Practice Variation in Team Decisions: Evidence from Physicians in Training*

David C. Chan†

October 4, 2017

Abstract

What causes professionals in organizations to practice differently in similar situations? This paper studies practice variation in team decisions, among physicians trainees, who frequently switch teams and are randomly assigned patients. I identify the effect of relative experience on influence in team decisions by a discontinuity in relative experience across the one-year tenure mark. The influence of senior trainees is larger for decisions requiring more discretion. Trainee effects fail to converge with training but show serial correlation consistent with learning and are unexplained by trainee characteristics or training histories. These findings suggest variation due to idiosyncratic learning and tacit knowledge.

JEL Codes: D83, L23, M53

*I am grateful to David Cutler, Joe Doyle, Bob Gibbons, and Jon Gruber for their guidance on this project from an early stage. 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, Jon Skinner, Doug Staiger, Chris Stanton, Chris Walters, and seminar audiences at Arizona, ASHEcon, Carnegie Mellon, Case Western Reserve University, Chicago Booth, Cornell Weill, Duke, Johns Hopkins, Maryland, MIT, NBER (Organizational Economics), North Carolina, Paris School of Economics, Queen's University, Rice, Stanford, Tulane, and WEAI 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.

1 Introduction

What causes professionals in organizations to practice differently in similar situations? In many industries, substantial variation in productivity across firms seems to be tied to variation in practices linked to key individuals inside the firms (Bertrand and Schoar, 2003; Bloom and Van Reenen, 2010; Syverson, 2011; Gibbons and Henderson, 2012). Within health care organizations, physicians treating similar patients appear to drive substantial variation in medical practice.[1] However, our understanding of the roots of practice variation in team settings is limited.

In this paper, I study practice variation in team decisions by applying a simple team-theoretic framework to the empirical setting of new physicians in residency training. A key insight of the team-theoretic literature is that knowledge is tacit, or embodied in individuals and not easily transferred, so that organizations must aggregate knowledge by distributing decision-making across agents (e.g., Cyert and March, 1963; Nelson and Winter, 1982; Garicano, 2000). I modify the canonical framework to allow for knowledge that is not only tacit but also *partial*: Rather than knowing how to solve a problem completely or not at all, agents may have some idea of the correct decision based on prior, related experience. This feature is realistic if the space of potential problems is large relative to any agent's experience, yet knowledge may be extrapolated to related problems that have not yet been encountered. In such a case, practice variation is a natural consequence.

The setting of residency training is well-suited to study this framework in health care for two substantive reasons. First, residency training is an intensive period of training designed to impart knowledge to physicians beyond facts, "developing habits, behaviors, attitudes, and values that will last a professional lifetime" (Ludmerer, 2014). Second, decision-making is explicitly and mechanically made in teams, which makes tractable the key stylized fact of team theory that has nevertheless received little empirical attention.

Specifically, I follow a diverse group of 802 physician trainees in the internal medicine residency program of a large hospital, 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,

[1] See Skinner (2012) for a review of the literature in health care, which dates at least to Wennberg and Gittelsohn (1973) in the US and Glover (1938) in the UK. This literature, focusing mostly on variation across regions, has received significant policy attention. For example, President Barack Obama paid special attention to this fact during US health care reform leading to the Affordable Care Act (e.g., Pear, 2009). Some policy analysts have conjectured that the US may save about a third of its health care spending, or at least 3% of the US GDP, by eliminating "unwarranted variation" (Kelly, 2009).

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 effect on team decisions of physician trainees at various points in their tenure. I define *practice variation* as the variance of these effects across trainees in a given tenure period. I then characterize how practice variation changes as a function of trainee tenure.

To understand this basic design, it is important to note that causal provider effects, and thus practice variation, reflect two objects of interest in team decisions, both 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). In this setting, I exploit a discontinuity in the relative experience of trainees at the end of their first year of training to separately identify the effect of influence on practice variation: Trainees are junior on the team before one year, whereas they are senior after one year. Under the assumption that trainee characteristics (e.g., knowledge, preferences, ability) are continuous across this mark, a discontinuous increase in the variation in trainee effects implies the contribution of influence by virtue of relative experience to practice variation. On the other hand, under sufficient learning, trainees should converge to a common judgment, so that practice variation may subsequently *decrease* in the two years remaining until the end of training.

The first finding of this study is that practice variation discontinuously increases across the one-year mark of training. Junior trainees ending their first year of training exhibit 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%. The discontinuity at the one-year mark of training, more than quadrupling practice variation, shows that the experience of a trainee relative to her teammate has a large effect on her influence in team decisions. Through the rest of residency, the variance of senior trainee effects does not generally decrease with greater tenure. That is, trainee effects do not converge as trainees gain more experience. This lack of convergence provides suggestive evidence for frictions in the transfer of knowledge, even in this intensive and mostly uniform period of residency training.

To further support the idea that influence due to relative experience is mediated by knowledge, I examine how practice variation across the one-year training discontinuity varies across a wide range

of decisions and settings. If knowledge plays a role in influence, then decisions relying on knowledge that accumulates with experience should show a larger influence effect, or a larger increase in practice variation, at the experience discontinuity. I find that practice variation discontinuously increases across the whole range, but that the increase is larger in practice domains with fewer clear rules and more discretion (e.g., laboratory, radiology, and blood transfusion, as opposed to medications). Practice variation profiles as a function of tenure are similar across patient groups with different expected mortality or expected total costs. However, the size of the senior-junior trainee influence gap is larger during the first half of each patient's hospital stay, when trainees know less about specific patients and must rely on knowledge generalized from prior experience.

Next, I compare how practice variation evolves with trainee tenure in each of the inpatient services of general medicine, cardiology, and oncology hosting the same group of trainees. The services are highly relevant for learning, since supervising physicians and the knowledge they use in patient care are distinct across services. Thus, I focus on whether practice variation narrows in the two years of training after the discontinuity, which would indicate that trainees converge to a common "best practice," because they learn to follow protocols or adopt well-accepted heuristics within a field of medicine. I find substantial convergence in diagnostic testing costs in the specialist-driven services of cardiology and oncology but not in general medicine; there is also convergence in overall costs in the cardiology service. In contrast to services, formal diagnoses have no bearing on how practice variation changes with tenure. This last finding is consistent with tacit knowledge in the sense that formal diagnoses are insufficient as labels for problems.

To assess learning more directly and to rule out intrinsic heterogeneity (e.g., different preferences or skills) as a key alternative source of persistent variation, I perform two additional sets of analyses. First, I estimate the correlation between effects of the same trainee in different periods of time. If practice variation is solely driven by unchanging heterogeneity, then there should be a high degree of within-trainee correlation between effects across all time periods, even if influence were to change the scale of practice variation at the one-year mark. Instead, I find that trainee effects are highly correlated in adjacent two-month periods, but are very weakly correlated between more distant periods. Second, 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) to examine whether any of these characteristics are correlated with effects on spending. When using a penalized

regression to guard against overfitting the data, I find that the sole predictive trainee characteristic is gender, which predicts a small fraction of the overall practice variation. Trainee effects, both on average and in distribution, are indistinguishable between high and low "quality" trainees, as defined by their position on the rank list and test scores. Likewise, trainees with above-median future incomes practice no differently than their peers with below-median future incomes that are on average about $160,000 or 40% lower.

Finally, I assess the extent to which trainee practices can be predicted from different training exposures within residency. Trainees who are randomly assigned to high- vs. low-spending supervising physicians are unaffected in their later spending. Comparing the main cohort of trainees with a group of trainees visiting from a residency program that is based in a hospital with 20% lower Medicare spending per inpatient, I also find no evidence that training in the other hospital – combined with differential selection of trainees into the two different residency programs – substantially shifts spending relative to the size of practice variation. This last set of results argues against practices as "schools of thought" that are predictably acquired (Phelps and Mooney, 1993). Instead, it is consistent with tacit knowledge as a foundation for practice variation, in that under tacit knowledge, practices (like knowledge) should also be difficult to transfer.

These findings contribute to several strands of literature. First, as noted above, these results link to a general team-theoretic literature based on the disaggregated nature of individual knowledge and its aggregation via organizations (e.g., Cyert and March, 1963; 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."

The "dispersion" of knowledge is not always due to some principal-agent problem. As Polanyi (1966) observed, much of productive knowledge, like skills, cannot easily be transferred:

> "We know more than we can tell [p. 4] ... The skill of a driver cannot be replaced by a thorough schooling in the theory of the motorcar [p. 20]."

This study provides a detailed empirical look at the acquisition and use of knowledge in an organization. Supporting the team-theoretic thesis, I find that decisions are driven by more experienced

agents, presumably with better knowledge. However, I also find that the influence of more experienced agents is pervasive across the whole range of decisions, which contradicts the "management by exception" prediction by Garicano (2000) that managers should attend to relatively rare problems. In the canonical team-theoretic framework, agents are assumed to learn to solve a specific problem fully or not at all, whereas this empirical result instead suggests partial knowledge, in which judgments even by the most knowledgeable agents will involve extrapolation and uncertainty. This characterization of knowledge seems naturally related to tacit knowledge, especially when direct experience is limited relative to the space of problems that could be encountered. When knowledge has this feature, as it does in health care, agents with more experience should influence a broad range of decisions, and hierarchies based on experience should be the rule.

Second, these results relate to a large literature documenting practice variation in health care.[2] Despite the size of this literature, little is empirically known about the mechanisms behind variation on the provider side. 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 wide practice variation among physician trainees in an environment *designed* to create homogeneity and, through multiple lens, suggests tacit and partial knowledge as a main foundation of practice variation.

Third, the empirical strategy of this paper is related to a series of papers beginning with Abowd et al. (1999) that decompose joint outcomes into effects from units (e.g., workers and firms, patients and hospital regions) that are observed in different pairings (Card et al., 2013; Finkelstein et al., 2016). Most closely related is work by Lazear et al. (2015), which decomposes outcomes into effects from workers and bosses. In this paper, however, I focus on the effect of becoming the senior member of a team in order to identify influence in the use of knowledge in team decisions. I make use of mechanical transitions to the senior role and a rich environment that allows estimation of trainee

[2]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.

effects separately within relatively short periods of tenure and for different types of decisions. I show that relative experience dramatically increases the size of physician effect variation, particularly for decisions that require discretion and general knowledge.

The remaining organization of this paper is as follows. Section 2 describes the institutional setting; Section 3 describes the data. Section 4 introduces a conceptual framework to understand team decisions, knowledge, and practice variation. Section 5 presents results on the evolution of practice variation as a function of trainee tenure. Section 6 examines how results differ across different types of decisions and settings. Section 7 examines evidence related to learning, intrinsic heterogeneity, and the impact of prior training on trainee practice. Section 8 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 gave rise to an increasing realization that physicians need to be intensively trained in the profession following graduation from medical school. The period between the two world wars witnessed even more rapid change, with large improvements in surgical technique; new fields such as radiology, pediatrics, and psychiatry; and advances in anesthesiology, blood transfusion, and intravenous fluid replacement. A review in 1943 listed 35 important classes of therapeutic agents, whereas only six such agents existed 30 years prior (Peter Bent Brigham Hospital, 1943, p. 26).

Against this backdrop, the first medical residency programs were founded in the 1890s, based on earlier European innovations, and came to dominate the training of American physicians in the period between the two world wars. By the late 1930s, specialty boards in the major fields of medicine had appeared, and the American Board of Medical Specialties adopted the requirement that, after 1942, physicians should have residency training of at least three years in duration to be certified as a specialist. "General practitioners" with one year of training in a "rotating internship" dwindled in numbers, and by 1975, the last general practitioner training program ceased to exist (American Medical Association, 1966).

Since the rise of medical residency as an institution, the basic principles of residency training

have remained relatively consistent despite continued changes in medical technology.[3] 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 scientific, pathophysiologic 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 Organization of Residency

Each patient is cared for by a team of trainees ("housestaff") comprised of a first-year junior trainee ("intern") and a second- or third-year senior trainee ("resident"). Residents are usually assigned to two interns at a time and therefore are responsible for twice the number of patients. Interns can thus devote more attention to each patient and usually are the first to examine patients and make judgments. Because residents have more experience, they are expected to know more and often engage in higher-level decision-making in patient care. There are however no other formal distinctions in decision rights or job responsibilities between interns and residents, including legal or regulatory ones. These 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.

In internal medicine, trainees from different programs and different "tracks" within a program often work together on the same clinical services. For example, a sizeable 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.[4] These plans are committed to prior to starting the

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

[4] In addition, tracks within a residency program include primary care, "short tracks" to fellowship training, research tracks such as genetics, and medicine-pediatrics or medicine-psychiatry combined programs.

internal medicine residency. In addition, medical trainees from another hospital and obstetrics-gynecology and emergency medicine trainees from the same hospital work alongside trainees from the main residency program in the same teams and on the same services.

Trainee schedules are arranged a year in advance to satisfy hospital programmatic requirements and broader regulations. Rotations include intensive care unit (ICU), outpatient, research, subspecialty (mostly outpatient) electives, and ward blocks. 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 Medical Knowledge

Paradoxically, the expansion of medical knowledge and the accompanying growth of new treatments, diagnostic tests, and modes of care can imply that less is known by a finite set of providers about the ideal care for any given patient. First, the more that is known globally, the smaller the proportion of knowledge that can be mastered by an individual provider. Second, the more treatment and diagnostic options are developed, the more choices there are for patient care. As choices multiply, the concept of "personalized medicine," which imagines ideal care that differs across individual patients, at once becomes more relevant and more difficult.

The gold standard of medical research – the randomized controlled trial – collects evidence about the effect of a discrete treatment compared to a finite set of alternative (often placebo) treatments on a set of patients with prespecified exclusion and inclusion criteria. As the number of treatment

and diagnostic combinations expands, and as the partition of relevant patient sets becomes finer, it becomes increasingly difficult to make inferences about care for a single patient, holding the set of experimental observations fixed. Recommendations for patient care require integration of evidence from multiple trials, accumulated patient care, and theories of pathophysiology. To date, this integration is an endeavor of expert human judgment.

For these reasons, measuring the formal informational content behind the space of medical decisions is intrinsically difficult. However, several features of medical practice suggest that it is low. First, the proportion of guidelines based directly on trial evidence rather than "expert opinion" has remained small (Shaneyfelt et al., 1999). Of supposedly evidence-based guidelines, several high-profile ones have been reversed (Rossouw et al., 2002; Jauhar, 2014). Second, despite a cottage industry devoted to developing *ad hoc* metrics of the quality of care, exceedingly few are agreed upon, much less accepted by organizations as goals for improvement (Kizer and Kirsh, 2012). Third, the proportion of decisions made without explicit guidelines, such as off-label prescribing, is high, and there are few policy efforts to restrain this proportion on quality grounds (Radley et al., 2006).

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.[5]

Detailed residency application information for each trainee includes demographics, medical school, USMLE test scores, membership in the Alpha Omega Alpha (AOA) medical honors society, other degrees, and position on the residency rank list. USMLE test scores represent a standardized measure of resident knowledge and ability. Position on the residency rank list represents desirability

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

to the residency program, according to both criteria that I observe and those assessed during the interview and potential recruitment process. Finally, I 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, psychiatry, or radiology after a preliminary intern year.

In addition to trainee characteristics predetermined relative to residency, I make use of observed specialty and subspecialty training after internal medicine residency to impute expected yearly future income in the five years immediately after this training. I use industry-standard survey data from the Medical Group Management Association (MGMA) to obtain average incomes at the specialty and subspecialty level. Expected future incomes range from $199,000 to $850,000. 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, reflecting that career paths for preliminary interns (e.g., future anesthesiologists, dermatologists, and radiologists) are usually more lucrative.

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. Clnical 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 attending physicians, since attending 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 that recent studies have used (Doyle et al., 2015; Skinner and Staiger, 2015; Chandra et al., 2016), and therefore provide new insight into welfare-relevant resource use.[6] Laboratory costs are based on 3.1 million physician laboratory orders; radiology costs are 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.

