NBER WORKING PAPER SERIES

INFORMATIONAL FRICTIONS AND PRACTICE VARIATION: EVIDENCE FROM PHYSICIANS IN TRAINING

David C. Chan, Jr

Working Paper 21855
http://www.nber.org/papers/w21855

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2016

Previously circulated as "Uncertainty, Tacit Knowledge, and Practice Variation: Evidence from Physicians in Training." 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, Emir Kamenica, Pat Kline, Jon Kolstad, Eddie Lazear, Frank Levy, Grant Miller, David Molitor, Jon Skinner, Doug Staiger, 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. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Informational Frictions and Practice Variation: Evidence from Physicians in Training
David C. Chan, Jr
NBER Working Paper No. 21855
January 2016, Revised April 2016
JEL No. D20,D83,I10,L23,L84,M11,M53,M54

ABSTRACT

Substantial practice variation across physicians for seemingly similar patients is an unresolved puzzle of great interest to economists and policy makers. This paper studies physicians in training to explore the behavioral foundations of practice variation. A discontinuity in the formation of teams reveals a large role for relative experience in the size of practice variation. Among the same physician trainees, convergence occurs in services driven by specialists, where there is arguably more explicit knowledge, but not on the general medicine service. Similarly, rich physician characteristics correlated with preferences and ability explain little if any variation.

David C. Chan, Jr
Center for Health Policy and
Center for Primary Care and Outcomes Research
117 Encina Commons
Stanford, CA 94305
and NBER
david.c.chan@stanford.edu

1 Introduction

There exists wide variation in health care utilization across the United States. The extensive and influential literature documenting this variation has largely focused on how care varies across regions, understandably because of unobserved patient sorting within regions to hospitals and physicians.[1] However, very little is known about the behavioral foundations of variation from individual agents. Using quasi-random assignment of patients to teams of physicians in training ("housestaff") at a large academic hospital, this paper aims to describe the evolution of practice variation among internal medicine housestaff in order to shed light on the broader behavioral foundations underlying practice variation.

While this study is necessarily limited to within an organization, it advances our understanding of practice variation in health care in two important ways. First, while the state of the art in the area-level variations literature is making progress in separating causes into two broad categories of patient "demand" factors and provider "supply" factors (Finkelstein et al., 2016), the empirical design in this paper holds fixed demand factors (by patient random assignment) as well as most supply factors, including financial incentives, capital endowments, market structure, and institutional rules and culture. This allows me to focus on variation across agents within an institution, which is intuitively important but not well understood. Second, residency training is an advantageous setting to study the dynamics of physician practice variation, as it is perhaps the most formative period in physicians' careers, when trainees begin with very little prior clinical experience and engage in intensive training in a controlled environment. Detailed patient-care and administrative data permit me to track variation attributable to individual housestaff, on a daily basis, as they transition to different team roles and rotate among different practice environments.

I first demonstrate clinically significant variation attributable to housestaff who are quasi-randomly assigned patients. Assignment of patients to housestaff within the same organization has causal effects on daily total spending, daily test spending, length of stay, and 30-day readmissions, and even 30-day mortality. While much of the regional variations literature has found

[1] See, e.g., Skinner (2012) for a recent review of the literature, which dates at least to Wennberg and Gittelsohn (1973).

no correlation between spending and outcomes, I find that housestaff who spend more have better outcomes, consistent with more recent evidence by Doyle Jr et al. (2015), which exploits random assignment to hospitals by ambulances. Although associations between spending and mortality are imprecise, reassigning patients from 10th-percentile to 90th-percentile housestaff in the spending distribution would lower readmissions by a 2.5 percentage points, eliminating about a fifth of readmissions.

Next, in a stylized framework with learning and team decisions, I conceptually show that practice variation does not necessarily decrease monotonically with experience. Rather, as physicians gain more precise beliefs about practicing medicine, they may also gain greater influence so that their beliefs count more in team decisions. I empirically examine this concept of influence within teams by exploiting a mechanical discontinuity in housestaff roles: Since patients are cared for by a team comprised of a first-year "intern" and a second- or third-year "resident," the relative experience of a housestaff changes discontinuously across the one-year mark. This separates the effect of influence from time-varying but plausibly continuous characteristics of the index housestaff, such as skills and cumulative learning. The standard deviation of spending effects across housestaff discontinuously increases from approximately a 20% difference in costs among year-end interns to a 70% difference in costs among beginning residents.

Further, I study the evolution of practice variation among the same set of housestaff but in different practice environments. Residents converge when practicing in specialist-driven services – cardiology and oncology – eliminating much of the variation by the end of residency, while the same residents show no convergence when practicing in general medicine. I argue that this division into specialist and generalist services represents a meaningful difference in the existence and use of explicit medical knowledge, but at a minimum this finding shows that the same housestaff with the same average training experiences develop different degrees of practice variation depending on the environment. This difference between specialist- and generalist-driven services is highly significant with systematic placebo tests randomizing service-block identities, and strikingly, this difference holds when matching patients based on formally coded diagnoses.

Finally, as a benchmark to the above results, I quantify practice variation predictable by housestaff characteristics and prior observable training experiences, using the following: (a)

detailed housestaff characteristics from confidential residency selection and administrative data (e.g., test scores, rank-list positions, honors); (b) precommitted career choices in which housestaff with different future plans are required to have the same initial training in internal medicine; and (c) histories of training experiences (e.g., whether the housestaff recently trained with a high-spending supervising physician) that are as good as randomly assigned. Housestaff characteristics predict in aggregate only a small portion of the large underlying spending variation. When using LASSO to avoid overfitting due to the large number of characteristics relative to the number of housestaff, the sole predictive characteristic is male sex, which predicts 4% less spending among residents, whereas the standard deviation of practice variation among residents is at least 15 times greater. Similarly, housestaff tenure does not significantly shift mean levels of spending or other outcomes (e.g., readmission and mortality), and a wide variety of training experiences have no true predictive power.

Together, these findings are most supportive of informational frictions as a major mechanism sustaining practice variation, in the following sense: The importance of relative influence suggests that knowledge is slowly gained and not easily passed from senior agents to junior agents.[2] Lack of convergence also suggests that the standard of practice is insufficiently defined or communicable in many cases, despite a setting with intensive training and close supervision. In contrast, intrinsic heterogeneity across physicians is unlikely to play a large role, to the extent that any of it is correlated to detailed physician characteristics related to skill and preferences. These findings, based on 3.2 million orders tracking physicians as they train, provide complementary evidence to surveyed physician beliefs (Cutler et al., 2013), especially when much of what physicians do may arise from knowledge, beliefs, or habits that are not easily elicited by asking them.

At least since Arrow (1963), the informational frictions in medical care have been well-known, and at least since Polanyi (1966), tacit knowledge – "knowledge," possibly including beliefs or habits, that is difficult to communicate – has been considered a significant barrier in the standardization of decisions and routines across workers (Nelson and Winter, 1982; Autor et al., 2003). However, tacit knowledge has not received much consideration in the economics literature

[2]In Appendix A-3, I discuss how this idea is consistent with possibility of social authority or hierarchy. Authority as a function of knowledge, even when used to describe a profession independent of any individual's knowledge, is explored in detail by Starr (2008), who documents the rise of the medical profession as specialized scientific knowledge expanded.

as a behavioral foundation for practice variation in health care, possibly because it is inherently difficult to measure and more generally due to a lack of micro-level data following significant variation in a controlled environment.[3] While differences in physician characteristics and training experiences may be larger across institutions, if informational frictions sustain variation within a set of housestaff under intense training in a single organization, such frictions are also likely to be larger across institutions.

The remaining organization of this paper is as follows. Section 2 describes the institutional setting; Section 3 describes the data. Section 4 presents a first look at meaningful variation across housestaff in several outcomes. Section 5 discusses variation across the discontinuity of relative experience, and Section 6 discusses convergence (or the lack thereof) in different environments. Section 7 describes results on the predictive power of housestaff characteristics and experience on outcomes. Section 8 discusses policy implications and concludes.

2 Institutional Setting

2.1 Medical Care by Physicians in Training

Since the Flexner Report in 1910, medical training has largely become standardized across the US (Flexner, 1910; Cooke et al., 2006). Each patient is cared for by a team of a first-year housestaff ("intern") and a second- or third-year housestaff ("resident"). Residents are usually assigned to two interns at a time and therefore are responsible for twice the number of patients. As a result, span of control considerations argue for *more* control by interns over their patients than residents do, as interns can devote more attention to each patient. There are no other formal distinctions in decision rights or job responsibilities between interns and residents, including legal or regulatory ones, but residents are expected to know more and often engage in higher-level decision-making in patient care. These housestaff teams are supervised by "attending" physicians and operate within a broader practice environment, which includes other health care

[3]See Cutler (2010) and Skinner (2012) for thoughtful reviews on potential causes of practice variation. Much of the conventional wisdom focuses on a lack of competition across firms, due for instance to health insurance and lack of quality measurement. The evidence in this paper would suggest that the issue goes beyond measuring "quality" and that for large areas of medicine, there is a lack of agreement on what constitutes best practices for a given patient.

workers (e.g., consulting physicians, pharmacists, and nurses), as well as institutional rules for deciding and implementing care.

Housestaff from different programs and different "tracks" within a program 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 internal medicine residency. Other residency programs include another internal-medicine residency from a different hospital, as well as obstetrics-gynecology and emergency medicine from the same hospital.

Housestaff 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 housestaff characteristics, although housestaff on certain tracks may be unavailable during certain times due to programmatic differences.[5] Scheduling does not consider the teams of intern, resident, and attending physicians that will be formed as a result. In fact, attending schedules are done independently, and neither housestaff 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 housestaff 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 housestaff changing rotations. When a housestaff replaces another one, she assumes the care of the entire list of patients from the other housestaff. Because housestaff blocks are generally two weeks in length and staggered for interns and residents, it is not uncommon for a patient to experience a change in either an intern or a resident.

[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.

[5]Housestaff are allowed to express preferences about vacation days, although these vacation days are few, about two weeks per year. Senior residents (third-year residents) may also express more general preferences about the timing of non-clinical blocks, such as research electives. For interns, schedules are assigned even prior to their arrival from medical school.

2.2 Medical Knowledge

Inpatient medical care is comprised of three services at this institution: cardiology, oncology, and general medicine. This organization represents the most common configuration of inpatient care across academic hospitals in the US. Of the 24 residency programs ranked by *US News & World Report* and shown in Table A-2, 22 and 19 programs have dedicated cardiology and oncology services, respectively. Gastroenterology, represented at 6 programs, is the next most common subspecialty service. A similar relationship among subspecialties exists in the universe of internal medicine programs recognized by ACGME (Table A-3). Specialist-driven services by definition are staffed by specialist attending physicians, who have several more years of training after internal medicine. In contrast, generalists are responsible for patients on general medicine services, who may optionally consult a specialist if they deem it helpful.

In recent decades, by important measures, medical knowledge has progressed in cardiology and oncology to a greater extent than for other diseases.[6] Table A-4 shows the number of original research articles appearing in the *New England Journal of Medicine* in the last ten years according to key disease specialty or subspecialty. Oncology and cardiology research papers are the most numerous by a substantial margin. Table A-5 reports current research funding by National Institute of Health (NIH) Institute or Center. Although Institutes often lump disease categories, the National Cancer Institute (NCI) with current funding of $6.7 billion and the National Heart, Lung, and Blood Institute (NHBLI) with current funding of $3.6 billion occupy the first and third positions for funding out of a list of 27 Institutes and Centers.

In this sense cardiology and oncology care have stronger "best practices" than other subspecialties of internal medicine. Differences in best practices can affect variation in two ways. First, strong best practices, embedded in attending physicians, ancillary staff, and institutional rules, constrain variation in housestaff decisions even if these housestaff have not yet fully internalized all information available at the institution. Second, environments with stronger best practices will be more conducive to learning. The fact that physicians need further subspecialty training to assume primary responsibility for cardiology and oncology patients, but not to treat pneumonia,

[6]The production and use of knowledge is in turn driven by government, academic, and industry priorities. For example, in some locations and in the past, tuberculosis wards were common but cease to exist today.

may also reflect a larger body of knowledge used to care for these patients.[7]

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 housestaff, 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.[8]

Detailed residency application information for each housestaff 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 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 housestaff 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.

I use scheduling data and past matches with supervising attending physicians and other housestaff to impute housesetaff experience over time. As described in Section 2, housestaff 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

[7]Several observers have noted that the increasing length of training in medicine seems to related to the growing role of scientific knowledge and technology in medicine (Ludmerer, 1988; Starr, 2008). For example, prior to the beginning of the 20th century, practitioners could become doctors in a matter of weeks. Although the first American residency was started at Johns Hopkins in 1889, teaching hospitals and residencies only grew to prominence in the 1920s, when sufficient technological knowledge (and, to a degree, urbanization) shifted care from patients' homes to hospitals.

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

types of patients and are equally likely to be assigned to high- or low-spending coworkers and attendings. In Appendix A-1, I present more formal analyses on conditional random assignment of housestaff physicians; I cannot reject the null that housestaff identities are jointly unrelated to patients types or other training experiences.

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 that are admitted for similar reasons (e.g., heart failure).[9]

For each patient-day, I observe total cost information, aggregated within 30 billing departments such as blood bank, various laboratory, nursing, nutrition, pharmacy, physical therapy, radiology. I also observed more detailed cost information specific to each of 3.2 million physician orders in laboratory and radiology (e.g., CT, MRI, nuclear medicine, ultrasound). Admission and discharge data allow me to impute length of stay in days and readmission rates. Finally, dates of death are provided via linkages with social security vital statistics data. While I study variation across housestaff in each of the outcome measures of daily total costs, daily test costs, length of stay, readmissions, and mortality in Section 4, for most of the paper I focus on daily test costs (a) because they are most closely controlled by housestaff, (b) because daily outcomes linked to a large number of physician orders allow for greater precision in measuring variation, and (c) because they capture variation in diagnostic approaches, the heart of medical decision-making, including in situations in which very little is known about patients.[10] The distribution of daily test costs is heavily right-skewed. I censor daily test cost observations greater than $800,

[9]These admitting services are more narrowly defined than the broad categories of cardiology, oncology, and general medicine. However, even within specific admitting service, attendings may have different types of patients (e.g., a vertically integrated HMO admits to the same service as the hospital's own attendings). Therefore, without hand-coding attendings to practice groups and conditioning on these groups, patients are not quasi-randomly assigned to attendings. Still, as described above, housestaff are quasi-randomly assigned to patients, other housestaff, and attendings.

[10]Medical spending has been the focus of much of the literature on practice variation (Fisher et al., 2003a,b) and is a key policy focus in its own right (Anderson et al., 2005). Test spending has particularly received increasing attention as the relative cost of tests has risen and now comprises a significant proportion of overall costs (Schroeder et al., 1974; Iwashyna et al., 2011).

which comprise 3% of the data; the resulting distribution is shown in Figure A-5.[11] The mean daily test cost is $124, while the median is $49 and the 90th percentile is $337. These daily costs aggregate to overall admission tests costs with a mean of $714.

4 Variation across Housestaff

I first examine variation attributable to housestaff, specifically residents, in each of the following outcomes: log total spending on day of admission, log test spending on day of admission, log length of stay, 30-day readmission, and 30-day mortality. For admission a at time t, associated with resident j and attending k, I specify outcome Y_{ajkt} based on patient and admission characteristics $\mathbf{X}_a$ (see Section 3), a set of time categories $\mathbf{T}_t$ for month-year combination and day of the week, and resident and attending identities:

$$Y_{ajkt} = g\left(P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)\right) + \xi_j + \varepsilon_{ajkt}, \tag{1}$$

where $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$ is a linear projection of outcomes Y_{ajkt} onto $\mathbf{X}_a$, $\mathbf{T}_t$, and k, using only within-housestaff variation, $g\left(\cdot\right)$ is a potentially flexible transformation of the projection, ξ_j is a resident random effect possibly correlated with $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$, and ε_{ajkt} is an error term that is normal for log spending and log length of stay or logistic for binary outcomes of readmission and mortality. For binary outcomes, Y_{ajkt} is a latent variable that determines the observed binary outcome $\tilde{Y}_{ajkt} = \mathbf{1}\left(Y_{ajkt} > 0\right)$. As a simple benchmark of variation in different outcomes, this exercise treats practice variation as fixed over time within residents, in contrast to the fuller statistical model in Sections 5 and 6 that nonparametrically allows for drift over time.

Although patients are *conditionally* as good as randomly assigned to housestaff (Appendix A-1), random assignment does not hold across time categories and admitting services. This motivates a treatment of resident effects that allows for correlation with $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$. I thus separate ξ_j into a correlated component u_j and an uncorrelated component v_j:

$$\xi_j = u_j + v_j,$$

[11] Results in this paper are robust to this censoring.

where $\text{Corr}\left(u_j, P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)\right) \neq 0$ and $\text{Corr}\left(v_j, P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)\right) = 0$. I restrict comparisons across housestaff to the uncorrelated component, v_j, for two reasons. First, comparing housestaff with different average $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$ is less likely to be valid because patients are not randomly assigned between these housestaff. Second, u_j is mechanically correlated across different outcome measures by the correlation between u_j and patient observed and unobserved characteristics captured by $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$.[12] Given this setup, which I describe in more detail in Appendix A-2, I calculate empirical Bayes predictions $\hat{v}_j$ (Searle et al., 1992) for each resident in each of the outcome measures.

Figure 1 shows distributions of resident effects in each outcome measure. Within the same institution and set of housestaff, reassigning patients from the 10th-percentile- to the 90th-percentile-$\hat{v}_j$ resident (among residents exposed to patients with the same $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$ on average) increases respective outcomes of total spending on admission, test spending on admission, and length of stay by about 20%. In dollar terms, respective reassignment from the 10th to 90th percentile resident increases total spending on admission from \$1,022 to \$1,245, and test spending on admission from \$135 to \$164.[13] Reassignment according to length of stay increases length of stay from 3.64 days to 4.35 days. Reassignment from the 10th to 90th percentile according to 30-day readmissions and 30-day mortality increases these events from 9.6% to 16.4%, and from 5.1% to 10.3%, respectively.

Figure 2 shows correlations between resident effects in spending and clinical outcomes. Consistent with Doyle Jr et al. (2015), which measures correlations between hospital effects on spending and outcomes, identified by random arrivals of patients by ambulance to hospitals, I find a similar positive relationship between spending and clinical outcomes across residents. Residents who spend more either in total or by testing have fewer 30-day readmissions. Reassigning

[12] Standard approaches (e.g., Abowd et al., 2008) to correlated random effects will mechanically assign u_j based on $P_Y\left(\mathbf{X}_a, \mathbf{T}_t, k\right)$ and therefore induce correlation in u_j between different outcome measures driven by projections of patient types onto outcomes. For example, patients in the bone marrow transplant service are more resource-intensive patients and also more likely to die. Comparing housestaff who spent more and less time on the bone marrow transplant service would thus induce a correlation in u_j between spending and mortality.

[13] Variation in spending on admission is less than variation in spending on all days, even when controlling for day of stay. This indicates that spending on the first day, even though it is generally higher, is less variable in log terms. I show results for variation in daily spending in Table 4 (Section 7), although that table and Figure 1 are not directly comparable since the former states variation in terms of the standard deviation of the directly estimated random effect distribution while the latter states variation in terms of shrunken empirical Bayes predictions $\hat{v}_j$. I focus on variation on day of admission in this section in order to assess the relationship between resident effects on admission-level outcomes and daily-level outcomes.

patients from the 10th to 90th percentile of $\hat{v}_j$ according to total spending would reduce readmissions by 2.5 percentage points, or eliminate about a fifth of readmissions. The relationship between spending and mortality is slightly negative but small and statistically insignificant.

5 Influence

This section examines the effect of relative influence in teams on the variation of housestaff effects. In Appendix A-3, I introduce a simple conceptual framework to consider decision-making in teams under uncertainty. While details are in the appendix, the intuition is straightforward: in a team environment under uncertainty, decisions will aggregate information from different agents' prior beliefs, and this aggregation depends on both the means of the priors as well as *relative* precisions of the priors. While agents learning from common data and underlying truth should converge to the same practices when making decisions individually, the information-aggregation feature of team decisions, which can be thought of as a foundation for *influence*, gives rise to the possibility that an agent's effect on variation may increase even as she learns.

