Haste or Waste?

Peer Pressure and the Distribution of Marginal Returns to Health Care

David Silver*

JOB MARKET PAPER

This version: January 2016

For the most recent version, please go to http://bit.ly/1RQ5pDn

Abstract

This paper estimates the within-physician marginal returns to healthcare in a large but understudied segment of the healthcare sector – the emergency department (ED). My empirical strategy exploits quasi-random assignment of physicians to coworker teams to generate instruments for case-level inputs based on workplace peer effects. I use time-stamped case-level data on millions of ED visits across New York State from 2005-2013 to infer time-varying coworker groups. I find that a physician's peers are influential in determining her pace of work. Peer effects have a variance one quarter to one third as large as physician effects within a hospital.

I use peer-induced variation in a physician's work pace to estimate the impacts of speeding a physician up on other inputs and on patient outcomes, namely 30-day mortality. I find robust evidence that physicians in fast-paced team environments ration care on other dimensions (tests and spending), causing increases in mortality among at-risk patients and cases with particularly vague symptoms. Among fast, low-spending physicians, marginal returns to time are high, whereas among slower physicians marginal returns are 0. At first glance, this is strong evidence of diminishing returns to treatment. However, the cross-physician relationship between intensity of care and patient outcomes is flat, suggesting that physicians operate on very different production functions, even within hospitals, and even within a single department of the hospital. Reallocation of time and testing away from slow physicians to fast physicians could produce efficiency gains. I discuss implications for increasingly popular physician-targeted incentives to cut back on wasteful care.

*Department of Economics, University of California, Berkeley. Email: dsilver@econ.berkeley.edu. I am indebted to my advisors, David Card, Pat Kline, Ben Handel, and Reed Walker, for input at all stages of this project. I also thank Eva Cheng, Will Dow, Hilary Hoynes, Jon Kolstad, Kaushik Krishnan, Jen Kwok, Enrico Moretti, Carl Nadler, Jesse Rothstein, Raffaele Saggio, Kevin Todd, Fabian Waldinger, Chris Walters, Heidi Williams, and Moises Yi, as well as participants in the UC Berkeley Labor Lunch, UC Berkeley Labor Seminar, and UC Berkeley IRLE Seminar, for helpful suggestions. This project would not be possible without the help of the staff at the New York Statewide Planning and Research Cooperative System (SPARCS). I also thank the New York State Department of Health Office of Quality and Patient Safety Bureau of Vital Statistics, and the New York City Department of Health and Mental Hygiene for their generous assistance in obtaining the data. I gratefully acknowledge fellowship support through the NBER (Alfred P. Sloan Foundation Grant #2011-6-22). All errors are my own.

1 Introduction

Many have suggested that incentivizing healthcare providers to cut back on spending would have little impact on patient health outcomes and could substantially decrease wasteful expenditures.[1] These arguments typically invoke wide cross-sectional variations in treatment and testing that are uncorrelated with patient outcomes as evidence of pervasive waste. However, cross-sectional relationships between health care inputs and patient outcomes may not reflect the marginal returns to care; high-cost health care providers likely differ in other uncontrolled or endogenous dimensions.[2] Direct assessments of how patient outcomes respond to health care providers cutting back on care are rare, as they require finding a plausibly exogenous within-provider source of input variation.

In this paper, I provide some of the first direct evidence on the within-physician marginal returns to care. This paper addresses two fundamental questions. First, are physicians providing valuable care on the margin? And second, which physicians are least productive on the margin? The answer to the first question has implications for whether policies should encourage physicians to cut back on care. The answer to the second question can inform the design of these physician-targeted policies, by revealing which physicians to target.

My paper addresses these questions in the context of the emergency department (ED), a large but understudied segment of the healthcare system. I use data from 137 hospital-based EDs covering 19 million cases in New York State, linked to Vital Statistics death files. I exploit the structure of the ED as a workplace to develop a research design for estimating within-physician marginal returns to care based on workplace peer effects. Emergency physicians (EPs) work shifts with one to three coworker EPs, but independently provide care to their assigned cases. EPs have centralized information about their peers' workloads, which facilitates mutual monitoring of whether physicians are "pulling their weight". A given physician may face different levels of pressure to perform across groups of coworkers. Detailed time stamps in the ED discharge data not only allow me to measure how fast a physician is working through cases, but also to reconstruct physicians' shifts and the identities of a physician's coworkers which change across (and often within) shifts.

In the first step of my analysis, I develop a novel, generalized approach to estimating the magnitude of

[1] See, for example, Skinner and Fisher (2010); MaCurdy et al. (2011); Berwick and Hackbarth (2012). Berwick and Hackbarth (2012) estimate that overtreatment contributed $158 to $226 billion in wasteful spending in 2011.