4 Conceptual Framework of Team Decisions

In this section, I sketch a simple model of decision-making in teams. The purpose of this model is to highlight the roles of knowledge and influence in the team decisions and to map these concepts to econometrically observable objects. As in the team-theoretic literature (e.g., Cyert and March, 1963; Radner, 1993; Garicano, 2000), I consider the organizational problem of using information dispersed across agents to make decisions.[7] In residency, a trainee *team* is composed of two junior trainees assigned to one senior trainee. The team is responsible for a set of patients, each of which is assigned to a single intern and the resident. Patient care decisions are made as follows:

1. Decisions to be made, indexed by d, arise exogenously and are linked to patients. Denote sets of decisions to be made as $\mathcal{D}_j$ and $\mathcal{D}_{j'}$, for interns j and j' respectively, and $\mathcal{D}_k = \mathcal{D}_j \cup \mathcal{D}_{j'}$ for resident k, who works with j and j'. A decision is defined by *action* a_d on the real line that ideally should match an unknown state θ_d.

2. Each agent (intern or resident) allocates bounded time and effort (or simply *effort*) within

[6] 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.

[7] Models in this literature abstract away from any moral hazard problem. A more complicated model allowing for heterogeneous preferences or conflicts of interest would have similar qualitative results as long as long as there is a common component to the decision that would be agreed upon by both agents under perfect information but must be made under incomplete information.

their set of decisions: $\sum_{d\in\mathcal{D}_h} e_{d,h} \leq 1$ for all $h \in \{j, j', k\}$. $e_{d,h}$ includes time and effort spent interviewing the patient, performing a physical exam, communicating to other agents, researching, and thinking done by h about the decision d.

3. Each agent forms a *judgment* about θ_d, which is a normal Bayesian prior distribution, summarized by mean $\mu_{d,h}$ and precision $\rho_{d,h}$. These judgment moments are a function of $e_{d,h}$ and *knowledge* ω_h. The judgment moments and effort are publicly observable, but knowledge is complex, multidimensional, and therefore tacit. There may also be external or public judgment about d, with mean 0 and precision P_d, for example due to attending physician beliefs, institutional rules, or other publicly accessible information.

4. Actions are taken. Each agent derives utility $u_h = -\sum_{d\in\mathcal{D}_h} (\theta_d - a_d)^2$.

4.1 Influence in Team Decisions

As is standard in team-theoretic environments, there is no misalignment of incentives nor disutility of supplying effort, so trainees supply effort inelastically, i.e., $\sum_{d\in\mathcal{D}_h} e_d^h = 1$. Conditional on judgments, the optimal action for decision d assigned to trainees j and k is

$$a_d^* = \frac{\rho_{d,i}\mu_{d,i} + \rho_{d,j}\mu_{d,j}}{\rho_{d,i} + \rho_{d,j} + P_d}. \tag{1}$$

This expression is a weighted average of the means of the judgments of the junior and senior trainee, weighted by the precisions of their respective judgments on d. Any public knowledge with mean 0 and precision P_d reduces the effect of either of the trainee's judgments on a_d^*.

The expression, $\rho_{d,h}/\left(\rho_{d,h} + \rho_{d,-h} + P_d\right)$, has a natural interpretation as the *influence* weight on a trainee h's mean judgment relative to that of her teammate $-h$ assigned to the same decision. The more precise her signal is relative to her teammate and the external practice environment, the greater her influence will be. With respect to the practice environment, a trainee's influence will be lower in a tighter practice environment with higher P_d. At the extreme, if external knowledge were perfect (i.e., if $P_d = \infty$), there should be no variation attributable to trainees. Similarly, if either trainee were to reach a perfect judgment with bounded effort (i.e., $\rho_{d,h} = \infty$), then the other trainee's judgment would carry no weight.

This decision-making structure is a generalization of the canonical setup of Garicano (2000) and earlier papers but allows for starkly different empirical results. In the canonical team-theoretic models, agents either know fully how to "solve a problem" or not at all (i.e., $\rho_{d,h} \in \{0, \infty\}$ for all d and h), and the organizational design problem becomes one of allocating future problems to agents to reduce the likelihood that they (perfectly) learn about problems they will only encounter rarely or waste effort on problems they do not know how to solve. Garicano (2000) thus predicts "management by exception," in which managers who have larger spans of control focus on problems that are "difficult" only because they are too rare to justify knowing how to solve by lower-tier workers. In contrast, if knowledge to solve problems is never perfect, and if instead expertise applies to a broad range of problems, then managers with more expertise may have a pervasive effect on team decisions despite larger spans of control.

In either case, hierarchy, influence, or "prestige" related to the effect on team decision-making is *endogenous* to knowledge.[8] This insight is similar to the one made in the seminal paper by Alchian and Demsetz (1972): Organizations exist because of a team process that induces efficiencies in information and production, and any notion of an intrinsically "superior authoritarian directive or disciplinary power" held by management is merely illusory. In residency, there is nothing that distinguishes junior from senior trainees, other than tenure and span of control. Senior trainees are mechanically drawn from the same pool of junior trainees. There is no selection into "management" based on quality nor any formal assignment of decision rights to senior trainees over junior trainees.[9]

4.2 Learning

Partial knowledge about how to make decisions has important implications for the process of learning. If the set of decisions were small, then with enough experience, agents should have increasingly perfect knowledge about how to make them. On the other hand, as the space of decisions becomes increasingly fine, less will be known about making any particular new decision based on the history

[8]The model here can be easily extended to a case in which the precision of judgments is not easily observed, but in which some expectation of it would be known and correlated with tenure.

[9]Similarly, senior trainees are not primarily responsible for the evaluation of junior trainees. Attending physicians ultimately have the most say in evaluation, which argues against career concerns in reputational cheap talk as a mechanism for greater influence of senior trainees (Scharfstein and Stein, 1990; Prendergast, 1993; Ottaviani and Sorensen, 2001). Another dynamic we do not capture here is "supervised learning," in which senior agents may allow junior agents to make their own decisions, even if they are wrong, so that they may learn faster (Lizzeri and Siniscalchi, 2008). This would imply that junior trainees should have *more* influence, all else equal.

of prior decisions and outcomes. Paradoxically, the subjective complexity of the world is driven by how much we know about it. For example, as more is known about how patients may differ from one another, and how drugs in seemingly similar classes have different and at times unexpected effects for different patients, the categorization of decisions becomes finer, and the application of prior knowledge becomes an act of extrapolation.

This property of knowledge is closely tied to ideas of human learning and tacit knowledge. In settings where knowledge can be prespecified as a set of rules, work can be automated, and human judgment would be marginalized (e.g., Autor et al., 2003). An intrinsic feature of human judgment is that it applies knowledge to problems not yet encountered, but this feature also makes knowledge tacit. The approach to training physicians in residency, described in Section 2, also suggests an attempt to instill tacit knowledge in the face of increasing medical complexity. Despite the proliferating space of decisions, residents are encouraged to spend more time on fewer patients and to contemplate how patient observations connect with underlying theories of pathophysiology. Departing from the standard team-theoretic framework, agents cannot *ex ante* be designed to learn how to solve any specific type of problem. Since experience and contemplation are both limited and idiosyncratic, training can only stimulate partial knowledge in any problem with probability 1. In this sense, training is general.

4.3 Empirical Moments

Consider the observed action a_d^* for decision d, assigned to trainees j and k, in Equation (1). The action can be decomposed into two parts:

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

with the first component due to j and the second due to k.

Decisions can be aggregated into empirical moments of practice variation across trainees, depending on the trainee and tenure. Enriching notation, consider the set, $\mathcal{D}_h^{\tau;c}$, of decisions to be made by trainee h of tenure τ, in a category c of decisions (e.g., treatment or diagnosis; for complex patients or for simple patients). The expected action for a decision in category c made by intern j

and resident k at time t can then be decomposed into components at the trainee level:

$$\begin{aligned} Y^*_{cjkt} &\equiv E\left[a^*_d \middle| d \in \mathcal{D}_j^{\tau(t,j);c} \cap \mathcal{D}_k^{\tau(t,j);c}\right] \\ &= E\left[a^*_{d,j} \middle| d \in \mathcal{D}_j^{\tau(t,j);c}\right] + E\left[a^*_{d,k} \middle| d \in \mathcal{D}_k^{\tau(t,k);c}\right], \end{aligned}$$

where tenure is written as a function of time and the trainee. From Equation (2), if $\rho_{d,-h}$ is (roughly) constant within $\mathcal{D}_i^{\tau;c}$, then the trainee effects are (roughly) additively separable. Each component expectation can be considered as a "trainee effect" or "practice style" and denoted as $\xi_h^{\tau;c} = E\left[a^*_{d,h} \middle| d \in \mathcal{D}_h^{\tau;c}\right]$. Trainee effects are further empirically identified by the fact that trainees change teams (Abowd et al., 1999).

Trainee effects only capture the persistent component of a trainee's judgment over decisions. To represent this, trainee h's judgment on d can be decomposed into two component normal distributions – a distribution that varies across decisions, perhaps formed by independent observation, $\mathcal{N}\left(\mu^o_{d,h}, 1/\rho^o_{d,h}\right)$, and a persistent distribution due to knowledge, $\mathcal{N}\left(\mu^\omega_{d,h}, 1/\rho^\omega_{d,h}\right)$ – such that $\mu_{d,h} = \left(\rho^o_{d,h}\mu^o_{d,h} + \rho^\omega_{d,h}\mu^\omega_{d,h}\right) / \left(\rho^o_{d,h} + \rho^\omega_{d,h}\right)$ and $\rho_{d,h} = \rho^o_{d,h} + \rho^\omega_{d,h}$. By construction, $E\left[\mu^o_{d,h} \middle| d \in \mathcal{D}_h^{\tau;c}\right] = 0$. Such independent observation by either trainee reduces practice variation from both trainees in the same way that external knowledge does, and the trainee effect is

$$\xi_h^{\tau;c} = E\left[\frac{\rho^\omega_{d,h}\mu^\omega_{d,h}}{\rho^\omega_{d,h} + \rho^\omega_{d,-h} + \rho^o_{d,h} + \rho^o_{d,-h} + P_d} \middle| d \in \mathcal{D}_h^{\tau;c}\right]. \tag{3}$$

In themselves, the trainee effects cannot separate knowledge from influence. However, in the empirical setting of residency, I can exploit the discontinuous increase in relative tenure at one-year when interns become residents. Since teams are always comprised of an intern and a resident, when a trainee's tenure passes the one-year mark, she will be assigned to a teammate who has one year less experience than her, while she previously worked with a teammate who had at least one year more experience. Comparing practice variation of h across this discontinuity in $\rho^\omega_{d,-h}$ identifies the effect of relative tenure on influence and also provides a measure of learning in terms of the effect of the change in $\rho^\omega_{d,-h}$ over training.[10] This discontinuity can be assessed for different categories of

[10]Strictly speaking, this requires an assumption that information from independent observation, or $\rho^o_{d,h} + \rho^o_{d,-h}$, remains continuous across this discontinuity. 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.

decisions. A discontinuous increase in influence across a broad range of decision categories suggests a pervasive effect of expertise.

Two other moments characterize learning among trainees. First, increasing serial correlation in effects within trainee over time measures the extent to which learning accumulates knowledge and stabilizes trainee effects. Second, convergence in practice variation, or decreasing variance of the distribution of trainee effects with tenure, measures the extent to which trainees reach agreement in their judgments. As trainees continue to learn, their mean judgments should converge, and the precision of their judgements should also increase. When a single trainee is fully responsible for making decisions, any learning of a common truth necessarily implies convergence in practice styles. In teams, concurrently increasing influence can obscure convergence under some learning, but weak or no convergence still suggests that knowledge remains partial and that learning is slow, since no variation should remain under complete knowledge. Appendix A-2 provides numerical examples of variation that does not converge with partial learning and team decisions. A lack of convergence combined with increasing serial correlation suggests that learning has slowed near the end of training. Theories in which learning is costly due to some form of informational frictions will have this feature.[11]

5 The Tenure Profile of Practice Variation

As a first analysis, this section examines the effects of trainees on team decisions as a function of each trainee's tenure. Random assignment of patients to trainees and frequent reshuffling of trainees to teams allows me to separate the effects of the junior and senior trainees on team decisions. In order to summarize a large number of team decisions, recorded as orders that span a rich space, I aggregate the direct costs associated with all of the team decisions made for each patient-day. Direct costs are provided by the hospital's accounting system for resource use due to physician orders (i.e., laboratory, radiology, pharmacy, blood transfusion) and nursing utilization.

[11] When knowledge to be learned is vast and experience (or time for contemplation) is limited, then learning will be incomplete. This intuition is also 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.

For each trainee, I allow for distinct effects in each training tenure period. Measuring how variation in trainee practice styles (i.e., the standard deviation of trainee effects) changes as the trainees gain in tenure is informative for two reasons. Primarily, since trainee teams are comprised of a first-year junior trainee and a second- or third-year senior trainee, comparing trainee effect variation before and after the one-year tenure mark reveals the component of variation due to the influence that trainees gain by being the senior member of the team. Identification of this influence effect relies on the plausible assumption that other trainee characteristics, including knowledge, are continuous across the one-year mark. Second, tracking practice variation due to senior trainees suggests whether trainees converge in their effects as they gain more experience, up until three years of training, when they are eligible to become fully licensed physicians.

Specifically, I model log total costs at the patient-day level as

$$Y_{it} = \mathbf{X}_i\beta + \mathbf{T}_t\eta + \xi_{j(i,t)}^{\tau(j(i,t),t)} + \xi_{k(i,t)}^{\tau(k(i,t),t)} + \zeta_{\ell(i,t)} + \nu_i + \varepsilon_{it}, \tag{4}$$

where i indexes the patient, and t indexes the day. $j\left(i,t\right)$, $k\left(i,t\right)$, and $\ell\left(i,t\right)$ refer to the junior trainee, senior trainee, and attending (supervising) physician, respectively, assigned to the patient i on day t. Equation (4) controls for patient and admission characteristics $\mathbf{X}_i$, and a set of time categories $\mathbf{T}_t$ for month-year combination, day of the week, and day of service relative to the admission day. Trainee effects – $\xi_{j(\cdot)}^{\tau(\cdot)}$ and $\xi_{k(\cdot)}^{\tau(\cdot)}$ for the junior and senior trainees, respectively – depend on the identity of the trainee and the tenure period $\tau\left(h,t\right)$ that day t falls in for housestaff $h \in \{j\left(i,t\right), k\left(i,t\right)\}$. Since patients are *not* randomly assigned to attending physicians, I model attending "effects" as fixed and treat these as nuissance parameters capturing both true effects and unobserved patient selection to attending physicians.[12] In some specifications, I also allow for daily costs to be correlated within patient by a patient effect ν_i.

The objects of interest in Equation (4) are measures of variation in trainee effects within tenure period. In this analysis, I impose trainee effects that are invariant within a discrete tenure period but otherwise assume no structure in the effects across tenure periods. 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

[12]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).

consider trainee effects as random in order to obtain unbiased estimates of the standard deviation of the tenure-specific distribution of trainee effects.[13]

In Appendix A-3, I detail a method akin to restricted maximum likelihood (REML) and similar to an approach by Chetty et al. (2014) that allows for a large number of fixed covariates, potentially correlated with the random effects, outside of the maximum likelihood estimation. Tenure-specific standard deviations of $\xi^{\tau}_{h\in\{j,k\}}$ are directly estimated by maximum likelihood. Patient effects, ν_i, if modeled, are also considered random and accounted for in 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; the estimated standard deviation of the distributions 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 (4) separately within bins of observations according to $\tau\left(j\left(i,t\right),t\right)$ or $\tau\left(k\left(i,t\right),t\right)$ and therefore impose no assumption on the structure of correlation between effects of the same trainee in different periods.

Figure 1 presents results for the estimated standard deviations of the trainee effect distributions within each tenure interval τ. In my baseline specification, I consider non-overlapping tenure intervals that are 60 days 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.[14] 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, the distribution of practice effects does not show any appreciable convergence: The standard deviation of the practice effect distribution remains above 20% throughout. By comparison, the standard deviation for patient effects, ν_i, ranges from 9% to 13%; including or omitting admission-level random effects does not significantly alter results.

