: This figure shows the standard deviation in a random effects model, as in Equation (1), of log daily total costs at each non-overlapping tenure interval estimated separately using diagnoses with (solid dots) and those without (hollow dots) published cataloged by the US Agency for Healthcare Research and Quality (`guidelines.gov`). Controls are the same as those listed in the caption for Figure 1. Trainees prior to one year in tenure are interns and become residents after one year in tenure; vertical lines denote the one-year tenure mark.

Figure A-7: Pseudo-cardiology Service



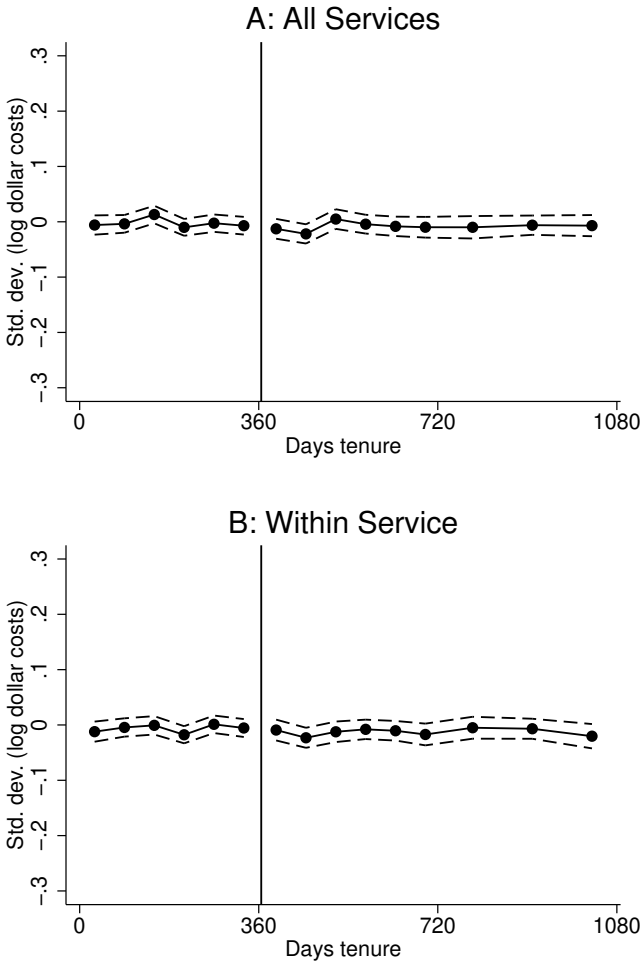
Note: This figure shows the tenure profiles of practice variation in log daily total costs (Panel A) and log daily diagnostic costs (Panel B) of a pseudo-cardiology service by ICD-9 codes. This service is constructed from general medicine observations, matching ICD-9 codes observed in cardiology. This procedure covers 97% of observations in the actual cardiology service. Each panels shows the standard deviation of trainee effects by tenure for actual services of cardiology (short-dashed line) and general medicine (long-dashed line), and for a pseudo-cardiology service (dot and solid line) comprised of patients in general medicine but matching ICD-9 code primary diagnoses in cardiology. Trainees prior to one year in tenure are interns and become residents after one year in tenure; vertical lines denote the one-year tenure mark. The flat values at 0 in Panel A indicates that no variation attributable to interns can be detected for total costs in cardiology. Estimation of Equation (1) includes admission-intern random effects to normalize higher variance in the (weighted) number of patients per intern in the pseudo-cardiology service (thus results are slightly different than in Figure 4, for example).