[2] For example, hospital-level treatment intensity is partly explained by factors such as expertise and overuse (Chandra and Staiger (2011)). Relatedly, econometric studies of production functions have long been concerned with the endogeneity of input choices, dating back at least to Marschak and Andrews (1944) and Mundlak (1961).

workplace peer effects, which I describe in more detail below. The primary assumption of this model is that peer groups that exert higher levels of pressure on a physician implicitly increase the shadow price of time spent on a given case. This increased shadow price leads the physician to cut back on the time she spends per case, and also leads her to cut back on time-costly diagnostic tests and other inputs into care. I estimate that a typical physician is induced to work 12.6% faster and spend 2.9% less on other inputs (notably CT scans and X-rays) when working in a two-standard deviation faster peer group environment. In this sense, peer pressure approximates an incentive placed on physicians to work faster and to reduce inputs, providing the necessary variation for estimating within-physician marginal returns to care.

Using peer-induced changes in a physician's inputs to care, I find meaningfully large within-physician returns to time in the ED. On average, if a physician speeds up her care by 10%, her spending decreases by 2.3%, and 30-day mortality among at-risk cases rises by .17 percentage points (a 4% increase in mortality, off a base rate of 4.2%). These effects are pronounced for non-specific, difficult-to-diagnose cases, where I estimate a 20% mortality increase from the same 10% physician speed-up. This effect is not driven by physicians taking care of healthier patients in fast-paced team environments; patient risk types are unrelated to peer-induced speed.

I document important heterogeneity in the marginal returns across physicians. In particular, slower, higher-spending physicians within hospitals have zero estimated marginal returns. High marginal returns are concentrated among the faster, lower-spending physicians in a hospital, for whom a 10% slowdown reduces mortality of at-risk patients by 7%. All physicians tend to cut back in spending at similar rates when induced to speed up, suggesting that slow physicians with zero marginal returns to time are over-testing and providing "flat-of-the-curve" care.

In contrast, faster, lower-spending physicians appear to have the same *average* outcomes as their slower, higher-spending counterparts.[3] Taken together with the within-physician results, this suggests that even in the same hospital (thus with the same organizational features and available technologies), physicians vary widely in their productivity. Fast physicians are able to achieve the same outcomes as their slow peers, while also displaying higher marginal returns.

Physician heterogeneity in marginal returns to care is informative for policies aiming to reduce spending and waiting times. For example, my results suggest that a two-sided policy incentivizing slow physicians to speed up and fast physicians to slow down could yield both lower costs and better outcomes.

[3]This result is in line with much of the cross-sectional area variations literature and notably with Doyle et al. (2010).

This paper makes three contributions. First, it constitutes one of the first evaluations of within-provider returns to health care. Within-provider marginal returns are of first-order importance for policies incentivizing physicians to cut back on spending. The majority of research studying returns to care uses across-provider variation in treatment intensity. Research in this vein of literature is concerned with breaking the link between unobserved patient-level characteristics and hospital choice that may bias estimates of a given provider's effect on patient outcomes (e.g. McClellan et al. (1994); McClellan and Staiger (2000); Cutler (2007); Doyle (2011); Doyle et al. (2010); Doyle et al. (2015)). In an important exception, Almond et al. (2010) provide regression-discontinuity evidence on the returns to increased medical inputs in at-risk babies who are just above or below a clinical threshold based on birthweight. Whereas this previous work provides estimates of marginal returns to care for the average hospital, my setting uncovers important heterogeneity in the marginal returns across physicians within hospitals.

Second, I use newly available and exceptionally detailed hospital discharge data. Observing the physician in charge of a case along with the date and time stamps of arrival and discharge allow me to reconstruct physician shifts and thus infer high-frequency time-varying groups of coworkers. I am also able to construct a measure of length of stay in hours. These data elements are typically only observed when working with electronic health records from a single institution, in which cases it is often difficult to observe health outcomes of interest beyond readmissions to the same hospital. In my setting, New York State provides linkage to their Vital Statistics files, allowing me to observe patient mortality, an unambiguously relevant health outcome.

To facilitate analysis of patient mortality, a rare outcome in emergency care, I incorporate machine learning methods (Breiman's (2001) Random Forests algorithm) to generate case-level risk scores for 30-day mortality. Random Forests substantially improve out-of-sample predictions relative to the standard logistic-regression alternatives and thus provide a more robust risk-scoring method. They do so in part by allowing for difficult-to-model interactions and nonlinearities in a patient's characteristics (e.g. interactions between their admitting complaint and their age), which are important in the setting of the ED. These are useful for two purposes. First, my marginal-returns analysis focuses on the subgroup of patients identified by these scores as at risk of mortality. Second, I use these scores for rigorous tests of sorting on observables.