Under the lens of the conceptual framework in Section 4, these findings suggest two contrasting features of learning. First, in a model in which knowledge drives influence, the fact that senior

[13]This approach is closely related to Bayesian shrinkage of fixed effects. However, standard Bayesian shrinkage procedures of single fixed effects (e.g., Morris, 1983) are not applicable here, because the two sets of effects – one for the junior and the other for the senior trainee – imply fixed effects of senior (junior) trainees that are inconsistent when shrinking the effects of the junior (senior) trainees, thus violating the requirement that fixed effects be consistent. I discuss this further and include simulations that demonstrate this point in Appendix A-3.

[14]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.

trainees have significantly more influence than junior trainees suggests that senior trainees have acquired a significant amount of knowledge relative to when they began as junior trainees. Second, if actions are driven by judgments (as opposed to preferences), the significant variation that remains implies that decisions are still made with partial knowledge, even among the most experienced trainees completing residency. As shown numerically in Appendix A-2, these contrasting features are consistent with a single model of partial learning and influence according to knowledge. However, outside of the model, influence may be exogenously granted to senior trainees for reasons unrelated to knowledge, and persistent variation may reflect intrinsic heterogeneity (e.g., heterogeneous ability or preferences). I will explore these threats to interpretation further in Sections 6.1 and 7, respectively.

6 Variation Profiles by Setting

6.1 Influence

The discontinuity in Figure 1 reflects the influence that trainee judgments have on team decisions. The standard team-theoretic framework considers the assignment of decisions to agents who either have full knowledge or no idea of the correct answer posed by the decision (Garicano, 2000). In decisions about which an agent has full knowledge, other agents on the same team should have no role in decision-making. In Section 4, I generalize this framework to allow for *partial knowledge*, in which decision-making draws from experience with related but not identical prior cases. In this general case, agents with partial knowledge jointly contribute to decision-making, in proportion to the precision of their judgments.

To shed empirical light on the role of knowledge in influence, I assess the discontinuity in practice variation across one-year tenure mark, when junior trainees become senior trainees, in different decisions and settings. If decision-making in a given category of decisions is fully *assigned* to either the junior or senior agents on the team, then practice variation in that category should be fully attributable to the respective junior or senior trainees. However, if experience leads to superior partial knowledge that is generally applicable, then decision-making should be shared broadly, and senior trainees should have greater influence across a broad range of decisions and settings. Finally, influence in team decisions may be exogenously granted for reasons uncorrelated to knowledge (i.e., reasons outside the team-theoretic framework). While I cannot directly disprove this possibility, I

will compare profiles of practice variation across decisions categories that may differ in the way that relevant knowledge is acquired and used in decision-making. Under the mechanism of knowledge, the effect of experience on influence should be greater when knowledge used in decision-making accrues to a greater degree with experience.

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 also shows moments of spending in these categories. In all decision categories, there is an increase in the standard deviation of trainee effects at the one-year mark. This implies that there is no category of decisions solely made by junior trainees, and further, senior trainees have larger influence in all decision categories.

Nonetheless, there is large variation in the size of the increase in practice variation at the one-year discontinuity. Diagnostic spending shows a large increase in trainee effect variation, from a standard deviation of 16% to one of 74% in the distributions of trainee effects before and after the one-year tenure mark. In contrast, the other major category of decisions under physician control – medication spending – shows relatively small variation and a small increase in trainee effect variation, from a standard deviation of 17% to one of 26%. Spending on blood transfusion exhibits an intermediate increase in trainee effect variation at one year of tenure, from 26% to 56%. Finally, nursing resources are understandably determined less by physician decisions but also show a moderate (relative) increase in trainee effect variation, from 8% to 16%, at the one-year mark.

Although I do not have a prior quantitative benchmark of the availability and acquisition of knowledge in various domains of medical practice, these results are qualitatively consistent with the idea that the influence related to experience is based on tacit knowledge. Medication decisions, conditional on diagnosis, can draw on relatively public knowledge, given the research orientation around pharmaceutical treatments for well-defined patient populations. On the other hand, diagnostic decisions are much less predefined. Much of clinical expertise relates to the process of diagnosis, and tradeoffs exist among speed, expense, convenience, and uncertainty of diagnostic results between modalities such as a traditional physical exam or an abdominal CT scan. Blood transfusion, despite its use for decades, remains a treatment choice with substantial discretion and uncertainty (Carson et al., 2016).

In Figure 3, I show practice variation profiles as a function of two other considerations. First,

I consider whether team decision-making differs according to the type of patient, for patients with above- vs. below-median expected mortality, and for patients with above- vs. below-median expected total spending, within each housestaff and tenure period. Second, I consider whether team decision-making depends whether a decision is made earlier or later in a patient's stay. As shown in the figure, practice variation profiles are similar between patient populations. However, there are significant differences between team decisions made earlier vs. later in each patient's stay. For decisions on days up to the middle of the stay, the standard deviation of practice variation is 4% just prior to the one-year training mark and 25% just after the one year. For decisions on days after the midpoint, the corresponding standard deviations before and after the one-year mark are 19% and 26%, respectively. These results suggest that junior trainees substantially increase their influence after more is specifically known about their patients. While caring for new patients must draw from general knowledge, patient-specific knowledge accumulates later in the patient's stay. Junior trainees, who on average care for only half of the team's patients and who interact more closely with patients, are particularly positioned to use this specific knowledge.

6.2 Convergence: Decreasing Practice Variation with Tenure

Practice variation thus far shows little convergence, or decrease with tenure, throughout residency training. In this subsection, I explore practice variation in each of the ward services in this residency program – cardiology, oncology, and general medicine – that the trainee team and patient are assigned to. The ward service is a relevant characterization of the learning environment, for several reasons. First, each service is staffed by a mutually exclusive set of supervising physicians, with homogeneous training and certification standards distinct from the other sets.[15] Second, patients are assigned to inpatient services through a triage process involving judgment, usually beginning with an evaluation of the patients' needs in the emergency department. In this process, for example, patients with chest pain deemed likely to have significant cardiac pathology are assigned to the cardiology, while those are unlikely to have such pathology are admitted to general medicine. Third, scientific knowledge and best practices are usually developed within subspecialty fields, and the division of

[15]Supervising physicians on the cardiology and oncology services have completed respective subspecialty fellowship training after internal medicine residency and are required to maintain subspecialty board certification, while supervising general medicine physicians have no subsequent subspecialty training.

medical care into cardiology, oncology, and general medicine is widespread in modern medicine.[16]

Each trainee rotates across each of the ward services, and the proportion of training in each service is designed to be roughly equal across trainees. Thus, comparing profiles of practice variation across services sheds light on differences in how knowledge is acquired – through the three years of residency training – and used for patient care in each of these services, holding the set of trainees fixed. In Figure 4, I show each of these profiles of trainee-effect variation over tenure for cardiology, oncology, and general medicine. Of the three services, cardiology is the only one that shows significant convergence in total costs through the end of training. Cardiology and oncology both show noticeable convergence in diagnostic costs. Such convergence in cardiology and to a lesser extent in oncology suggests learning toward a common "best practice," which could be driven by knowledge more commonly agreed upon by supervising physicians in a service.[17] I assess the robustness of these findings with systematic placebo tests and find that differences in convergence across services are highly significant by randomization inference (see Appendix A-4 for details).

Relatedly, I assess the degree to which knowledge can be encoded in formal diagnoses. 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."[18] I further find no greater convergence in care for patients with more common diagnostic codes within each service (Figure A-8) or for patients with a formal diagnosis linked to a guideline catalogued by the US Agency for Healthcare Research and Quality (`guidelines.gov`) (Figure A-9). Finally, I replicate 97% of the diagnostic-code makeup of the cardiology service using patients from general medicine with the same ICD-9 codes as in cardiology and weighting them appropriately. I find no convergence in these patients from general medicine but with diagnostic codes in common with cardiology (Figure A-10). These results strongly suggest that learning, as measured by convergence, is not captured by formally coded diagnoses, an idea consistent with tacit knowledge.

[16]Table A-2 details the existence of subspecialty ward services in the top internal medicine residency programs, and Table A-3 provides similar evidence across all internal medicine residency programs accredited by the American Council for Graduate Medical Education (ACGME) and accessed at `www.acgme.org`.

[17]Interestingly, cardiology and oncology have substantially more research activity than other specialties or internal medicine subspecialties. In Tables A-4 and A-5, I show counts of *New England Journal of Medicine* articles by medical subspecialty and dollars of research funding by the National Institutes of Health, respectively.

[18]Table A-6 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.

7 Idiosyncratic Learning

The usual lack of convergence in physician practice styles suggests that, despite intensive training designed to create homogeneity, there remains significant variation in physician judgments. In this section, I examine more directly the idea that practice variation is based upon tacit knowledge. If knowledge cannot be easily passed among agents and instead must be gained from experience and contemplation, the limited set of experiences that a single physician can have implies decision-making by heuristics based on idiosyncratic experiences.

7.1 Serial Correlation

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). In the case of unchanging heterogeneity, physician practice styles should be constantly and highly correlated 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.

As in Section 5, the model of trainee effects on team spending decisions remains as specified in Equation (4). However, while I previously treated the same trainee in different time periods as distinct, as a parameter of interest in this section, I explicitly model the correlation between effects of the same trainee in different periods. Details are described in Appendix A-3.2.

In Figure 5, 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.

7.2 Trainee Characteristics

The second method I use to assess the role of intrinsic heterogeneity in practice variation exploits 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 3, trainees with above-median future incomes will earn substantially more than their peers with below-median future incomes.

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

$$Y_{it} = \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{aijkt}, \tag{5}$$

where $\text{Characteristic}^m_h$ is an indicator for whether the junior (or senior) trainee h has the characteristic m, ζ_{-h} is a fixed effect for the other senior (or junior) trainee $-h$, and ζ_ℓ is a fixed effect for attending ℓ.[19] The coefficient of interest, α_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 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 (5):

$$Y_{it} = \sum_m \alpha_m \text{Characteristic}^m_{h(i,t)} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{6}$$

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 $\tilde{Z}_h = Z_h / \sqrt{\text{Var}\left(Z_h\right)}$ with standard deviation 1. Second, I regress daily total spending on this

[19] In principle, I could include trainee characteristics as mean shifters in the baseline random effects model in Equation (4). However, since characteristics are generally insignificant predictors of variation, results of (residual) variation attributable to trainees are unchanged.

normalized score:

$$Y_{it} = \alpha \tilde{Z}_{h(i,t)} + \mathbf{X}_i \beta + \mathbf{T}_t \eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}. \tag{7}$$

In addition, I evaluate the predictive power of housestaff 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 (5).

Table 3 shows results for Equation (7) and a subset of results for Equation (5). Considering characteristics individually in Equation (5), 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 *a priori* unlikely given that these are the characteristics that the residency program makes acceptance decisions based on,[20] 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 6, the distributions of trainee effects throughout training are not meaningfully different between groups of trainees separated by their test scores, rank list positions, or future earnings. This finding implies that trainees who differ significantly along meaningful dimensions still practice similarly not only on average but also in terms of variation over time. That is, trainees evaluated with higher test scores, more desirable rankings, or higher future earnings do not exibit lower variation or higher convergence over training.

[20] 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%.

7.3 Acquired Skill or 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.[21]

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

$$Y_{it} = \alpha_m \text{Experience}^m_{h(i,t),t} + \mathbf{X}_i\beta + \mathbf{T}_t\eta + \zeta_{-h(i,t)} + \zeta_{\ell(i,t)} + \varepsilon_{it}, \tag{8}$$

where $\text{Experience}^m_{h,t}$ is an indicator for whether housestaff 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 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 (5). This effect includes both differences in selection (i.e., intrinsic heterogeneity) into the different program and in training experiences across the programs. Table 3 shows that visiting trainees do not have significantly different spending effects, either as interns or residents.[22]

[21] The related concept of "schools of thought," in which physicians may have systematically different training experiences, has been proposed as a mechanism for geograhic 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.

[22] 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

Overall, these results indicate that summary measures of trainee experience are also poor predictors of practice and outcomes, especially relative to the large variation across trainees. The results fail to support the "schools of thought" mechanism put forward by Phelps and Mooney (1993) as a major source of practice variation, at least within an organization with *ex ante* uniform training experiences but nonetheless large practice variation. Instead, they provide additional support for the idea that learning is mostly idiosyncratic.

8 Discussion and Conclusion

I follow physicians in residency training as they acquire professional knowledge and make decisions in teams. Exploiting a discontinuity in relative experience, I find that physicians who are senior on the team have greater influence in decision-making. This influence effect is greater in diagnostic decisions, for which knowledge is less explicit, and earlier in the stay of each patient, when decisions must draw on general rather than patient-specific knowledge. While this effect of influence implies important knowledge that accrues with experience, there is usually little convergence in practice, although this depends on specialty-driven learning environments. The evidence suggests idiosyncratic learning and partial knowledge as key underpinnings of practice variation.

These findings have implications for the widespread variation in health care spending that features prominently in policy discussions. Previously proposed policy levers for reducing variation include physician financial incentives, reporting, and patient cost-sharing (see Skinner, 2012, for a summary). The problem with these levers is twofold. First, the focus on spending and on the overall *level* of care obscures the specific decisions that must be made in patient care. When caring for a specific patient, physicians do not ask, "How much should I spend?" but rather "What is the best next thing to do?" For example, Abaluck et al. (2016) find that any welfare loss from physicians ordering too many CT scans on average are small compared to the misallocation of CT scans, holding the overall level fixed. Second, any objective system of process measurement for reporting or pay presumes that the problem can be sufficiently described and that the correct choice is known and agreed upon, for a meaningful share of hypothetical physician decisions. As some commentators have noted (e.g., Jauhar, 2014), this view is inconsistent with the training process to become a

their mean practice styles can be dwarfed by variation within training program.

physician and with the practice of medicine itself.

More generally, the qualities of (i) tacit knowlege, (ii) gradual on-the-job training, and (iii) influential (but not blindly followed) advice from more experienced practitioners seem relevant to a broad range of professions, including managers, academics, and judges. In this sense, informational frictions may underlie persistent practice variation in other industries (e.g., Syverson, 2011; Gibbons and Henderson, 2012). The resulting fundamental problem is that practitioners have limited experience relative to the set of decisions they could face and therefore partial knowledge that must be generalized in decision-making. The natural solution to such a problem would appear to be informational, through either technology or organizational practice. For example, innovations in machine learning show some promise in reducing the dimensionality of decisions, with some medical applications (James et al., 2013; Chen et al., 2017), although much work on causal inference and transparency remains. Alternatively, organizational innovations such as "continuous quality improvement" and "lean management" appear to gather, codify, and disseminate information to practitioners at the local level. Evaluating the extent to which these innovations can reduce practice variation and improve outcomes is an area of promising future research.

References

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2016): "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, 106, 3730–3764.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333.

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

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2006): "Learning and Disagreement in an Uncertain World," Working Paper 12648, National Bureau of Economic Research.

Alchian, A. A. and H. Demsetz (1972): "Production, information costs, and economic organization," *The American Economic Review*, 62, 777–795.

American Medical Association (1966): "Citizens Commission on Graduate Medical Education," in *The Graduate Education of Physicians*, Chicago, IL: American Medical Association.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Bartel, A. P., N. Beaulieu, C. Phibbs, and P. W. Stone (2014): "Human Capital and Productivity in a Team Environment: Evidence from the Healthcare Sector," *American Economic Journal: Applied Economics*, 6, 231–259.

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

Bertrand, M. and A. Schoar (2003): "Managing with Style: The Effect of Managers on Firm Policies," *The Quarterly Journal of Economics*, 118, 1169–1208.

