Learning on the Job: Evidence from Physicians in Training∗

David C. Chan†

March 10, 2018

Abstract

Learning on the job creates a tradeoff in team decisions: Workers with less knowledge have less to contribute to team decisions, but experiential learning may require that trainees also have a stake in decisions to learn. This paper studies learning and influence in team decisions among physicians 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, which 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 receive much more influence than statically efficient relative to supervising physicians, possibly to improve 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, 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.


1 Introduction

In a broad range of professions, important knowledge is acquired on the job. In these settings, on-the-job training often involves work in teams with more experienced co-workers. The decisions made in teams have real consequences but also serve as an opportunity for inexperienced workers to learn. How learning occurs on the job therefore has implications for how knowledge should be used in teams to make decisions. 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 learning and influence in team decisions among physicians in training in a large academic hospital. The process of training physicians is a particularly salient setting to study this general issue. 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 of training in the medical profession, there exists substantial variation in practices and beliefs across physicians (Cutler et al., 2013; Currie and MacLeod, 2017), 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.

Further, medical residency provides an empirically tractable setting for extracting and studying the 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, which is immediately after medical school. For every patient, decision-making is explicitly made in a team comprised of 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 and exploit detailed administrative data assigning physician trainees to teams caring for patients. Team decisions are measured 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. I focus on practice variation as the standard deviation of the distribution of these 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). Variation in judgments across providers should decrease as a group of providers gains more complete knowledge. But at the same time, influence in team decisions should accrue with knowledge, serving as a force to increase practice variation. I first assess the role of influence in reduced form, by exploiting a discontinuity in the relative experience at the end of trainees’ first year of training: Trainees in their first year have relatively less experience than their teammate, while they have 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, via a change in relative experience, on practice variation.

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

---

1The strategy I employ is similar to that used in a number of papers starting with Abowd et al. (2008), 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 that supports intrinsic heterogeneity (e.g., different preferences or skills) or learning by copying others as alternative sources of practice variation.  

In order to study more directly 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 receive less influence than optimal (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), in which trainees are given a stake in decisions so that they may gain experiential knowledge.

The structural model maps primitives of information — specifically, the knowledge held by trainees when they start residency, rates of learning in different phases of residency, the external information brought by the hospital supervisory structure, and deviations from static efficiency at a given year of training — onto previously estimated moments of practice variation at each tenure level in residency. 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, at the end of three years of residency, physician trainees possess knowledge with 9.10 units of precision, which implies practice variation with a standard deviation of $9.10^{1/2} \approx 33\%$ if they were to make decisions on their own. Remarkably, more than 90% of this knowledge

---

2I 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 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.
is acquired in the second year of training, or the first half of training in the role of senior trainee. Trainees begin residency with essentially no useful knowledge, yet learning in the second year is about 30 times faster than in the first year. Influence is approximately efficiently allocated between junior and senior trainees. However, the entire supervisory hospital structure — including the supervising physician, nurses, pharmacists, consultants, and the computer order entry system — contributes information to decision-making that is slightly less than 40% of the knowledge of a fresh residency graduate. This last estimate 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 result is consistent with supervising physicians (and the rest of the hospital) granting autonomy to trainees so that they may learn.

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. 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,” a dictum 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). My results suggest that, on its own, seeing prompts relatively little learning, but that doing — specifically 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.”

I show how knowledge is aggregated across agents in a team via influence. Unlike predictions in the canon-

---

3 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 Lizzieri and Siniscalchi (2008), which develops a theory of parental sheltering, experiential learning has largely been overlooked.
ical 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 little knowledge. However, agents who are still in training but not the most junior drive more decision-making than is statically efficient, relative to supervising physicians with the same workload, plausibly because experiential learning is an organizational priority.

Third, these results relate to a large literature documenting practice variation in health care. 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 that experts often and widely disagree (Cutler et al., 2013). This paper highlights informational mechanisms behind wide practice variation in a training environment designed to create homogeneity. The experiential nature of learning may explain why heterogeneity persists despite guideline promulgation. The evidence also suggests that trainees largely cease to learn before the end of training, which implies that extending the period of training would not diminish practice variation.

The remaining organization of this paper is as follows. Section 2 describes the institutional setting and the 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. Section 6 discusses policy implications for practice variation and concludes.

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 a realization that physicians need to be intensively trained in the profession following graduation from medical school. In the period between the two world wars, medicine underwent large improvements in surgical

---

4 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. 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.
technique; new fields such as radiology, pediatrics, and psychiatry; and advances in anesthesiology, blood transfusion, and intravenous fluid replacement. Residency programs came to dominate the training of physicians.

Since this period, the basic principles of residency training have remained relatively consistent despite continued changes in medical technology. First, trainees are to be given independence and responsibility to make clinical decisions on their own, with this responsibility granted in a “graded” fashion consistent with their knowledge. Second, the emphasis is on delving deeply into the cases of relatively few patients, with close and immersive observation per patient. Rather than memorize facts in an increasingly rich world of information, trainees are encouraged to relate their patients’ presentations to underlying principles, to think critically about clinical observations and evidence, and to engage in debate about diagnosis and treatment choices. Finally, much of training is informal, with trainees emulating 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. Of note, 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, and teams are 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 every resident. Interns are thus assigned half the number of patients as residents and can thus devote more attention to each patient and usually are the first to examine patients and make judgments. Aside from differences in experience and span of control, the intern and resident roles have no other formal distinctions in decision rights or in regulatory considerations. Trainee teams are supervised by “attending” physicians and operate within a broader practice environment, including other health care workers (e.g., consulting physicians, pharmacists, and nurses), institutional rules, and information systems that influence decisions.

Trainees on the same teams may come from different predetermined career “tracks,” other programs

---

5See 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.
A sizable number of interns only plan to spend one year in the internal medicine residency ("preliminary" interns, as opposed to the standard "categorical" interns), subsequently proceeding to other residency programs, such as anesthesiology, radiology, or dermatology. Trainee schedules are arranged a year in advance. This study focuses on inpatient ward rotations, which are comprised of cardiology, oncology, and general medicine services. Per residency administration, preferences are not collected about rotations, 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 done independently, and neither trainee nor attending scheduling is aware of each other’s results in advance.

Patients arriving at the hospital are assigned to interns and residents by algorithm, which distributes patients in a rotation among trainees that are “on-call” and have not reached the maximum number of patients. Patients who remain admitted for more than one day may also be mechanically transferred between trainees changing rotations. When a trainee replaces another one, she assumes the care of the entire list of patients from the other trainee. Because trainee blocks are generally two weeks in length and staggered for interns and residents, many patients to 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 – the 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 48,185 admissions, equivalent to 220,117 patient-day observations. Corresponding to these admissions are 724 unique interns, 410 unique residents, and 540 unique attendings. Of the trainees, 516 interns and 347 residents are from the same-hospital internal medicine residency, with the remainder visiting from another residency program within the same hospital or from the other hospital. There is no unplanned attrition across years of residency.\(^6\)

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 precommitted career “tracks” for each trainee physician, including special tracks (e.g., primary care, genetics), the standard “categorical” internal medicine track, and tracks into another residency such as anesthesiology, dermatology, [obstetrics-gynecology and emergency medicine] and another hospital. A sizable number of interns only plan to spend one year in the internal medicine residency (“preliminary” interns, as opposed to the standard “categorical” interns), subsequently proceeding to other residency programs, such as anesthesiology, radiology, or dermatology. Trainee schedules are arranged a year in advance. This study focuses on inpatient ward rotations, which are comprised of cardiology, oncology, and general medicine services. Per residency administration, preferences are not collected about rotations, 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 done independently, and neither trainee nor attending scheduling is aware of each other’s results in advance.