The detailed nature of my data allow me to test and control for various types of sorting of physicians to shifts, cases, and teams that could lead to concerns about identifying variation. Anecdotally, emergency physicians are expected to have general skills and be able to treat any patient at any time. Physicians

are typically expected to work a variety of shifts, but may have preferences over their mixture of night shifts, weekends, or holidays. Consistent with this, I find that physicians do sort across shifts. However, conditional on an expansive set of hospital-time effects, I find very little evidence of sorting of patients to physicians on a range of case-level observables, including the risk scores discussed above.

Third, I contribute methodologically to the large literature estimating workplace peer effects (see Mas and Herbst (2015) for a recent review). My estimation approach allows for a novel variance components analysis that directly quantifies the importance of peers in the workplace in determining a worker's behavior. In particular, I exploit high-frequency variation in a physician's coworker group to estimate a worker-by-coworker group match effects model for case length of stay. The within-physician variance in these match effects serves as a summary measure of the importance of workplace peer effects. Using a split-sample technique to remove variance due to noise and correlated day-level shocks, I estimate that workplace peer effects in a physician's pace of care have a variance that is roughly one-quarter to one-third the variance of the physician effects. My approach allows me to relate my estimated match effects to a number of peer group characteristics. Peer effects conform to previously established results, namely that workers work faster when surrounded by fast-paced peers (Falk and Ichino (2006); Mas and Moretti (2009)). Whereas previous literature has focused on specific dimensions of peer groups that influence individual choices or outcomes, my approach captures the net extent of workplace peer effects.

The paper proceeds as follows. Section 2 discusses trends and institutional details of emergency care. In Section 3, I present a conceptual framework linking workplace peer effects to speed and testing. Section 4 introduces my data and provides decriptive statistics. In Section 5 I discuss my empirical methods for identifying the effects of coworkers on physician behavior and provide results on workplace peer effects. Section 6 turns to the main goal of the paper: estimating the within-physician marginal returns to emergency care. Section 7 discusses my findings in light of the fact that slower and faster physicians produce similar outcomes. Section 8 concludes.

2 Background

2.1 Trends and Institutional Details

Emergency department wait times and crowding constitute one of the largest policy concerns facing emergency care and have garnered national attention in recent years (IOM (2006)). The federal government

(Centers for Medicare and Medicaid [CMS]) has recognized the need for "Timely and Effective Care" in the ED.[4] CMS has begun to publish direct measures of hospital wait times and has included length of stay as a new measure of hospital quality. Consequently, the time it takes a physician to work through her cases has become an increasingly important performance indicator for EDs and management, as it contributes directly to wait times and is thought to impact patient satisfaction (Thompson et al. (1996)). However, there is a lack of rigorous research investigating the impacts of incentivizing speedier care on patient outcomes. Unless physicians operate with considerable slack, they are likely to be forced to cut back on other potentially valuable inputs in order to speed up, possibly inducing costly errors or undertreatment.[5]

Increased focus on length of stay is motivated by a large increase in the volume of patients seeking emergency care over the past two decades. For example, in the state of New York emergency department visits rose from around 5.5 million visits in 2005 to nearly 7 million visits in 2013. In order to hold constant the number of cases per hour under a physician's responsibility, physician-hours would have needed to increase by 30 percent. In reality they only increased by 20 percent, as shown in Figure 1. As a result, the number of cases the typical emergency physician treated per hour on a shift increased about 10 percent, from 1.6 cases per hour to about 1.75 cases per hour.

2.2 The Emergency Department as a Workplace

Emergency physicians (EPs) spend the majority of their clinical hours caring for patients in the ED, where they are tasked with evaluating, diagnosing, treating, and planning follow-up care for all arriving patients, and have considerable autonomy in doing so.[6] There are typically one to four physicians on duty at a time in a medium-sized ED (20,000-50,000 annual discharges), and EPs provide full-time (24/7) coverage of the ED.

Each arriving patient is assigned to one physician, who is responsible for directing the care of that patient, from ordering tests and treatments, to making decisions about follow-up care. Additionally, the physician

[4] See https://www.medicare.gov/hospitalcompare/Data/Measures.html

[5] The prevalence of missed diagnosis of strokes and mini-strokes (transient ischemic attacks [TIA]), as well as missed diagnoses of heart attacks (acute myocardial infarctions [AMI]) in the ED has been documented in the medical literature (see, e.g., Thomas et al. (2000); Pope et al. (2000); Wilson et al. (2014); Kachalia et al. (2007)). However, the causes of these errors are not well-understood.