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

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

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

Carson, J. L., G. Guyatt, N. M. Heddle, B. J. Grossman, C. S. Cohn, M. K. Fung, T. Gernsheimer, J. B. Holcomb, L. J. Kaplan, L. M. Katz, N. Peterson, G. Ramsey, S. V. Rao, J. D. Roback, A. Shander, and A. A. R. Tobian (2016): "Clinical Practice Guidelines From the AABB: Red Blood Cell Transfusion Thresholds and Storage," *The Journal of the American Medical Association*, 316, 2025.

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

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016): "Healthcare Exceptionalism? Productivity and Allocation in the U.S. Healthcare Sector," *American Economic Review*, 106, 2110–2144.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie (1987): "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, 40, 373–383.

Chen, J. H., M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman (2017): "Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets," *Journal of the American Medical Informatics Association*, 24, 472–480.

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

Cooper, Z., S. Craig, M. Gaynor, and J. Van Reenen (2015): "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured," Tech. Rep. w21815, National Bureau of Economic Research, Cambridge, MA.

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

Cutler, D., J. Skinner, A. D. Stern, and D. Wennberg (2013): "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," Working Paper 19320, National Bureau of Economic Research.

Cyert, R. and J. March (1963): *A Behavioral Theory of the Firm*, Oxford: Blackwell.

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010): "Returns to physician human capital: Evidence from patients randomized to physician teams," *Journal of Health Economics*, 29, 866–882.

Doyle, J. J., J. A. Graves, J. Gruber, and S. Kleiner (2015): "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, 123, 170–214.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey (1998): "Comorbidity Measures for Use with Administrative Data," *Medical Care*, 36, 8–27.

Ellison, G. and D. Fudenberg (1993): "Rules of thumb for social learning," *Journal of Political Economy*, 101, 612–643.

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

Finkelstein, A., M. Gentzkow, and H. Williams (2016): "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *Quarterly Journal of Economics*, 131, 1681–1726.

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

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

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

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

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

Glover, A. (1938): "The Incidence of Tonsillectomy in School Children," *Proceedings of the Royal Society of Medicine*, 31, 1219–1236.

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

Jacob, B. A. and L. Lefgren (2007): "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers," *The Quarterly Journal of Economics*, 122, 1603–1637.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning*, vol. 103 of *Springer Texts in Statistics*, New York, NY: Springer New York, dOI: 10.1007/978-1-4614-7138-7.

Jauhar, S. (2014): "Don't Homogenize Health Care," *The New York Times*.

Kane, T. J. and D. O. Staiger (2002): "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," in *Brookings Papers on Education Policy*, ed. by D. Grissmer and H. F. Ladd, Washington, DC: Brookings Institution Press, 235–283.

Kelly, R. (2009): "Where can $700 billion in waste be cut annually from the U.S. healthcare system?" Tech. rep., Thomson Reuters.

Kizer, K. W. and S. R. Kirsh (2012): "The Double Edged Sword of Performance Measurement," *Journal of General Internal Medicine*, 27, 395–397.

Lazear, E. P., K. L. Shaw, and C. Stanton (2015): "The Value of Bosses," *Journal of Labor Economics*, 33.

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

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

Molitor, D. (2017): "The evolution of physician practice styles: Evidence from cardiologist migration," *American Economic Journal: Economic Policy*, Forthcoming.

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

Nelson, R. R. and S. G. Winter (1982): *An Evolutionary Theory of Economic Change*, Harvard University Press.

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

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

Pear, R. (2009): "Health Care Spending Disparities Stir a Fight," *The New York Times*.

Peter Bent Brigham Hospital (1943): "Annual Report," Tech. rep., Boston, MA.

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

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

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

Radley, D. C., S. N. Finkelstein, and R. S. Stafford (2006): "Off-label Prescribing Among Office-Based Physicians," *Archives of Internal Medicine*, 166, 1021.

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

Rogerson, R., R. Shimer, and R. Wright (2005): "Search-Theoretic Models of the Labor Market: A Survey," *Journal of Economic Literature*, 43, 959–988.

Rossouw, J. E., G. L. Anderson, R. L. Prentice, A. Z. LaCroix, C. Kooperberg, M. L. Stefanick, R. D. Jackson, S. A. A. Beresford, B. V. Howard, K. C. Johnson, J. M. Kotchen, J. Ockene, and Writing Group for the Women's Health Initiative Investigators (2002): "Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial," *The Journal of the American Medical Association*, 288, 321–333.

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

Shaneyfelt, T. M., M. F. Mayo-Smith, and J. Rothwangl (1999): "Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature," *The Journal of the American Medical Association*, 281, 1900–1905.

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

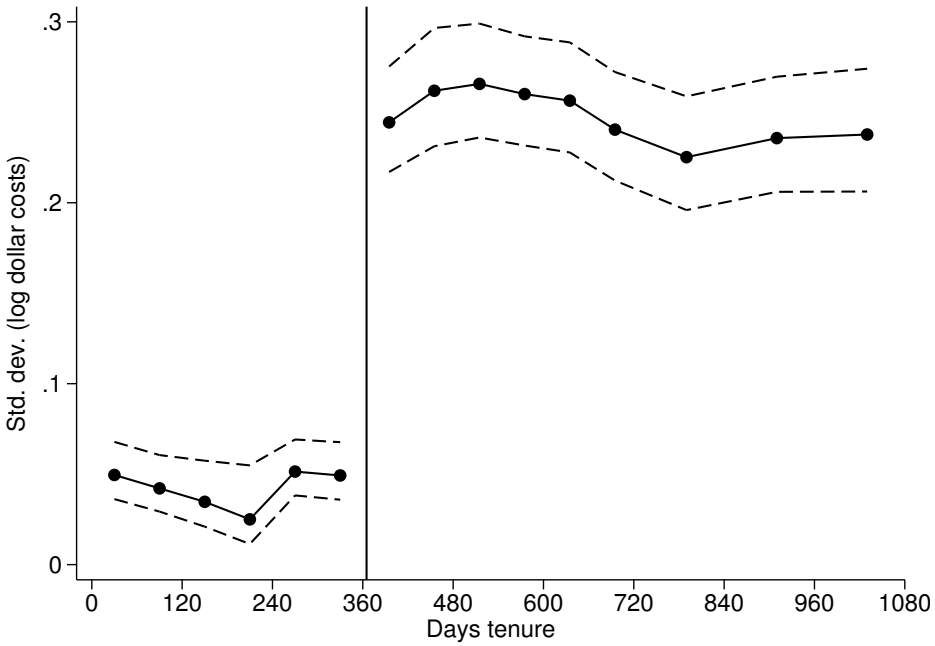
Skinner, J. and D. Staiger (2015): "Technology Diffusion and Productivity Growth in Health Care," *Review of Economics and Statistics*, 97, 951–964.

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

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

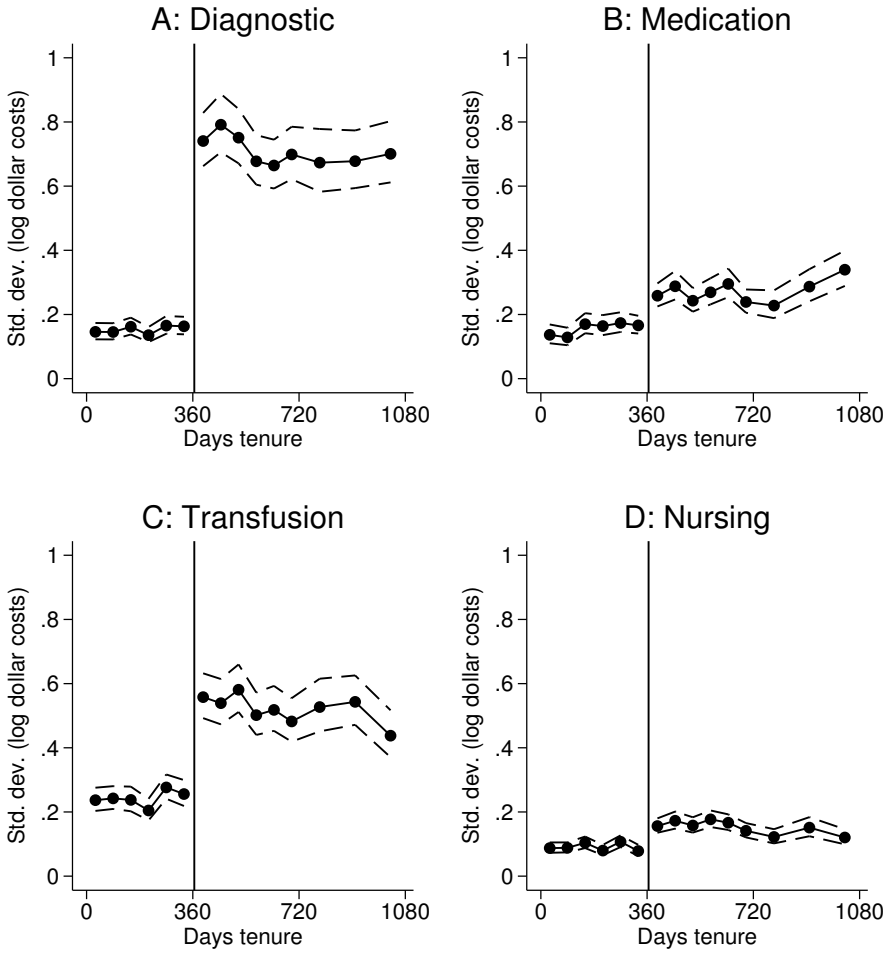
Wennberg, J. and A. Gittelsohn (1973): "Small area variations in health care delivery," *Science*, 182, 1102–1108.

Figure 1: Profile of Practice Variation by Tenure



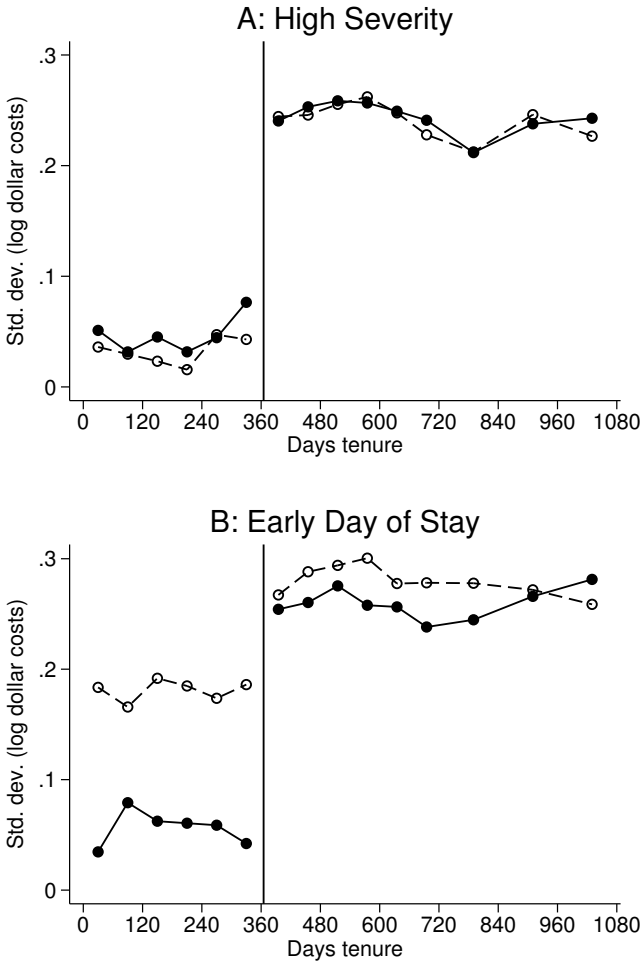
Note: This figure shows the standard deviation of random trainee effects in a model, stated in Equation (4), of log daily total costs at each non-overlapping tenure period. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. The model controls for patient 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



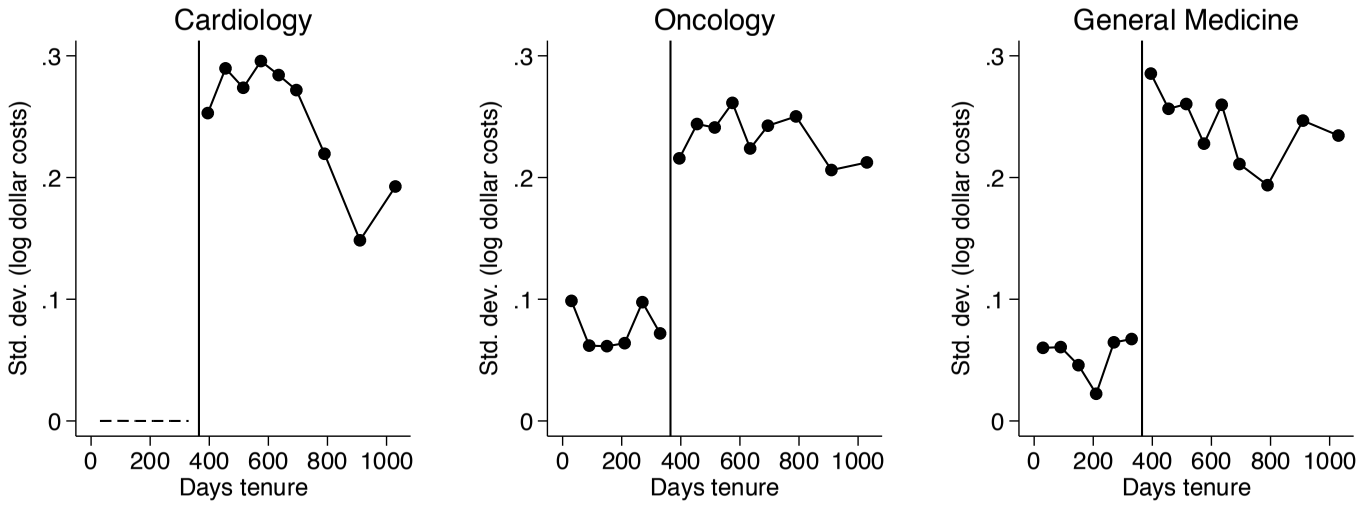
Note: This figure shows the standard deviation of random trainee effects in a model, stated in Equation (4), of log daily costs at each non-overlapping tenure period. Each panel shows a different spending category. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed 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



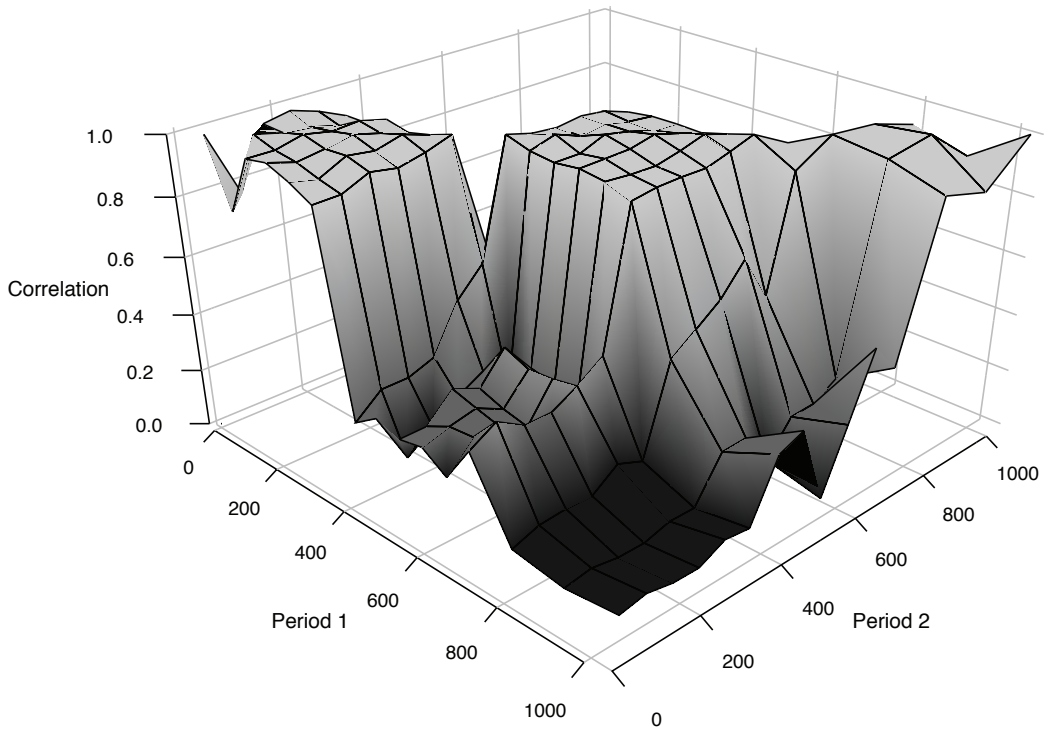
Note: This figure shows the standard deviation of random trainee effects in a model, stated in Equation (4), of log daily total costs at each non-overlapping tenure period. Panel A estimates the model separately in two samples of patients with above- (solid dots) vs. below-median (hollow dots) expected 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