Patients arriving at the hospital are assigned to interns and residents by algorithm, which distributes patients in a rotation among trainees that are “on-call” and have not reached the maximum number of patients. Patients who remain admitted for more than one day may also be mechanically transferred between trainees changing rotations. When a trainee replaces another one, she assumes the care of the entire list of patients from the other trainee. Because trainee blocks are generally two weeks in length and staggered for interns and residents, many patients to experience a change in either the intern or the resident on the team.

\(^6\)In two cases, interns with hardship or illness in the family were allowed to redo intern year.
psychiatry, or radiology after a preliminary intern year. In addition to trainee characteristics predetermined relative to residency, I use observed physician specialty after training to impute expected yearly future income in the five years immediately after this training, based on industry-standard survey data from the Medical Group Management Association (MGMA). 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.\(^7\)

I use scheduling data and past matches with supervising attending physicians and other trainees. As described in Section 2, trainees do not choose most of their learning experiences, at least in terms of their clinical rotations and in what order, peers and supervising physicians, and patients seen on the wards. Table 1 shows that interns and residents, respectively, 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 (e.g., heart failure). 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, that I further group into categories of diagnostic (laboratory and radiology), medication, blood bank, and nursing spending. The costs I observe are based on the hospital’s accounting of resource utilization due to physician actions, as opposed to measures of Medicare reimbursement used in recent studies (Doyle et al., 2015; Skinner and Staiger, 2015; Chandra et al., 2016), and therefore provide new insight into welfare-relevant resource use.\(^8\) Laboratory costs are based on 3.1 million physician laboratory orders; radiology costs are

\(^7\)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.

\(^8\)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.
based on 268,065 orders ordered tests in CT, MRI, nuclear medicine, and ultrasound; and medication costs are based on 3.4 million 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 in swaying team decisions; and (iii) intrinsic heterogeneity in ability or preferences that may result in different actions even under the same information. In this analysis, I will focus on practice variation across the one-year tenure mark, when trainees experience a discontinuity in relative experience: Immediately before one year, trainees are junior, with at least a year less experience, relative to their trainee teammate; immediately after one year, they are senior relative to their trainee teammate. Under the assumption 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

\[ Y_{it} = X_i \beta + T_t \eta + \xi_{j(i,t)}^{\tau(i,t)} + \xi_{k(i,t)}^{\tau(i,t)} + \zeta_{\ell(i,t)} + v_i + \epsilon_{it}, \]  

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 — \( \xi_{j(i,t)}^{\tau(h,t)} \) and \( \xi_{k(i,t)}^{\tau(h,t)} \) 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 \( X_i \); time categories \( T_t \) (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 by a patient effect \( v_i \).
Conditional on service and time categories, trainee effects in Equation (1) are identified by quasi-random assignment of patients to trainees (see Appendix A-1.1) and frequent and quasi-random reshuffling of trainees to teams (see Appendix A-1.2).\footnote{This 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).} Since patients are \textit{not} randomly assigned to attending physicians, I model attending “effects” as fixed and treat these as nuisance parameters capturing both true effects and unobserved patient selection to attending physicians.\footnote{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).}

Because observations per trainee are finite, direct 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. 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 each other. The former assumption is an approximation; practice variation estimates are still meaningful even if the 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.

Figure 1 presents results for the estimated standard deviations of the trainee effect distributions within each tenure interval \(\tau\). In my baseline specification, I consider non-overlapping tenure intervals that are 60 days in length for the first two years of residency, and 120 days in length for the third year, since third-year trainees have fewer inpatient days.\footnote{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.} 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 that senior influence plays in 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 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 receive more influence than 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 increase at the one-year discontinuity does depend on the type of decision. Diagnostic spending shows a large increase in practice variation, from a standard deviation of 16% to one of 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 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 more well-described in publicly accessible sources of knowledge, but diagnostic decisions draw more on a process of clinical reasoning that would be difficult to prespecify and reference for trainees who have never encountered a patient presentation before.

In Figure 3, I show that the practice variation profile is largely similar for patients with high or 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 different ward services — cardiology, oncology, and general medicine.
— that represent different subspecialties of medicine. The same trainees rotate across these ward 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 formal diagnostic codes. 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

Finally, I directly evaluate two alternative mechanisms behind practice variation, both of which have clear policy implications. In the first mechanism, intrinsic physician heterogeneity in skills or preferences are the root of practice variation. In the second mechanism, physicians primarily learn from others, as opposed to learning 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 prior evidence supporting them has been weak, and they do not seem consistent with wide practice variation among a highly selected group of trainees in an intensive residency program.

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 are largely unpredictive of trainee effects. In Figure 5, I show the distribution of trainee effects in each tenure period throughout residency for high- vs. low-ranked trainees, and for trainees with high vs. low future income. I describe these analyses further in Appendix A-3.1 and present more exhaustive results

---

12 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.

13 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.
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, the trainee effects should be correlated across distant periods. Instead, I find that trainee effects are strongly correlated between two adjacent tenure periods but only weakly correlated between distant tenure periods (Appendix A-3.2 provides further details).

I evaluate the mechanism of learning from others (or “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 that a trainee is exposed to. 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 shows that the prior experience with supervising physicians — defined as the average supervising-physician spending effect for all prior patients and for prior patients in the last prior two, four, and six months — are unpredictive of current spending decisions, throughout the course of training. 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 (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) external information, including from the attending physician, that acts as a single supervisory “agent.” Each decision $d$ can be summarized perfectly by an unknown parameter $\theta_d$. If $\theta_d$ 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 $\theta_d$. 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 $d$.

\footnote{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\%.}
2. There may also be external or public knowledge about $d$. Some of this knowledge is held by the attending physician, but other sources derive from the hospital supervisory structure of nurses, consultants, and protocols. Each agent may also collect information about the decision, which I assume to be independent of prior knowledge. 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

$$a_d^* = \frac{\rho_{d,i}\mu_{d,i} + \rho_{d,j}\mu_{d,j}}{\rho_{d,i} + \rho_{d,j} + P_d}. \quad (2)$$

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

The optimal weights on judgments in Equation (2),

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

have a natural interpretation as the influence of trainee $h$ on the action $a_d^*$. The more precise is the signal from her prior knowledge, relative to her teammate and any external information, the greater her influence will be. In the limit, if either her teammate or the external environment had 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.

In principle, however, influence may not always be optimally allocated. For example, junior trainees may exercise less influence than optimal if they are overly deferential to seniors, 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 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 only consider influence defined by $g_{d,h;\rho_{d,h}}$ as statically efficient, in that they optimize the decision at hand. In estimation, I allow for deviations
from static efficiency as
\[ \hat{a}_d = \frac{\tilde{\rho}_{d,i} \mu_{d,i} + \tilde{\rho}_{d,j} \mu_{d,j}}{\tilde{\rho}_{d,i} + \tilde{\rho}_{d,j} + \Pi_d}. \] (3)

\( \tilde{\rho}_{d,h} = \rho_{d,h} + \delta(h) \) is an effective “precision” that equals the true precision of \( h \)'s knowledge adjusted by \( \delta(h) \), depending on the tenure of \( h, \tau_h \). \( \Pi_d = P_d + \Delta_d \) is similarly an effective “precision” for external sources of information. That is, even though supervising physicians and the broader supervisory structure may have precision \( P_d \) of information about \( d \), this information may be underweighted \( \Delta_d < 0 \) or over-weighted by \( \Delta_d > 0 \).