To estimate the importance of influence in practice variation in a team environment, I exploit the discontinuous change in roles at the end of the first year of training. In particular, a housestaff near the end of the first year still has at least one year less experience than the other teammate, while the same housestaff at the beginning of the second year has one year *more* experience than the other teammate. This allows me to focus on a discontinuous change in *relative influence*, while holding everything else fixed about the index housestaff that is plausibly continuous. The institutional setting of residency has the dual advantages of no differences in formal roles that mechanically increase resident influence and no unobserved selection into senior roles.[14] Nevertheless, the "influence effect" may still be considered a reduced-form combination of (a) true differences in the relative quality of information for a given housestaff at the discontinuity and (b) perceived differences that can include things like hierarchy and prestige due to seniority in medicine. As discussed in Appendix A-3, the latter phenomenon of hierarchy can

[14]Moreover, as mentioned in Section 2, two interns are usually assigned to a resident, and as a result, interns have more per patient clinical interactions and greater control over orders. These institutional facts suggest that, if information were equal, interns should have *more* influence than residents in the care of a given patient. As such, an observed increase in influence at the first-year mark may be viewed as a lower bound of the effect of more precise information on influence.

be thought of as consistent with the framework of team decision-making under uncertainty, in which informational frictions prevent agents from objectively knowing or even communicating each other's true informational content.

For a patient being treated on day t of admission a by intern i, resident j, and attending k, I specify log daily test costs as

$$Y_{aijkt} = \mathbf{X}_a\beta + \mathbf{T}_t\eta + \xi_i^{\tau(i,t)} + \xi_j^{\tau(j,t)} + \zeta_k + \nu_a + \varepsilon_{aijkt}. \tag{2}$$

Equation (2) includes patient and admissions characteristics $\mathbf{X}_a$, 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. I allow for attending fixed effects, ζ_k.[15]

The parameters of interest in Equation (8) characterize distributions of time-varying random effects, $\xi_i^{\tau(i,t)}$ and $\xi_j^{\tau(j,t)}$ for intern i and resident j, respectively, at discrete tenure interval $\tau(\cdot, t)$ that is function of the housestaff and time. $\xi_i^{\tau(i,t)}$ and $\xi_j^{\tau(j,t)}$ is constant within each tenure interval and housestaff, but for this analysis I impose no structure across tenure intervals for the same housestaff. As described in Appendix A-5, I employ a method akin to restricted maximum likelihood (REML) and similar to an approach by Chetty et al. (2014) that allows random effects to be correlated with fixed covariates without further modeling of the correlation. Tenure-specific standard deviations of $\xi_{h\in\{i,j\}}^{\tau(\cdot)}$ are then directly and jointly estimated by maximum likelihood. These empirical estimates of practice variation are unbiased even in finite samples.[16] Finally, in some specifications I allow for shocks at the admission level, ν_a, allowing some patients, even controlling for patient observables, to randomly result in more test costs than others.

Figure 3 presents results for the estimated standard deviations of the distributions of housestaff effects within each tenure interval τ. In my baseline specification, I consider non-overlapping

[15] 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, 2011). I do not focus on practice variation among attending physicians for two practical reasons: First, unlike housestaff physicians, they are not randomly assigned patients. Second, they are only variably observed in the data, with many attendings working only a few weeks a year.

[16] It is well-known that in finite samples fixed effect estimates of $\xi_{h\in\{i,j\}}^{\tau(\cdot)}$ would include measurement error and therefore would have a distribution with greater variance than the underlying distribution of true effects. However, because I necessarily specify two sets of effects, one for the intern and the other for the resident, there are two complications to the standard Bayesian shrinkage procedure (e.g., Morris, 1983) which result in biased estimates of the distribution that I confirm in simulations. This is discussed further in Appendix A-5.

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 housestaff have fewer inpatient days.[17] I find large and significant variation in housestaff effects during all intervals of time. A standard-deviation increase in the intern effect, $\xi_i^{\tau(i,t)}$, increases test spending by about 20%. A standard-deviation increase in the resident effect, $\xi_j^{\tau(j,t)}$, increases spending by about 70%. In comparison, the standard deviation for admission-level effects, ν_a, is 40%; including or omitting admission-level random effects does not significantly alter results. Given the large qualitative heterogeneity across patients in inpatient care, it is notable that residents alone are responsible for more variation in spending than unobserved patient characteristics.

Physician effects are determined by both individual beliefs and relative influence. However, under the assumption that housestaff beliefs are continuous over time, the discontinuity at the one-year tenure mark identifies the change in influence due to a discontinuous increase in relative tenure, from being at least one year less experienced to being one year more experienced than the teammate. The change in spending-effect variation indeed is highly discontinuous, tripling in standard deviation across the one-year tenure mark. This implies a large effect of relative influence on the size of physician spending variation.

6 Learning: Persistence and Convergence

In this section, I examine housestaff learning, based on two main sources of evidence. First, I study the serial correlation of housestaff effects across adjacent time periods, as a measure of persistence. Because correlation should be invariant to changes in scale, it measures persistence in a way that is conceptually distinct from changes in influence. Increasing persistence only reflects that physicians are settling on choices similar to their past choices, and these choices may be different from those of other physicians.[18]

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

[18] In the conceptual framework in Section A-3, particularly in Equation (A-8), this persistence may be most literally thought of as persistence of beliefs m_h. The development of persistent but heterogeneous practices is consistent with housestaff ceasing to learn a common practice. However, unchanging heterogeneity alone may also represent heterogeneous preferences or skills. These two sources can be separated somewhat by the time course of correlation (e.g., high correlation from the beginning suggests intrinsic heterogeneity). I explore intrinsic heterogeneity further in Section 7.

Second, I study the convergence of housestaff effects with tenure, separately in the different practice environments of specialist and generalist services. Convergence – defined as a decrease in the variation of housestaff effects with tenure – implies that housestaff become more like one another in their effects and is a more direct test of learning to practice a common standard. I compare convergence (or the lack thereof) of housestaff effects in specialist services versus the general medicine service, consistent with greater knowledge in the specialist services. I rule out an alternative mechanism under which differences in learning occur because cardiology and oncology have a higher concentration of diagnoses. Interestingly, I also show that convergence seems unrelated to formal diagnoses, which suggests that the information separating patients into services is mostly informal and tacit.

6.1 Persistence of Housestaff Effects

I study the serial correlation across estimated housestaff effects across tenure intervals. The model for housestaff effects remains specified in Equation (2), but the estimation procedure now includes two periods and specifies a parameter in the variance-covariance matrix of housestaff-tenure effects that allows for this correlation. Details are described in Appendix A-5.2. This procedure can yield estimates of the correlation between effects in any two tenure periods, but I am particularly interested in the serial correlation between two adjacent periods.

Figure 4 shows correlation estimates between each tenure interval and the previous interval. Estimates are less precise than the standard deviation across housestaff effects within each tenure period (Figure 3).[19] The overall lower precision is not surprising given that correlation estimates require observing the *same* housestaff across different periods. It is also important to have a sufficient number of observations per housestaff in each period, for a sufficient number of housestaff, because the correlation depends on the relative values of effects across housestaff both within period and across periods. By contrast, measuring the standard deviation across housestaff effects only requires more than one observation per housestaff within period in order

[19]This figure additionally shows results based on a Bayesian refinement, discussed in Appendix A-6, that also uses correlations between non-adjacent periods. Results and similar with or without the refinement. This perhaps reflects a general consistency in estimation correlations both between adjacent periods and between non-adjacent periods. Alternatively, Proposition A-4 in Appendix A-6 also states that the informativeness of these auxiliary correlations can be low if they are close to 0.

to decompose the variance components due to housestaff and patient-days.

Nonetheless, central estimates are all above 0 and are generally increasing with tenure. That is, a higher-spending housestaff is always more likely than not to be higher-spending in the next period. Many of the central estimates are economically significant, using correlations estimated by Chetty et al. (2014) for teacher value-added as a reference. At the same time, the upper limit of the 95% credible interval of the Bayesian posterior rules out extremely high serial correlations for almost all of the tenure periods. Only one of the fourteen periods has an upper limit greater than 0.70. This suggests that some non-trivial learning continues to occur throughout training and is inconsistent with pure intrinsic heterogeneity as the sole explanation for practice style variation. I will explore intrinsic heterogeneity correlated with rich observable characteristics further in Section 7.

6.2 Convergence to Best Practices

As described in Section 2, I consider specialist-directed services of cardiology and oncology as taking place in an environment with stronger best practices relative to general medicine. By definition, these services are driven by attendings with greater specialized knowledge. Further, this pattern of organizing inpatient care is common across most academic hospitals in the US and in the production of knowledge by research. As the baseline analysis of convergence, I therefore estimate Equation (2) for each of the three ward services of cardiology, oncology, and general medicine. As in Section 5, this yields the standard deviation of housestaff effect distributions by tenure, now separately for each of the ward services.

In Figure 5, I show each of these profiles of housestaff-effect variation over tenure for cardiology, oncology, and general medicine. Housestaff effects significantly converge in cardiology and oncology, but for the same residents, there is no evidence of convergence in their practice patterns in general medicine. The standard deviation of spending variation steadily declines from 85% in cardiology and 75% in oncology, at the beginning of second year (as residents), to 37% in cardiology and 53% in oncology by the end of training. Convergence in specialist services suggests that housestaff significantly learn toward a best practice in these environments, in which there is qualitatively more information. In contrast, variation remains largely unchanged in the

general medicine service, in which care is directed by generalists and is less amenable to the use of specialized knowledge.

Merging cardiology and oncology services into a single "specialist service," I quantify a rate of convergence in spending effects among residents of about a 16% percentage-point decrease in the standard deviation of housestaff effects per year. In other words, given a standard deviation of 74% at the beginning of the second year (when interns become residents), this is equivalent to a relative decrease of 43% of this standard deviation over the next two years.[20] Randomizing over 10,000 placebo combinations of housestaff-service-months (of about 1.27×10^{970} combinations) yields a range of placebo convergence estimates of $[-0.073, 0.085]$, suggesting that the actual estimate -0.160 is extremely significant (see Figure 6). Details are given in Appendix A-7.

6.2.1 Decomposing Experience Leading to Convergence

Using variation in the order of housestaff training experiences, I explore the contribution of general versus specific experience on cardiology or oncology in determining convergence in these respective services. This distinction is informative for understanding the pathways through which learning takes place for the care of patients on these services, for example distinguishing the information being learned (routines for cardiology patients) vs. the teachers *per se* (cardiologists). Convergence according to specific experience suggests that learning occurs via direct experience with patients and attending physicians on the respective cardiology and oncology services. Convergence according to general experience is still consistent with stronger best practices for patients on specialist-driven services, but that learning towards these best practices is not limited but possibly even complemented by experiences outside of these services.

In order to exploit variation in housestaff training over time in the random effects framework, I decompose the set of observations into subsets representing deciles of specific experience "orthogonal" to general experience, and vice versa. I construct linear boundaries between subsets by estimating linear quantile regressions of specific experience (i.e., number of days on service s that resident j has had by day t, $\tau_s(j,t)$) on general experience (i.e., days of tenure $\tau(j,t)$), and vice versa, over housestaff-day observations in service s. Figure 8 shows the variation in

[20] The standard deviation during the first tenure period of the second year is 69%, but the linearized projection of the trend over the next two years implies a standard deviation of 74% for this tenure period.

specific and general experience, for cardiology and oncology, with overlaid decile boundaries. This representation of experience is most informative when there is large variation in training experiences (i.e., specific experience is not perfectly predicted by general experience).[21]

I then estimate the distribution of resident-tenure effects in Equation (2) for each orthogonal decile of specific experience or general experience.[22] Figure 9 shows plots of estimated resident effect standard deviations using observations in each of these deciles. Practice in cardiology shows clear reductions in variation along increasing deciles of general and specific experience. Results for oncology are less clear; convergence perhaps is stronger with increases in general experience. These results decompose convergence in the specialist-driven services into two mechanisms. First, at least for cardiology, convergence specifically occurs via experience on the same service. Second, general experience, independent of time spent on cardiology or oncology rotations, also fosters adoption of the best practices for patients on the specialty services. For example, by exposure to a spectrum of cardiovascular disease and care in outpatient, emergency department, and general inpatient care, trainees may learn more about how to handle patients with well-defined cardiovascular disease on inpatient cardiology wards. This pathway appears present in both cardiology and oncology and is consistent with a cohesive learning environment with knowledge spillovers (albeit asymmetric ones) across internal medicine services.

6.2.2 Best Practices as Encoded by Organization

Given convergence with general experience, it is natural to ask whether convergence reflecting stronger best practices can be predicted by coded diagnoses. First, I explore whether convergence may occur in cardiology and oncology because these services have a higher concentration of diagnoses by constructing pseudo-services within general medicine that include the three most common Major Diagnostic Categories (MDC) of circulatory, respiratory, and digestive (see Table A-6 for summary statistics). I find no difference in convergence between these pseudo-services

[21] Intuitively, measures that are strongly positively correlated will result in a large proportion of overlapping observations in sets but in reverse order, e.g., a large proportion of observations in the first-decile set of one measure being in the last-decile set of the other measure. This therefore will bias finding convergence with increasing deciles in *both* measures, regardless of arbitrary actual positive effects of both measures on convergence. Of course, if measures are perfectly correlated, then defining orthogonal deciles will be impossible.

[22] As before, I impose no relationship between ξ_h^τ and $\xi_h^{\tau'}$ for $\tau \neq \tau'$, but because ξ_h^τ and $\xi_h^{\tau'}$ may now both be in the same estimation sample (i.e., in the same orthogonal decile), I explicitly consider ξ_h^τ and $\xi_h^{\tau'}$ as separate random effects.

(Figure A-7). Relatedly, there is no greater convergence in care for patients with more common diagnostic codes within service (Figure A-8).

Second, I examine whether stronger best practices can be identified by specific diagnoses, linked to published guidelines in the national guideline repository maintained by the US Agency for Healthcare Research and Quality (`guidelines.gov`). Roughly half of the diagnoses coded in *all* services are linked to a published guideline. As shown in Figure A-9, there is no difference in practice convergence, within service, for patients with and without diagnoses linked to guidelines. This null finding suggests that guideline existence is an imperfect representation of true best practices, and that coded diagnoses, despite their potential richness and widespread use as the foundation for reimbursement (and research), are an imperfect measure of care-relevant patient conditions. Finally, I replicate 97% of the diagnostic-code makeup of the cardiology service using patients from general medicine, by selecting patients with ICD-9 codes in common with cardiology and weighting them appropriately. I find no convergence in these patients from general medicine but with diagnostic codes in common with cardiology (Figure 7).

These findings are consistent with the complexity of information not only in characterizing best practices but in identifying the patients themselves for which best practices are applicable. Although it may be surprising that potentially rich administrative diagnostic codes are uninformative for predicting convergence, closer examination reveals that codes used in practice are quite coarse. For example, the most common formal diagnosis in both cardiology and general medicine is "Chest pain, not otherwise specified."[23] Further, the strong difference in convergence between specialist and generalist services suggests that much more information is used in assigning patients in practice, and that this assignment is meaningful.

7 Housestaff Characteristics and Experience

In Sections 5 and 6, I show that practice variation depends to a large degree on team roles and the practice environment, despite a fixed cohort of housestaff with broadly similar experiences in the same training program. This suggests that information-based mechanisms are important

[23]Table A-7 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.

drivers of practice variation. Given the traditional emphasis on human capital and intrinsic heterogeneity (e.g., ability) (e.g., Doyle Jr et al., 2010; Fox and Smeets, 2011; Bartel et al., 2014), a natural comparison is to examine predicted differences in spending according to housestaff characteristics and experience. I use rich data on housestaff characteristics and quasi-experimental variation in training experiences to address this question in detail, and I find that mean effects of numerous housestaff characteristics and measures of experience are either insignificant or an order of magnitude less important than the mechanisms of relative influence and potential convergence. This suggests that traditional concepts of intrinsic heterogeneity and human capital are less valuable predictors than informational mechanisms in understanding variation in health care practice.

7.1 Housestaff Characteristics

In the same training program, I observe predetermined and unusually detailed characteristics that are likely correlated with differences in preferences and abilities.[24] For example, USMLE scores directly measure medical knowledge as a medical student; position on the residency rank lists reflects overall desirability; and residency tracks reflect important career decisions and lifestyle preferences, such as a decision to become a radiologist rather than a primary care physician. In addition to housestaff in the main residency program, I observe both interns and residents from an internal medicine residency based in another hospital. For these outside-hospital housestaff, I can evaluate the effect of their presence on medical teams. This effect includes both differences in selection into the different program and in training experiences across the programs (the outside residency is nationally recognized but lower ranked, and the outside hospital is known to be more cost-conscious).

Separately for each of these characteristics, and for interns and residents, I assess the relationship these characteristics and daily test spending in regressions of this type:

$$Y_{aijkt} = \alpha_m Characteristic_h^m + \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_{-hk} + \varepsilon_{aijkt}, \tag{3}$$

[24] Previous studies have investigated the effect of coarse measures of observable physician characteristics (e.g., gender) and training experiences (e.g., place of medical school or residency) in a single regression (e.g., Epstein and Nicholson, 2009). A challenge with this approach is that housestaff may select into different experiences. However, these studies have also been unable to find any significant predictors of physician practice styles.

where $Characteristic_h^m$ equals 1 if housestaff $h \in \{i, j\}$ had characteristic (or made track choice) m prior to starting residency, and ζ_{-hk} is a fixed effect for the other housestaff $-h$ and attending k.[25] The coefficient of interest is α_m, which is the causal effect of a patient being assigned to a housestaff with characteristic m, includes effects that may be directly related to m as well as effects due to any unobserved traits correlated with m.

I also evaluate the combined predictive effect of housestaff characteristics in two steps. First, I regress outcomes on all direct housestaff characteristics, with continuous characteristics like position on rank list entered linearly, along with the other admission and time regressors in Equation (3):

$$Y_{aijkt} = \sum_m \alpha_m Characteristic_h^m + \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_{-hk} + \varepsilon_{aijkt}. \tag{4}$$

This yields a predicted score Z_h for each housestaff h, $Z_h = \sum_m \hat{\alpha}_m Characteristic_h^m$, which I normalize to $\tilde{Z}_h = Z_h / \sqrt{\text{Var}\,(Z_h)}$ with standard deviation 1. Similar to Equation (3), I then regress daily test spending on this normalized score:

$$Y_{aijkt} = \alpha\tilde{Z}_h + \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_{-hk} + \varepsilon_{aijkt}. \tag{5}$$

Finally, I evaluate Equation (3) 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 (3).

Table 2 shows results for Equation (5) and a subset of results for Equation (3). Considering characteristics individually in Equation (3), only two characteristics are statistically significant: male sex and high USMLE test score. Male interns have 2% lower daily spending costs, significant

[25] In principle, I could include housestaff characteristics as mean shifters in the baseline random effects model in Equation (2). However, since characteristics are generally insignificant predictors of variation, results of (residual) variation attributable to housestaff are unchanged.

at the 10% level; male residents have 4% lower daily spending costs, significant at the 5% level. A high USMLE score predicts 3% lower daily spending, significant at the 10% level, for residents. Table 2 also considers the mean effect of having housestaff from the other residency program, an effect that could be due to selection (i.e., intrinsic heterogeneity) or differences in learning experiences across the two programs. While other-program interns do not have significantly different mean spending effects, other-program residents spend 17% less, but the latter is only significant at the 10% level because of relatively few housestaff from the other program.

A standard-deviation change in the overall predictive score changes costs by about 2% for both interns and residents. By comparison, using the same characteristics to predict whether a housestaff 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%. LASSO selected no intern characteristic as significant and selected only resident male sex as significant. In this sense, the overall predictive score is likely to be an overestimate of variation due to intrinsic heterogeneity.

Overall, these results show that intrinsic heterogeneity, to the extent that it is correlated with any of the rich pre-residency characteristics and choices I observe, explains relatively little compared to the size of variation that depends on influence and learning. The single characteristic selected by LASSO, male sex, has a larger effect for residents than for interns, which supports the idea of increasing influence, but effects are an order of magnitude less than the variation across housestaff in any tenure period.