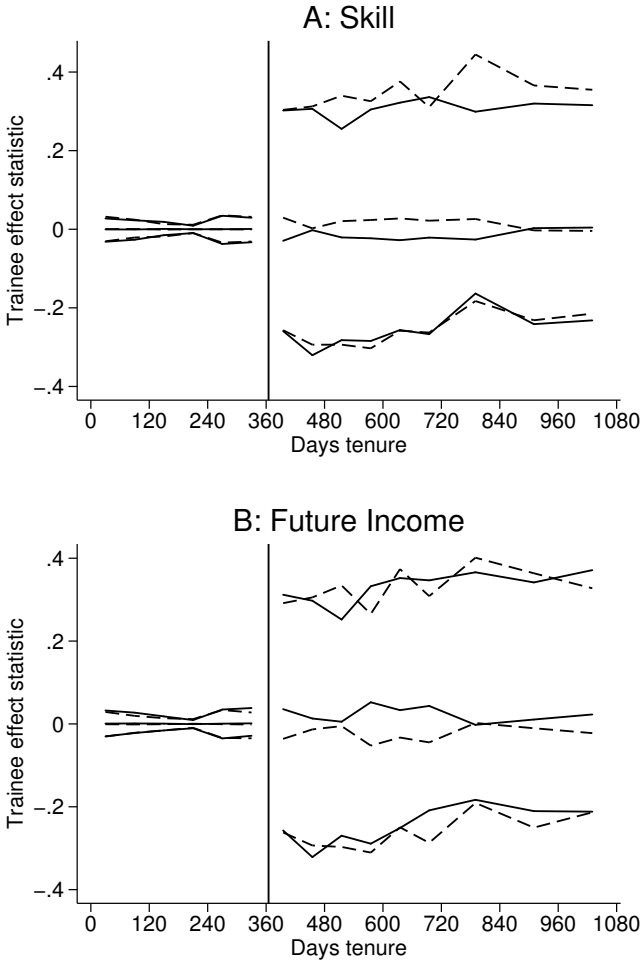
Note: This figure shows the standard deviation in a random effects model, as in Equation (4), of log daily total costs at each non-overlapping two-month tenure interval. Each panel shows results estimated from within a service of cardiology, oncology, or general medicine. The dashed line prior to 365 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 A-6 shows the corresponding figure for log daily diagnostic costs.

Figure 5: Serial Correlation of Trainee Random Effects



Note: This figure shows the serial correlation between random effects within trainee between two tenure periods. Details of the estimation routine are given in Appendix A-3.2. The random effect model of log daily total costs is given in Equation (4). 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 6: Practice Style Distribution by Trainee Type



Note: This figure shows the patient-day-weighted 90th percentile, mean, and 10th percentile of the practice style (trainee effect) distribution according to trainee type. The unconditional distribution in each tenure period is normalized to have mean 0. Panel A shows the distribution for high-skill trainees (solid lines) relative to low-skill trainees (dashed lines), where "skill" is defined as position on the rank list more favorable than median when defined, and above-median USMLE test score when position on the rank list is missing. Panel B shows the distribution for trainees with above-median expected future income relative (solid lines) to those with below-median future income (dashed lines), where future income is based on known subsequent subspecialty training (if any) and imputed with national average yearly income in the first five years of practice after training. The average yearly future incomes of above- and below-median junior trainees are $424,000 and $268,000, respectively; the respective yearly future incomes for senior trainees are $409,000 and $249,000 (junior trainees include "preliminary interns," described in Section 2, who generally move on to more lucrative specialties). Practice styles are calculated as the Best Linear Unbiased Predictor (BLUP) posterior mean from the random effects model specified in Equation (4), 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.

Table 1: Exogenous Assignment for Trainees with Above or Below Average Spending

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

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

Table 2: Summary Statistics of Spending in Categories and Services

	Log daily total costs				
	(1) Radiology	(2) Laboratory	(3) Medication	(4) Transfusion	(5) Nursing
Cardiology					
5th percentile	0	11	4	0	189
10th percentile	0	16	14	0	244
Median	0	34	67	16	658
Mean	54	51	113	32	661
90th percentile	125	103	233	56	1,075
95th percentile	375	145	417	87	1,212
Oncology					
5th percentile	0	3	0	0	192
10th percentile	0	13	13	0	256
Median	0	34	94	12	673
Mean	66	58	155	77	681
90th percentile	248	124	350	204	1,033
95th percentile	423	212	542	411	1,270
General Medicine					
5th percentile	0	8	2	0	160
10th percentile	0	12	10	0	205
Median	0	35	69	14	561
Mean	66	62	99	38	577
90th percentile	234	139	210	48	959
95th percentile	385	222	286	95	1,130

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

Table 3: Effect of Trainee Characteristics on Spending

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

Note: This table reports results for some regressions of the effect of indicators of some trainee characteristics, including other hospital status, and a normalized predictive score (with standard deviation 1) based on *all* observed trainee characteristics. Panel A shows results for interns; Panel B shows results for residents. Columns (1) to (5) are regressions of the form in Equation (5), 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 7.2). 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 (7), 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 4: Effect of Trainee Experience on Spending

	Log daily total costs				
	(1) Number of days	(2) Number of patients	(3) Number of attendings	(4) Attending spending	(5) Attending spending
Panel A: Interns					
Effect of trainee with	0.001	0.000	-0.002	-0.006	0.001
measure above median	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Observations	181,874	181,874	181,874	155,523	129,636
Adjusted R^2	0.088	0.088	0.088	0.089	0.090
Panel B: Residents					
Effect of trainee with	0.003	0.003	-0.005	0.008	0.013**
measure above median	(0.008)	(0.007)	(0.007)	(0.005)	(0.005)
Observations	199,934	199,934	199,934	182,017	174,534
Adjusted R^2	0.090	0.090	0.090	0.087	0.087
Measure and median within service	Y	Y	Y	N	Y

Note: This table reports results for some regressions of the effect of indicators of trainee experience. Panel A shows results for interns; Panel B shows results for residents. Regressions are of the form in Equation (5), 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.

Appendix (for Online Publication per Referees / Editor)

A-1 Random Assignment

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

A-1.1 Assignment of Patients to Trainees

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

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

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

In column (1) of Table A-1, I show F-statistics and the corresponding p-values for the null hypothesis that $\left(\zeta^{\tau<T}_j, \zeta^{\tau>T}_k\right)_{j\in\mathcal{J},k\in\mathcal{K}} = \mathbf{0}$. I perform the regression (A-1) separately each of the following patient characteristics X_a as a dependent variable: patient age, a dummy for male gender, and a dummy for white race.[23] I also perform (A-1) using as dependent variables the linear prediction of log admission total spending based on patient age, race, and gender. I fail to find joint statistical significance for any of these tests.

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

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

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

[23] 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 (4) does not qualitatively affect results.

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 that $(\gamma_1, \gamma_2) = \mathbf{0}$. Column (2) includes all trainee characteristics in Z_h; column (3) excludes position on the rank list, since this information is missing for a sizeable 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 housestaff spending effects that are fixed within junior or senior role, using this regression:

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

where Y_a is log total spending for admission a, and other variables are defined similarly as in Equation (A-1). Figure A-1 shows kernel density plots of the age distributions for patients assigned to interns and residents, respectively, each of which compare trainees with practice styles above and below the mean. Figure A-2 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:

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

where $r \equiv \mathbf{1}(\tau > T)$ is an indicator for whether the fixed effect for trainee h was calculated while h was a junior trainee ($r = 0$) or a senior trainee ($r = 1$). As in Equation (A-1), I assume no relationship between $\hat{\zeta}_h^{\tau<T}$ and $\hat{\zeta}_h^{\tau>T}$. Each observation in Equation (A-4) corresponds to an admission a, but where error terms are clustered at the level of the intern-resident-attending team, since there are multiple observations for a given team. $\hat{\zeta}_\ell$ is the estimated fixed effect for attending k.[24] Estimates for γ_h and γ_ℓ are small, insignificant, and even slightly negative.

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

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

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

A-2 Variation over Time under Example Learning Parameters

This appendix further explores the implications of the conceptual framework in Section 4, in which decision-making is modeled in a team-theoretic environment, along a continuous action space, for two agents with normal priors. While this framework is not meant to be taken literally (e.g., actions may not be continuous, decision-making may not be strictly team-theoretic), this appendix provides further intuition and numerical examples in this framework for how learning could lead to persistent practice variation.

A-2.1 Analytical Evaluation

Consider the standard deviation of tenure-specific trainee effects ξ_h^τ. In the simple case in which judgment is based only on knowledge and not observation (and ignoring categories of decisions), Equation (3) can be stated as

$$\xi_h^\tau = E\left[\frac{\rho_{d,h}\mu_{d,h}}{\rho_{d,h}+\rho_{d,-h}+P_d}\middle| d \in \mathcal{D}_h^\tau\right].$$

The tenure of $-h$ depends on τ in a dependable fashion: If h is a junior trainee, then the tenure of $-h$ is one or two years greater than τ; if h is a senior trainee, then the tenure of $-h$ is one or two years less than τ and less than one year. Assume that $\rho_{d,h}$ is only a function of h's tenure, denoting it as $\rho(\tau)$, and assume $P_d = P$ for all d. Also denote the difference in tenure between the teammate $-h$ and h as Δ.

For Bayesian judgments to be consistent, it is natural that

$$\mathrm{Var}_h\left(E\left[\mu_{d,h}\middle| d \in \mathcal{D}_h^\tau\right]\right) = \frac{1}{\rho(\tau)}.$$

Then, denoting the standard deviation of $\{\xi_h^\tau\}$ as a function of τ as $\sigma(\tau)$,

$$\sigma(\tau) = \frac{\rho(\tau)^{1/2}}{\rho(\tau)+\rho(\tau+\Delta)+P}. \tag{A-5}$$

$\sigma(\tau)$ can be thought of a profile of practice variation across trainees over different tenure periods, akin to the profiles empirically estimated in the paper (e.g., Figure 1). The extent to which $\sigma(\tau)$ decreases with τ can be thought of as convergence, and the extent to which $\rho(\tau)$ increases with τ can be thought of as the rate of learning.

As a first observation, practice variation profiles may be scaled by any arbitrary constant, holding

relative shape of the profile constant:

Proposition A-1. *Consider a practice variation profile, $\sigma(\tau)$, that exists under a learning function $\rho(\tau)$ and external prior P. Any practice variation profile that takes the form $\tilde{\sigma}(\tau) = \kappa\sigma(\tau)$ with constant κ also exists.*

Proof. The learning profile $\tilde{\rho}(\tau) = \rho(\tau)/\kappa^2$ and external prior $\tilde{P} = P/\kappa^2$ yield the desired practice variation profile $\kappa\sigma(\tau)$ under Equation (A-5). □

Scaling both the learning profile and the external prior by a constant preserves the "influence" that each agent has relative to each other and to the external practice environment. Uniformly smaller (larger) practice variation $\sigma(\tau)$ imply uniformly larger (smaller) judgment precisions $\rho(\tau)$. Under no forgetting, $\rho(\tau)$ is weakly increasing in τ.

Next, consider the discontinuity in practice variation across the one- and two-year tenure marks. For simplicity of notation, consider τ in years.

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

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

Proof. Consider the conservative case that interns only work with second-year residents in their last month. Then

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

and

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

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

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

Finally, consider the derivative of variation with respect to tenure:

$$\sigma'(\tau) = \frac{\frac{1}{2}\rho(\tau)\rho'(\tau)(\rho(\tau)+\rho(\tau+\Delta)+G) - \rho(\tau)^{1/2}(\rho'(\tau)+\rho'(\tau+\Delta))}{(\rho(\tau)+\rho(\tau+\Delta)+P)^2}.$$

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

$$\sigma'(\tau) < 0 \Leftrightarrow \rho(\tau) > \frac{\rho'(\tau)}{2\rho'(\tau+\Delta)+\rho'(\tau)}\left(\rho(\tau+\Delta)+P\right). \tag{A-6}$$

This condition highlights that convergence is not supported at all τ under all learning profiles $\rho(\tau)$. In particular, if the precision of the index trainee's subjective prior $\rho(\tau)$ is less than the combined precision of the peer's subjective prior $\rho(\tau+\Delta)$ and the external practice environment's precision P, then convergence may not be supported, particularly if $\rho'(\tau)$ is large relative to $\rho'(\tau+\Delta)$. The intuition for this is related to influence. For small $\rho(\tau)$ relative to $\rho(\tau+\Delta)+P$, the trainee has relatively low influence, and increases in $\rho(\tau)$ may increase variation primarily by increasing influence. This is especially true if most of the learning occurs in the index trainee's cohort as opposed to the peer's cohort, or $\rho'(\tau) \gg \rho'(\tau+\Delta)$, because learning by the peer reduces influence. However, regardless of the size of $\rho'(\tau)$, a sufficient condition for convergence is $\rho(\tau) > \rho(\tau+\Delta)+G$. Given that $\rho(\cdot)$ is monotonically increasing, this suggests that convergence is more likely with residents than with interns.

In order to make further observations, I consider a piecewise linear function for the learning profile $\rho(\tau)$.

Proposition A-3. *Assume that $\rho(\tau)$ takes a piecewise linear form, such that*

$$\rho(\tau) = k_0 + k_1 \min(\tau, 1) + k_2 \max(\tau - 1, 0). \tag{A-7}$$

For any $\rho(\tau)$ that satisfies the form (A-7), conditional on some $\Delta > 0$ (i.e., $\tau < 1$), there exists a unique point $\tau^*_{\Delta>0}$ such that $\sigma'(\tau) > 0$ for all $\tau < \tau^*_{\Delta>0}$, and $\sigma'(\tau) < 0$ for all $\tau > \tau^*_{\Delta>0}$. Similarly, conditional on some $\Delta < 0$ (i.e., $\tau > 1$), there exists a unique point $\tau^*_{\Delta<0}$ such that $\sigma'(\tau) > 0$ for all $\tau < \tau^*_{\Delta<0}$, and $\sigma'(\tau) < 0$ for all $\tau > \tau^*_{\Delta<0}$. The specific forms that $\tau^*_{\Delta>0}$ and $\tau^*_{\Delta<0}$ take are

$$\tau^*_{\Delta>0} = \frac{P + k_1 + k_2(\Delta - 1) - 2k_0k_2/k_1}{k_1 + k_2}; \tag{A-8}$$

$$\tau^*_{\Delta<0} = \frac{P + k_1\Delta - 2k_1(k_0 + k_1)/k_2}{k_1 + k_2} + 1. \tag{A-9}$$

Proof. State the convergence condition in Equation (A-6) as a criterion function $\mathcal{P}(\tau;\Delta)$ in which convergence occurs if and only if $\mathcal{P}(\tau;\Delta) > 0$:

$$\mathcal{P}(\tau;\Delta) = \rho(\tau)\left(2\rho'(\tau+\Delta)+\rho'(\tau)\right) - \rho'(\tau)\left(\rho(\tau+\Delta)+P\right),$$

Under any $\rho(\tau)$ of the form (A-7), $\mathcal{P}(\tau;\Delta)$ is monotonically increasing in τ, which implies a single solution to $\mathcal{P}(\tau^*_\Delta;\Delta) = 0$ conditional on Δ. To arrive at the specific functions that $\tau^*_{\Delta>0}$ and $\tau^*_{\Delta<0}$ take in Equations (A-8) and (A-9), plug Equation (A-7) into $\mathcal{P}(\tau^*_\Delta;\Delta) = 0$ and solve for τ^*_Δ. □

Note that $\tau^*_{\Delta>0}$ in Equation (A-8) may be less than 0 or greater than 1. In the former case, there is convergence for all $\tau \in [0,1]$ (the entire first year); in the latter case, there is divergence (variation is increasing) for all $\tau \in [0,1]$. If $\tau^*_{\Delta>0} \in (0,1)$, then variation in practice styles first increases then decreases. Similarly, practice variation may be increasing over the tenure period as a resident $\tau \in [1,3]$, decreasing over the entire period, or first increasing then decreasing.[25] As noted above, and by comparing (A-8) and (A-9), convergence is more likely and occurs earlier during the period as resident than during the period as intern.