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., \( \rho_{d,h} = \rho(h, t(d)) \). 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 knowledge is fixed within tenure blocks. Further, if \( P_d = P \), then we can state influence between a pair of trainees in terms of the tenure-dependent precision of knowledge.\(^{15} \) For all \( d \) by any pair of trainees with tenures \( \tau_h \) and \( \tau_{-h} \), 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 \( \Pi_d = \Pi \), effective influence would be implied by the weights

\[ g(\tau_h; \tau_{-h}) = \frac{\tilde{\rho}(\tau_h)}{\tilde{\rho}(\tau_h) + \tilde{\rho}(\tau_{-h}) + \Pi}. \] (4)

Given tenure-specific influence weights, we can state trainee effects in terms of influence and expected judgments averaged over decisions. Define the set of decisions \( D^T_h \) involving trainee \( h \) in tenure block \( \tau \).

For intern \( j \) and resident \( k \) in respective tenure blocks \( \tau_j \) and \( \tau_k \), the expected action in Equation (3) for

\(^{15} \)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.
these decisions can be decomposed into two parts corresponding to their effects:

\[
E_d \left[ \hat{a}_d \mid d \in D_j^T \cap D_k^T \right] = E_d \left[ \hat{a}_d \mid \tau_j, \tau_k \right] \\
= g \left( \tau_j ; \tau_k \right) E_d \left[ \mu_{d,j} \mid \tau_j \right] + g \left( \tau_k ; \tau_j \right) E_d \left[ \mu_{d,k} \mid \tau_k \right] .
\]

The first equality holds because patients (and decisions) are randomly assigned; the second equality holds because influence is constant conditional on the tenure of the junior and senior trainees. In the second line, the first term corresponds to the intern trainee effect, or \( \xi^T_j \), and the second term corresponds to the resident trainee effect, or \( \xi^T_k \), 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 \( d \):

\[
Var_h \left( \mu_{d,h} \mid \tau_h = \tau \right) = 1 / \rho (\tau) .
\]

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 \( Var_h \left( \mu_{d,h} \mid \tau_h = \tau \right) \). Using a simple variance components model, one may use a convenient formulation

\[
Var_h \left( \xi^T_h \right) = Var_h \left( E_d \left[ \mu_{d,h} \mid \tau_h = \tau \right] \right) = \kappa / \rho (\tau),
\]

for some \( \kappa \in (0,1) \). The observed standard deviation of trainee effects for trainees with tenure \( \tau_h \), working with teammates with tenure \( \tau_{-h} \), is then

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

It is easy to see that the practice variation profile with respect to tenure is scaled by \( \sqrt{\kappa} \). That is, the practice variation profile retains the same shape across tenure periods, but is smaller in scale with lower \( \kappa \). 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, in precision terms, will similarly be scaled by \( 1 / \kappa \). Ratios of knowledge and learning at different points in training will remain identical, regardless of \( \kappa \).

\[\text{To arrive at this, consider } \mu_{d,h} = \eta_d + \bar{\mu}_h + \tilde{\mu}_{d,h}, \text{ where } Var_h \left( \bar{\mu}_h \mid \tau_h = \tau \right) = \kappa / \rho (\tau), \text{ and } E_h \left[ \tilde{\mu}_{d,h} \mid \tau_h = \tau \right] = 0 \text{ for all } h \text{ and } \tau_h, \text{ and } Var_h \left( \tilde{\mu}_{d,h} \mid \tau_h = \tau \right) = (1 - \kappa) / \rho (\tau). \text{ This model implies that learning occurs at the same rate for the trainee mean judgment (the trainee effect) and deviations from the trainee mean judgment.}\]
4.2 Interpretation and Identification

In the efficient benchmark, as the precision of knowledge increases for a 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 then 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 \( \kappa \), 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-17 and provides numerical examples.

Given that trainee teams have a clear structure (i.e., senior trainees with tenure \( \tau_k \) work with junior trainees who have tenure \( \tau_k \mod 1 \)), 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 the discontinuity in practice variation at the one-year tenure mark, to a lesser extent the potential discontinuity at the two-year tenure mark (see Proposition A-1 in Appendix A-4), and 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 then decreasing with respect to tenure, but according to the model, it should never be decreasing then increasing within a continuous segment.

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, \( \kappa \), would also reduce the scale of practice variation. Because I cannot separately identify \( \kappa \) and precision in \( \rho(\tau) \) and \( P \), I normalize \( \kappa = 1 \) and attribute smaller scale to
larger $\rho(\tau)$ and $P$.\textsuperscript{17}

Because trainee effects measure judgments that are common across decisions, they reflect prior knowledge that the trainee brings to the decision, independent of any information she may gather in the process. That is, $\rho(\tau_h)$ only counts the precision of her prior knowledge and does not include the precision of information she may have gathered for the case. The latter information is instead included $P$. 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. While I consider the distribution of this “outside” information as having mean 0 in the simple model, this assumption is inconsequential, as it is orthogonal to trainee knowledge.\textsuperscript{18}

Finally, I allow for deviations of influence from the efficient benchmark. To do so, I impose two restrictions on learning on the job. First, the precision of knowledge changes continuously through learning on the job. Second, the precision of knowledge does not decrease with tenure. Given these restrictions, I model the “effective” precision of the trainee as $\tilde{\rho}(\tau_h) = \rho(\tau_h) + \delta(\tau_h)$, where $\delta(\tau_h)$ is a step function with respect to completed years of training, representing the deviation from using true precision $\rho(\tau_h)$. Trainees who have completed a full year of training (and are now senior on the team) may discontinuously receive more or less effective precision in influence; similarly, trainees who have completed the second year of training may receive an additional discontinuous change in effective precision. This formulation models deviations from efficient influence, $g^*(\tau_h; \tau_{-h})$, between trainees. Any deviation from efficient influence between the trainee team and the supervisory hospital structure is implicitly captured by the estimated $\hat{\Pi}$. Under the assumption that the precision of outside information must be at least as large as that of trainee knowledge at the end of training, or $P \geq \rho(\tau_h = 3)$, I identify a lower bound on this deviation between trainees and supervisors if $\hat{\Pi} < \hat{\rho}(3)$.

\textsuperscript{17}In principle, one could separately identify $k$ 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.

\textsuperscript{18}The “judgment” of the outside information can be viewed as captured by all terms other than the trainee effects in the regression Equation (1), including the error term.
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 $\tau_h$ working with teammates of tenure $\tau_{-h}$. 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.

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

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

(6)

where $\rho_0$ represents the precision of knowledge before starting residency, $\rho_1$ is the yearly rate of learning in the first year of residency as a junior trainee, and $\rho_2$ and $\rho_3$ are analogous rates of learning in the second and third years, respectively, as a senior trainee.