7.2 Housestaff Experience

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

$$Y_{aijkt} = \alpha_m \mathbf{1}\left(Experience^m_{h,\tau(h,t)-1} < \text{Median}^m_{\tau(h,t)-1}\right) + \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_h + \zeta_{-hk} + \varepsilon_{aijkt}, \quad (6)$$

where the coefficient of interest α_m is on whether the measure $Experience^m_{h,\tau(h,t)-1}$ is above

median, where both the measure and the median are calculated using observations before the tenure period associated with the index observation. I also consider service-specific measures, $Experience^{ms}_{h,\tau(h,t)-1}$, calculated using observations within service s (e.g., the number of patients seen on cardiology service) and evaluated against a service-specific median. Patient characteristics $\mathbf{X}_a$ and time indicators $\mathbf{T}_t$ are the same as used in previous regressions. In my baseline specification, I control for the identities of the housestaff as ζ_h and the peer-attending combination as ζ_{-hk} separately, although whether I include ζ_h at all or include a fixed effect for intern-resident-attending ζ_{ijk} does not qualitatively influence results, consistent with random assignment of housestaff to patients and peers (Appendix A-1). Results from Equation (6) are shown in Table 3 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 of these measures as significant.

I also consider the effect of resident tenure on outcomes of test daily spending, total daily spending, length of stay, 30-day readmissions, and 30-day mortality for each of the ward services. Because I also control for month-year interactions, I study this as the effect of having a third-year housestaff, as opposed to having a second-year housestaff, as the resident:

$$Y_{aijkt} = \alpha \mathbf{1}\left(\tau\left(j,t\right) > 2 \text{ years}\right) + \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_{ik} + \varepsilon_{aijkt}. \tag{7}$$

The coefficient α is small and insignificant for all of these outcomes. Table 4 lists results along with counterfactuals for switching to a resident one standard deviation above or below in housestaff-effect distribution for the relevant outcome.

Overall, these results indicate that summary measures of housestaff experience are also poor predictors of practice and outcomes, especially relative to the large variation across housestaff. In this setting with the distinctive advantage that housestaff are as good as randomly assigned to training experiences, I am able to reject that formal differences in training are responsible for any significant subsequent variation in housestaff behavior. This fails to support the view of formal "schools of thought," at least within an organization with largely uniform training experiences,

but nonetheless in an environment with large practice variation. Rather, it is consistent with the view, as previously suggested in Section 6.2.2, that summary measures of experience, even with (administratively) rich data, are likely to be impractical representations of the lessons to be learned via specific experiences.

8 Discussion and Conclusion

The fact that there exists persistent variation in medical care has attained tremendous prominence in policy discussions.[26] However, the behavioral foundations of such variation in medical care, and indeed in closely related variation in other industries (e.g., Syverson, 2011), remain poorly understood. Although the scope of this paper is necessarily limited to studying variation *within* an organization, its empirical setting is well-suited to capture two important facts in health care delivery that have been largely overlooked in the empirical literature on practice variation: Medical care is delivered in teams within organizations, and physician practice patterns must be learned. I find learning-related mechanisms with large effects on variation: influence given to residents with greater experience and convergence depending on the strength of best practices. These channels dwarf the contributions of intrinsic heterogeneity, human capital, and learned practice styles (i.e., "schools of thought") from individual supervising physicians.[27]

While this paper is the first, to my knowledge, to empirically show evidence of the contribution of informational frictions in the evolution of medical practice variation, these findings are consistent with original thinking and evidence in the practice variation literature. It has long been suspected that practice variation arises because of a lack of consensus on how medical technology should be used. Jack Wennberg and colleagues indeed document that there exists

[26]For example, it is well-known that President Barack Obama paid special attention to this fact during US health care reform leading to the Affordable Care Act (e.g., Pear, 2009). An article about health care spending variation by Atul Gawande in 2009 in the *New Yorker* was dubbed by David Brooks of the *New York Times* as the most influential essay of the year. The existence of medical spending variation has led influential policymakers, such as Peter Orszag, to conclude that $700 billion (or over 30%) of health care spending could simply be eliminated without any ill effects and has led some to propose penalizing areas with higher-than-average per capita spending (see, e.g., Roy, 2010, and Jauhar, 2014, in the popular press for references to these suggestions and "contrarian views" against them).

[27]Although intrinsic heterogeneity and differences in training could play a larger role in variation across institutions, informational frictions are also likely to be greater across institutions could explain much of regional variation. Moreover, from a welfare perspective, variation within institutions is no less relevant than variation across institutions.

larger variation in surgical procedures where there is more disagreement (Wennberg et al., 1980; McPherson et al., 1982; Wennberg et al., 1982). This view accords more generally with Polanyi's (1958) thesis that knowledge is difficult to communicate and therefore highly personal. As Nelson and Winter (1982) observe, there is a connection between the tacit nature of knowledge across individuals and the transferrability of practices or "routines" across organizations, and Gibbons and Henderson (2012) discuss how this translates to persistent performance differences across seemingly similar enterprises, in which managers may fail to perceive, understand, or implement solutions to problems due to these frictions.

The notion that practice variation is a symptom of informational frictions has important policy implications. For example, in team decision-making, if there are differences in relative influence due to informational frictions in which agents cannot evaluate the quality of specific recommendations, then educational interventions may be more effective in targeting agents who have greater influence on teams, even if they have less to learn. That is, when informational quality is imperfectly observed among agents, the same recommendation may have a greater impact on patient care when internalized by an agent who is more influential. More importantly, for most of medical care, simply selecting or incentivizing physicians and institutions to spend the "right" amount of resources in medical care is unlikely to be feasible, for the essential reason that we do not know what the "right" level of care is for a given patient. Rather, this paper suggests a high degree of knowledge about "best practices" required to generate convergence, even during an intense period of training at a high-quality academic institution. The billions of yearly NIH spending in cardiology and oncology is a lower bound on the societal resources required to support such knowledge, and the fact that formal diagnoses bear less relevance to convergence than human triage decisions with discretion imply that universal algorithms, in the absence of more knowledge, would be blunt and likely counterproductive means to reign in variation.

References

Abowd, John M., Francis Kramarz, and Simon Woodcock, "Econometric Analyses of Linked Employer-Employee Data," in Laszio Matyas and Patrick Sevestre, eds., *The Econo-*

metrics of Panel Data, number 46. In 'Advanced Studies in Theoretical and Applied Econometrics.', Springer Berlin Heidelberg, January 2008, pp. 727–760.

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz, "Learning and Disagreement in an Uncertain World," Working Paper 12648, National Bureau of Economic Research October 2006.

Alchian, Armen A. and Harold Demsetz, "Production, information costs, and economic organization," *The American Economic Review*, 1972, *62* (5), 777–795.

Anderson, G. F., P. S. Hussey, B. K. Frogner, and H. R. Waters, "Health spending in the United States and the rest of the industrialized world," *Health Affairs*, 2005, *24* (4), 903–914.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, December 1963, *53* (5), 941–973.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, November 2003, *118* (4), 1279–1333.

Bartel, Ann P., Nancy Beaulieu, Ciaran Phibbs, and Patricia W. Stone, "Human Capital and Productivity in a Team Environment: Evidence from the Healthcare Sector," *American Economic Journal: Applied Economics*, April 2014, *6* (2), 231–259.

Bartling, Bjorn, Ernst Fehr, and Holger Herz, "The Intrinsic Value of Decision Rights," *Econometrica*, November 2014, *82* (6), 2005–2039.

Bates, Douglas, Martin Machler, Ben Bolker, and Steve Walker, "Fitting Linear Mixed-Effects Models using lme4," *Journal of Statistical Software*, October 2015, *67* (1), 1–48. arXiv: 1406.5823.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, May 2014, *28* (2), 29–50.

Caplin, Andrew and Mark Dean, "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, July 2015, *105* (7), 2183–2203.

Chamberlain, Gary, "Panel Data," in Zvi Griliches and M.D. Intrilligator, eds., *Handbook of Econometrics*, Vol. Chapter 22, Amsterdam: North Holland, 1984, pp. 1248–1318.

Charlson, Mary E., Peter Pompei, Kathy L. Ales, and C. Ronald MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, 1987, *40* (5), 373–383.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Cooke, Molly, David M. Irby, William Sullivan, and Kenneth M. Ludmerer, "American Medical Education 100 Years after the Flexner Report," *New England Journal of Medicine*, 2006, *355* (13), 1339–1344.

Csikszentmihalyi, Mihaly, *Flow: The Psychology of Optimal Experience*, Harper & Row, 1990.

Cutler, David, "Where Are the Health Care Entrepreneurs?," *Issues in Science and Technology*, 2010, *27* (1), 49–56.

_ , Jonathan Skinner, Ariel Dora Stern, and David Wennberg, "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," Working Paper 19320, National Bureau of Economic Research August 2013.

Cyert, Richard and James March, *A Behavioral Theory of the Firm*, Oxford: Blackwell, 1963.

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey, "Comorbidity Measures for Use with Administrative Data," *Medical Care*, January 1998, *36* (1), 8–27.

Ellison, G. and D. Fudenberg, "Rules of thumb for social learning," *Journal of Political Economy*, 1993, *101* (4), 612–643.

Epstein, A. J. and S. Nicholson, "The formation and evolution of physician treatment styles: an application to cesarean sections," *Journal of Health Economics*, 2009, *28* (6), 1126–1140.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams, "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *Quarterly Journal of Economics*, 2016, *Forthcoming.*

Fisher, Elliott S., David E. Wennberg, Therese A. Stukel, Daniel J. Gottlieb, F. L. Lucas, and Etoile L. Pinder, "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care," *Annals of Internal Medicine*, February 2003, *138* (4), 273–287.

_ , _ , _ , _ , _ , and _ , "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine*, February 2003, *138* (4), 288–298.

Flexner, Abraham, *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, 1910.

Fox, Jeremy T. and Valerie Smeets, "Does Input Quality Drive Measured Differences in Firm Productivity?," *International Economic Review*, November 2011, *52* (4), 961–989.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, October 2000, *108* (5), 874–904.

_ and Esteban Rossi-Hansberg, "Organization and Inequality in a Knowledge Economy," *The Quarterly Journal of Economics*, November 2006, *121* (4), 1383–1435.

Gawande, Atul, "The Cost Conundrum," *The New Yorker*, June 2009.

Gibbons, Robert and Rebecca Henderson, "What do managers do? Exploring persistent performance differences among seemingly similar enterprises," in Robert Gibbons and John

Roberts, eds., *The Handbook of Organizational Economics*, Princeton, NJ: Princeton University Press, 2012, pp. 680–732.

Gilks, W. R. and P. Wild, "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 1992, *41* (2), 337.

Iwashyna, T. J., A. Fuld, D. A. Asch, and L. M. Bellini, "The impact of residents, interns, and attendings on inpatient laboratory ordering patterns: A report from one university's hospitalist service," *Academic Medicine*, 2011, *86* (1), 139.

Jacob, Brian A. and Lars Lefgren, "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers," *The Quarterly Journal of Economics*, November 2007, *122* (4), 1603–1637.

Jauhar, Sandeep, "Don't Homogenize Health Care," *The New York Times*, December 2014.

Jr, J. J. Doyle, S. M. Ewer, and T. H. Wagner, "Returns to physician human capital: Evidence from patients randomized to physician teams," *Journal of Health Economics*, 2010, *29* (6), 866–882.

Jr, Joseph J. Doyle, John A. Graves, Jonathan Gruber, and Samuel Kleiner, "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, February 2015, *123* (1), 170–214.

Kane, Thomas J. and Douglas O. Staiger, "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," in David Grissmer and Helen F. Ladd, eds., *Brookings Papers on Education Policy*, Washington, DC: Brookings Institution Press, January 2002, pp. 235–283.

Ludmerer, Kenneth M., *Learning to Heal: The Development of American Medical Education*, Perseus Books Group, January 1988.

McPherson, Klim, John E. Wennberg, Ole B. Hovind, and Peter Clifford, "Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New

England, England, and Norway," *New England Journal of Medicine*, November 1982, *307* (21), 1310–1314.

Molitor, David, "The evolution of physician practice styles: Evidence from cardiologist migration," Technical Report, Massachusetts Institute of Technology 2011.

Morris, Carl N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, March 1983, *78* (381), 47–55.

Nelson, Richard R. and Sidney G. Winter, *An Evolutionary Theory of Economic Change*, Harvard University Press, 1982.

Patterson, H. D. and R. Thompson, "Recovery of inter-block information when block sizes are unequal," *Biometrika*, December 1971, *58* (3), 545–554.

Pear, Robert, "Health Care Spending Disparities Stir a Fight," *The New York Times*, June 2009.

Polanyi, Michael, *Personal Knowledge: Towards a Post-Critical Philosophy*, University of Chicago Press, 1958.

_ , *The Tacit Dimension*, New York: Doubleday Press, 1966.

Prendergast, Canice, "A Theory of Yes Men," *The American Economic Review*, September 1993, *83* (4), 757–770.

Radner, Roy, "The Organization of Decentralized Information Processing," *Econometrica*, 1993, *61* (5), 1109–46.

Rogerson, Richard, Robert Shimer, and Randall Wright, "Search-Theoretic Models of the Labor Market: A Survey," *Journal of Economic Literature*, December 2005, *43* (4), 959–988.

Roy, Avik, "The Dartmouth Atlas and Obamacare," *National Review*, June 2010.

Scharfstein, David S. and Jeremy C. Stein, "Herd Behavior and Investment," *The American Economic Review*, June 1990, *80* (3), 465–479.

Schroeder, S. A., A. Schliftman, and T. E. Piemme, "Variation among physicians in use of laboratory tests: relation to quality of care," *Medical Care*, 1974, *12* (8), 709–713.

Searle, S. R., G. Casella, and C. E. McCulloch, *Variance Components*, Wiley New York, 1992.

Skinner, Jonathan, "Causes and Consequences of Regional Variations in Healthcare," in Mark V Pauly, Thomas G McGuire, and Pedro Barros, eds., *Handbook of Health Economics*, Vol. 2, San Francisco: Elsevier, 2012, pp. 49–93.

Starr, Paul, *The Social Transformation Of American Medicine: The Rise Of A Sovereign Profession And The Making Of A Vast Industry*, Basic Books, August 2008.

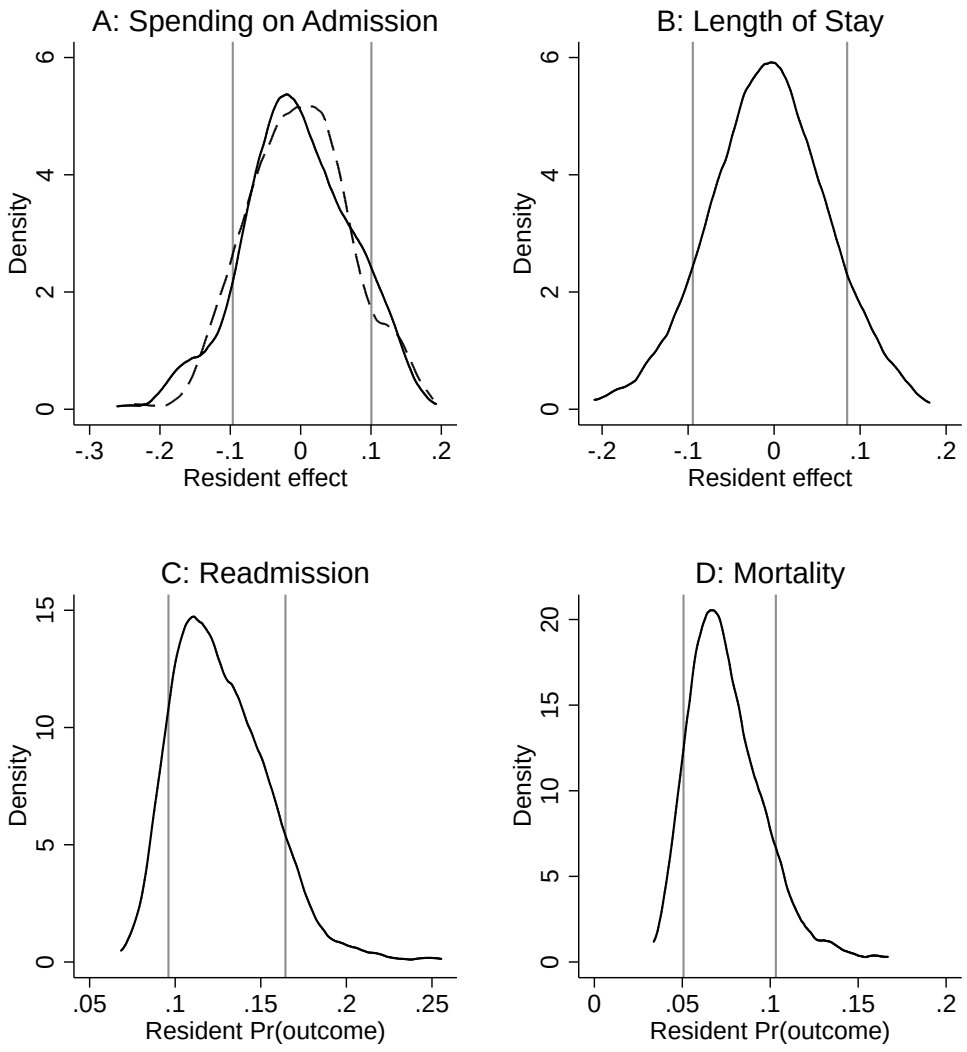
Syverson, Chad, "What Determines Productivity?," *Journal of Economic Literature*, June 2011, *49* (2), 326–365.

Wennberg, J. and A. Gittelsohn, "Small area variations in health care delivery," *Science*, 1973, *182* (4117), 1102–1108.

Wennberg, J. E., J. P. Bunker, and B. Barnes, "The need for assessing the outcome of common medical practices," *Annual Review of Public Health*, 1980, *1*, 277–295.

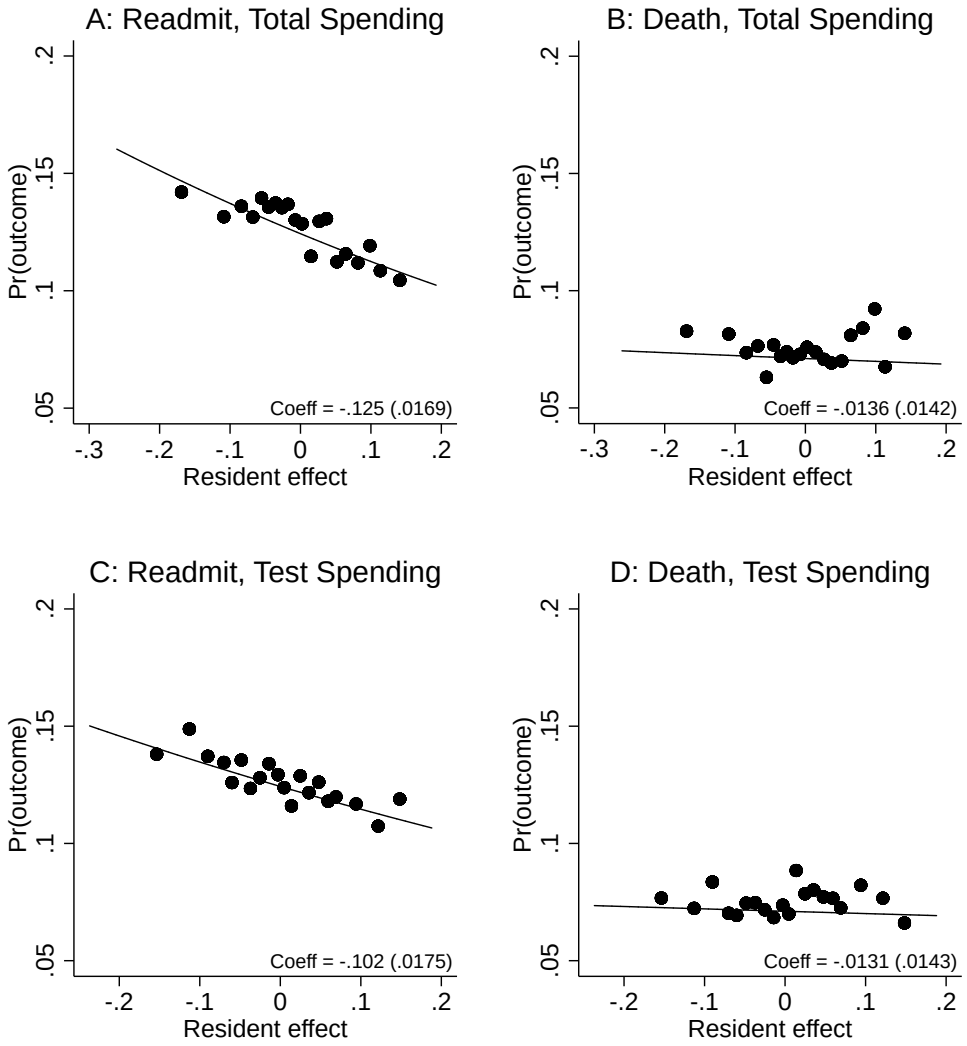
Wennberg, John E., Benjamin A. Barnes, and Michael Zubkoff, "Professional uncertainty and the problem of supplier-induced demand," *Social Science & Medicine*, January 1982, *16* (7), 811–824.