A-2.2 Numerical Examples

Figure A-5 presents a few numerical examples of variation profiles under different learning profiles described by functions of the piecewise linear form in Equation (A-7). The three parameters of interest are k_0, or initial precision; k_1, or the rate of increase in the precision during the first year as a junior trainee; and k_2, or the rate of increase during the subsequent two years as a senior trainee. The precision of judgments at the end of training is $\rho(3) = k_0 + k_1 + 2k_2$. 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. Given Proposition A-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 k_0, k_1, k_2, and P by some constant.

I discuss each panel of Figure A-5 in turn:

- Panel A considers equal $k_0 = k_1 = k_2 = 0.2$, 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 to years as resident.

- Panel B imposes no resident learning ($k_2 = 0$) 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 k_2 by increasing the ratios of k_0 and k_1 to k_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 k_1 relative to k_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

[25] This is ensured even across $\tau = 2T$ because $\tau^*_{-T} > \tau^*_{-2T}$.

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 (k_1/k_2) is similar, but learning during training is reduced relative to knowledge gained prior to training (k_0) and to the external practice environment (P).

A-3 Statistical Model of Trainee Effects

In this appendix, I introduce a statistical model to estimate the standard deviation $\sigma(\tau)$ of trainee effects ξ_h^τ in discrete tenure period τ and the correlation $\rho(\tau_1, \tau_2)$ between trainee effects $\xi_h^{\tau_1}$ and $\xi_h^{\tau_2}$ in two discrete periods τ_1 and τ_2. Random assignment of patients to trainee, conditional on time categories, allows me to estimate trainee effects.[26] Finite observations per trainee-period means that effects will be estimated with error, which implies that standard deviations of unshrunken effects will overstate the true $\sigma(\tau)$. Further, correlations of fixed effect estimates of $\xi_h^{\tau_1}$ and $\xi_h^{\tau_2}$ will be generally understate true correlations, and comparing the relative magnitudes of correlations between two pairs of periods will be invalid.

Standard Bayesian shrinkage procedures to adjust for finite-sample overestimates of $\sigma(\tau)$ (e.g., Morris, 1983),[27] however, deal with a single effect entering the right-hand side of each observation. In this setting, I must deal with two effects – one for the intern and one for the resident – for which I want to estimate distributions. Having two sets of effects results in two complicating issues: First, it is possible that all trainees may not form a single connected set, so effects must be first demeaned within connected set. Second, more importantly, shrinking one set of effects requires a relatively precise mean to shrink toward; this requirement is violated because the effects of the other set are equally problematic, which results in biased estimates of the underlying distribution. Even without this complication, Bayesian shrinkage does not resolve the issue of biased estimates of $\rho(\tau_1, \tau_2)$, since errors in estimates of $\xi_h^{\tau_1}$ and $\xi_h^{\tau_2}$ are not eliminated but only shrunken.[28]

[26] 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.

[27] Recent examples of papers that have used this procedure include Kane and Staiger (2002), Jacob and Lefgren (2007), and Chandra et al. (2016).

[28] Chetty et al. (2014) develop a method of moments approach of predicting unbiased teacher effects that accounts for drift in effects over time and actually estimates the covariance between effects in different periods. However, a crucial assumption they make is that effects follow a stationary process, which is obviously not true among trainees because of both learning and influence.

I therefore adopt a random effects approach in which I simulaneously estimate both distributions of intern and resident effects by maximum likelihood. First, similar in spirit to Chetty et al. (2014) and closely related to the idea of restricted maximum likelihood (REML) (Patterson and Thompson, 1971), I create the differenced outcome $\tilde{Y}_{it} = Y_{it} - \left(\mathbf{X}_i\hat{\beta} + \mathbf{T}_t\hat{\eta} + \hat{\zeta}_{\ell(t)}\right)$, 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 $\mathbf{X}_i$, $\mathbf{T}_t$, and ζ_ℓ.[29] Note that $E\left[\tilde{Y}_{it}\right] = 0$. In practice, given random assignment of attending physicians and patients to trainees, conditional on schedules, I am only concerned with correlations between trainee effects and $\mathbf{T}_t$, but differencing out projections due to $\mathbf{X}_i$ and ζ_ℓ simplifies computation and avoids the incidental parameters problem in the later maximum-likelihood stage. In the next two subsections I will describe in turn how I calculate $\sigma(\tau)$ and $\rho(\tau_1, \tau_2)$. In simulated data (not shown), I confirm that Bayesian shrinkage results in inaccurate estimates of these moments and that the statistical method outlined in this appendix yield close estimates of the true moments of the data generating process, regardless of the number of observations per intern or residents.

A-3.1 Standard Deviation of Trainee Effects

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

$$\tilde{Y}_{it} = \xi_{j(i,t)}^{\tau(j(i,t),t)} + \xi_{k(i,t)}^{\tau(k(i,t),t)} + \varepsilon_{it}, \tag{A-10}$$

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

$$\tilde{Y}_{it} = \xi_{j(i,t)}^{\tau(j(i,t),t)} + \xi_{k(i,t)}^{\tau(k(i,t),t)} + \nu_i + \varepsilon_{it}, \tag{A-11}$$

where ν_i is a random effect for the patient admission.[30] Because trainees are assigned conditionally randomly to each other and to patients, $\xi_{j(i,t)}^{\tau(j(i,t),t)}$, $\xi_{k(i,t)}^{\tau(k(i,t),t)}$, and ν_i are uncorrelated with each other. Assuming ξ_j^τ, $\xi_k^{\tau'}$, and ν_i are normally distributed, their standard deviations $\sigma_{\xi,\tau}$, $\sigma_{\xi,\tau'}$, and σ_ν are the parameters of interest in the following maximum-likelihood estimation, done in separate samples selected on τ.

Equations (A-10) and (A-11) can be stated in vector form:

$$\tilde{\mathbf{Y}} = \mathbf{Z}\mathbf{u} + \varepsilon, \tag{A-12}$$

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

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

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

Let N_τ be the number of trainees with some tenure interval τ (e.g., 1 to 60 days) and N^τ_{-h} be the corresponding teammates observed in the sample. Then in the case that (A-12) represents (A-10), $\mathbf{Z}$ is an $n \times (N_\tau + N^-_\tau)$ selection matrix for trainees with tenure τ and their peers, and $\mathbf{u}$ is an $(N_\tau + N^-_\tau) \times 1$ stacked vector of trainees and peer random effects. The variance-covariance matrix of $\mathbf{u}$ is diagonal:

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

Similarly, in the case that (A-12) represents (A-11), $\mathbf{Z}$ is an $n \times (N_\tau + N^-_\tau + N_i)$ selection matrix for trainees of tenure τ, teammates, and patient admissions, and $\mathbf{u}$ is an $(N_\tau + N^-_\tau + N_i) \times 1$ stacked vector of intern, resident, and admission random effects, where N_i is additionally the number of admissions in the sample. The diagonal variance-covariance matrix of $\mathbf{u}$ is

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

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

$$L = -\frac{1}{2} \left\{ n \log (2\pi) + \log |\mathbf{V}| + \tilde{\mathbf{Y}}' \mathbf{V}^{-1} \tilde{\mathbf{Y}} \right\}. \tag{A-13}$$

I estimate (A-10) or (A-11) by maximum likelihood, for each τ separately. Holding the tenure of h fixed at τ, the tenure of the other teammate will possibly be a mixture if τ is less than one year (i.e., corresponds to junior trainees). I thus focus on $\sigma(\tau) \equiv \sigma_{\xi,\tau}$ and not $\sigma_{\xi,\tau'}$, although results do not qualitatively depend on this.

A-3.2 Correlation of Trainee Effects

To estimate $\rho(\tau_1, \tau_2)$, I augment models in (A-10) and (A-11) to account for two separate tenure periods τ_1 and τ_2 across which trainee effects may be correlated. Although I observe each trainee across 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 τ_1 and τ_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-10), I consider

$$\tilde{Y}_{it} = \xi^{p(i,t)}_{h(i,t)} + \xi_{-h(i,t)} + \varepsilon_{it}, \tag{A-14}$$

which features the function $p(i,t) \in \{\tau_1, \tau_2\}$. This specifies that effects of trainees in the tenure periods of interest (τ_1 and τ_2) may be drawn from two separate distributions depending on the tenure period τ_1 or τ_2 corresponding to observation t, while effects of the teammates are pooled into

a single distribution not dependent on tenure. The analog for Equation (A-11) is

$$\tilde{Y}_{it} = \xi_{h(i,t)}^{p(i,t)} + \xi_{-h(i,t)} + \nu_i + \varepsilon_{it}. \tag{A-15}$$

As above, both (A-14) and (A-15) can be written in the vector form of (A-12). When representing (A-14) as (A-12), the selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_\tau^-)$, since it now maps observations onto one of two random effects, depending on whether $p(i,t) = \tau_1$ or $p(i,t) = \tau_2$, for each trainee h observed in both τ_1 and τ_2 tenure periods. The stacked vector of random effects $\mathbf{u}$ is similarly of size $(2N_\tau + N_\tau^-) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

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

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

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

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

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

Representing (A-15) as (A-12) is a similar exercise. The selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_\tau^- + N_i)$, and the vector of random effects $\mathbf{u}$ is of size $(2N_\tau + N_\tau^- + N_i) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

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

where $\mathbf{G}_\tau$ is the same as before.

The log likelihood is the same as in Equation (A-13), but using revised definitions of $\mathbf{G}$ that allow for covariance between random effects of the same trainees across tenure periods. The correlation parameter of interest $\rho(\tau_1, \tau_2)$ is estimated from $\hat{\mathbf{A}}$ and is constrained to be between -1 and 1. 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-4 Systematic Placebo Tests

I consider the statistical significance for convergence in (i) cardiology for log daily total costs and (ii) the specialist services (i.e., cardiology and oncology) for log daily diagnostic costs by performing the following thought experiments, in which "treatment" observations are defined as those in cardiology and specialist services, respectively, and "control" observations are the remaining observations. If there is no difference in true convergence between treatment and control, then randomly assigning actual months for each resident to a placebo treatment or control service should result in similar convergence in these placebo services over time for a large proportion of these placebo tests. On the other hand, if very few of these placebo tests result in convergence similar to that observed in the actual treatment assignment, then this suggests statistical significance.

I implement these placebo tests as follows:

1. Defining a service as either "treatment" or "control," count the number of residents in a specialist service during each month t. Call this number N_t^{treat}. The proportion of residents in cardiology, oncology, and general medicine during each month us shown in Figure A-7.

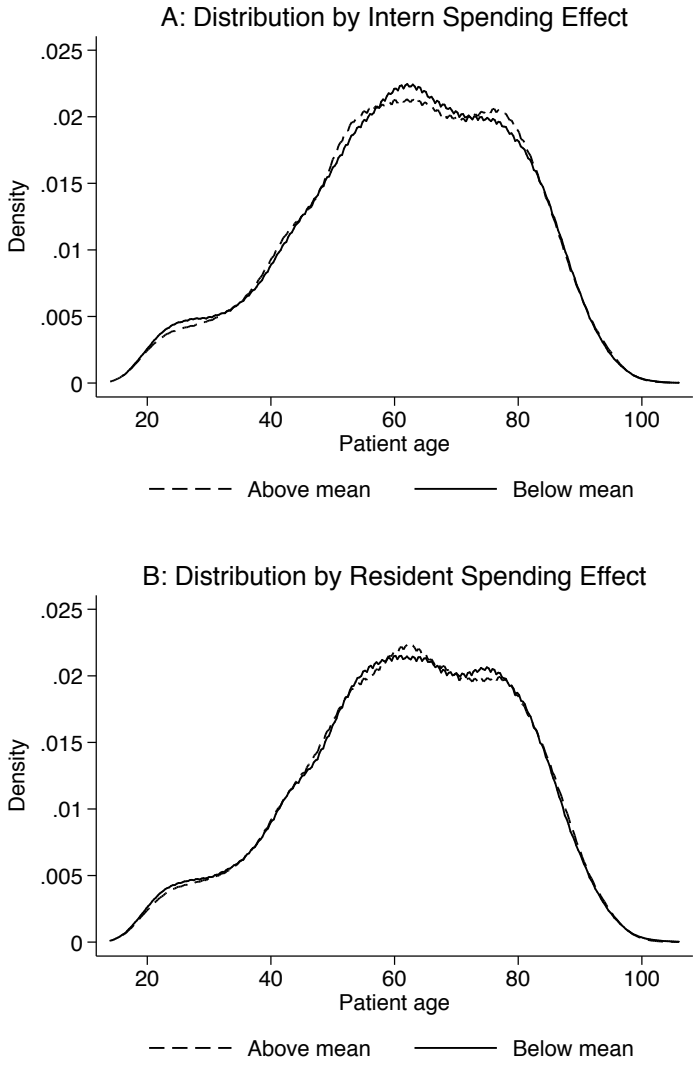
2. For each resident-month-service block of observations in each month t, randomly choose N_t^{treat} blocks and designate observations belonging to these blocks as pseudo-treatment observations.

3. Using pseudo-treatment service observations, estimate the standard deviation in resident spending distribution, as described in Appendix A-3, for each 60-day tenure period within two years of tenure and each 120-day tenure period in the third year.

4. Estimate the rate of convergence by regressing $\hat{\sigma}_{\xi,\tau}$ on the midpoint in days tenure of a tenure period τ (e.g., the first 60-day tenure period has a midpoint of 30 days tenure), for tenure periods after intern year, weighting by the number of patient-days during each tenure period. The yearly rate of convergence is the coefficient on days tenure multiplied by 365.

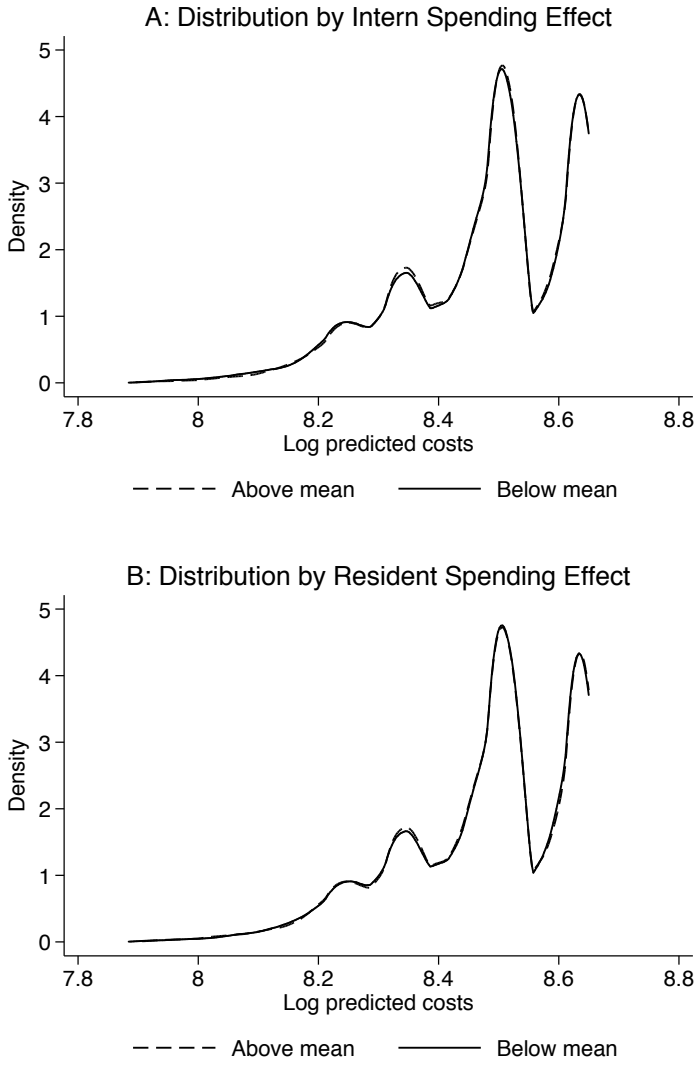
5. Repeat for 10,000 times steps 2 to 4, collecting the yearly rate of convergence for each run. The proportion of placebo runs producing a rate of convergence greater than the true rate of convergence is the p-value under randomization inference.