Figure A-8: Serial Correlation of Trainee Random Effects



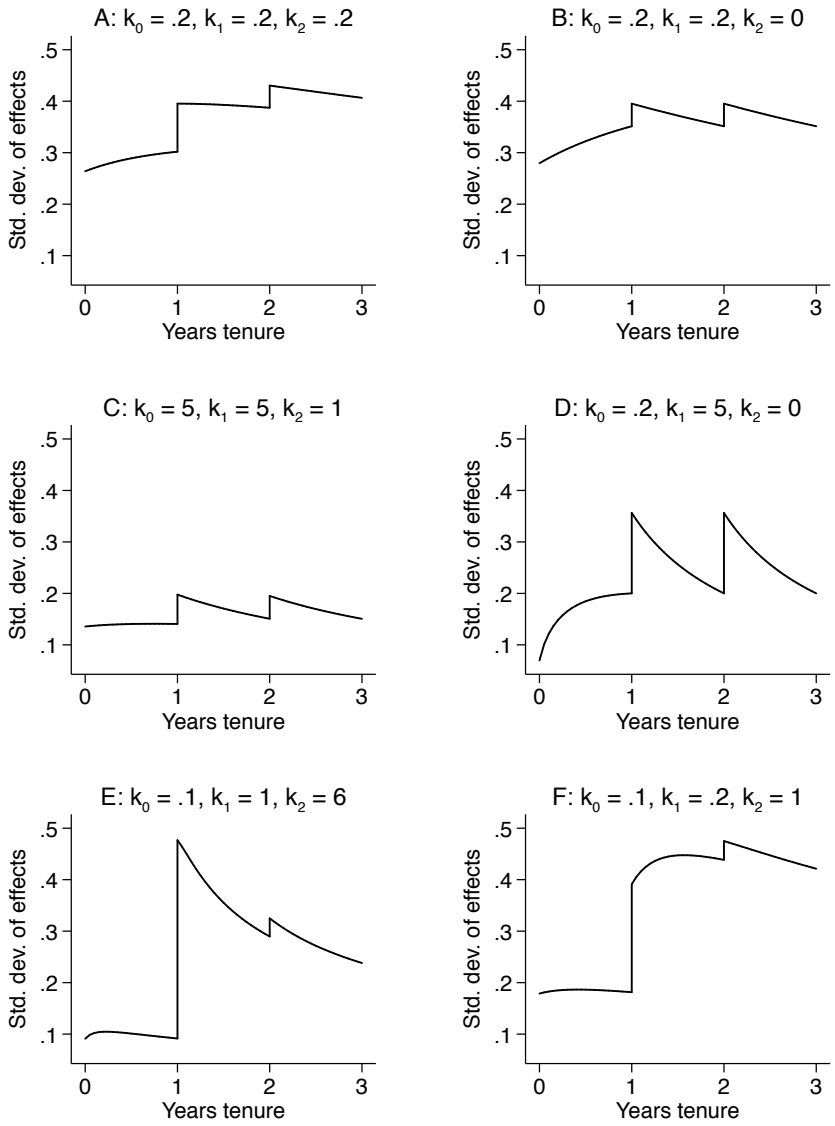
Note: This figure shows the serial correlation between random effects within trainee between two tenure periods. Details of the estimation routine are given in Appendix A-2.2. The random effect model of log daily total costs is given in Equation (1). The model controls are as stated for Figure 1. Trainees prior to one year in tenure are junior trainees and become senior trainees after one year in tenure

Figure A-9: Effect of High Exposure to Senior-trainee Spending



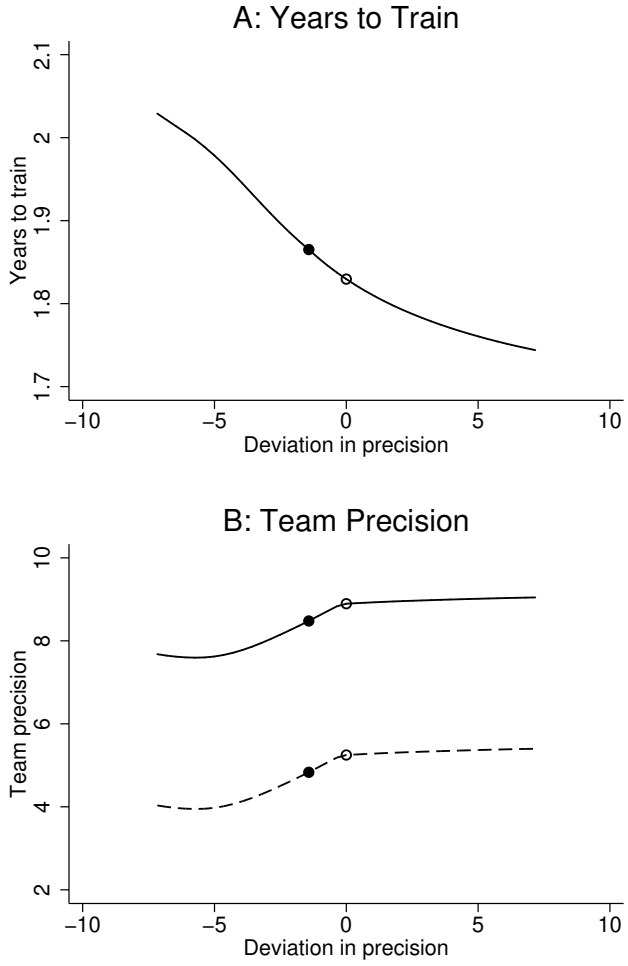
Note: This figure shows the effect of high prior exposure to senior-trainee spending. This exposure measure is discussed in further detail in Appendix A-3.3 and in Table A-3 and reflects the average spending effects of senior trainees that a given trainee was matched to in the past as a junior trainee. The tenure-specific effect of having high prior exposure to spending is estimated as in Equation (A-14). 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-3); 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-3). 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 6.

Figure A-10: 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 (7), 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-16). 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-4.

Figure A-11: 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 static efficiency. Lines are plotted for counterfactual $\delta_2^{\Delta}/\rho^{\Delta}(2) \in [-1,1]$. Further details are given in Appendix A-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	$F(1055, 46364) = 0.98$ $p = 0.655$	$F(20, 16069) = 0.68$ $p = 0.848$	$F(18, 37494) = 0.79$ $p = 0.711$
Male	$F(1055, 46364) = 1.01$ $p = 0.389$	$F(20, 16069) = 1.18$ $p = 0.256$	$F(18, 37494) = 1.26$ $p = 0.201$
White	$F(1055, 46364) = 1.02$ $p = 0.356$	$F(20, 16069) = 0.79$ $p = 0.734$	$F(18, 37494) = 0.92$ $p = 0.558$
Predicted spending	$F(1055, 46364) = 0.98$ $p = 0.685$	$F(20, 16069) = 0.79$ $p = 0.731$	$F(18, 37494) = 1.08$ $p = 0.368$

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: 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	0.090	0.090	0.091	0.089	0.089	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	0.095	0.095	0.088	0.088	0.094	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-11), 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 (details are given in Section A-3.1). 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 for Equation (A-13), where the regressor 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.