Figure 1: Distributions of Resident Effects



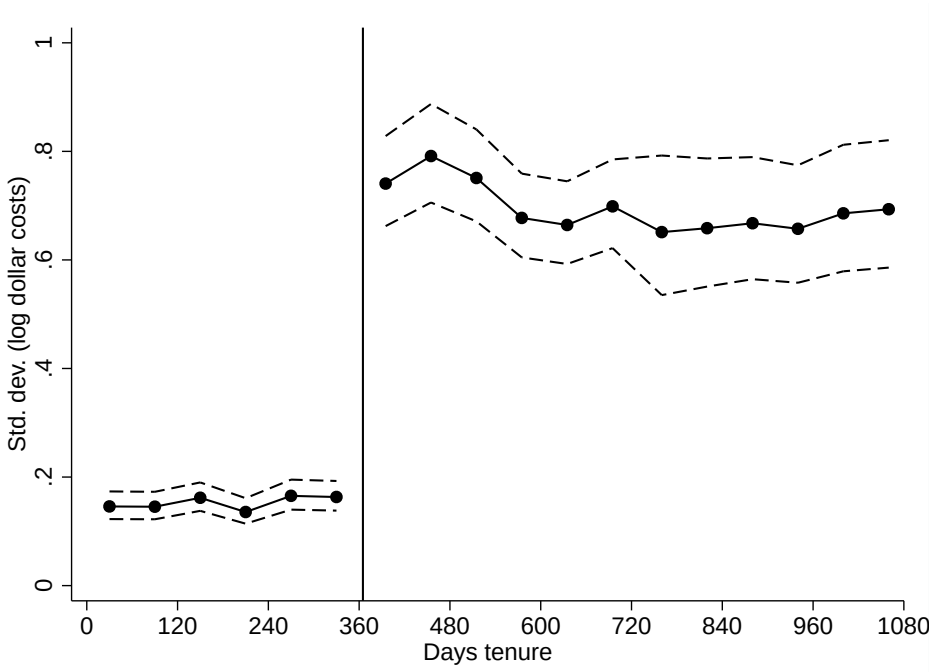
Note: This figure shows distributions of empirical Bayes predictions (BLUPs) of resident random effects for outcomes of spending on day of admission (Panel A), length of stay (Panel B), 30-day readmission (Panel C), and 30-day mortality (Panel D). Random effects are modeled according to Equation (1), allowing for correlation with patient and admission characteristics, time categories, and attending identities. The empirical Bayes predictions are of the component of the random effects that is orthogonal to a projection of these characteristics onto the outcome. More details are described in Section 4 and Appendix A-2. Random effects are represented directly on the x-axis for continuous outcomes of (log) spending and (log) length of stay in Panels A and B, while they are transformed into probabilities for the average patient for binary outcomes of readmission and mortality in Panels C and D. Panel A shows distributions for both total spending (solid line) and test spending (dashed line). The vertical gray lines represent 10th and 90th percentiles of the distribution (total spending in Panel A); because BLUPs are shrunken 10th and 90th percentiles are less dispersed than implied by standard deviations of the random effects distributions (Table 4).

Figure 2: Effects on Spending and Clinical Outcomes



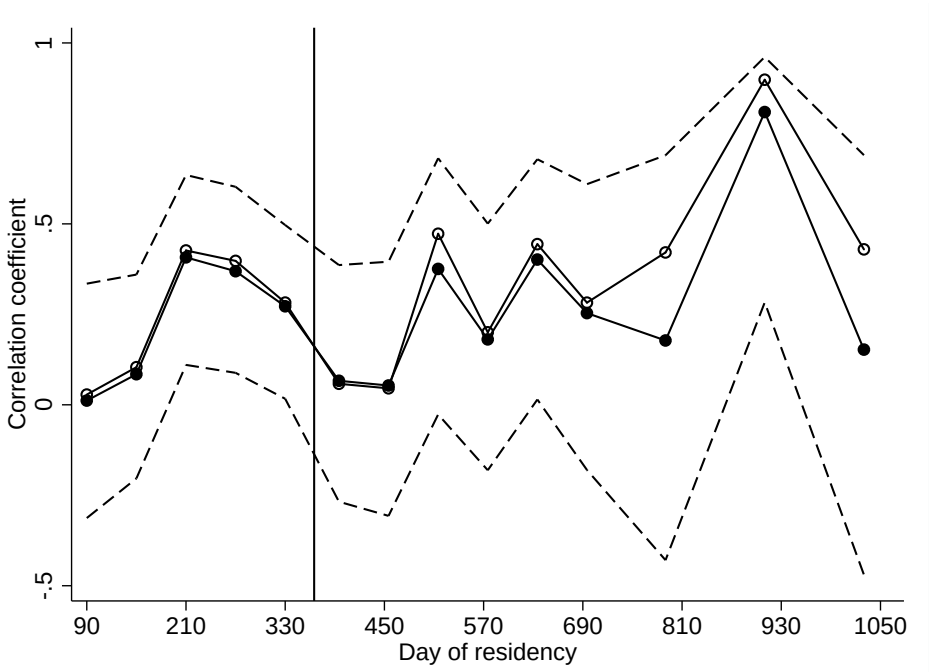
Note: This figure shows the relationship between empirical Bayes predictions of resident random effects for spending and for clinical outcomes. Each panel is a binned scatterplot of average random effects for spending within vigintiles on the x-axis and the corresponding random effects clinical outcomes among the same residents in the spending vigintile on the y-axis. Clinical outcome random effects are transformed into probabilities for the average patient, as in Figure 1. The figure considers two measures of spending: total spending on day of admission (Panels A and B) and test spending on day of admission (Panels C and D). Two clinical outcomes are 30-day readmissions (Panels A and C) and 30-day mortality (Panels B and D). Regression lines are also plotted, with y-coordinates of the lines transformed to probabilities for the average patient by a logistic transformation. Coefficients (and standard errors in parentheses) correspond to a linear regression fit of the transformed clinical probabilities on spending random effects. More details on how the random effects are calculated are described in Section 4 and Appendix A-2.

Figure 3: Standard Deviation of Housestaff Random Effects by Tenure



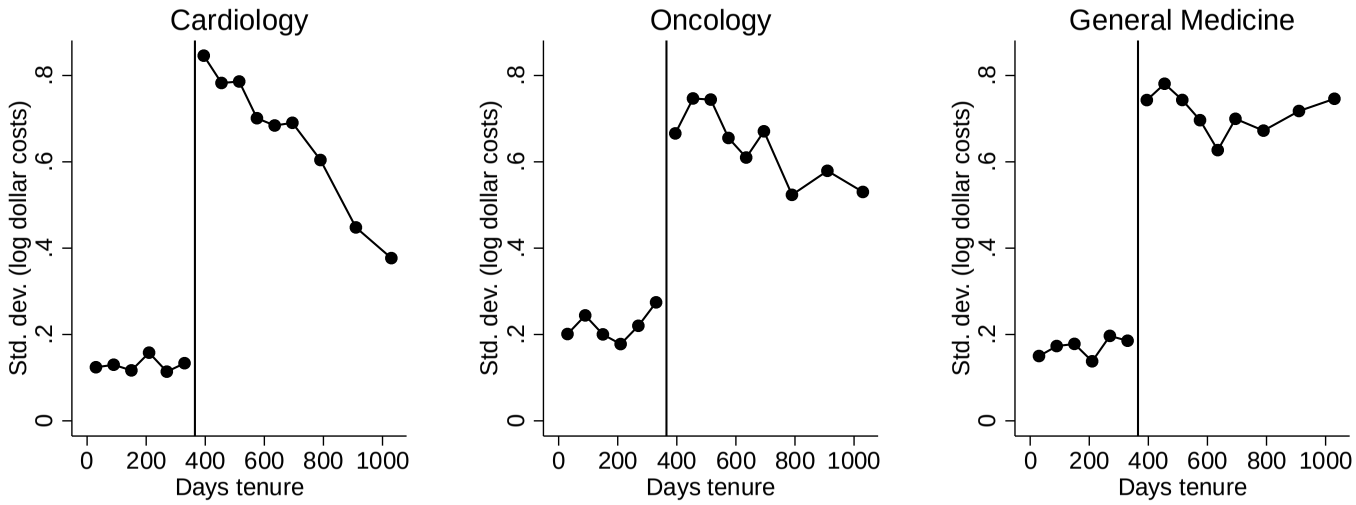
Note: This figure shows the standard deviation in a random effects model of log daily test costs shown in Equation (2) at each non-overlapping two-month tenure interval. Point estimates are shown as connected dots; 95% confidence intervals are shown as dashed lines. The model controls for patient and 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"). Housestaff prior to one year in tenure are interns and become residents after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 4: Serial Correlation of Housestaff Random Effects over Tenure



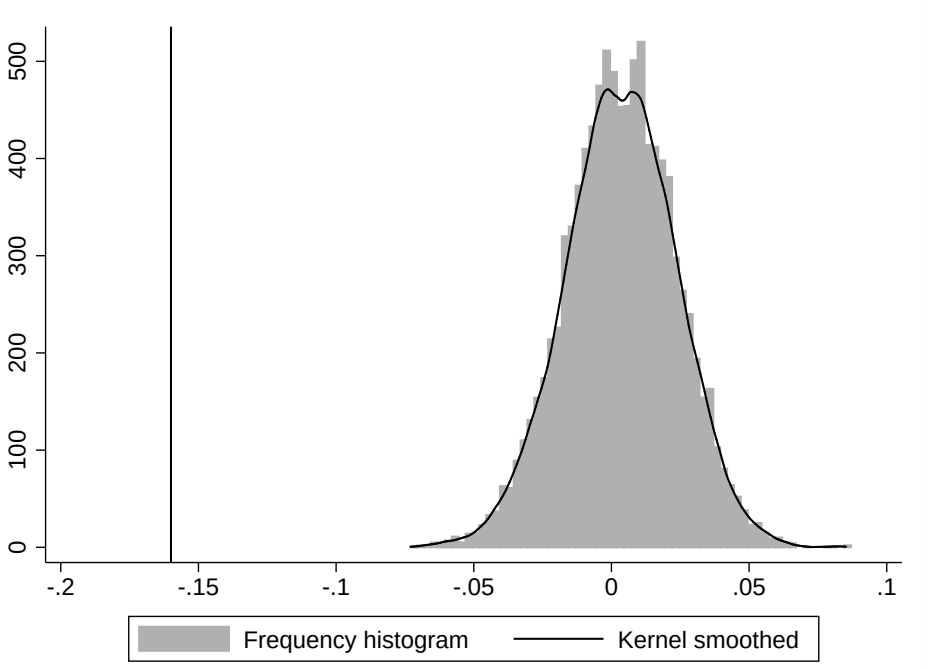
Note: This figure shows the serial correlation between random effects within housestaff in a given tenure period and the previous tenure period. Hollow dots show directly estimated correlations from maximum likelihood of data from the two tenure periods (details in Appendix A-5.2). Solid dots show posterior correlations from a Bayesian refinement procedure that includes both the directly estimated correlation and information from other correlations between non-adjacent periods (details in Appendix A-6). The dashed lines are the 95% credible interval for the posterior correlations. The 95% confidence interval for the directly estimated correlations are slightly larger but otherwise similar and are omitted from this figure for simplicity. The random effect model of log daily test costs is first estimated as in Equation (2), as described in the notes for Figure 3. Housestaff prior to one year in tenure are interns and become residents after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure 5: Housestaff-effect Variation by Tenure in Each Service



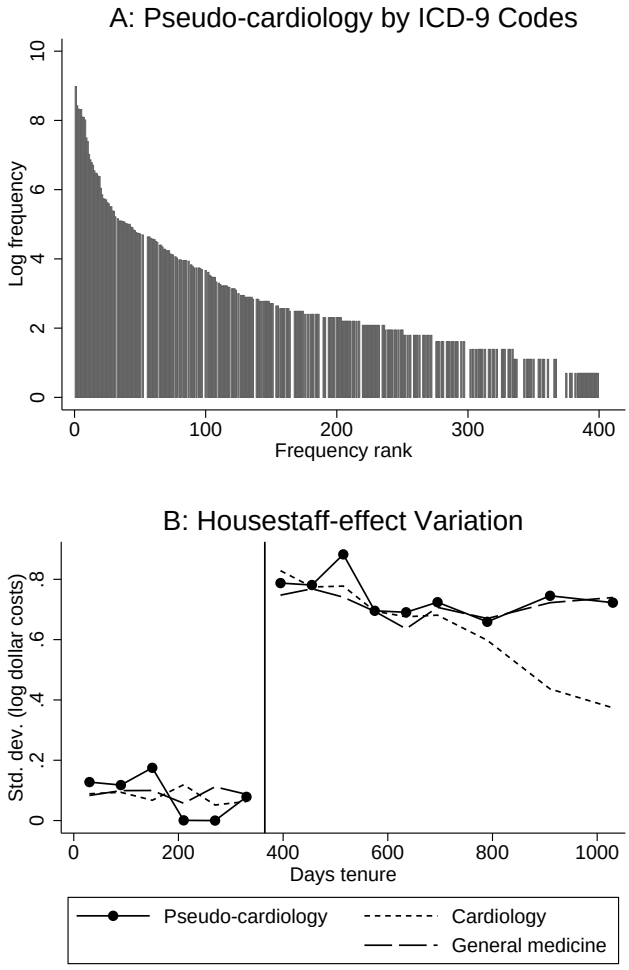
Note: Similar to Figure 3, this figure shows the standard deviation in a random effects model, as in Equation (2), of log daily test costs at each non-overlapping two-month tenure interval but for each service of cardiology, oncology, and general medicine. Controls are the same as those listed in the caption for Figure 3. Housestaff 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 6: Systematic Placebo Tests for Specialist-service Convergence



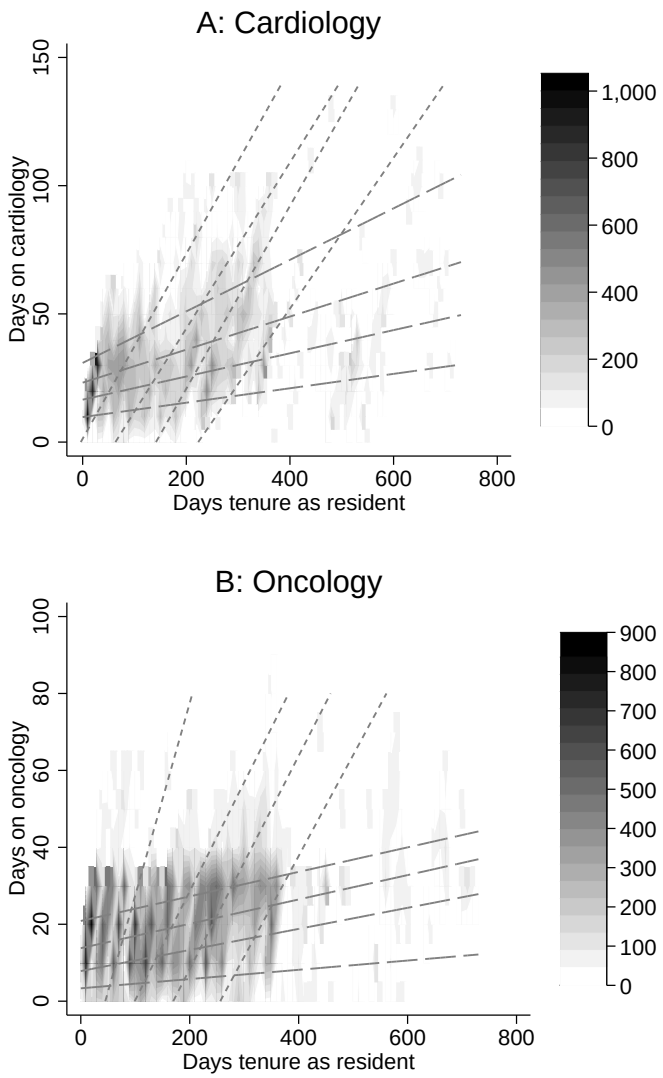
Note: This figure shows 10,000 random placebo tests for convergence in the specialist services. Merging cardiology and oncology yields an actual estimate of -0.160, or a 16% percentage point decrease per year in the standard deviation of spending effects of residents over the two years of the resident role, shown by the vertical line. In each of 10,000 placebo tests, I randomize combinations of housestaff-month-service to a placebo specialist service, matching the number of housestaff-month-services assigned to specialist services in each month of tenure. I estimate the same random effects model of log daily test costs shown in Equation (2) for the placebo specialist service and estimate the rate of placebo convergence using estimated housestaff effects in this placebo specialist service. Estimates for convergence are shown as a frequency histogram with a kernel-smoothed overlay.

Figure 7: Pseudo-cardiology Service



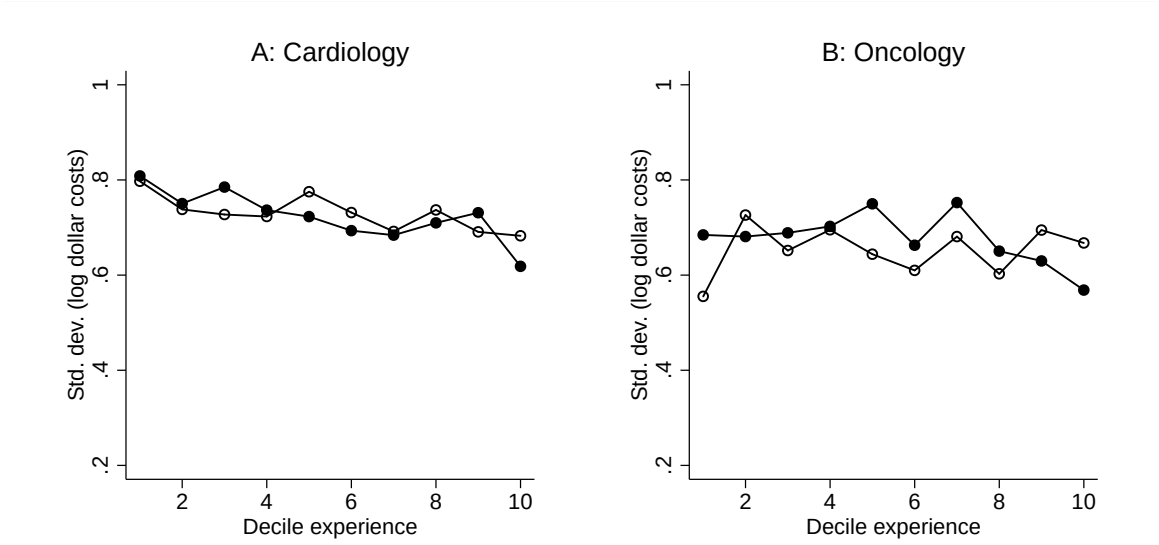
Note: This figure shows the construction of a pseudo-cardiology service by ICD-9 codes (Panel A) and housetaff-effect variation by tenure in this service (Panel B). This service is constructed from general medicine observations, matching ICD-9 codes observed in cardiology. This procedure covers 97% of observations in the actual cardiology service. Panel A shows ICD-9 codes ranked by frequency in cardiology; gray bars represent ICD-9 codes matched with observations in general medicine. Eight of 410 ICD-9 codes have only one observation and are therefore not shown with a non-zero log frequency. Panel B shows the standard deviation of housestaff 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 (2) includes admission-intern random effects to normalize higher variance in the number of patients per intern in the pseudo-cardiology service (thus results are slightly different than in Figure 5, for example. Housestaff 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 8: Orthogonal Quantiles of General and Specific Experience



Note: This figure shows orthogonal deciles of general and specific experience in cardiology (Panel A) and oncology (Panel B), as described in Section 6.2.1. Days on cardiology (i.e., cardiology wards and coronary care units, including at affiliated hospitals) are considered specific experience for cardiology; days on oncology (i.e., oncology wards and bone marrow transplant service) are considered specific experience for oncology. Overall tenure as a resident is considered general experience. Numbers of observations in each 10×10 day bin are shown as densities. Quintile (rather than decile) boundaries are plotted for visual simplicity: Short-dashed lines illustrate orthogonal quintiles of general experience; long-dashed lines illustrate orthogonal quintiles of specific experience.