Figure A-1: Patients Age by Housesetaff Spending Effect



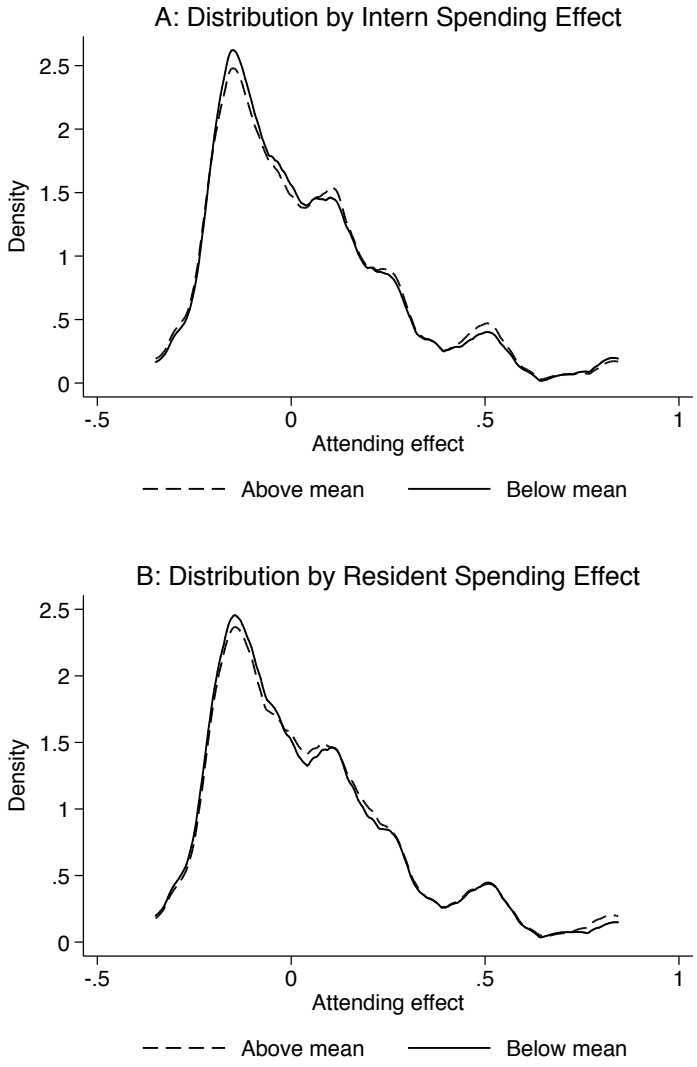
Note: This figure shows the distribution of the age of patients assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed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

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



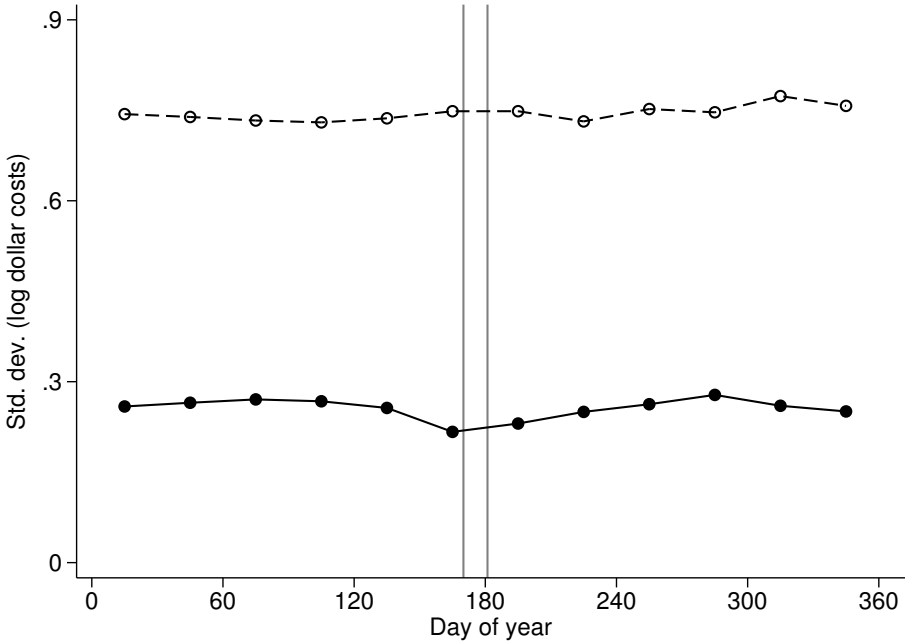
Note: This figure shows the distribution of predicted log costs (based on patient age, race, and gender) for patients assigned interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee spending effects, not conditioning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

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



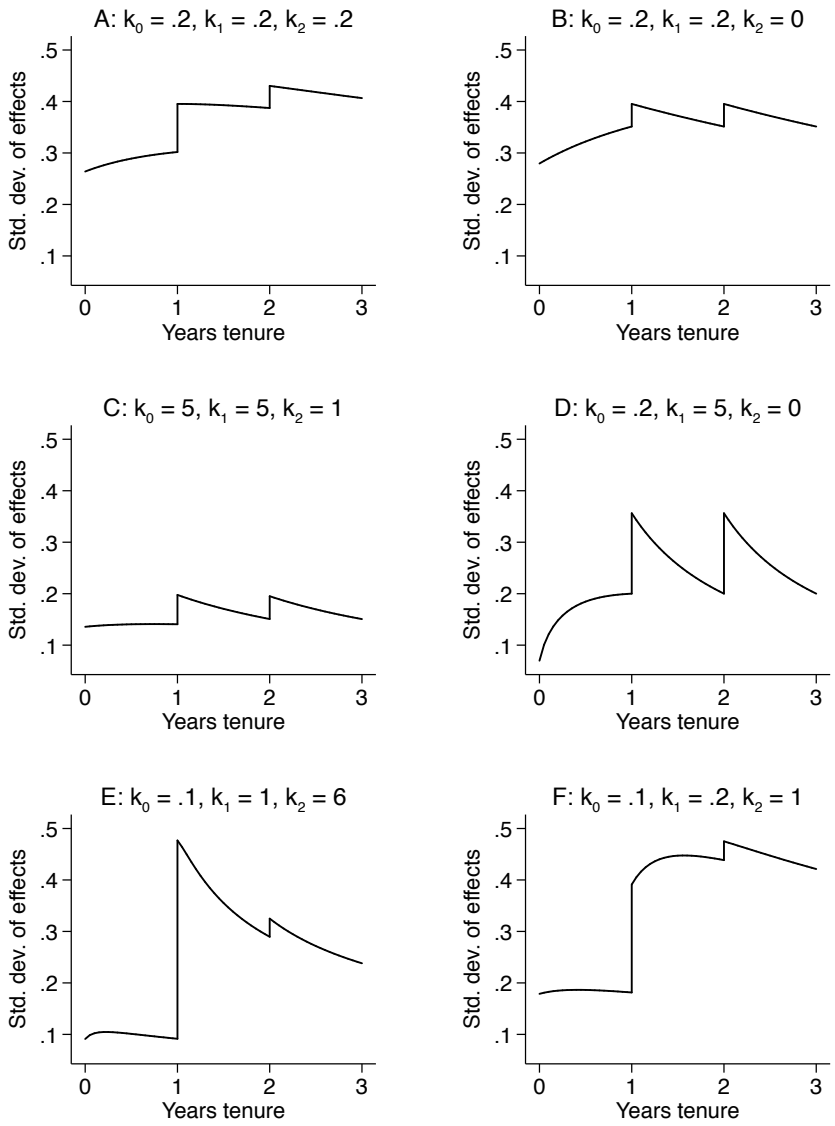
Note: This figure shows the distribution of spending fixed effects for attendings assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Trainee and attending spending effects, not conditioning by tenure, are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

Figure A-4: Trainee-associated and Residual Variation by Day of Year



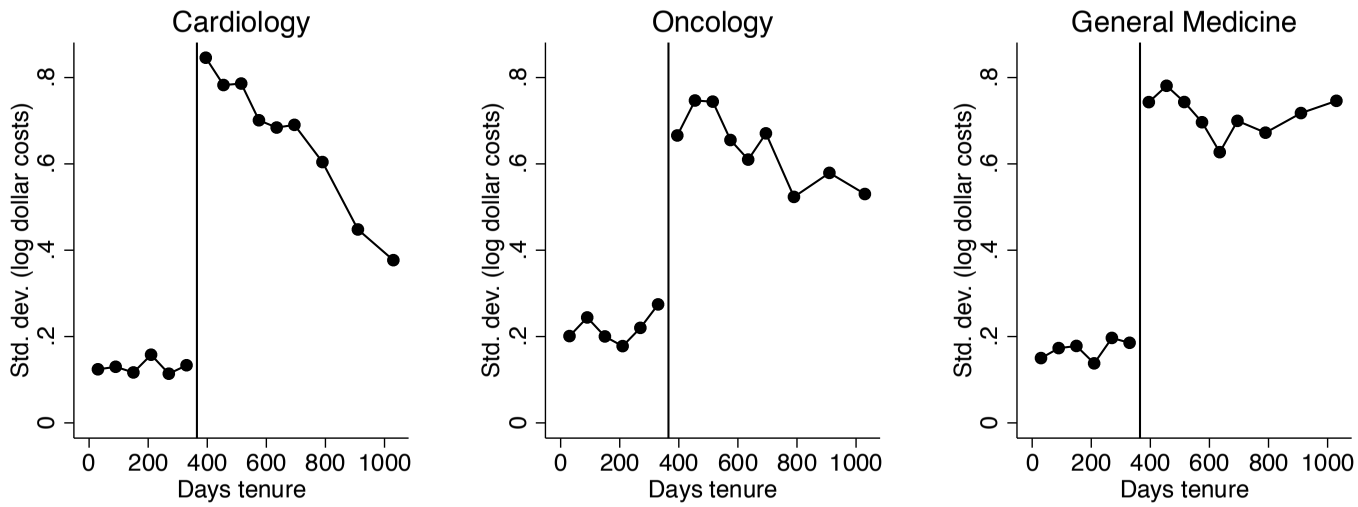
Note: This figure shows the standard deviation of random effects due to junior and senior trainee teams (solid dots) and the standard deviation of the residual (hollow dots) in 30-day periods by day of the year. Residual variation can be interpreted as variation due to independent observation. The two vertical gray lines indicate when new junior trainees begin residency on July 19 and when senior trainees advance a year on July 28 (i.e., becoming a new second-year senior trainee, becoming a third-year trainee, or completing residency). The model is similar to Equation (4), 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: Numerical Examples of Variation Profiles



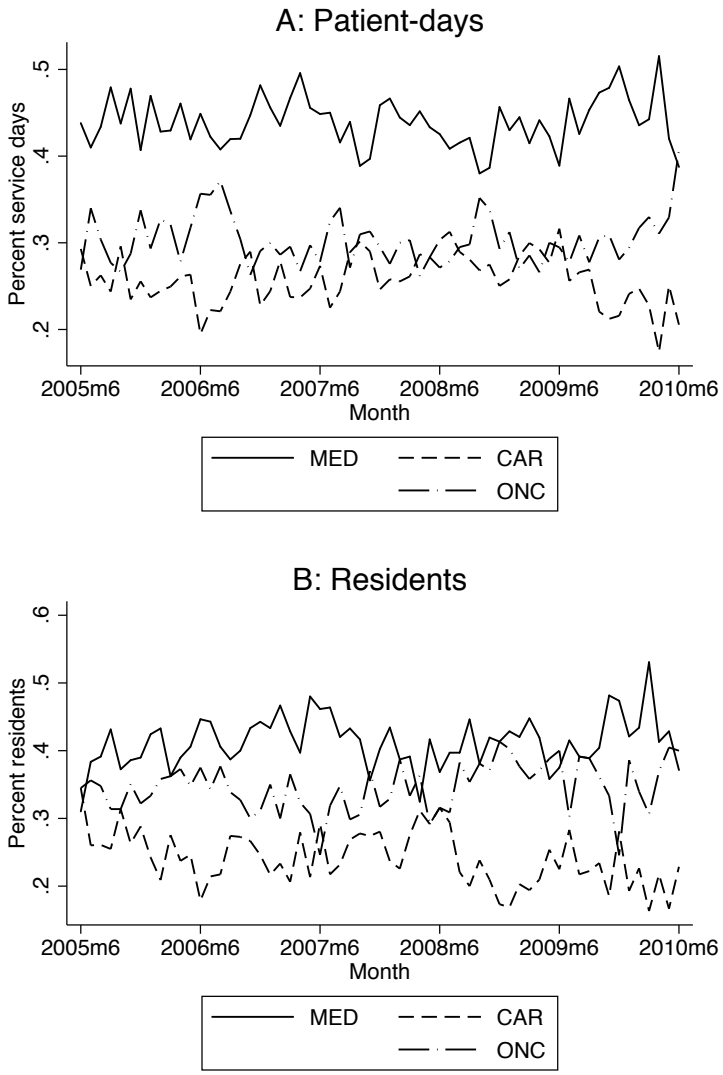
Note: This figure shows variation profiles of the expected standard deviation of trainee effects over tenure, $\sigma(\tau)$, differing by the underlying profile of learning over tenure. Learning is parameterized as a piecewise linear function $g(\tau)$ that describes how the precision of subjective priors increases over tenure. In particular, this figure considers piecewise linear functions of the form (A-7), parameterized by k_0, k_1, and k_2. 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-5). 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-2.

Figure A-6: Practice Variation Profile by Service (Diagnostic Spending)



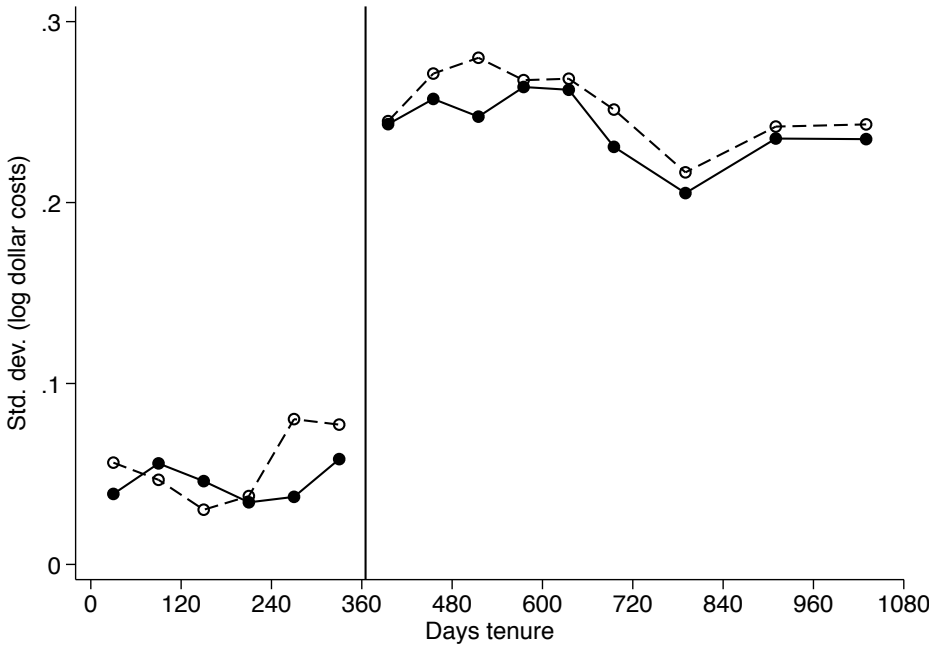
Note: This figure shows the standard deviation in a random effects model, as in Equation (4), of log daily diagnostic costs at each non-overlapping two-month tenure interval. Each panel shows results estimated from within a service of cardiology, oncology, or general medicine. No significant positive 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 A-6 shows the corresponding figure for log daily total costs.