Table A-3: 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	5.73%	6.11%	4.90%	5.16%
61-120	5.54%	5.94%	4.98%	5.23%
121-180	4.95%	5.81%	4.53%	5.11%
181-240	4.82%	5.85%	3.81%	4.55%
241-300	4.34%	5.61%	3.87%	4.86%
301-365	4.05%	5.33%	3.47%	4.56%
366-425	3.90%	6.34%	4.15%	5.94%
426-485	4.05%	6.66%	4.27%	6.41%
486-545	3.59%	6.21%	3.65%	5.21%
546-605	3.43%	5.69%	3.89%	5.95%
606-665	3.65%	6.60%	4.33%	6.63%
666-730	3.76%	6.43%	3.75%	5.61%
731-850	3.81%	5.35%	2.41%	3.99%
851-970	3.85%	6.33%	2.68%	4.36%
971-1095	3.35%	4.23%	1.94%	3.22%
Panel B: Exposure to Spending by Senior Trainees				
0-60	18.73%	19.15%	20.61%	20.70%
61-120	19.73%	20.19%	22.74%	22.92%
121-180	19.39%	20.93%	21.40%	23.22%
181-240	19.44%	20.38%	21.97%	23.36%
241-300	18.78%	19.87%	21.79%	23.42%
301-365	17.52%	17.93%	19.84%	20.24%

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 best linear unbiased prediction (BLUP) in 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-4: 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	0.000	-0.002	-0.006	0.001
measure above median	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Observations	181,874	181,874	181,874	155,523	129,636
Adjusted R^2	0.088	0.088	0.088	0.089	0.090
Panel B: Residents					
Effect of trainee with	0.003	0.003	-0.005	0.008	0.013**
measure above median	(0.008)	(0.007)	(0.007)	(0.005)	(0.005)
Observations	199,934	199,934	199,934	182,017	174,534
Adjusted R^2	0.090	0.090	0.090	0.087	0.087
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-11), 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-5: Top Diagnostic Codes by Service

Cardiology		Oncology		General Medicine	
ICD-9	Description	ICD-9	Description	ICD-9	Description
786.50	Chest pain NOS	*162.9*	Malignant neoplasm of bronchus/lung NOS	*786.50*	Chest pain NOS
428.0	Congestive heart failure NOS	202.80	Other lymphoma unspecified site	780.2	Syncope and collapse
410.90	Acute myocardial infarction NOS	*174.9*	Malignant neoplasm of breast NOS	*486*	Pneumonia, organism NOS
414.9	Chronic ischemic heart disease NOS	*171.9*	Malignant neoplasm of soft tissue NOS	578.9	Gastrointestinal hemorrhage NOS
411.1	Intermediate coronary syndrome	*203.00*	Multiple myeloma without remission	*786.09*	Respiratory abnormality NEC
427.31	Atrial fibrillation	*780.6*	Fever	*789.00*	Abdominal pain unspecified site
427.1	Paroxysmal ventricular tachycardia	*183.0*	Malignant neoplasm of ovary	*428.0*	Congestive heart failure NOS
428.9	Heart failure NOS	*153.9*	Malignant neoplasm of colon NOS	*410.90*	Acute myocardial infarction NOS
780.2	Syncope and collapse	*276.51*	Dehydration	*577.0*	Acute pancreatitis
425.4	Primary cardiomyopathy NEC	205.00	Acute myeloid leukemia without remission	*496*	Chronic airway obstruction NEC
786.09	Respiratory abnormality NEC	157.9	Malignant neoplasm of pancreas NOS	*276.51*	Dehydration
427.89	Cardiac dysrhythmias NEC	*486*	Pneumonia, organism NOS	300.9	Nonpsychotic mental disorder NOS
996.00	Malfunctioning cardiac device/graft NOS	*185*	Malignant neoplasm of prostate	682.9	Cellulitis NOS
427.32	Atrial flutter	*789.00*	Abdominal pain unspecified site	*599.0*	Urinary tract infection NOS
413.9	Angina pectoris NEC/NOS	*150.9*	Malignant neoplasm of esophagus NOS	*285.9*	Anemia NOS

Note: This table lists the top 15 primary admission diagnoses, by ICD-9 codes, in order of descending frequency, for each of the ward services of cardiology, oncology, and general medicine. Italicized ICD-9 codes denote codes that are linked to guidelines on `guidelines.gov`. "NOS" = "Not Otherwise Specified"; "NEC" = "Not Elsewhere Classified."