I also specify potential deviations from statically efficient influence. Between trainees, I specify level shifts in effective precision that may occur at the completion of the first or second year of training:

$$
\tilde{\rho}(\tau) = \rho(\tau) + \delta_1 1(\tau \geq 1) + \delta_2 1(\tau \geq 2).
$$

(7)

$\delta_1$ represents the deviation in effective precision based on the title of “senior trainee” alone, and $\delta_2$ represents a potentially additional deviation in effect precision when trainees enter the third year of training. $\delta_1$ and $\delta_2$ capture deviations from efficient influence between trainees. Deviations from efficient influence between the trainee team and supervisors are implicitly captured by $\hat{\Pi}$.

Using nonlinear least squares, I estimate the parameter vector $\theta = (\rho_0, \rho_1, \rho_2, \rho_3, \delta_1, \delta_2, \Pi)$ that minimizes the objective

$$
Q(\theta) = \frac{1}{||\mathcal{T}||} \sum_{(\tau_h, \tau_{-h}) \in \mathcal{T}} (\hat{\sigma}(\tau_h, \tau_{-h}) - \sigma(\tau_h, \tau_{-h}; \theta))^2,
$$

19
where \( \hat{\sigma}(\tau_h, \tau_{-h}) \) are empirical estimates of practice variation from the first step; \( \sigma(\tau_h, \tau_{-h}; \theta) \) are model-implied practice variation from Equation (5) given \( \theta \); and \( \mathcal{T} \) is the set of tenure-period pairs corresponding to the index and peer groups of trainees, respectively.

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. To account for uncertainty in the first-step estimates, I calculate 95% confidence intervals by bootstrapping, described in further detail in Appendix. I 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 \( \rho_0, \delta_1, \) and \( \Pi \) 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 the 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, external information \( \Pi \) 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 \( \Pi = 3.7. \) Finally, practice variation is relatively flat in the third year, despite previously beginning to converge in the second year. This implies close to no learning in the third year (\( \rho_3 = 0 \) as the baseline estimate, and \( \rho_3 = 0.36 \) as the mean bootstrapped estimate).
With respect to influence relative to the statically efficient benchmark, I estimate that $\delta_1 = 0.23$. Although this represents a large increase in effective precision relative to knowledge at the end of first year ($\rho_0 + \rho_1 = 0.24$) but a small increase relative to the learning that occurs in the second year ($\delta_1 / \rho_2 \cdot 365 \text{ days} = 11 \text{ days’ worth of second-year learning}$). Thus, a reasonable view is that influence is close to statically efficient between junior and senior trainees. Third-year trainees have less influence than statically efficient, with $\delta_2 = -1.4$, although this parameter is imprecisely estimated and small relative to $\rho_2$. Because supervising physicians are not randomly assigned, I do not have estimates of practice variation for them to infer their knowledge. However, under the assumption that supervising physicians (and indeed the entire supervisory structure in the hospital) should embody at least as much knowledge as a graduating trainee. Thus, $\Pi = 3.7$ represents a strikingly low amount of effective external information, less than half of the knowledge of a graduating trainee.

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. Interestingly, I cannot reject the hypothesis that there is no learning involved in prescription decisions. Finally, consistent with similar practice variation profiles in Figure 3, learning and influence is similar regardless of patient severity but differ substantively between decisions made earlier vs. later in a patient’s stay. There is substantially more learning for decisions earlier in the stay.

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 share features with previously experienced patients (Polanyi, 1966).
Exploiting a discontinuity in relative experience, I find reduced-form evidence that physicians with more experience receive 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 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 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).  

Although these questions are mostly outside of 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).  

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.

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 vs. comparative advantage.
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 large enough to justify policy (Bloom et al., 2013, 2014).

References


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) vs. below-median (hollow dots) expected 30-day mortality. Panel B estimates the model separately in two samples of days before (solid dots) vs. 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.
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.
Table 1: Exogenous Assignment for Trainees with Above or Below Average Spending

<table>
<thead>
<tr>
<th>Patient characteristics</th>
<th>Interns</th>
<th></th>
<th></th>
<th>Residents</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below-median spending</td>
<td>Above-median spending</td>
<td>Below-median spending</td>
<td>Above-median spending</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.483 (0.500)</td>
<td>0.482 (0.500)</td>
<td>0.484 (0.500)</td>
<td>0.482 (0.500)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White race</td>
<td>0.707 (0.455)</td>
<td>0.705 (0.456)</td>
<td>0.703 (0.457)</td>
<td>0.709 (0.454)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black race</td>
<td>0.161 (0.367)</td>
<td>0.156 (0.363)</td>
<td>0.156 (0.363)</td>
<td>0.161 (0.368)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted log total costs</td>
<td>8.477 (0.142)</td>
<td>8.478 (0.139)</td>
<td>8.498 (0.140)</td>
<td>8.477 (0.140)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physician teammates</th>
<th>Interns</th>
<th></th>
<th></th>
<th>Residents</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Above-median-spending residents</td>
<td>0.504 (0.500)</td>
<td>0.495 (0.500)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above-median-spending attendings</td>
<td>0.486 (0.500)</td>
<td>0.509 (0.500)</td>
<td>0.484 (0.500)</td>
<td>0.510 (0.500)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. 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

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cardiology</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th percentile</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>189</td>
</tr>
<tr>
<td>10th percentile</td>
<td>0</td>
<td>16</td>
<td>14</td>
<td>0</td>
<td>244</td>
</tr>
<tr>
<td>Median</td>
<td>0</td>
<td>34</td>
<td>67</td>
<td>16</td>
<td>658</td>
</tr>
<tr>
<td>Mean</td>
<td>54</td>
<td>51</td>
<td>113</td>
<td>32</td>
<td>661</td>
</tr>
<tr>
<td>90th percentile</td>
<td>125</td>
<td>103</td>
<td>233</td>
<td>56</td>
<td>1,075</td>
</tr>
<tr>
<td>95th percentile</td>
<td>375</td>
<td>145</td>
<td>417</td>
<td>87</td>
<td>1,212</td>
</tr>
</tbody>
</table>

| **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

<table>
<thead>
<tr>
<th>Service</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>General Medicine</td>
<td>Cardiology</td>
<td>Oncology</td>
</tr>
<tr>
<td>Knowledge parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior to training ($\rho_0$)</td>
<td>0.035</td>
<td>0.286</td>
<td>0.002</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.206)</td>
<td>(0.007)</td>
<td>(0.484)</td>
</tr>
<tr>
<td>First year ($\rho_1$)</td>
<td>0.241</td>
<td>0.309</td>
<td>0.000</td>
<td>0.822</td>
</tr>
<tr>
<td></td>
<td>(0.129)</td>
<td>(0.308)</td>
<td>(0.000)</td>
<td>(1.189)</td>
</tr>
<tr>
<td>Second year ($\rho_2$)</td>
<td>7.365</td>
<td>7.910</td>
<td>5.549</td>
<td>4.136</td>
</tr>
<tr>
<td></td>
<td>(2.627)</td>
<td>(2.927)</td>
<td>(3.439)</td>
<td>(3.575)</td>
</tr>
<tr>
<td>Third year ($\rho_3$)</td>
<td>0.411</td>
<td>0.062</td>
<td>18.272</td>
<td>0.417</td>
</tr>
<tr>
<td></td>
<td>(1.242)</td>
<td>(0.704)</td>
<td>(23.060)</td>
<td>(0.498)</td>
</tr>
<tr>
<td>Influence parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deviation after first year ($\delta_1$)</td>
<td>0.164</td>
<td>0.860</td>
<td>0.223</td>
<td>0.608</td>
</tr>
<tr>
<td></td>
<td>(0.238)</td>
<td>(0.809)</td>
<td>(0.140)</td>
<td>(1.855)</td>
</tr>
<tr>
<td>Deviation after second year ($\delta_2$)</td>
<td>-0.844</td>
<td>0.034</td>
<td>-1.438</td>
<td>-0.975</td>
</tr>
<tr>
<td></td>
<td>(1.171)</td>
<td>(1.174)</td>
<td>(2.767)</td>
<td>(1.591)</td>
</tr>
<tr>
<td>External information ($\Pi$)</td>
<td>3.571</td>
<td>4.378</td>
<td>3.082</td>
<td>3.954</td>
</tr>
<tr>
<td></td>
<td>(0.503)</td>
<td>(1.121)</td>
<td>(0.511)</td>
<td>(1.698)</td>
</tr>
<tr>
<td>Likelihood ratio test $p$-value</td>
<td>0.003</td>
<td>0.022</td>
<td>0.000</td>
<td>0.793</td>
</tr>
</tbody>
</table>