Figure A-7: Service Days and Residents on Ward Services over Time



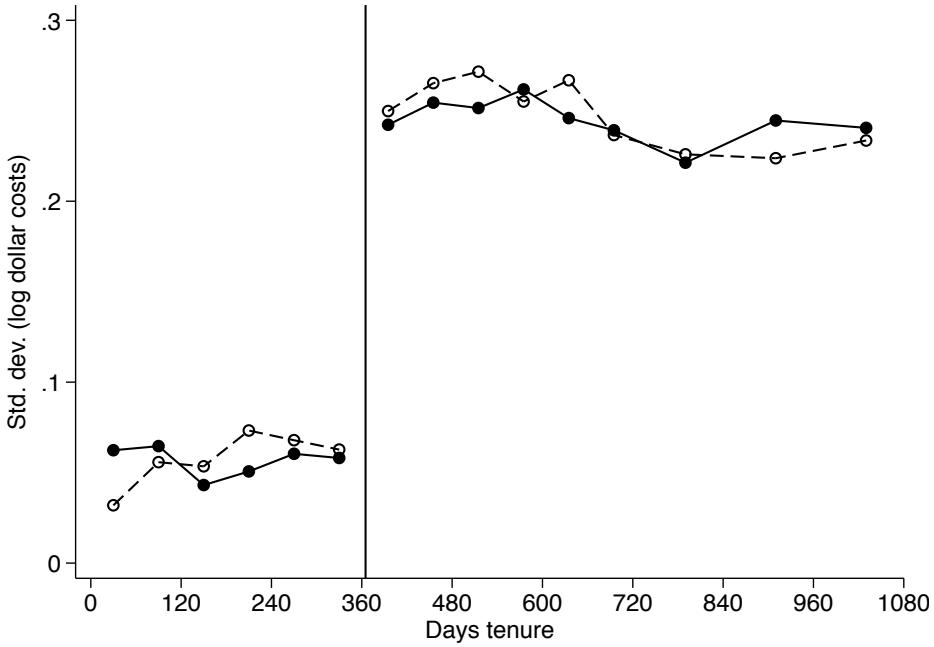
Note: This figure shows the percentage of patient-days (Panel A) and residents on service (Panel B) during each month in the data for each service of general medicine, cardiology, and oncology. Residents may be counted in more than one service if they spent time in more than one service in the same month.

Figure A-8: Practice Variation Profile by ICD-9 Code Frequency



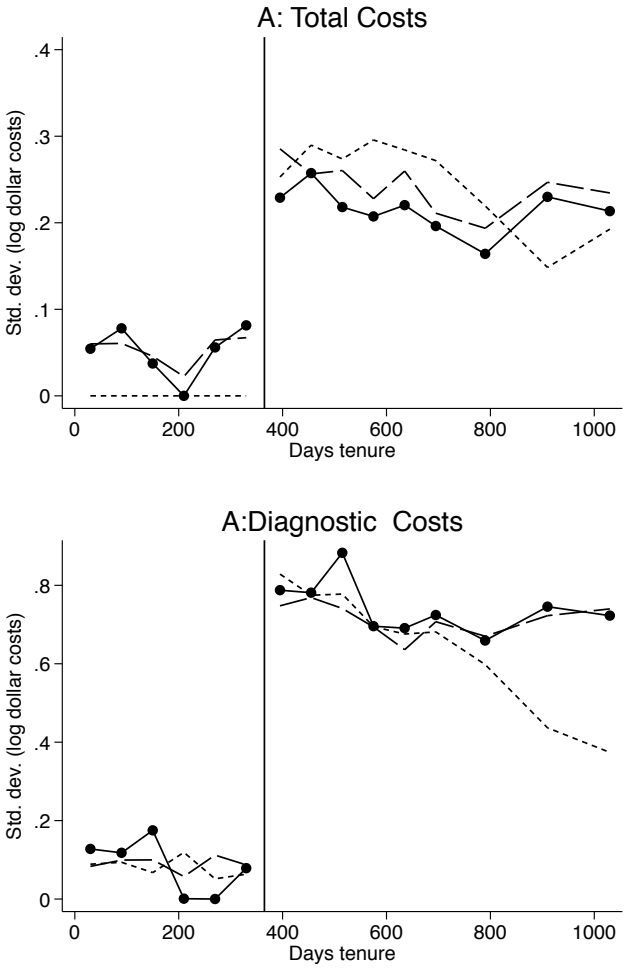
Note: This figure shows the standard deviation in a random effects model, as in Equation (4), of log daily total costs at each non-overlapping tenure interval estimated separately using observations with relatively common ICD-9 diagnostic codes (within service) (solid dots) and those with uncommon diagnoses (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-9: Practice Variation Profile by Guideline Existence



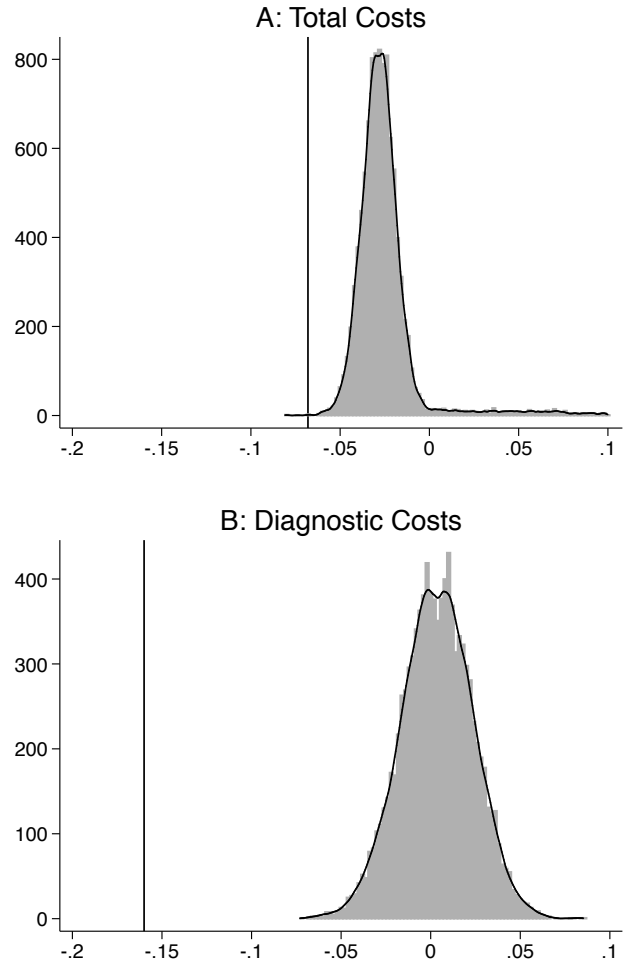
Note: This figure shows the standard deviation in a random effects model, as in Equation (4), of log daily total costs at each non-overlapping tenure interval estimated separately using diagnoses with (solid dots) and those without (hollow dots) published catalogued by the US Agency for Healthcare Research 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-10: Pseudo-cardiology Service



Note: This figure shows the tenure profiles of practice variation in log daily total costs (Panel A) and log daily diagnostic costs (Panel B) of a pseudo-cardiology service by ICD-9 codes. This service is constructed from general medicine observations, matching ICD-9 codes observed in cardiology. This procedure covers 97% of observations in the actual cardiology service. Each panels shows the standard deviation of trainee effects by tenure for actual services of cardiology (short-dashed line) and general medicine (long-dashed line), and for a pseudo-cardiology service (dot and solid line) comprised of patients in general medicine but matching ICD-9 code primary diagnoses in cardiology. Estimation of Equation (4) 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-11: Systematic Placebo Tests for Convergence



Note: This figure shows 10,000 random placebo tests for convergence in trainee effects on log daily total costs in the cardiology service (Panel A) and on log daily diagnostic costs in the cardiology and oncology services combined (Panel B). The vertical line in each panel shows the actual measure of convergence. In Panel A, the measure is -0.068, or a 6.8% percentage point decrease per year in the standard deviation of spending effects of senior trainees on log daily total costs in cardiology. In Panel B, I consider convergence on observations from cardiology and oncology; among these observations, the measure is -0.160, or a 16% percentage point decrease per year in the standard deviation of spending effects of senior trainees. In each placebo test, I randomize observations to a pseudo-cardiology service (Panel A) or a pseudo-specialist service (Panel B), matching the actual number of trainee-months assigned to each service. I then estimate the same random effects model of log daily test costs shown in Equation (4) for the placebo specialist service and estimate the rate of placebo convergence using estimated trainee effects in this placebo specialist service. Corresponding placebo tests are shown in each panel as a histogram with a kernel-smoothed overlay. Fifteen out of 10,000 tests exceeded the actual value of -0.068 in Panel A (p-value = 0.0015); no placebo test exceeded the actual value of -0.160 in Panel B (p-value = 0). Details are given in Appendix A-4.

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

	Independent variables		
	Trainee identities	Trainee characteristics	
Patient characteristic	(1)	(2)	(3)
Age	$F(1055, 46364) = 0.98$ $p = 0.655$	$F(20, 16069) = 0.68$ $p = 0.848$	$F(18, 37494) = 0.79$ $p = 0.711$
Male	$F(1055, 46364) = 1.01$ $p = 0.389$	$F(20, 16069) = 1.18$ $p = 0.256$	$F(18, 37494) = 1.26$ $p = 0.201$
White	$F(1055, 46364) = 1.02$ $p = 0.356$	$F(20, 16069) = 0.79$ $p = 0.734$	$F(18, 37494) = 0.92$ $p = 0.558$
Predicted spending	$F(1055, 46364) = 0.98$ $p = 0.685$	$F(20, 16069) = 0.79$ $p = 0.731$	$F(18, 37494) = 1.08$ $p = 0.368$

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

Table A-2: Core Rotations for Most Recognized Internal Medicine Residencies

	Ward rotations							
Residency program	MED	CAR	ONC	GI	PULM	RENAL	ID	RHEUM
Massachusetts General Hospital	✓							
Johns Hopkins University	✓	✓	✓					
Brigham and Women's Hospital	✓	✓	✓					
University of California, San Francisco	✓	✓		✓				
Mayo Clinic	✓	✓	✓	✓	✓	✓	✓	
Duke University Hospital	✓	✓	✓		✓			
Washington University	✓	✓	✓					
University of Pennsylvania	✓	✓	✓					
New York Presbyterian (Columbia)	✓	✓	✓				✓	
Northwestern University	✓	✓	✓	✓				
University of Michigan	✓	✓	✓	✓	✓			
University of Washington	✓	✓	✓					
University of Texas Southwestern	✓	✓						
Cleveland Clinic	✓	✓	✓	✓		✓		
Mount Sinai Hospital	✓							
Stanford University	✓	✓	✓					
Vanderbilt University	✓	✓	✓					
New York Presbyterian (Cornell)	✓	✓	✓					✓
University of Chicago	✓	✓	✓					
Emory University	✓	✓						
UCLA Medical Center	✓	✓	✓					
Beth Israel Deaconess Medical Center	✓	✓	✓					
Yale-New Haven Medical Center	✓	✓	✓	✓		✓	✓	
New York University	✓	✓	✓					
Total Counts (out of 24)	24	22	19	6	3	3	3	1

Note: This table shows core ward organ-based medical rotations for the 24 highly recognized internal medicine residency programs reported by *US News & World Report*, ordered by nominations in a survey of internists and residency program directors. The identities of core rotations were obtained by browsing each residency program's website. Abbreviations: general medicine (MED), cardiology (CAR), hematology/oncology (ONC), gastroenterology (including liver) (GI), pulmonary (PULM), nephrology (RENAL), infectious disease (ID), and rheumatology (RHEUM). I exclude rotations in palliative care and geriatrics, as these are not traditional organ-based subspecialties, and in neurology, as it is a specialty outside of internal medicine. Total counts are shown in the last row.

Table A-3: Core Rotations in Universe of Internal Medicine Residencies

Ward Rotations	Program count
General Medicine (MED)	310
Cardiology (CAR)	131
Hematology / Oncology (ONC)	85
Nephrology (RENAL)	34
Gastroenterology, including Hepatology (GI)	28
Pulmonology (PULM)	27
Infectious Disease (ID)	22
Rheumatology (RHEUM)	7
Endocrinology (ENDO)	3

Note: This table shows core ward medical rotations in the universe of internal medicine residency programs accredited by the American Council for Graduate Medical Education (ACGME), accessed at `www.acgme.org`. Of the 345 programs listed in the website, 310 programs had curricula detailing core ward rotations. Core ward rotations are defined as required rotations on ward services.

Table A-4: *New England Journal of Medicine* Research Articles by Specialty

Specialty / subspecialty	Internal medicine	Article count
Hematology / Oncology	Y	596
Cardiology	Y	562
Genetics	N	476
Infectious Disease	Y	453
Pulmonary / Critical Care	Y	329
Pediatrics	N	285
Endocrinology	Y	283
Gastroenterology	Y	257
Neurology / Neurosurgery	N	245
Surgery	N	228
Primary Care / Hospitalist	Y	179
Nephrology	Y	158

Note: This table reports the number of research papers appearing in the last ten years in the *New England Journal of Medicine*, by specialty or subspecialty as categorized by the journal. Specialties or subspecialties are also categorized as being within internal medicine or not. A training path in clinical genetics is possible from internal medicine, but genetics can also be pursued from pediatrics, obstetrics-gynecology, and other specialties. The *New England Journal of Medicine* has the highest impact factor, 51.7, out of all medical journals; only five other medical journals have double-digit impact factors, with the second-highest of 39.1 belonging to the *Lancet*, and the third-highest of 30.0 belonging to the *Journal of the American Medical Association*. Articles counted as research papers are "scientific reports of the results of original clinical research." Other categories, as defined at http://www.nejm.org/page/author-center/article-types, include reviews, clinical cases, perspective, commentary, and other.

Table A-5: Research Funding by National Institutes of Health (NIH) Institute or Center

NIH Institute or Center	Grants open	Funding (millions)
National Cancer Institute (NCI)	9,872	$6,670
National Institute of Allergy and Infectious Diseases (NIAID)	7,271	$5,433
National Heart, Lung, and Blood Institute (NHLBI)	6,294	$3,591
National Institute of General Medical Sciences (NIGMS)	6,268	$2,614
National Institute of Diabetes and Digestive And Kidney Diseases (NIDDK)	4,971	$2,397
Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD)	3,295	$1,814
National Institute of Neurological Disorders And Stroke (NINDS)	4,639	$1,753
National Institute of Mental Health (NIMH)	3,650	$1,500
National Institute on Drug Abuse (NIDA)	2,809	$1,229
National Institute on Aging (NIA)	2,749	$1,220
National Institute of Environmental Health Sciences (NIEHS)	1,504	$1,091
Office of the Director (OD)	820	$756
National Eye Institute (NEI)	1,798	$733
National Human Genome Research Institute (NHGRI)	623	$627
13 Other Institutes and Centers	8,564	$4,259

Note: This table lists the top fourteen Institutes and Centers of the National Institutes of Health (NIH), ordered by current funding as defined by funds to currently open grants. Grants open and current funding (in millions of dollars) are both listed. For brevity, the thirteen other Institutes and Centers are not listed individually but are aggregated in the last line.

Table A-6: Top Diagnostic Codes by Service

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

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