Figure 9: Convergence by Orthogonal Deciles of General and Specific Experience



Note: This figure plots the standard deviation of resident spending effects estimated by Equation (2), but decomposing experience into deciles of general and specific components, as described in Section 6.2.1. Controls are the same as those listed in the caption for Figure 3. Each estimation sample is defined by an "orthogonal decile" of general (solid dots) or specific (hollow dots) experience, which are deciles of general (or specific) experience orthogonal to linear quantile predictions based on specific (general) experience. The set of observations comprising each decile is illustrated in Figure 8. Panel A shows results in cardiology; Panel B shows results in oncology. See notes in Figure 8 for how general and specific experience are defined.

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

	Interns		Residents	
	Below-median test spending	Above-median test spending	Below-median test spending	Above-median test spending
Patient characteristics				
Age	62.11 (16.90)	62.13 (16.86)	62.07 (16.82)	62.15 (16.93)
Male	0.484 (0.500)	0.482 (0.500)	0.489 (0.500)	0.478 (0.500)
White race	0.706 (0.455)	0.703 (0.457)	0.708 (0.455)	0.702 (0.457)
Black race	0.161 (0.367)	0.159 (0.365)	0.157 (0.364)	0.162 (0.368)
Charlson comorbidity index	2.87 (2.79)	2.87 (2.79)	2.84 (2.77)	2.90 (2.81)
Diagnostic-related Group (DRG) weight	1.25 (0.86)	1.25 (0.84)	1.27 (0.85)	1.24 (0.84)
Supervising physicians				
Above-median-spending residents	0.500 (0.501)	0.500 (0.501)	N/A	N/A
Above-median-spending attendings	0.503 (0.501)	0.502 (0.501)	0.501 (0.501)	0.502 (0.501)

Note: This table shows evidence of exogenous assignment for housestaff with below-median or above-median averaged spending effects. Average spending effects, not conditioning by tenure, are estimated as fixed effects by a regression of log test 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: Effect of Housestaff Characteristics on Spending

	Log daily test costs				
	(1) Male	(2) High USMLE	(3) Highly ranked	(4) Other hospital	(5) Overall score
Panel A: Interns					
Effect of housestaff	-0.021	-0.003	0.011	0.007	0.019
with characteristic	(0.012)	(0.013)	(0.018)	(0.025)	(0.006)
Observations	186,694	185,497	131,418	220,074	190,640
Adjusted R^2	0.166	0.166	0.166	0.165	0.165
Sample characteristic mean	0.596	0.258	0.234	0.055	N/A
Panel B: Residents					
Effect of housestaff	-0.039	-0.013	0.002	-0.169	0.022
with characteristic	(0.016)	(0.020)	(0.028)	(0.095)	(0.008)
Observations	206,802	199,715	129,508	220,074	206,802
Adjusted R^2	0.180	0.180	0.178	0.178	0.180
Sample characteristic mean	0.564	0.235	0.213	0.060	N/A

Note: This table reports results for some regressions of the effect of indicators of some housestaff characteristics, including other hospital status, and a normalized predictive score (with standard deviation 1) based on *all* observed housestaff characteristics. Panel A shows results for interns; Panel B shows results for residents. Columns (1) to (4) are regressions of the form in Equation (3), where the coefficient of interest is on an indicator for a group of housestaff identified by either pre-residency characteristics or whether the housestaff is from the other academic hospital. The effect of many other characteristics of interest (or groups) were estimated as insignificant and omitted from this table for brevity. Column (5) is reports results for Equation (5), where the regressor of interest is a normalized predictive score based on age, sex, minority status, housestaff 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 housestaff 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 3: Effect of Housestaff Experience on Spending

	Log daily test costs				
	(1) Number of days	(2) Number of patients	(3) Number of attendings	(4) Attending spending	(5) Attending spending
Panel A: Interns					
Effect of housestaff with measure above median	-0.004 (0.016)	-0.016 (0.016)	-0.017 (0.016)	-0.009 (0.013)	0.014 (0.058)
Observations	182,166	182,166	182,166	155,762	129,863
Adjusted R^2	0.172	0.172	0.172	0.170	0.192
Panel B: Residents					
Effect of housestaff with measure above median	-0.034 (0.035)	-0.050 (0.030)	-0.20 (0.039)	0.040 (0.036)	-0.025 (0.054)
Observations	200,276	200,276	200,276	182,329	174,834
Adjusted R^2	0.181	0.181	0.181	0.181	0.187
Measure and median within service	Y	Y	Y	N	Y

Note: This table reports results for some regressions of the effect of indicators of housestaff experience. Panel A shows results for interns; Panel B shows results for residents. Regressions are of the form in Equation (3), where the coefficient of interest is on an indicator for a group of housestaff identified whether their measure (e.g., number of days) is above the median within a 60-day tenure interval (across all housestaff). 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 housestaff 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 housestaff 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 housestaff 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 housestaff 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: Mean Effect of Resident Tenure and Variation across Residents

	(1) Daily log test spending	(2) Daily log total spending	(3) Log length of stay	(4) 30-day readmit	(5) 30-day mortality
Mean resident tenure effect regression					
Third-year resident	0.0057	0.0035	0.0072	0.0028	0.0002
	(0.0070)	(0.0042)	(0.0060)	(0.0038)	(0.0027)
Observations	219,727	219,727	48,175	47,874	48,175
Adjusted R^2	0.138	0.087	0.271	0.046	0.193
Counterfactual outcomes (none are log)					
Mean outcome	\$123.75	\$1,279.57	3.996	0.124	0.071
Third-year resident	\$124.45	\$1,284.07	4.024	0.127	0.071
	(\$0.87)	(\$5.35)	(0.024)	(0.004)	(0.003)
1 s.d. increase in resident effect	\$210.81	\$1,563.90	4.346	0.168	0.107
	(\$4.08)	(\$11.62)	(0.019)	(0.003)	(0.003)
1 s.d. decrease in resident effect	\$72.64	\$1,046.93	3.674	0.091	0.047
	(\$1.40)	(\$7.77)	(0.016)	(0.003)	(0.002)

Note: In the top panel, this table reports results of regressions of various outcomes on having a third-year (as opposed to a second-year resident), as defined by Equation (7). Total spending includes imputed costs, such as physician and nurse salaries and operating costs. The third-year coefficient is insignificant in all of the models. In the bottom panel, mean (non-logged) outcomes are reported, a counterfactual for having a third-year resident (assuming that a second-year resident was previously responsible for the mean outcome), and counterfactuals for switching the resident for another one who has a spending effect one standard deviation higher or lower in the relevant outcomes. Distributional counterfactuals are generated by random-effect models, discussed in Appendix A-2. The random effect models are linear for daily log spending and log length of stay; they are logistic for readmissions and mortality. Random-effect models are estimated for the entire sample, assuming constant spending effects within the two years in the role of resident. Therefore, variation in spending effects is less than in baseline Equation (2) which allows tenure-specific spending effects. Standard errors are shown in parentheses.

Appendix (for Online Publication per Referees / Editor)

A-1 Quasi-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 housestaff. Section A-1.2 presents the assignment of housestaff to supervising physicians.

A-1.1 Assignment of Patients to Housestaff

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

$$X_a = \mathbf{T}_t\eta + \mu_s\left(a \in Service_s\right) + \zeta_i^{\tau<T} + \zeta_j^{\tau>T} + \zeta_k + \varepsilon_{aijtk}, \tag{A-1}$$

where X_a is some patient characteristic or linear combination of patient characteristics for the patient at a unique admission a at time t, being cared for by intern i, resident j, and attending k on the day of admission. $\mathbf{T}_t$ is a set of time categories, including the day of the week and the month-year interaction; μ_s is a fixed effect that corresponds to the admitting service s (e.g., "heart failure service" or "oncology service"). $\zeta_i^{\tau<T}$, $\zeta_j^{\tau>T}$, and ζ_k are fixed effects for the intern i, resident j, and attending k, respectively. For simplicity, I do not impose any relationship between the fixed effect of a housestaff as an intern and the fixed effect of the same housestaff as a resident. I then test for the joint significance of the fixed effects $\left(\zeta_i^{\tau<T}, \zeta_j^{\tau>T}\right)_{i\in\mathcal{I},j\in\mathcal{J}}$.

In column (1) of Table A-1, I show F-statistics and the corresponding p-values for the null hypothesis that $\left(\zeta_i^{\tau<T}, \zeta_j^{\tau>T}\right)_{i\in\mathcal{I},j\in\mathcal{J}} = \mathbf{0}$. 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 sex, and a dummy for white race.[28] I also perform (A-1) using as dependent variables the linear prediction of log admission test 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 housestaff characteristics in regressions of this form:

$$X_a = \mathbf{T}_t\eta + \mu_s\left(a \in Service_s\right) + \gamma_1 Z_i + \gamma_2 Z_j + \zeta_k + \varepsilon_{aijkt}. \tag{A-2}$$

Equation (A-2) is similar to Equation (A-1), except for the use of a vector of housestaff characteristics Z_i and Z_j for intern i and resident j, respectively, to test whether certain types of residents are more likely to be assigned certain types of patients. Housestaff characteristics include the following: position on the rank list; USMLE Step 1 score; sex; age at the start of training; and dummies for foreign medical school, rare medical school, AOA honor society membership, PhD or another graduate degree, and racial minority.

[28] I do not test for balance in patient diagnoses, because these are discovered and coded by physicians potentially endogenous. Including or excluding them in the baseline specification of Equation (2) does not qualitatively affect results.

Columns (2) and (3) of Table A-1 show F-statistics and the corresponding p-values for the null hypothesis that $(\gamma_1, \gamma_2) = \mathbf{0}$. Column (2) includes all housestaff characteristics in Z_h; column (3) excludes position on the rank list, since this information is missing for a sizeable proportion of housestaff. 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 distribution of patient age and the predicted test costs across patients admitted to interns and residents with high or low test spending effects, which previously I estimate in a regression of this form:

$$Y_{aijkt} = \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_i^{\tau<T} + \zeta_j^{\tau>T} + \zeta_k + \varepsilon_{aijkt}, \tag{A-3}$$

where Y_{aijkt} is log test spending, $\mathbf{X}_a$ is a set of admission characteristics as described in Section 3, $\mathbf{T}_t$ is a set of time categories, and intern, resident, and attending fixed effects denoted similarly as in Equation (A-1). Figure A-2 shows kernel density plots of the age distributions for patients assigned to interns and residents, respectively, each of which compare housestaff with practice styles above and below the mean. Figure A-3 plotting the distribution of predicted spending for patients assigned to housestaff 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 housestaff 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 Housestaff to Other Providers

To test whether certain types housestaff are more likely to be assigned to certain types of housestaff and attending physicians, I perform the following regressions:

$$\hat{\zeta}_h^r = \gamma_h\hat{\zeta}_{-h}^{1-r} + \gamma_k\hat{\zeta}_k + \varepsilon_{ijka}, \tag{A-4}$$

where $r \equiv \mathbf{1}(\tau > T)$ is an indicator for whether the fixed effect for housestaff h was calculated while h was an intern ($r = 0$) or a resident ($r = 1$). 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. $\hat{\zeta}_k$ is the estimated fixed effect for attending k.[29] Estimates for γ_h and γ_k 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 housestaff with above- or below-mean

[29] I use two approaches to get around the reflection problem due to the first-stage joint estimation of $\zeta_i^{(0)}$, $\zeta_j^{(1)}$, and ζ_k (Manski, 1993). First, I perform (A-4) using "jack-knife" estimates of fixed effects, in which I exclude observations with $-h$ and k 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 (i, j, k).

spending practice styles. In Figure A-4, the practice-style distribution for attendings is similar for those assigned to high- vs. low-spending housestaff. 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 Time-Fixed Practice Variation

This appendix discusses in more detail the approach I take in Section 4, where I estimate practice variation attributable to residents in several outcome measures. Unlike the fuller statistical model in Appendix A-5, used in Sections 5 and 6, I focus on variation due to the resident and consider this variation to be fixed over time. However, this approach allows for both continuous and binary outcomes, whereas the model in Appendix A-5 is restricted to continuous outcomes.

Consider an outcome Y_{aijkt} observed for admission a at time t under intern i, resident j, and attending k. In the first step, I calculate a linear projection of the outcome $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$ using patient-admission characteristics $\mathbf{X}_a$, a vector of time categories $\mathbf{T}_t$, and the attending identity k, using only within-housestaff variation. That is, I estimate the regression

$$Y_{aijkt} = \mathbf{X}_a\beta + \mathbf{T}_t\eta + \zeta_k + \nu_{ij} + \varepsilon_{aijkt},$$

and calculate $P_Y(\mathbf{X}_a, \mathbf{T}_t, k) = \mathbf{X}_a\hat{\beta} + \mathbf{T}_t\hat{\eta} + \hat{\zeta}_k$. This linear projection is easy to calculate and avoids incidental parameter problems in nonlinear models for binary outcomes.

Unlike the approach in Appendix A-5, I cannot simply difference $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$ from Y_{aijkt} and estimate a model restricted maximum likelihood (REML) model, because Y_{aijkt} may be a binary variable. I therefore estimate a random effects model that controls for patient-admission characteristics, time categories, and attending identities via this linear projection:

$$Y_{ajkt} = g\left(P_Y(\mathbf{X}_a, \mathbf{T}_t, k)\right) + \xi_j + \varepsilon_{ajkt}, \tag{A-5}$$

omitting i in the subscript for the outcome. $g(\cdot)$ is a potentially flexible nonlinear function that may be used if Y_{ajkt} is a binary outcome determined by a nonlinear model.[30] ξ_j is a resident effect, and ε_{ajkt} is an independent error term that is normal for continuous outcomes of log spending and log length of stay and logistic for binary outcomes of readmission and mortality.

Because I control for $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$ directly in Equation (A-5), I need to explicitly allow for correlations between ξ_j and $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$. I allow for this in a correlated random effects approach (Abowd et al., 2008) by considering two components of ξ_j – u_j that is correlated with $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$, and v_j that is uncorrelated with u_j and $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$ – and by modeling u_j as a

[30] I operationalize $g(\cdot)$ as a linear function of cubic splines. Results are insensitive to whether I allow for $g\left(P_Y(\mathbf{X}_a, \mathbf{T}_t, k)\right)$ or simply take the projection linearly as $\gamma P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$. In my baseline results, I use the latter approach.

projection of the empirical expectation of $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$ conditional on j. Define this empirical expectation as

$$\hat{P}_{Y|j} = \frac{\sum \mathbf{1}(j(a) = j) P_Y(\mathbf{X}_a, \mathbf{T}_t, k)}{\sum \mathbf{1}(j(a) = j)},$$

where $\mathbf{1}(j(a) = j)$ is an indicator for whether an admission a, for which there exists a projection $P_Y(\mathbf{X}_a, \mathbf{T}_t, k)$, is associated with resident j. I then estimate Equation (A-5) as

$$Y_{ajkt} = g(P_Y(\mathbf{X}_a, \mathbf{T}_t, k)) + \delta \hat{P}_{Y|j} + v_j + \varepsilon_{ajkt}. \tag{A-6}$$

The term $\delta \hat{P}_{Y|j}$ absorbs the component u_j.[31] In applications of correlated random effects that are interested in the variance of ξ_j, both components u_j and v_j are used in this calculation, i.e., $\widehat{\text{Var}}(\xi_j) = \widehat{\text{Var}}\left(\hat{\delta}\hat{P}_{Y|j}\right) + \widehat{\text{Var}}(v_j)$. However, part of Section 4 aims to compare the effects on different outcomes among residents who are as good as randomly assigned patients. Therefore, I focus on empirical Bayes predictions of v_j formed by estimates of (A-6) (Searle et al., 1992). This exercise compares resident effects that are orthogonal to projections of average patient characteristics, time categories, and attending identities, and more closely approximates the conditional random assignment design that one obtains under the REML approach in Sections 5 and 6 (Appendix A-5).

A-3 Conceptual Framework of Team Decisions

A-3.1 Influence in Team Decisions

Consider a simple team-theoretic environment of decision-making (e.g., Cyert and March, 1963; Radner, 1993; Garicano, 2000), in which team members use the information they have to make the best decision for caring for a particular patient.[32] The team must take an action a to match an unknown state θ, and will receive utility

$$u(a; \theta) = -(\theta - a)^2. \tag{A-7}$$

The team responsible for the care of a patient is comprised of two housestaff agents, a first-year "intern" i and a second- or third-year "resident" j. These two agents also operate within a practice environment, including other supervising ("attending") and consulting physicians, institutional rules (e.g., they are required to get consultant approval to order expensive tests in certain cases), and known standards of practice at the institution and more broadly.

[31] I also consider versions of Equation (A-6) that allow for cubic splines of $\delta \hat{P}_{Y|j}$, and results are robust to these nonlinear transformations.

[32] Although the experimental literature has shown that agents may have intrinsic utility for influence (Bartling et al., 2014), I abstract from heterogeneous preferences or specialization at the individual physician level to highlight the simple mechanism that more-experienced agents should have greater influence in the absence of moral hazard. However, the intuition should follow in more complicated settings as long as there is a common component to the decision that is agreed upon by both agents, and there is incomplete information about that component.

The intern has a normal prior subjective distribution of θ, with mean m_i and precision g_i, dropping reference to time for simplicity. The resident also has a normal subjective distribution of θ, with mean m_j and precision g_j. Finally, I model the practice environment by another "prior" with mean 0 and precision G. If agents $h \in \{i, j\}$ can communicate (m_h, g_h) and know G, the utility-maximizing action is

$$a^* = \frac{g_i m_i + g_j m_j}{g_i + g_j + G}. \tag{A-8}$$

This framework illustrates that the "best guess" or mean of each housestaff's belief is weighted by a factor akin to *influence* on the team and within the practice environment, $g_h / (g_h + g_{-h} + G)$.

The more precise her signal is relative to her teammate and the practice environment, the greater her influence will be. Because teams are always comprised of an intern and a resident, when a housestaff'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. This discontinuous decrease in g_{-h} results in a discontinuous increase in her influence (and the variation in medical care attributable to newly minted residents relative to seasoned interns), even if m_h and g_h are continuous across time. With respect to the practice environment, a housestaff's influence will be lower in a tighter practice environment with higher G. At the extreme, if care were dictated by attending physicians or guidelines, there should be no variation attributable to housestaff.

A-3.2 Learning and Convergence

Next consider convergence due to learning, or the process by which housestaff beliefs change over time. The key intuition is that the rate of learning may depend on the the amount and accessibility of knowledge to be learned, because learning requires accessing outside knowledge and incorporating it to future clinical practice.[33] This intuition appeals to a broad literature on search theory (see e.g., Rogerson et al. 2005, for a review), which allows physician learning to slow down or stop if the search costs of learning exceed the benefits.[34] I model this in reduced-form as a precision function $g_h = g(\tau; \mathcal{K})$ that depends on the tenure τ (or experience) of housestaff h and implicitly on the practice environment $\mathcal{K}$ in which the housestaff learns. Under classical Bayesian learning, the distribution of subjective means m_h conditional on tenure τ has mean 0 and standard deviation $g(\tau; \mathcal{K})^{-1/2}$.

[33] Although θ is known perfectly *ex post* in the setup in Section A-3.1, one may consider θ to be imperfectly observed (e.g., observed with some noise), imperfectly remembered, or most importantly imperfectly informative for future patients, who will be different, in the absence of devoting some cost to learning.

[34] 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. In this section I am agnostic about the mechanism of learning, except that agents increase the precision of their beliefs with experience. One intriguing possibility, that seems consistent with some of the numerical results in Appendix A-4, is that housestaff learn more as residents because they get feedback on decisions that they influence, an idea explored in psychology (Csikszentmihalyi, 1990).

Thus, restating Equation (A-8) as

$$a^* = a_i^* + a_j^* = \frac{g_i m_i}{g_i + g_j + G} + \frac{g_j m_j}{g_i + g_j + G}, \tag{A-9}$$

the standard deviation $\sigma(\tau; \mathcal{K})$ of experience-specific housestaff effects $a_{h,\tau}^*$ can be stated as

$$\sigma(\tau; \mathcal{K}) = \frac{g(\tau; \mathcal{K})^{1/2}}{g(\tau; \mathcal{K}) + g(\tau + \Delta; \mathcal{K}) + G}, \tag{A-10}$$

where the index cohort $\{h\}$ has tenure τ and the cohort $\{-h\}$ of the other team member has tenure $\tau + \Delta$, where Δ may be positive or negative. At time t relative to the beginning of the academic year, intern tenure is t, and resident tenure is $t + T$ or $t + 2T$, where T is one year, for second- or third-year residents, respectively.