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 (6): $\rho_0$ is precision of knowledge prior to training; $\rho_1$, $\rho_2$, and $\rho_3$ are increases in precision (learning) in the first, second, and third years, respectively. Influence parameters $\delta_1$ and $\delta_2$ are deviations from the statically efficient benchmark in terms of effective precision as a function of completed years of training, as given in Equation (7). Specifically, a trainee who has completed one year of training receives influence that is $\delta_1$ 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 $\delta_2$ relative to the efficient benchmark. $\Pi$ is the effective precision of external information, including knowledge from external supervisors or information produced by the trainees. The likelihood ratio test $p$-value compares the estimated model against a restricted model of no learning (i.e., only $\rho_0$, $\delta_1$, and $\Pi$ are non-zero). Standard errors of the parameter estimates, generated by bootstrap, are displayed in parentheses.
Table 4: Model Parameter Estimates by Spending Category

<table>
<thead>
<tr>
<th>Service</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diagnostic</td>
<td>Medication</td>
<td>Transfusion</td>
<td>Nursing</td>
</tr>
<tr>
<td>Knowledge parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior to training ($\rho_0$)</td>
<td>0.652</td>
<td>0.253</td>
<td>1</td>
<td>0.477</td>
</tr>
<tr>
<td></td>
<td>(0.277)</td>
<td>(0.200)</td>
<td>(1.755)</td>
<td></td>
</tr>
<tr>
<td>First year ($\rho_1$)</td>
<td>0.442</td>
<td>0.461</td>
<td>3.798</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.272)</td>
<td>(0.278)</td>
<td>(1.173)</td>
<td></td>
</tr>
<tr>
<td>Second year ($\rho_2$)</td>
<td>0.212</td>
<td>0.309</td>
<td>13.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.172)</td>
<td>(0.236)</td>
<td>(4.654)</td>
<td></td>
</tr>
<tr>
<td>Third year ($\rho_3$)</td>
<td>0.332</td>
<td></td>
<td>2.864</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.471)</td>
<td></td>
<td>(3.099)</td>
<td></td>
</tr>
<tr>
<td>Influence parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deviation after first year ($\delta_1$)</td>
<td>3.559</td>
<td>0.393</td>
<td>0.730</td>
<td>−1.817</td>
</tr>
<tr>
<td></td>
<td>(1.123)</td>
<td>(0.346)</td>
<td>(0.099)</td>
<td>(2.207)</td>
</tr>
<tr>
<td>Deviation after second year ($\delta_2$)</td>
<td>−0.334</td>
<td></td>
<td></td>
<td>−8.315</td>
</tr>
<tr>
<td></td>
<td>(0.563)</td>
<td></td>
<td></td>
<td>(3.182)</td>
</tr>
<tr>
<td>External information (\Pi)</td>
<td>0.281</td>
<td>0.734</td>
<td>3.745</td>
<td>4.746</td>
</tr>
<tr>
<td></td>
<td>(0.498)</td>
<td>(0.229)</td>
<td>(0.190)</td>
<td>(1.356)</td>
</tr>
<tr>
<td>Likelihood ratio test $p$-value</td>
<td>0.009</td>
<td>0.001</td>
<td>N/A</td>
<td>0.022</td>
</tr>
</tbody>
</table>

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 $\rho_0$, $\delta_1$, and $\Pi$ 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 of the parameters, generated by bootstrap, are displayed in parentheses.
Table 5: Model Parameter Estimates by Patient Severity and Day of Stay

<table>
<thead>
<tr>
<th>Service</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early</td>
<td>Late</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

**Knowledge parameters**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to training ($\rho_0$)</td>
<td>0.081</td>
<td>0.027</td>
<td>0.126</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>(0.154)</td>
<td>(0.047)</td>
<td>(0.115)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>First year ($\rho_1$)</td>
<td>0.328</td>
<td>0.202</td>
<td>0.223</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>(0.210)</td>
<td>(0.137)</td>
<td>(0.209)</td>
<td>(0.193)</td>
</tr>
<tr>
<td>Second year ($\rho_2$)</td>
<td>5.919</td>
<td>7.215</td>
<td>6.672</td>
<td>6.605</td>
</tr>
<tr>
<td></td>
<td>(3.221)</td>
<td>(2.600)</td>
<td>(2.498)</td>
<td>(3.382)</td>
</tr>
<tr>
<td>Third year ($\rho_3$)</td>
<td>0.276</td>
<td>2.893</td>
<td>0.038</td>
<td>1.942</td>
</tr>
<tr>
<td></td>
<td>(0.921)</td>
<td>(3.588)</td>
<td>(0.273)</td>
<td>(2.834)</td>
</tr>
</tbody>
</table>

**Influence parameters**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation after first year ($\delta_1$)</td>
<td>0.221</td>
<td>0.174</td>
<td>0.272</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>(0.575)</td>
<td>(0.220)</td>
<td>(0.428)</td>
<td>(0.243)</td>
</tr>
<tr>
<td>Deviation after second year ($\delta_2$)</td>
<td>−0.549</td>
<td>−0.093</td>
<td>−1.243</td>
<td>0.380</td>
</tr>
<tr>
<td></td>
<td>(0.859)</td>
<td>(2.189)</td>
<td>(1.176)</td>
<td>(2.294)</td>
</tr>
<tr>
<td>External information ($\Pi$)</td>
<td>3.647</td>
<td>3.464</td>
<td>3.701</td>
<td>3.549</td>
</tr>
<tr>
<td></td>
<td>(0.995)</td>
<td>(0.517)</td>
<td>(0.850)</td>
<td>(0.508)</td>
</tr>
<tr>
<td>Likelihood ratio test $p$-value</td>
<td>0.151</td>
<td>0.000</td>
<td>0.021</td>
<td>0.182</td>
</tr>
</tbody>
</table>

**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 $\rho_0$, $\delta_1$, and $\Pi$ are non-zero). Standard errors of the parameters, generated by bootstrap, 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:

\[ X_a = T_{t(a)} \eta + \mu_{s(a)} + \zeta_{\tau^T j(a)} + \zeta_{\tau^T k(a)} + \zeta_{\ell(a)} + \varepsilon_a, \tag{A-1} \]

where \( a \) is a patient admission and \( X_a \) is some patient characteristic or linear combination of patient characteristics for the patient in admission \( a \), described in Section 2.3. \( t(a) \) refers to the day of admission, \( s(a) \) is the service of admission, \( j(a) \) is the junior trainee, \( k(a) \) is the senior trainee, and \( \ell(a) \) is the supervising physician. \( T_{t(a)} \) is a set of time categories for the admission day, including the day of the week and the month-year interaction; \( \mu_s \) is a fixed effect that corresponds to the admitting service (e.g., “heart failure service” or “oncology service”). \( \zeta_{\tau^T j} \), \( \zeta_{\tau^T k} \), and \( \zeta_{\ell} \) 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 \( (\zeta_{\tau^T j}, \zeta_{\tau^T k})_{j \in J, k \in K} \).

In Column 1 of Table A-1, I show F-statistics and the corresponding p-values for the null hypothesis that \( (\zeta_{\tau^T j}, \zeta_{\tau^T k})_{j \in J, k \in K} = 0 \). 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.\(^{21}\) 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:

\[ X_a = T_{t(a)} \eta + \mu_{s(a)} + \gamma_1 Z_{j(a)} + \gamma_2 Z_{k(a)} + \zeta_{\ell(a)} + \varepsilon_a. \tag{A-2} \]

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.

\(^{21}\)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 \((\gamma_1, \gamma_2) = 0\). 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:

\[
Y_a = X_a \beta + T_{t(a)} \eta + \zeta_{T, h(a)}^T + \zeta_{T, k(a)}^T + \zeta_{T, \ell(a)}^T + \varepsilon_a,
\]

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 plotting 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 trainees are more likely to be assigned to certain types of 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:

\[
\hat{\sigma}_r \hat{\zeta}_{h(a)} = \gamma_h \hat{\sigma}_r - \hat{\zeta}_{h(a)} + \gamma \hat{\zeta}_{\ell(a)} + \varepsilon_a,
\]

where \(r \equiv 1(\tau > T)\) 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 \(\hat{\zeta}_{r < T}^h\) and \(\hat{\zeta}_{r > T}^h\). 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 \(k\).\(^{22}\) Estimates for \(\gamma_h\) and \(\gamma_\ell\) 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- vs. 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

\(^{22}\)I use two approaches to get around the reflection problem due to the first-stage joint estimation of \(\zeta_{l,j}^0\), \(\zeta_{l,k}^1\), and \(\zeta_{l}\) (Manski, 1993). First, I perform (A-4) using “jack-knife” estimates of fixed effects, in which I exclude observations with \(-h\) and \(\ell\) to compute the \(\hat{\zeta}_h\) estimate that I use with \(\hat{\zeta}_{r - h}\) and \(\hat{\zeta}_{h}\). 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 \((j,k,\ell)\).
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 \( \xi_h^\tau \) in discrete tenure period \( \tau \) and the correlation \( \rho(\tau_1, \tau_2) \) between trainee effects \( \xi_h^{\tau_1} \) and \( \xi_h^{\tau_2} \) in two discrete periods \( \tau_1 \) and \( \tau_2 \). Random assignment of patients to trainee, conditional on time categories, allows me to estimate trainee effects.\(^{23}\) Finite observations per trainee-period means that effects will be estimated with error, which implies that standard deviations of unshrunk effects will overstate the true \( \sigma(\tau) \). Further, correlations of fixed effect estimates of \( \xi_h^{\tau_1} \) and \( \xi_h^{\tau_2} \) will be 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 \( \tilde{Y}_{it} = Y_{it} - \left( X_i \hat{\beta} + T_t \hat{\eta} + \hat{\zeta}_\ell(t) \right) \), where \( \hat{\beta}, \hat{\eta}, \) and \( \hat{\zeta}_\ell \) are estimated by using variation within trainee pairs and discrete tenure periods. This allows random trainee effects to be correlated with \( X_i, T_t, \) and \( \zeta_\ell \).\(^{24}\) Note that \( E[\tilde{Y}_{it}] = 0 \). In practice, given random assignment of attending physicians and patients to trainees, conditional on schedules, I am only concerned with correlations between trainee effects and \( T_t \), but differencing out projections due to \( X_i \) and \( \zeta_\ell \) simplifies computation and avoids the incidental parameters problem in the later maximum-likelihood stage. In the next two subsections I will describe in turn 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 \( \tau \),

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

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

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

---

\(^{23}\)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.

\(^{24}\)An alternative albeit slightly more involved approach involves estimating “correlated random effects,” as described by Chamberlain (1984) and Abowd et al. (2008).
where \( \nu_i \) is a random effect for the patient admission.\(^{25}\) Because trainees are assigned conditionally randomly to each other and to patients, \( \xi_{j(i,t)}^{\tau}, \xi_{k(i,t)}^{\tau}, \) and \( \nu_i \) are uncorrelated with each other. Assuming \( \xi_j^{\tau}, \xi_k^{\tau}, \) and \( \nu_i \) are normally distributed, their standard deviations \( \sigma_{\xi,\tau}, \sigma_{\xi,\tau'}, \) and \( \sigma_{\nu} \) are the parameters of interest in the following maximum-likelihood estimation, done in separate samples selected on \( \tau. \)

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

\[
\tilde{Y} = Zu + \varepsilon, \tag{A-7}
\]

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

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

\[
\text{Var } u = G = \begin{bmatrix}
\sigma_{\xi,\tau}^2 I_{N_{\tau}} & 0 \\
0 & \sigma_{\xi,\tau'}^2 I_{N_{\tau}^T}
\end{bmatrix}.
\]

Similarly, in the case that (A-7) represents (A-6), \( Z \) is an \( n \times (N_{\tau} + N_{\tau}^T + N_i) \) selection matrix for trainees of tenure \( \tau \), teammates, and patient admissions, and \( u \) is an \( (N_{\tau} + N_{\tau}^T + N_i) \times 1 \) 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 \( u \) is

\[
\text{Var } u = G = \begin{bmatrix}
\sigma_{\xi,\tau}^2 I_{N_{\tau}} & 0 & 0 \\
0 & \sigma_{\xi,\tau'}^2 I_{N_{\tau}^T} & 0 \\
0 & 0 & \sigma_{\nu}^2 I_{N_i}
\end{bmatrix}.
\]

Using the definition \( V = ZGZ' + \sigma_{\varepsilon}^2 I_{nt} \), the log likelihood function under either of the above specifications is

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

I estimate (A-5) or (A-6) by maximum likelihood, for each \( \tau \) separately. Holding the tenure of \( h \) fixed at \( \tau \), the tenure of the other teammate will possibly be a mixture if \( \tau \) 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.

\(^{25}\)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 \( \nu_{ai} \) as an intern-admission effect. While it is easier to estimate a specification with \( \nu_{ai} \), I will describe this specification for ease of explication. In practice, results are materially unaffected by whether I use \( \nu_{ai} \) or \( \nu_{ai}, \) 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 $\tau_1$ and $\tau_2$ across which trainee effects may be correlated. Although I observe each trainee across their entire training, I only observe a subset of these trainees in each 60-day or 120-day tenure period, and the number of trainees observed in two different tenure periods is even smaller. Because trainees that I do not observe in both $\tau_1$ and $\tau_2$ do not contribute to the estimate of $\rho(\tau_1, \tau_2)$, I only include in the estimation sample observations associated with a trainee observed in both tenure periods.