Define convergence as a reduction in $\sigma(\tau; \mathcal{K})$ with time, i.e., as $\partial\sigma(\tau; \mathcal{K})/\partial\tau < 0$ within academic years. Unlike in settings where there is a single decision-maker and $G = 0$, $g'(\tau) > 0$ is not sufficient for convergence. First, convergence in variation attributable to a decision-maker is muted when that decision-maker's influence is limited. Second, as long as influence is limited (i.e., there are other agents with information), increasing $g(\tau)$ may primarily increase influence and therefore even widen variation. I explore these implications further and provide numerical examples in Appendix A-4.

A-3.3 Tacit Knowledge and Hierarchy

One reason why learning can only increase with tenure is that knowledge cannot be costlessly passed from senior to junior housestaff. That is, knowledge is tacit, or equivalently, there are informational frictions in the transfer of knowledge that can have important implications for practice variation, even with identical preferences and no systematic differences in experiences (i.e., "schools of thought") across agents in the same cohort. A natural extension of this is that agents $h \in \{i, j\}$ cannot fully communicate their beliefs, in particular g_h, even in the specific case of the decision at hand. This leads to the idea of *hierarchies*, in which agents may simply weight m_h based on characteristics of h such as tenure or perceived expertise that are imperfectly correlated with g_h.

This particularly applies to hierarchies in which there is no asymmetric decision rights or formal mechanism for one agent to reward or punish another.[35] In this sense, hierarchies are

[35] Another explanation for why subordinate agents contribute less information about θ than managers do involves *moral hazard*, outside of the team-theoretic framework: Agents exert private effort to gather information, managers are principals who incentivize agents to exert effort but only can assess this by gauging $\hat{m}_i$ relative to m_j, where j is now the principal, and agents can observe m_j (Scharfstein and Stein, 1990; Prendergast, 1993). However, in this and many other settings, senior team members (i.e., residents) are not principals and cannot provide incentives. Only attending physicians perform housestaff evaluations, with only weak career implications. If attending physicians assess intern effort by comparing $\hat{m}_i$ to m_j, this must still be founded upon learning, in which $g_j > g_i$.

analogous to directives within a firm as discussed by Alchian and Demsetz (1972): They are not defined by "some superior authoritarian directive or disciplinary power," but rather they arise endogenously given the underlying information structure and production process, and given the need to settle on a routine that is most likely to yield efficient outcomes. Of note, absent learning and tacit knowledge, Garicano (2000) and Garicano and Rossi-Hansberg (2006) predict "management by exception" under which managers with larger *spans of control* (i.e., the resident, who is assigned to two interns) should have *less* influence over average daily decisions. The prediction here does not contradict their important insight but highlights an additional mechanism, due to learning, in which senior agents can have greater influence under tacit knowledge.

A-4 Variation over Time under Example Learning Parameters

This appendix further explores the implications of the conceptual framework in Section A-3, 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-4.1 Analytical Evaluation

Consider the standard deviation of experience-specific housestaff effects $a^*_{h,\tau}$, originally stated in Equation (A-10):

$$\sigma(\tau) = \frac{g(\tau)^{1/2}}{g(\tau) + g(\tau + \Delta) + G}, \tag{A-11}$$

omiting reference to the learning environment $\mathcal{K}$ for brevity. $\sigma(\tau)$ can be thought of a profile of practice variation across housestaff over different tenure periods, akin to the profiles empirically estimated in the paper (e.g., Figure 3). $g(\tau)$ is the precision of a housestaff's subjective prior, given that the housestaff has tenure τ, and can be thought of as related to learning over τ: Greater $g(\cdot)$ reflects greater knowledge; greater $g'(\cdot)$ reflects faster learning. In the standard case, assume that $g'(\cdot) > 0$, i.e., there is no "forgetting." Δ is the tenure difference between housestaff of tenure τ and other housestaff whom this group works with. Finally, recall that G reflects the strength of the external practice environment, or the precision of the "prior" that includes attending physicians and institutional rules, which I will refer to as the "external prior."

A few observations about practice variation and learning can be made. First, note that the scale and the shape of the practice variation profiles can be separately rationalized.

Proposition A-1. *Consider a practice variation profile, $\sigma(\tau)$, that exists under a learning profile $g(\tau)$ and external prior G. Then $\kappa\sigma(\tau)$ also exists for any constant κ.*

Proof. The learning profile $g(\tau)/\kappa^2$ and external prior G/κ^2 yield the desired practice variation profile $\kappa\sigma(\tau)$ under Equation (A-11). □

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. However, variation across agents in their mean beliefs will be increased (or decreased) as they all have subjective prior distributions smaller (or greater) precisions.

Next, consider the discontinuity in practice variation across the one- and two-year tenure marks. Recall that at the beginning of the academic year in June, new interns (first-year housestaff) arrive, and experienced interns proceed to the role of resident. Housestaff train for a total of three years, so that in June there are both residents with one year of training and two years of training.

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

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

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

$$\frac{\sigma(T^+)}{\sigma(T^-)} = \frac{g(T) + g(2T) + G}{g(T) + g(0) + G},$$

and

$$\frac{\sigma(2T^+)}{\sigma(2T^-)} = \frac{g(2T) + g(T) + G}{g(2T) + g(0) + G}.$$

Since $g(\cdot)$ is monotonically increasing, $g(0) < g(T) < g(2T)$, which yields our result. □

Because there is a change in the tenure of the other housestaff 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 $\tau_h = T$ is always larger than the increase at $\tau_h = 2T$ for two reasons, both related to the monotonic increase in precision with tenure: First, housestaff at $\tau_h = T$ have less precise subjective priors than those at $\tau_h = 2T$, so any decrease in the relative tenure of their peer housestaff increases their influence by more. Second, the decrease in the relative tenure of the peer is greater at $\tau_h = T$ (from $\tau_{-h} = 2T$ to $\tau_{-h} = 0$) than at $\tau_h = 2T$ (from $\tau_{-h} = T$ to $\tau_{-h} = 0$). I will show below in the numerical examples that, within this framework, this difference in the discontinuous increases at $\tau_h = T$ and at $\tau_h = 2T$ can be quite large, and that the discontinuity at $\tau_h = 2T$ can be quite trivial. Of course, there are other reasons for a negligible discontinuity at $\tau_h = 2T$, including discrete decisions and rules of thumb, such as titles of "resident" and "intern" meaning more than actual tenure within titles.

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

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

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

$$\sigma'(\tau) < 0 \Leftrightarrow g(\tau) > \frac{g'(\tau)}{2g'(\tau+\Delta) + g'(\tau)} (g(\tau+\Delta) + G). \tag{A-12}$$

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

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

$$g(\tau) = k_0 + k_1 \min(\tau, T) + k_2 \max(\tau - T, 0). \tag{A-13}$$

For any $g(\tau)$ that satisfies the form (A-13), conditional on some $\Delta > 0$ (i.e., $\tau < T$), 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., $\tau > T$), 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

$$\tau^*_{\Delta>0} = \frac{G + k_1 T + k_2(\Delta - T) - 2k_0 k_2/k_1}{k_1 + k_2}; \tag{A-14}$$

$$\tau^*_{\Delta<0} = \frac{G + k_1 \Delta - 2k_1 (k_0 + k_1 T)/k_2}{k_1 + k_2} + T. \tag{A-15}$$

Proof. State the convergence condition in Equation (A-12) as a criterion function $\mathcal{G}(\tau; \Delta)$ in which convergence occurs if and only if $\mathcal{G}(\tau; \Delta) > 0$:

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

Under any $g(\tau)$ of the form (A-13), $\mathcal{G}(\tau;\Delta)$ is monotonically increasing in τ, which implies a single solution to $\mathcal{G}(\tau^*_\Delta;\Delta) = 0$ conditional on Δ. To arrive at the specific functions that $\tau^*_{\Delta>0}$ and $\tau^*_{\Delta<0}$ take in Equations (A-14) and (A-15), plug Equation (A-13) into $\mathcal{G}(\tau^*_\Delta;\Delta) = 0$ and solve for τ^*_Δ. □

Note that $\tau^*_{\Delta>0}$ in Equation (A-14) may be less than 0 or greater than T. In the former case, there is convergence for all $\tau \in [0,T]$ (the entire intern year); in the latter case, there is divergence (variation is increasing) for all $\tau \in [0,T]$. If $\tau^*_{\Delta>0} \in (0,T)$, then variation in practice styles first increases then decreases. Similarly, practice variation may be increasing over the tenure period as a resident $\tau \in [T,3T]$, decreasing over the entire period, or first increasing then decreasing.[36] As noted above, and by comparing (A-14) and (A-15), convergence is more likely and occurs earlier during the period as resident than during the period as intern.

A-4.2 Numerical Examples

Figure A-1 presents a few numerical examples of variation profiles under different learning profiles described by functions of the piecewise linear form in Equation (A-13). The three parameters of interest are k_0, or the precision of subjective beliefs before starting training; k_i, or the rate of increase in the precision during intern year; and k_j, or the rate of increase during the subsequent two years as a resident. I normalize the scale of time with $T = 1$, so that k_i and k_j also represent increases in the precision *per year*, and the precision of beliefs at the end of training is $g(3T) = k_0 + k_i + 2k_j$. I also normalize $G = 1$, 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_i, k_j, and G by some constant.

I discuss each panel of Figure A-1 in turn:

- Panel A considers equal $k_0 = k_i = k_j = 0.2$, which are relatively small compared to $G = 1$. 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 $t = T$ than at $t = 2T$. Variation increases in intern year and decreases but only slightly in the next to years as resident.

- Panel B imposes no resident learning ($k_j = 0$) and presents the limiting case in which discontinuous increases in variation at $t = T$ and $t = 2T$ are the same. Variation is still

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

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_j by increasing the ratios of k_0 and k_i to k_j. The scale of variation is smaller than in Panel B, which reflects that precision in housestaff beliefs are now larger. A rescaled version with smaller precisions (and smaller G) would reveal larger relative increases in variation at the discontinuities.

- Panel D examines increasing k_i 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 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 $\tau = T$ and $\tau = 2T$ 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 $\tau = T$, while the increase is negligible at $\tau = 2T$. 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_i/k_j) is similar, but learning during training is reduced relative to knowledge gained prior to training (k_0) and to the external practice environment (G).

A-5 Statistical Model of Housestaff Effects

In this appendix, I introduce a statistical model to estimate the standard deviation $\sigma(\tau)$ of housestaff effects $a^*_{h,\tau}$ in discrete tenure period τ and the correlation $\rho(\tau_1, \tau_2)$ between housestaff effects a^*_{h,τ_1} and a^*_{h,τ_2} in two discrete periods τ_1 and τ_2. Random assignment of patients to housestaff, conditional on time categories, allows me to estimate housestaff effects.[37] Finite observations per housestaff-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 estimates $\hat{a}^*_{h,\tau_1}$and $\hat{a}^*_{h,\tau_2}$ 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),[38] however, deal with a single effect entering the right-hand side of each

[37] I do not strictly require conditional random assignment of patients to housestaff 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.

[38] Recent examples of papers that have used this procedure include Kane and Staiger (2002), Jacob and Lefgren (2007), and **?**.

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 housestaff 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 a^*_{h,τ_1} and a^*_{h,τ_2} are not eliminated but only shrunken.[39]

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 $\tilde{Y}_{aijkt} = Y_{aijkt} - \left(\mathbf{X}_a\hat{\beta} + \mathbf{T}_t\hat{\eta} + \hat{\zeta}_k\right)$, where $\hat{\beta}$, $\hat{\eta}$, and $\hat{\zeta}_k$ are estimated by using variation within housestaff pairs and discrete tenure periods. This allows random housestaff effects to be correlated with $\mathbf{X}_a$, $\mathbf{T}_t$, and ζ_k.[40] Note that $E\left[\tilde{Y}_{aijkt}\,|a,k,t\right] = 0$ for all a, k, and t. In practice, given quasi-random assignment of attending physicians and patients to housestaff, conditional on schedules, I am only concerned with correlations between housestaff effects and $\mathbf{T}_t$, but differencing out projections due to $\mathbf{X}_a$ and ζ_k 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-5.1 Standard Deviation of Housestaff Effects

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

$$\tilde{Y}_{aijkt} = \xi^{\tau}_{h} + \xi^{\tau+\Delta}_{-h} + \varepsilon_{aijkt}, \tag{A-16}$$

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

$$\tilde{Y}_{aijkt} = \xi^{\tau}_{h} + \xi^{\tau+\Delta}_{-h} + \nu_a + \varepsilon_{aijkt}, \tag{A-17}$$

[39] 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 housestaff because of both learning and influence.

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

where ν_a is an admission effect.[41] Because housestaff are assigned conditionally randomly to each other and to patients, ξ_h^τ, $\xi_{-h}^{\tau+\Delta}$, and ν_{ai} are uncorrelated with each other. Assuming ξ_h^τ, $\xi_{-h}^{\tau+\Delta}$, and ν_a are normally distributed, their standard deviations $\sigma_{\xi,\tau}$, $\sigma_{\xi,\tau+\Delta}$, and σ_ν are estimated by the standard maximum-likelihood method.

Equations (A-16) and (A-17) can be stated in vector form:

$$\tilde{\mathbf{Y}} = \mathbf{Z}\mathbf{u} + \varepsilon, \tag{A-18}$$

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

Let N_h be the number of housestaff with tenure τ and N_{-h} be the corresponding peers observed in the sample. Then in the case that (A-18) represents (A-16), $\mathbf{Z}$ is an $n \times (N_\tau + N_{\tau+\Delta})$ selection matrix for housestaff with tenure τ and their peers, and $\mathbf{u}$ is an $(N_\tau + N_{\tau+\Delta}) \times 1$ stacked vector of housestaff and peer random effects. The variance-covariance matrix of $\mathbf{u}$ is diagonal:

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

Similarly, in the case that (A-18) represents (A-17), $\mathbf{Z}$ is an $n \times (N_\tau + N + N_a)$ selection matrix for intern i, resident j, and admission a, and $\mathbf{u}$ is an $(N_i + N_j + N_a) \times 1$ stacked vector of intern, resident, and admission random effects, where N_a is additionally the number of admissions in the sample. The diagonal variance-covariance matrix of $\mathbf{u}$ is

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

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

$$L = -\frac{1}{2}\left\{ n \log(2\pi) + \log|\mathbf{V}| + \tilde{\mathbf{Y}}'\mathbf{V}^{-1}\tilde{\mathbf{Y}} \right\}. \tag{A-19}$$

I thus estimate (A-16) or (A-17) by maximum likelihood, for each τ separately. Although each estimation yields results for both $\sigma_{\xi,\tau}$ and $\sigma_{\xi,\tau+\Delta}$, the parameter of interest for a given τ is $\sigma_{\xi,\tau} \equiv \sigma(\tau)$. Note that for τ corresponding to interns, the peer housestaff are residents who may have tenure one or two years greater than τ, and the distribution of $\xi_{-h}^{\tau+\Delta}$ should not be interpreted as tenure-specific. For τ corresponding to residents, $\sigma_{\xi,\tau+\Delta}$ is estimated for only part of the sample of interns working with residents of tenure τ.

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

A-5.2 Correlation of Housestaff Effects

To estimate $\rho(\tau_1, \tau_2)$, I augment models in (A-16) and (A-17) to account for two separate tenure periods τ_1 and τ_2 across which housestaff effects may be correlated. Although I observe each housestaff across their entire training, I only observe a subset of these housestaff in each 60-day or 120-day tenure period, and the number of housestaff observed in two different tenure periods is even smaller. Because housestaff 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 housestaff observed in both tenure periods.

Specifically, in place of Equation (A-16), I consider

$$\tilde{Y}_{aijkt} = \xi_h^{\tau(h,t)} + \xi_{-h}^{\tau+\Delta} + \varepsilon_{aijkt}, \tag{A-20}$$

which features the function $\tau(h,t) \in \{\tau_1, \tau_2\}$. This specifies that effects of housestaff 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 peer housestaff (with tenure $\tau + \Delta$) are pooled into a single distribution. The analog for Equation (A-17) is

$$\tilde{Y}_{aijkt} = \xi_h^{\tau(h,t)} + \xi_{-h}^{\tau+\Delta} + \nu_a + \varepsilon_{aijkt}. \tag{A-21}$$

As above, both (A-20) and (A-21) can be written in the vector form of (A-18). When representing (A-20) as (A-18), the selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_{\tau+\Delta})$, since it now maps observations onto one of two random effects of the index housestaff h, depending if $\tau(h,t) = \tau_1$ or $\tau(h,t) = \tau_2$. The stacked vector of random effects $\mathbf{u}$ is similarly of size $(2N_\tau + N_{\tau+\Delta}) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

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

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 housestaff 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-21) as (A-18) is a similar exercise. The selection matrix $\mathbf{Z}$ is of size $n \times (2N_\tau + N_{\tau+\Delta} + N_a)$, and the vector of random effects $\mathbf{u}$ is of size $(2N_\tau + N_{\tau+\Delta} + N_a) \times 1$. The variance-covariance matrix of $\mathbf{u}$ is

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

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

The log likelihood is the same as in Equation (A-19), but using revised definitions of $\mathbf{G}$ that allow for covariance between random effects of the same housestaff 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-6 Bayesian Refinement of Serial Correlation Estimates

Appendix A-5.2, describes a procedure to estimate the correlation between housestaff effects in any two tenure periods. While I am most interested in evaluating how serial correlation between two adjacent periods changes through training, there is valuable information in the correlation between non-adjacent periods that relates to these parameters of interest. This is particularly the case since I only observe a subset of housestaff practicing in any given pair of periods. The efficient method of incorporating all of this information would be to jointly estimate all correlations at once, but given the computational burden of estimating a crossed random effects model and the large number of observations in the full sample, I am required to keep the specification simple and sample restricted.[42]

Given this, I develop a methodology to refine estimates of the correlation between housestaff effects in adjacent periods based on estimates of other correlations between effects in non-adjacent periods. To be more notationally concrete, assume that τ is an integer from 1 (the first tenure period) to $\tau_{\max} = 15$ (the last tenure period), and denote the set $\mathcal{T} = \{1, \ldots, \tau_{\max}\}$. In this approach, I first infer prior distributions of $\rho(\tau, \tau+1)$ based on other correlations from non-adjacent periods and then use these prior distributions and the maximum-likelihood estimate $\hat{\rho}(\tau, \tau+1)$ described in Appendix A-5.2 to compute a posterior distribution.

[42]Many crossed random effects models without any correlation parameters were computationally feasible until a few years ago when sparse matrix methods became available, which some statistical packages such as Stata have yet to incorporate. With 15 periods, the fully specified model would have 105 correlations to estimate jointly. The fully specified model Further, computational issues are considered important even for "moderately large" datasets, defined as having between 10,000 to 100,000 observations (Bates et al., 2015), while the full dataset of this study has more than 200,000 patient-day observations.

The first step is to use estimates of correlations between non-adjacent periods as information on a correlation $\rho(\tau, \tau+1)$ for some τ. The insight here is that if, for some $\tau' \notin \{\tau, \tau+1\}$, correlations $\rho(\tau, \tau')$ and $\rho(\tau+1, \tau')$ are known, then this information would place bounds on admissible values of $\rho(\tau, \tau+1)$.

Proposition A-4. *Consider random variables X, Y, and Z, such that $Corr(X,Y) = \gamma$ and $Corr(Y,Z) = \varphi$. Then $Corr(X,Z)$ satisfies*

$$\gamma\varphi - \sqrt{(1-\gamma^2)(1-\varphi)^2} \leq Corr(Y,Z) \leq \gamma\varphi + \sqrt{(1-\gamma^2)(1-\varphi^2)}.$$

Proof. Without loss of generality, assume that $E[X] = E[Y] = E[Y] = 0$ and $\text{Var}(X) = \text{Var}(Y) = \text{Var}(Z) = 1$. If these conditions do not hold, we can renormalize the random variables without changing the correlation between them. Consider the projection of Z on X and Y:

$$Z = \alpha X + \beta Y + U, \tag{A-22}$$

where $\text{Corr}(X,U) = 0$ and $\text{Corr}(Y,U) = 0$. In addition, consider the projection of Y on X:

$$Y = \gamma X + V, \tag{A-23}$$

where $\text{Corr}(X,V) = 0$. Observe that the coefficient on X in this projection is indeed γ: $\text{Corr}(X,Y) = \text{Corr}(X, \gamma X + V) = \gamma\text{Corr}(X,X) = \gamma$. Next, substituting (A-23) into (A-22) gives

$$Z = (\alpha + \beta\gamma) X + U + BV. \tag{A-24}$$

Therefore, $\text{Corr}(X,Z) = \text{Corr}(X, (\alpha+\beta\gamma)X + U + BV) = \alpha + \beta\gamma$ since $\text{Corr}(X,U) = 0$ and $\text{Corr}(X,V) = 0$. Hence, we have $\varphi = \alpha + \beta\gamma$, or equivalently, $\alpha = \varphi - \beta\gamma$.