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

$$\tilde{Y}_{it} = \xi_{ph(i,t)} + \xi_{-h(i,t)} + \epsilon_{it}, \quad (A-9)$$

which features the function $p(i,t) \in \{\tau_1, \tau_2\}$. This specifies that effects of trainees in the tenure periods of interest ($\tau_1$ and $\tau_2$) may be drawn from two separate distributions depending on the tenure period $\tau_1$ or $\tau_2$ 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

$$\tilde{Y}_{it} = \xi_{ph(i,t)} + \xi_{-h(i,t)} + v_i + \epsilon_{it}. \quad (A-10)$$

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 $Z$ 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 $\tau_1$ and $\tau_2$ tenure periods. The stacked vector of random effects $u$ is similarly of size $(2N_\tau + N' - \tau) \times 1$. The variance-covariance matrix of $u$ is

$$\text{Var} u = G = \begin{bmatrix} G_\tau & 0 \\ 0 & \sigma^2 \xi_{\tau} I_{N' - \tau} \end{bmatrix},$$

where $G_\tau$ is a $2N_\tau \times 2N_\tau$ block-diagonal matrix of the form

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

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

$$\text{Var} \begin{bmatrix} \xi_{\tau_1}^h \\ \xi_{\tau_2}^h \\ \xi_{-h} \\ \epsilon_{it} \end{bmatrix} = A, \text{ for all } h.$$

Representing (A-10) as (A-7) is a similar exercise. The selection matrix $Z$ is of size $n \times (2N_\tau + N' - N_i)$,
and the vector of random effects $\mathbf{u}$ is of size $(2N_{\tau} + N_{i} - 1) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

$$
\text{Var} \, \mathbf{u} = G = \begin{bmatrix}
G_{\tau} & 0 & 0 \\
0 & \sigma_{\zeta,\tau}^{2}I_{N_{\tau}} & 0 \\
0 & 0 & \sigma_{\zeta,\ell}^{2}I_{N_{i}}
\end{bmatrix},
$$

where $G_{\tau}$ is the same as before.

The log likelihood is the same as in Equation (A-8), but using revised definitions of $G$ 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 $\hat{A}$ and is constrained to be between $-1$ and $1$. Standard errors of the correlation estimate are calculated by a likelihood ratio test comparing the likelihood of models fit while holding the correlation fixed but varying all other parameters with the globally optimal fit (i.e., they do not depend on any assumption about the distribution of $\hat{\rho}(\tau_{1}, \tau_{2})$).

### 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:

$$
Y_{it} = \alpha_{m} \text{Characteristic}_{h(i,t)}^{m} + X_{i}\beta + T_{t}\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \epsilon_{aijkt},
$$

where Characteristic$_{h(i,t)}^{m}$ is an indicator for whether the junior (or senior) trainee $h$ has the characteristic $m$, $\zeta_{-h}$ is a fixed effect for the other senior (or junior) trainee $-h$, and $\zeta_{\ell}$ is a fixed effect for attending $\ell$.\textsuperscript{26} The coefficient of interest, $\alpha_{m}$, quantifies the predictive effect a trainee with characteristic $m$ on patient spending decisions. I also evaluate the combined predictive effect of trainee characteristics in two steps. First, I

\textsuperscript{26}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.
regress outcomes on all direct trainee characteristics, with continuous characteristics like position on rank list entered linearly, along with the other admission and time regressors in Equation (A-11):

\[ Y_{it} = \sum_m \alpha_m \text{Characteristic}^{m}_{h(i,t)} + X_i \beta + T_t \eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \epsilon_{it}. \]  

(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:

\[ Y_{it} = \alpha \tilde{Z}_{h(i,t)} + X_i \beta + T_t \eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \epsilon_{it}. \]  

(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 yet other unmeasured and orthogonal characteristics that are more relevant for practice variation, this seems \textit{a priori} unlikely given that these are the characteristics that the residency program makes acceptance decisions based on,\textsuperscript{27} and given 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.

\textbf{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

\textsuperscript{27}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\%. 

A-7
across time periods, regardless of the amount of time between the periods. However, if patients are incorporating new knowledge and evolving in their practice styles, then adjacent time periods should exhibit higher correlation in trainee effects than distant time periods. Appendix A-2.2 describes details of the 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 as there is 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 is 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 a broader practice environment.28

To explore the potential effect of learning from others in greater detail, I estimate supervising physician “effects” by shrinking their observed fixed effects and 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 with 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

28The 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.
spending range from 17.5% to 23.4%.

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

\[ Y_{it} = \sum_{\tau < 1} \alpha_\tau I (\tau (j (i, t), t) = \tau) \cdot \text{HighSpendingExposure}_{j(i,t),t}^m + \sum_{\tau \geq 1} \alpha_\tau I (\tau (k (i, t), t) = \tau) \cdot \text{HighSpendingExposure}_{k(i,t),t}^m + X_i \beta + T_t \eta + \zeta_{\ell(i,t)} + \varepsilon_{it}, \]  

(A-14)

where, as in Equation (1), \( j (i, t) \) is the junior trainee, \( k (i, t) \) is the senior trainee, and \( \tau (j (i, t), t) \) and \( \tau (k (i, t), t) \) are the relevant tenure periods of the junior and senior trainees at \( t \). \( \text{HighSpendingExposure}_{j(i,t),t}^m \) and \( \text{HighSpendingExposure}_{k(i,t),t}^m \) 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 \( \tau \). 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

\[ Y_{it} = \alpha_m \text{Experience}_{h(i,t),t}^m + X_i \beta + T_t \eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}, \]  

(A-15)

where \( \text{Experience}_{h(i,t),t}^m \) 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, separately 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 residents.\(^{29}\)

Overall, these results indicate that summary measures of trainee experience are also poor predictors of

\(^{29}\)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.
practice and outcomes, especially relative to the large variation across trainees. The results fail to support the “learning from others” as a major source of practice variation, at least within an organization with ex ante uniform training experiences but nonetheless large practice variation.

A-4 Practice Variation under Example Learning Parameters

A-4.1 Analytical Evaluation

This appendix makes some 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:

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

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

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

**Proof.** Assume that interns work with second-year residents in $\lambda$ 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 $\tau$, $\rho(0) \leq \rho(1) \leq \rho(2) \leq \rho(3)$, which yields our result. \qed

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

A-10
\( \tau_h = 1 \) (from \( \tau_{-h} = 2 \) to \( \tau_{-h} = 0 \)) than at \( \tau_h = 2 \) (from \( \tau_{-h} = 1 \) to \( \tau_{-h} = 0 \)). I will 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 \( \Delta \) is fixed in a continuous portion of practice variation (i.e., not at the one- or two-year discontinuities). Applying the quotient rule to \( \sigma(\tau_h, \tau_{-h}) = \sigma(\tau_h, \tau_h + \Delta) \),

\[
\sigma'(\tau_h) \equiv \frac{\partial \sigma(\tau_h, \tau_h + \Delta)}{\partial \tau_h} = \frac{\frac{1}{2} \rho(\tau_h)^{-1/2} \rho'(\tau_h) (\rho(\tau_h) + \rho(\tau_{-h}) + P) - \rho(\tau)^{1/2} (\rho'(\tau) + \rho'(\tau_{-h}))}{(\rho(\tau) + \rho(\tau_{-h}) + P)^2}.
\]

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,” as 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 also 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 trainees 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 also involve practice variation only increasing or only decreasing, but not decreasing then increasing with tenure.

**A-4.2 Numerical Examples**

Figure A-10 presents a few numerical examples of variation profiles under different learning profiles described by functions of the piecewise linear form in Equation (6). The three parameters of interest are \( \rho_0 \), or initial knowledge; \( \rho_1 \), 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 external 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 \( \rho_0 \), \( \rho_1 \), \( \rho_2 \), 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 external 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 \( \rho_2 \) by increasing the ratios of \( \rho_0 \) and \( \rho_1 \) to \( \rho_2 \). 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 \( \rho_1 \) relative to \( \rho_0 \), so that more learning occurs in the first year of training as opposed to 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 with working with a 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 \((\rho_1/\rho_2)\) is similar, but learning during training is reduced relative to knowledge gained prior to training \((\rho_0)\) and to the external practice environment \((P)\).
Figure A-1: Patients Age by Housestaff Spending Effect

A: Distribution by Intern Spending Effect

B: Distribution by Resident 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 0.080 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 catalogued 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 panel 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. 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). 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-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.
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 (6), parameterized by $\rho_0$, $\rho_1$, 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.
Table A-1: Tests of Joint Significance of Trainee Identities and Characteristics

<table>
<thead>
<tr>
<th>Patient characteristic</th>
<th>Independent variables</th>
<th>Trainee identities</th>
<th>(1)</th>
<th>Trainee characteristics</th>
<th>(2)</th>
<th>Trainee characteristics</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1055,46364) = 0.98</td>
<td></td>
<td>F(1055,46364) = 1.01</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td>p = 0.655</td>
<td></td>
<td>p = 0.389</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td>F(20,16069) = 0.68</td>
<td></td>
<td>F(20,16069) = 1.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p = 0.848</td>
<td></td>
<td>p = 0.256</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td></td>
<td></td>
<td>F(18,37494) = 0.79</td>
<td></td>
<td>F(18,37494) = 0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p = 0.79</td>
<td></td>
<td>p = 0.558</td>
<td></td>
</tr>
<tr>
<td>Predicted spending</td>
<td></td>
<td></td>
<td></td>
<td>F(18,37494) = 1.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p = 0.685</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>High</td>
<td>Highly</td>
<td>High</td>
<td>Other</td>
<td>Overall</td>
</tr>
<tr>
<td></td>
<td>USMLE</td>
<td>ranked</td>
<td>future</td>
<td>income</td>
<td>hospital</td>
<td>score</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Panel A: Interns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of trainee with</td>
<td>-0.001</td>
<td>0.002</td>
<td>0.010*</td>
<td>0.007*</td>
<td>0.017*</td>
<td>0.003</td>
</tr>
<tr>
<td>characteristic</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.004)</td>
<td>(0.010)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Observations</td>
<td>186,398</td>
<td>185,201</td>
<td>131,247</td>
<td>215,678</td>
<td>219,727</td>
<td>190,331</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.090</td>
<td>0.090</td>
<td>0.091</td>
<td>0.089</td>
<td>0.089</td>
<td>0.090</td>
</tr>
<tr>
<td>Sample characteristic</td>
<td>0.596</td>
<td>0.258</td>
<td>0.234</td>
<td>0.415</td>
<td>0.055</td>
<td>N/A</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Panel B: Residents</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of trainee with</td>
<td>-0.013***</td>
<td>0.010**</td>
<td>-0.004</td>
<td>-0.001</td>
<td>0.013</td>
<td>0.004*</td>
</tr>
<tr>
<td>characteristic</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.011)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Observations</td>
<td>206,455</td>
<td>199,371</td>
<td>129,281</td>
<td>218,376</td>
<td>219,727</td>
<td>206,455</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.095</td>
<td>0.095</td>
<td>0.088</td>
<td>0.088</td>
<td>0.094</td>
<td>0.090</td>
</tr>
<tr>
<td>Sample characteristic</td>
<td>0.564</td>
<td>0.235</td>
<td>0.214</td>
<td>0.332</td>
<td>0.060</td>
<td>N/A</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Tenure period (days)</th>
<th>Differences Between High and Low Exposure</th>
<th>All services</th>
<th>Within service</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All prior</td>
<td>Prior 3 months</td>
</tr>
<tr>
<td>0-60</td>
<td>5.73%</td>
<td>6.11%</td>
<td>4.90%</td>
</tr>
<tr>
<td>61-120</td>
<td>5.54%</td>
<td>5.94%</td>
<td>4.98%</td>
</tr>
<tr>
<td>121-180</td>
<td>4.95%</td>
<td>5.81%</td>
<td>4.53%</td>
</tr>
<tr>
<td>181-240</td>
<td>4.82%</td>
<td>5.85%</td>
<td>3.81%</td>
</tr>
<tr>
<td>241-300</td>
<td>4.34%</td>
<td>5.61%</td>
<td>3.87%</td>
</tr>
<tr>
<td>301-365</td>
<td>4.05%</td>
<td>5.33%</td>
<td>3.47%</td>
</tr>
<tr>
<td>366-425</td>
<td>3.90%</td>
<td>6.34%</td>
<td>4.15%</td>
</tr>
<tr>
<td>426-485</td>
<td>4.05%</td>
<td>6.66%</td>
<td>4.27%</td>
</tr>
<tr>
<td>486-545</td>
<td>3.59%</td>
<td>6.21%</td>
<td>3.65%</td>
</tr>
<tr>
<td>546-605</td>
<td>3.43%</td>
<td>5.69%</td>
<td>3.89%</td>
</tr>
<tr>
<td>606-665</td>
<td>3.65%</td>
<td>6.60%</td>
<td>4.33%</td>
</tr>
<tr>
<td>666-730</td>
<td>3.76%</td>
<td>6.43%</td>
<td>3.75%</td>
</tr>
<tr>
<td>731-850</td>
<td>3.81%</td>
<td>5.35%</td>
<td>2.41%</td>
</tr>
<tr>
<td>851-970</td>
<td>3.85%</td>
<td>6.33%</td>
<td>2.68%</td>
</tr>
<tr>
<td>971-1095</td>
<td>3.35%</td>
<td>4.23%</td>
<td>1.94%</td>
</tr>
</tbody>
</table>

Panel A: Exposure to Spending by Supervising Physicians

Panel B: Exposure to Spending by Senior Trainees

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

<table>
<thead>
<tr>
<th></th>
<th>Log daily total costs</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Number of days</td>
<td>(2) Number of patients</td>
<td>(3) Number of attendings</td>
<td>(4) Attending spending</td>
<td>(5) Attending spending</td>
</tr>
<tr>
<td><strong>Panel A: Interns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of trainee with measure above median</td>
<td>0.001</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.006</td>
<td>0.001</td>
</tr>
<tr>
<td>Observations</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.088</td>
<td>0.088</td>
<td>0.088</td>
<td>0.089</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Panel B: Residents</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of trainee with measure above median</td>
<td>0.003</td>
<td>0.003</td>
<td>-0.005</td>
<td>0.008</td>
<td>0.013**</td>
</tr>
<tr>
<td>Observations</td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.090</td>
<td>0.090</td>
<td>0.090</td>
<td>0.087</td>
<td>0.087</td>
</tr>
</tbody>
</table>

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

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

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.”