Now we are ready to bound $\text{Corr}(Y,Z) = 0$:

$$\text{Corr}(Y,Z) = \text{Corr}(\gamma X + V, (\alpha+\beta\gamma)X + U + BV) = \gamma(\alpha+\beta\gamma) + \beta\text{Var}(V),$$

using $\text{Corr}(V,U) = \text{Corr}(Y - \gamma X, U) = 0$, since $\text{Corr}(X,U) = \text{Corr}(Y,U) = 0$. In addition, the variance of V can be found from (A-23):

$$1 = \gamma^2 + \text{Var}(V).$$

Hence,

$$\text{Corr}(Y,Z) = \gamma(\alpha+\beta\gamma) + \beta\left(1-\gamma^2\right) = \alpha\gamma + \beta. \tag{A-25}$$

Substituting $\alpha = \varphi - \beta\gamma$ derived above gives

$$\text{Corr}(Y,Z) = \varphi\gamma + \beta\left(1-\gamma^2\right).$$

Since γ and φ are fixed, we only need to bound β to bound $\text{Corr}(Y, Z)$. We will use (A-24), which can be written as

$$Z = \varphi X + U + \beta V,$$

since $\varphi = \alpha + \beta\gamma$. So taking the variance of both sides,

$$1 = \varphi^2 + \text{Var}(U) + \beta^2 \text{Var}(V).$$

We have prviously seen that $\text{Var}(V) = 1 - \gamma^2$, and we know that $\text{Var}(U) \geq 0$. Thus,

$$|\beta| \leq \sqrt{\frac{1-\varphi^2}{1-\gamma^2}}.$$

Substitiuting this inequality into (A-25) produces our result.[43] □

Proposition A-4 would produce sharp bounds for $\rho(\tau, \tau+1)$ if $\rho(\tau, \tau')$ and $\rho(\tau+1, \tau')$, for some $\tau' \notin \{\tau, \tau+1\}$, were known with certainty (and at least one of these correlations is nonzero). However, in practice, both $\rho(\tau, \tau')$ and $\rho(\tau+1, \tau')$ will also be estimated with error. I therefore create prior distributions that generally cover the entire support to create "prior distributions" of $\rho(\tau, \tau+1)$, given data between τ and τ' and between $\tau+1$ and τ'.

These prior distributions and the subsequent Bayesian refinement process will be in a transformed inverse hyperbolic tangent space, which conveniently transforms some correlation $\rho \in [-1, 1]$ to $\tilde{\rho} = \tanh^{-1}\rho \in (-\infty, \infty)$. I characterize estimates of $\rho(\tau, \tau')$ and $\rho(\tau+1, \tau')$ as normal distributions in this transformed space. In particular, let $\hat{\rho}_{0.5}(\tau, \tau')$ denote the maximum-likelihood central estimate, and let $\hat{\rho}_{0.025}(\tau, \tau')$ and $\hat{\rho}_{0.975}(\tau, \tau')$ denote the respective 95% lower and upper confidence limits of $\rho(\tau, \tau')$, as described in Appendix A-5.2. Then switching to a Bayesian framework, I consider $\tilde{\rho}(\tau, \tau')$ as a normally distributed random variable with density:

$$f_{\tilde{\rho}(\tau,\tau')}(x) = \phi\left(x - \tilde{\mu}(\tau, \tau') / \tilde{\sigma}(\tau, \tau')\right), \tag{A-26}$$

where $\phi(\cdot)$ is the normal probability density function and

$$\begin{aligned} \tilde{\mu}(\tau, \tau') &= \tanh^{-1}\hat{\rho}_{0.5}(\tau, \tau'); \\ \tilde{\sigma}(\tau, \tau') &= \frac{\tanh^{-1}\hat{\rho}_{0.975}(\tau, \tau') - \tanh^{-1}\hat{\rho}_{0.025}(\tau, \tau')}{2 \cdot 1.96}. \end{aligned}$$

Now consider the bounds on $\rho(\tau, \tau+1)$ implied by $\rho(\tau, \tau')$ and $\rho(\tau+1, \tau')$ from Proposition

[43] I am grateful to Denis Chetverikov for showing me this result.

A-4. With some abuse of notation, define the lower and upper "bounds," respectively, as

$$\begin{aligned}\rho^{LB}\left(\tau,\tau+1|\tau'\right) &= \rho\left(\tau,\tau'\right)\rho\left(\tau+1,\tau'\right)-\sqrt{\left(1-\rho\left(\tau,\tau'\right)^2\right)\left(1-\rho\left(\tau+1,\tau'\right)^2\right)}, \text{ and}\\ \rho^{UB}\left(\tau,\tau+1|\tau'\right) &= \rho\left(\tau,\tau'\right)\rho\left(\tau+1,\tau'\right)+\sqrt{\left(1-\rho\left(\tau,\tau'\right)^2\right)\left(1-\rho\left(\tau+1,\tau'\right)^2\right)}.\end{aligned}$$

Because both $\rho\left(\tau,\tau'\right)$ and $\rho\left(\tau+1,\tau'\right)$ are estimated with error, I use the central estimates of these correlations, $\hat{\rho}_{0.5}\left(\tau,\tau'\right)$ and $\hat{\rho}_{0.5}\left(\tau+1,\tau'\right)$, to calculate $\hat{\rho}^{LB}_{0.5}\left(\tau,\tau+1|\tau'\right)$ and $\hat{\rho}^{UB}_{0.5}\left(\tau,\tau+1|\tau'\right)$. I then transform these to $\tilde{\mu}^{LB}\left(\tau,\tau+1|\tau'\right)$ and $\tilde{\mu}^{UB}\left(\tau,\tau+1|\tau'\right)$ via the inverse hyperbolic tangent. In order to compute $\tilde{\sigma}^{LB}\left(\tau,\tau+1|\tau'\right)$ and $\tilde{\sigma}^{UB}\left(\tau,\tau+1|\tau'\right)$, I use the delta method, assuming that $\text{Cov}\left(\hat{\rho}\left(\tau,\tau'\right),\hat{\rho}\left(\tau+1,\tau'\right)\right)=0$.[44] I construct a "prior distribution" from the parameters of $\tilde{\rho}^{LB}\left(\tau,\tau+1|\tau'\right)$ and $\tilde{\rho}^{UB}\left(\tau,\tau+1|\tau'\right)$. Note that $\Pr\left(\tilde{\rho}^{LB}<x\right)=\Phi\left(\left(x-\tilde{\mu}^{LB}\right)/\tilde{\sigma}^{LB}\right)$, where $\Phi\left(\cdot\right)$ is the normal cumulative distribution function, and where I have omitted the argument $\left(\tau,\tau+1|\tau'\right)$ for simplicity. Similarly, $\Pr\left(\tilde{\rho}^{UB}>x\right)=\Phi\left(\left(x-\tilde{\mu}^{UB}\right)/\tilde{\sigma}^{UB}\right)$. If $\tilde{\rho}^{LB}$ and $\tilde{\rho}^{UB}$ were known with certainty (i.e., $\tilde{\sigma}^{LB}=\tilde{\sigma}^{UB}=0$), then this prior distribution would have a very simple probability density function:

$$f_{\tilde{\rho}(\tau,\tau+1|\tau')}\left(x\right)\propto\begin{cases}1, & x\in\left[\tilde{\rho}^{LB}\left(\tau,\tau+1|\tau'\right),\tilde{\rho}^{UB}\left(\tau,\tau+1|\tau'\right)\right]\\ 0, & \text{otherwise}\end{cases}.$$

In the presence of uncertainty, I elaborate this density function to

$$f_{\tilde{\rho}(\tau,\tau+1|\tau')}\left(x\right)\propto\begin{cases}\Phi\left(\left(x-\tilde{\mu}^{LB}\right)/\tilde{\sigma}^{LB}\right), & x\leq x_c\\ 1-\Phi\left(\left(x-\tilde{\mu}^{UB}\right)/\tilde{\sigma}^{UB}\right), & x>x_c\end{cases}, \tag{A-27}$$

where $x_c=\left(\tilde{\sigma}^{LB}\tilde{\mu}^{UB}+\tilde{\sigma}^{UB}\tilde{\mu}^{LB}\right)/\left(\tilde{\sigma}^{LB}+\tilde{\sigma}^{UB}\right)$ is chosen to ensure that $f_{\tilde{\rho}(\tau,\tau+1|\tau')}\left(x\right)$ is continuous.

I am now at a point where I can state the posterior distribution, which I denote as $f_{\tilde{\rho}(\tau,\tau+1|\mathcal{T})}$ as a function of the maximum likelihood estimate in (A-26) and the prior distributions in (A-27):

$$f_{\tilde{\rho}(\tau,\tau+1|\mathcal{T})}\left(x\right)\propto f_{\tilde{\rho}(\tau,\tau+1)}\left(x\right)\cdot\prod_{\tau'\notin\{\tau,\tau+1\}}f_{\tilde{\rho}(\tau,\tau+1|\tau')}\left(x\right). \tag{A-28}$$

It can be shown that this function is log-concave. Thus, I am conveniently able to evaluate moments of the posterior distribution, including its mean and 95% credible interval using adaptive rejection sampling (Gilks and Wild, 1992). I finally transform these moments back to the domain of $[-1,1]$ with the hyperbolic tangent function in order to present them as estimates of the

[44] This covariance is unknown because I estimate $\rho\left(\tau,\tau'\right)$ and $\rho\left(\tau+1,\tau'\right)$ separately. In order to estimate the covariance, I would need to estimate them jointly, but of course in such a model, I would also estimate $\rho\left(\tau,\tau+1\right)$. Therefore, bounds would not be necessary with such an approach. The main difficulty with this approach is computational feasibility.

correlation $\rho(\tau, \tau+1|\mathcal{T})$.

A-7 Systematic Placebo Tests

I consider the statistical significance for convergence in the specialist services (i.e., cardiology and oncology) relative to general medicine by performing the following thought experiment. If there is no difference in true convergence between specialist and generalist services, then randomly assigning actual months for each resident on either specialist or generalist services to a placebo specialist or generalist 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 specialist services, then this suggests statistical significance.

I implement these placebo tests as follows:

1. Defining a service as either "specialist" or "generalist," count the number of residents in a specialist service during each month t. Call this number N_t^{spec}. The proportion of residents in cardiology, oncology, and general medicine during each month us shown in Figure A-6.

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

3. Using pseudo-specialist service observations, estimate the standard deviation in resident spending distribution, as described in Appendix A-5, 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 number of possible placebo tests in the procedure above is quite large. For example, consider a representative month in which there are 30 resident-month blocks in the specialist service ($N_t^{\text{spec}} = 30$) out of a total of 55 resident-month-service blocks ($N_t = 55$). The number of random combinations in that month alone, such that we assign exactly 30 resident-month-service blocks to the pseudo-specialist service is

$$\text{Combinations for } t = \frac{55!}{30! \times (55-30)!} = 3.09 \times 10^{15}.$$

Performing this calculation for each of the 62 months in the data and multiplying together yields a total number of combinations of 1.27×10^{970}.

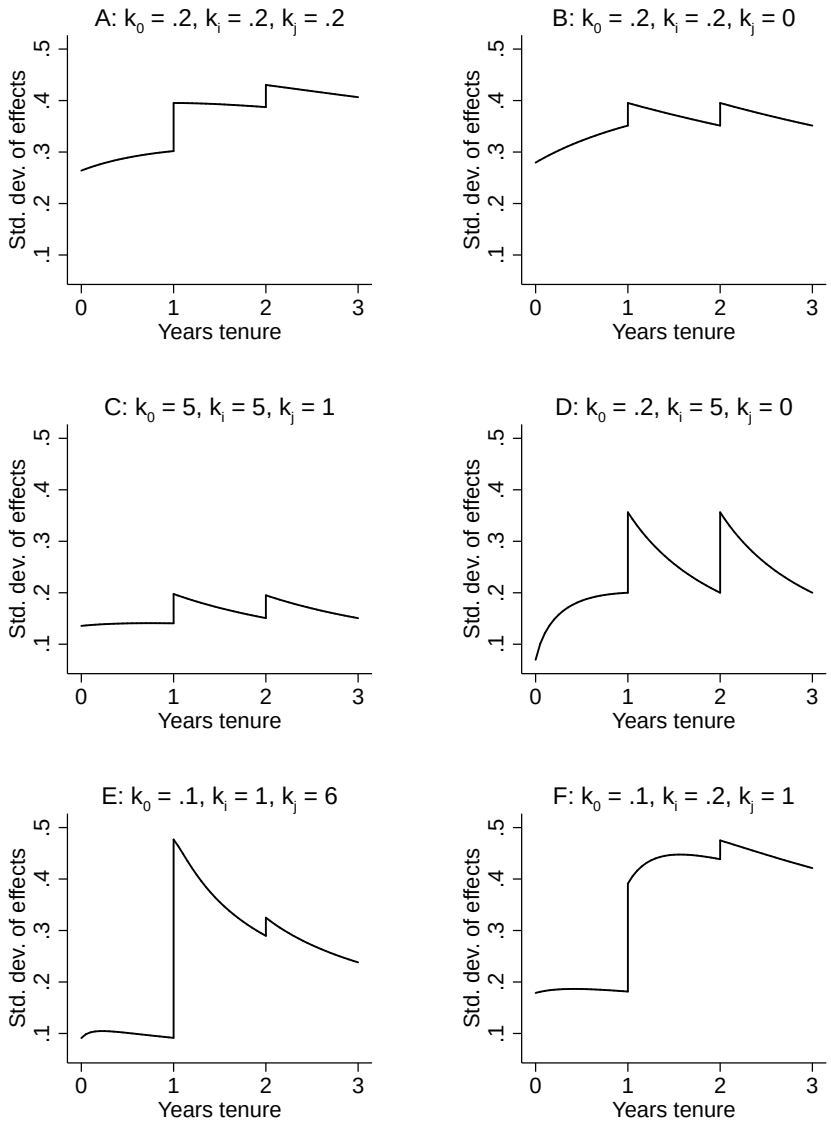
A-8 Additional Results

In this appendix, I describe the following additional appendix tables and figures:

- Figure A-1 shows numerical examples of variation profiles of the standard deviation of housestaff effects over tenure, depending on the underlying learning function, in which the precision of subjective priors is parameterized as a piecewise linear of tenure, $g(\tau)$, as discussed in Appendix A-4.
- Figure A-2 shows distributions of age of patients assigned to high- and low-spending interns and residents.
- Figure A-3 shows distributions of predicted spending (based on patient age, race, and sex) assigned to high- and low-spending interns and residents.
- Figure A-4 shows distributions of attending spending effects for attendings assigned to high- and low-spending interns and residents.
- Figure A-5 shows the distribution of test costs across patient-days.
- Figure A-6 describes the number of observations in terms of patient-days and residents on service for each service across months.
- Figure A-7 shows variation in housestaff effects by tenure for two pseudo-services constructed from the general medicine service. These pseudo-services are constructed by Major Diagnostic Categories (MDCs), separating highly diagnosis-concentrated MDCs into one pseudo-service and leaving the remaining MDCs in the other. The purpose of this is to test the idea that convergence results from more concentrated services. Table A-6 describes summary statistics of both the actual services (cardiology, oncology, and general medicine), as well as these two pseudo-services.
- Figure A-8 shows variation in housestaff effects by tenure, dividing patients in each service by whether they have a primary ICD-9 code (administrative code for diagnosis) that is more or less common than the median observation in each service.
- Figure A-9 shows variation in housestaff effects by tenure, dividing patients in each service by whether there exists a published guideline for a patient's primary ICD-9 code. Guidelines and their linkages to ICD-9 codes are collected from the national guideline repository at `guidelines.gov`.

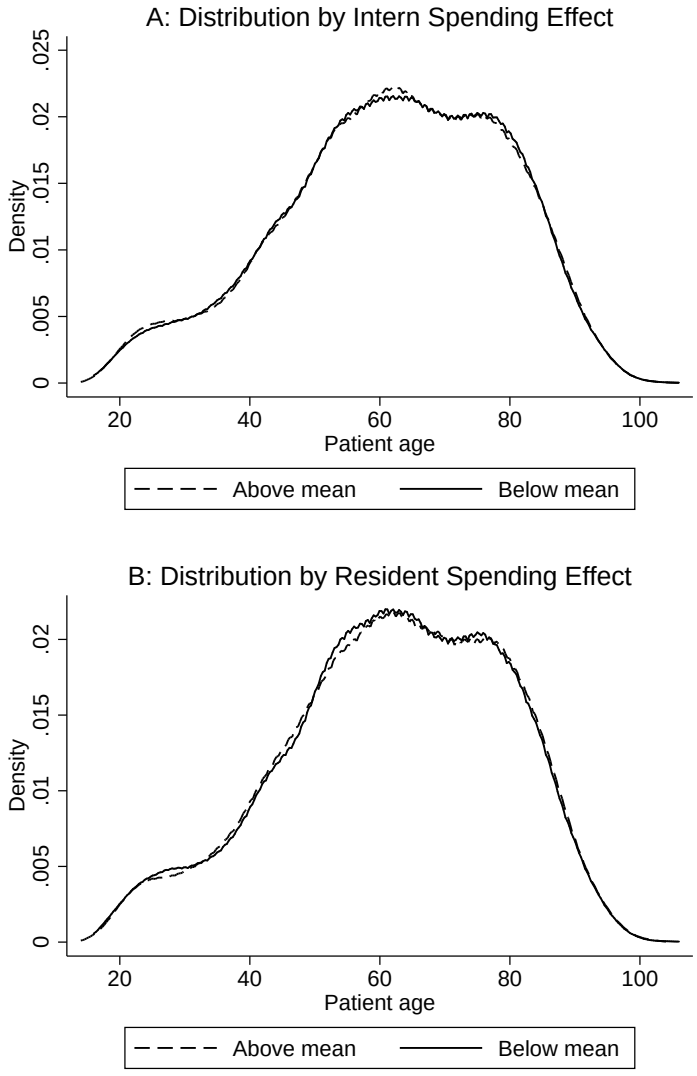
- Table A-1 presents F-statistics testing for the joint significance housestaff identities and housestaff characteristics, as described by Equations (A-1) and (A-2) and in Appendix A-1.
- Table A-2 lists core rotations in the top 24 recognized internal residency programs, as a measure of the organization of medical care in academic hospitals.
- Table A-3 presents the number of core rotations in the universe of US internal medicine residencies, according to the American Council for Graduate Medical Education (ACGME).
- Table A-4 presents the number research papers in the last ten years in the *New England Journal of Medicine*, as a measure of major research activity in different specialties.
- Table A-5 presents the amount of research funding by National Institutes of Health (NIH) Institute or Center,a s a measure of prioritized major research activity in different specialties.
- Table A-6 presents summary statistics for patients admitted to the three ward services (cardiology, oncology, and general medicine), as well as the two pseudo-services constructed from general medicine. Numbers of admissions, MDCs, and ICD-9 coces are also presented, as well as the concentration of MDCs and ICD-9 codes within each service.
- Table A-7 lists the top 15 ICD-9 codes in each service, as well as whether there exists a guideline linked to that diagnostic code in the `guidelines.gov` national repository.

Figure A-1: Numerical Examples of Variation Profiles



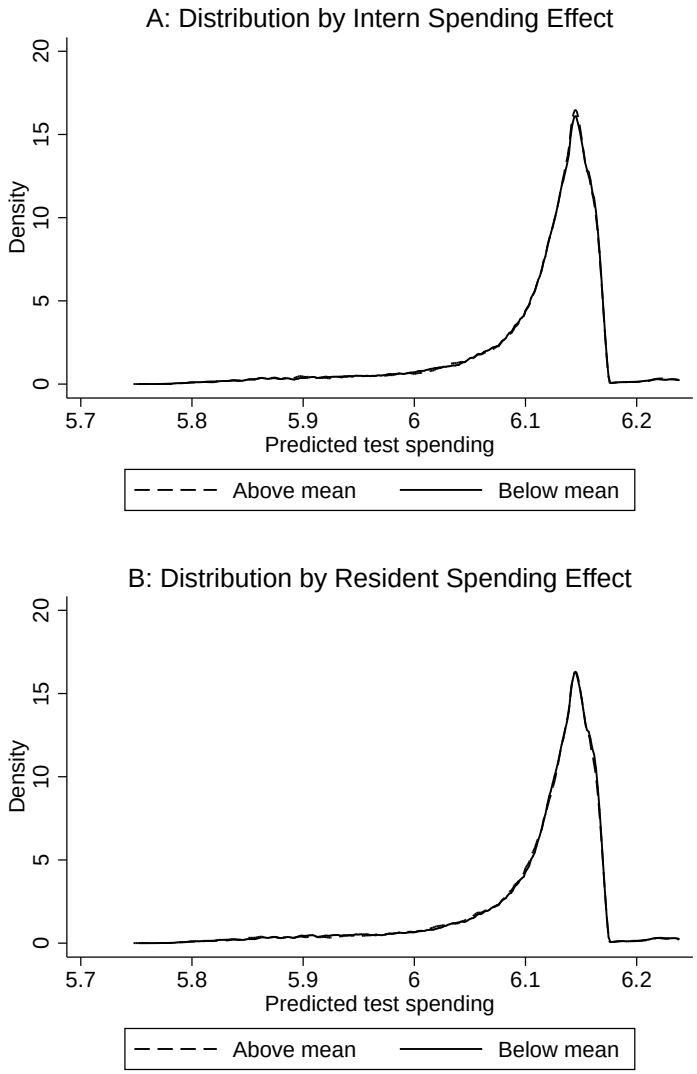
Note: This figure shows variation profiles of the expected standard deviation of housestaff effects over tenure, $\sigma(\tau)$, differing by the underlying profile of learning over tenure. Learning is parameterized as a piecewise linear function $g(\tau)$ that describes how the precision of subjective priors increases over tenure. In particular, this figure considers piecewise linear functions of the form (A-13), parameterized by k_0, k_i, and k_j. Each panel considers a different set of parameters of $g(\tau)$. Given $g(\tau)$, I calculate the expected standard deviation of housestaff effects over tenure using Equation (A-11). I assume that interns are equally likely to work with second-year residents and third-year residents. These profiles are discussed further in Appendix A-4.

Figure A-2: Patients Age by Housesetaff Spending Effect



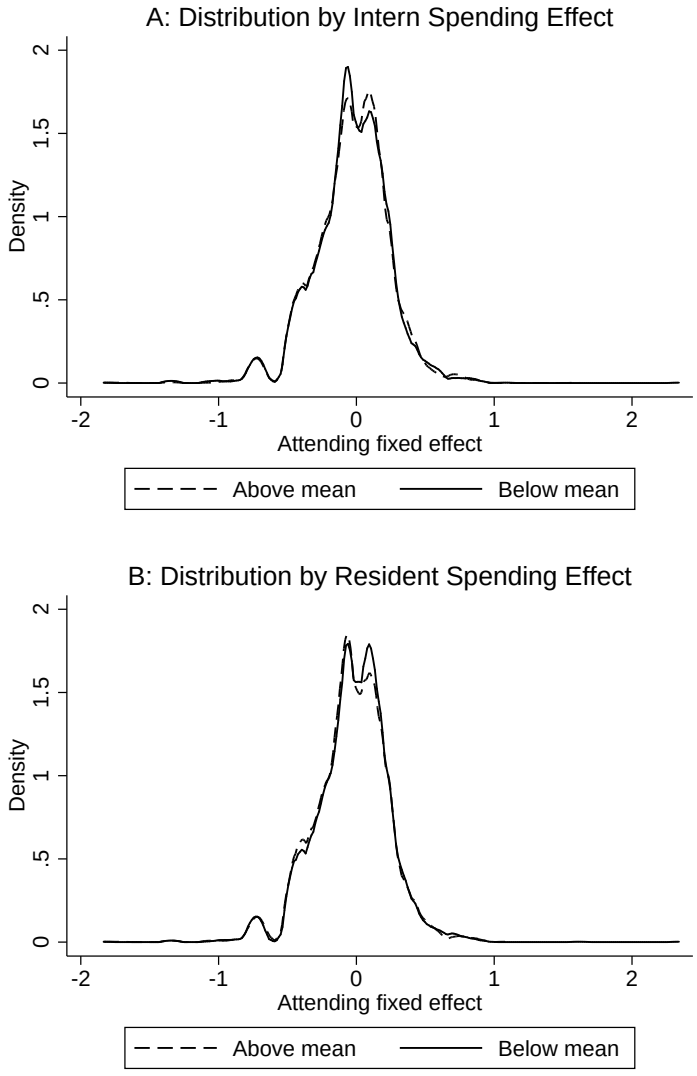
Note: This figure shows the distribution of the age of patients assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Housestaff spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log test spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.995 and 0.635 for interns (Panel A) and residents (Panel B), respectively.

Figure A-3: Demographics-predicted Spending by Housestaff Spending Effect



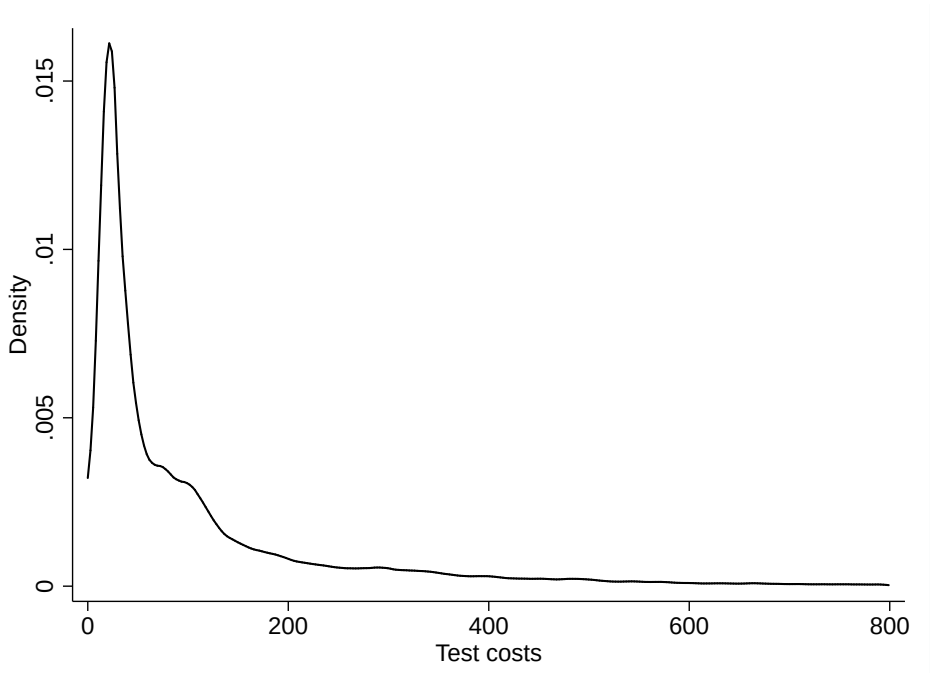
Note: This figure shows the distribution of predicted test spending (based on patient age, race, and gender) for patients assigned interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Housestaff spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log test spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.892 and 0.447 for interns (Panel A) and residents (Panel B), respectively.

Figure A-4: Attendings Spending Effects by Housestaff Spending Effect



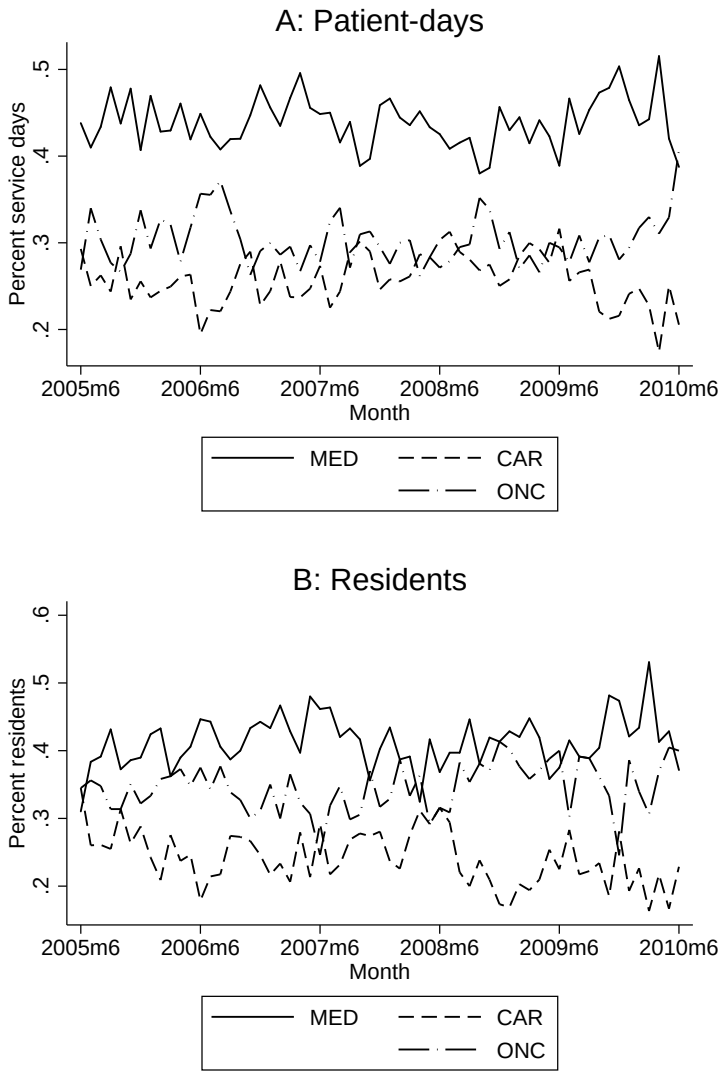
Note: This figure shows the distribution of spending fixed effects for attendings assigned to interns with above- or below-average spending effects (Panel A) and residents with above- or below-average spending effects (Panel B). Housestaff and attending spending effects, not conditioning by tenure, are estimated by Equation (A-3) as fixed effects by a regression of log test spending on patient characteristics and physician (intern, resident, and attending) identities. Kolmogorov-Smirnov statistics testing for the difference in distributions yield p-values of 0.443 and 0.069 for interns (Panel A) and residents (Panel B), respectively.

Figure A-5: Distribution of Daily Test Spending



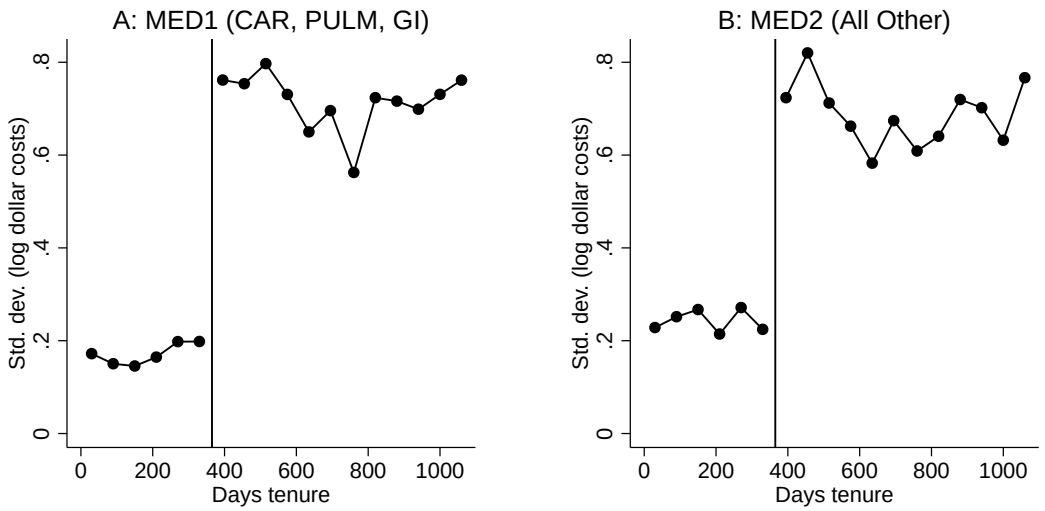
Note: This figure shows the density daily test costs. The distribution is shown up to $800 per day.

Figure A-6: Service Days and Residents on Ward Services over Time



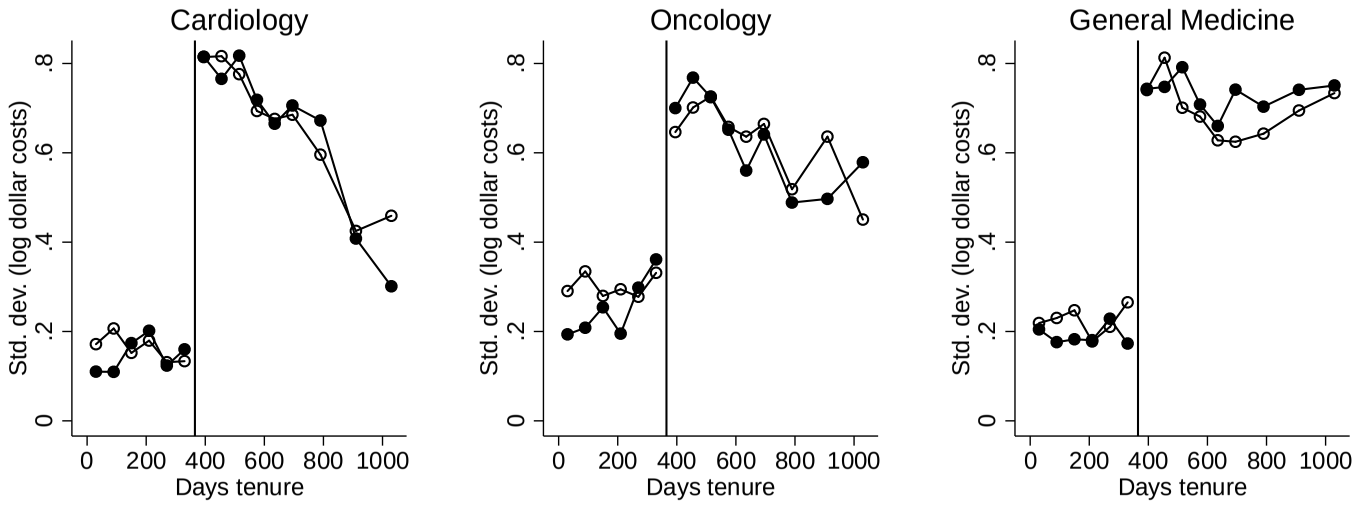
Note: This figure shows the percentage of patient-days (Panel A) and residents on service (Panel B) during each month in the data for each service of general medicine, cardiology, and oncology. Residents may be counted in more than one service if they spent time in more than one service in the same month.

Figure A-7: Housestaff-effect Variation by Tenure in Pseudo-services



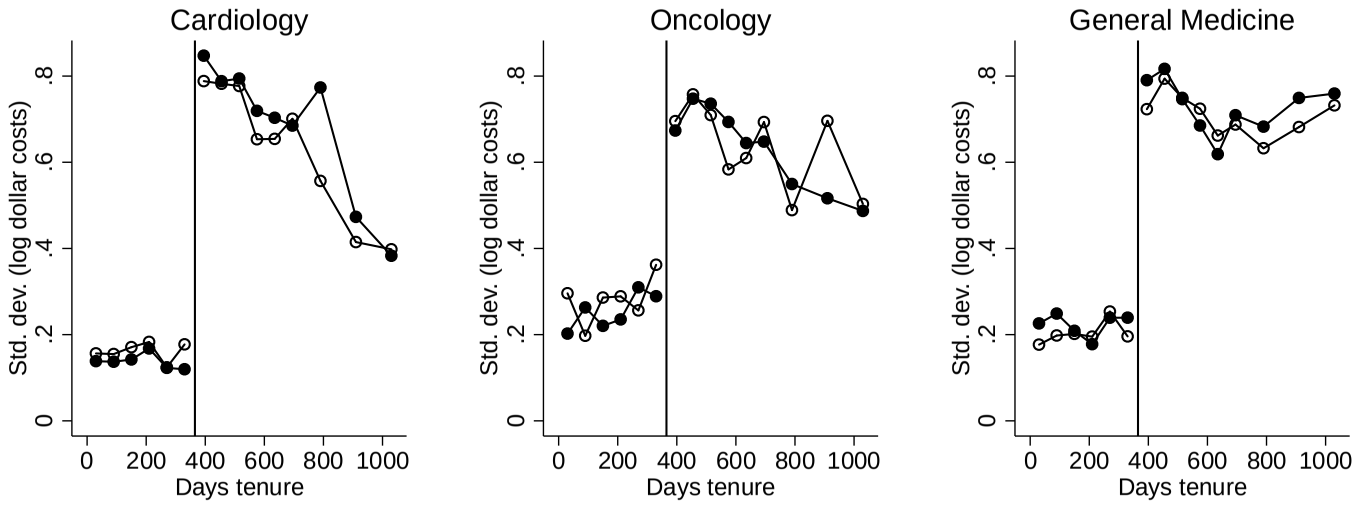
Note: This figure shows the standard deviation of test-spending effects over housestaff tenure in two pseudo-services formed from general medicine admissions. These pseudo-services are meant to create a difference in diagnostic concentration. MED1 includes the most common Major Diagnostic Categories (MDCs) of Circulatory System (MDC 5), Respiratory System (MDC 4), and Digestive System (MDC 6), roughly equivalent to cardiology, pulmonology, and gastroenterology; MED2 includes all other MDCs. Summary statistics for these two pseudo-services are given in Table A-6. The random effects model is still Equation (2), estimated at non-overlapping two-month tenure intervals. 95% confidence intervals are omitted for simplicity. Controls are the same as those listed in the caption for Figure 3. Housestaff prior to one year in tenure are interns and become residents after one year in tenure; a vertical line denotes the one-year tenure mark.

Figure A-8: Housestaff-effect Variation in Each Service by ICD-9 Code Frequency



Note: This figure shows the standard deviation in a random effects model, as in Equation (2), of log daily test costs at each non-overlapping tenure interval but for each service and for relatively common (within service) ICD-9 diagnostic codes (solid dots) and uncommon diagnoses (hollow dots). Controls are the same as those listed in the caption for Figure 3. Housestaff 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: Housestaff-effect Variation in Each Service by Guideline Existence



Note: This figure shows the standard deviation in a random effects model, as in Equation (2), of log daily test costs at each non-overlapping tenure interval but for each service and for diagnoses with (solid dots) and without (hollow dots) published guidelines. Controls are the same as those listed in the caption for Figure 3. Housestaff prior to one year in tenure are interns and become residents after one year in tenure; vertical lines denote the one-year tenure mark.

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

	Independent variables		
	Housestaff identities	Housestaff 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 housestaff characteristics: housestaff's position on the rank list; USMLE Step 1 score; sex; age at the start of training; and dummies for whether the housestaff 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 housestaff 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 test 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: Ward Service Summary Statistics

	Actual services			Pseudo-services	
	CAR	ONC	MED	MED1	MED2
Mean admission characteristics					
Patient age	63.71	59.25	62.79	64.76	60.67
DRG weight	2.44	2.24	1.69	1.64	1.75
Test costs	$613.61	$855.38	$687.18	$634.70	$743.75
All costs	$9,703.80	$7,544.00	$5,303.48	$5,071.63	$5,553.42
Length of stay (days)	3.89	4.69	3.66	3.47	3.87
30-day readmission	0.089	0.218	0.090	0.089	0.091
30-day mortality	0.031	0.175	0.034	0.032	0.036
Observations					
Admission count	12,485	22,711	12,989	11,784	10,927
MDC count	23	24	23	3	21
ICD-9 count	440	1101	623	602	897
Concentration					
MDC HHI	0.740	0.117	0.103	0.347	0.101
ICD-9 HHI	0.055	0.019	0.025	0.038	0.013

Note: This table shows summary statistics for actual services – cardiology (CAR), oncology (ONC), and general medicine (MED) – and for "pseudo-services" formed based on Major Diagnostic Categories (MDC) from the general medicine service. The pseudo-service MED1 includes Circulatory System (MDC 5), Respiratory System (MDC 4), and Digestive System (MDC 6); MED2 includes all other MDCs. Summary statistics include mean admission characteristics (patient age, DRG weight) and outcomes (costs, length of stay, readmission, and mortality), counts (Numbers of admissions, MDCs, and ICD-9 codes), and Herfindahl-Hirschman Indices (HHI).

Table A-7: 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."