[6] Unlike other specialties, emergency physicians are disproportionately self-employed (32 percent), while another 19.8 percent are independent contractors, and the remaining 48.2 percent are employees (Association (2003)). Nurses and other midlevels combine efforts to administer much of the physician-ordered care. Roughly one third of employed physicians work for group practices or free-standing centers, while about two thirds work directly for hospitals. Hospitals that do not directly employ or contract with individual physicians contract their emergency department physician staffing out to independent, self-managed physician groups, contract management groups (CMGs). Multi-hospital contract management groups (CMGs) owned by non-physicians have become increasingly prevalent, especially in staffing small, rural emergency departments.

of record is solely responsible for any negative outcomes of a case, including claims of malpractice.[7]

Physicians work shifts. Shift scheduling is based loosely on physician preferences, e.g. night-shift, holiday, weekend, and vacation preferences. Shifts are typically scheduled far in advance (see, e.g., Chan (2015)). Predetermined schedules are often cited as an attractive characteristic of the specialty.[8]Shifts vary in length, typically between 6 and 12 hours, as shown in Figure 2.[9] Many shifts begin in the morning between 6am and 9am, as patient arrival rates begin to rise. Overnight shifts typically begin in the evening between 6pm and 8pm. Figure 3 provides the hours distribution of shift beginnings alongside the typical patient flows from one of the hospitals in my sample. Physicians work a mixture of morning, afternoon, and night shifts, so that staffing levels coincide roughly with average patient flows.

Figure 4 provides a simple hypothetical shift schedule. This figure illustrates the natural variation in a physician's coworker group that occurs due to scheduling. A given physician will, over the course of a week of shifts, work with a number of different peer groups. Over a longer time horizon, a physician will repeatedly work with many of those groups. As a result of these staggered shifts, a physician's peer group changes over time, even sometimes within a shift. This workplace organization allows me to track how a physician's work practices vary across different team configurations.

3 Conceptual Framework: Workplace Interactions

The nature of emergency care is ideal for studying workplace peer effects. While physicians are autonomous and individually responsible for their patients' care, the load of incoming cases is shared across on-duty physicians. In this sense, one physician's pace of care has direct implications for another physician's workload. If a physician's colleagues are working slowly, she can either pick up the slack herself or pressure her colleagues to do more.

The typical ED is small enough that the two to four physicians on duty interact in the common spaces away from patient beds. Interaction with colleagues paired with the fact that most EDs have central boards listing current cases and their physician assignments makes mutual monitoring and social pressure particularly feasible. Physicians may feel guilt or shame if their colleagues are taking a disproportionate share

[7]The annual incidence of malpractice claims against emergency physicians is around 7-8%, which corresponds to the average incidence across all specialties (Jena et al. (2011)).

[8] See http://meded.ucsf.edu/ume/career-information-emergency-medicine and http://www.medscape.com/viewarticle/750482.

[9]I infer shift length from spells of cases that a physician works, as described in Appendix A.

of the workload, as discussed in the seminal work of Kandel and Lazear (1992).

In this section, I briefly lay out a simple model of peer effects in this setting, in the spirit of Kandel and Lazear (1992) (henceforth KL). This framework demonstrates how peer pressure translates into a physician speeding up, and doing so by cutting time-costly inputs. Most importantly, I show how this setup links directly to an instrumental variables framework for estimating the marginal returns to care.

3.1 Peer Pressure on Pace of Care

The KL model presumes that individuals in the workplace are privately interested and choose an effort level to maximize their individual utilities. Coworkers can exert peer pressure on each other to conform to an effort norm. In partnerships where the social benefits of individual effort exceed the private benefits, peer pressure can reduce free-riding behavior. This framework has been used widely in the empirical literature on peer effects in the workplace (see, e.g., Falk and Ichino (2006); Mas and Moretti (2009); Bandiera et al. (2010)).

Following KL, I specify a case-level utility function for worker i of the form:

$$U_i = f_i(\mathbf{t}_i) - C_i(T_i) - P_i(T_i, a_i; \mathbf{T}_{-i}, \mathbf{a}_{-i})$$

Here, T_i represents the time a physician allocates to the case, while $\mathbf{t}_i = (t_{i,1}, \ldots, t_{i,K})$ represents the time spent on each of K time-costly inputs into patient care, so that the physician's faces the constraint $\sum_{k=1}^{K} t_{i,k} = T_i$. $f_i(\cdot)$ represents the monetary and non-monetary rewards to physician i, most notably including the quality of care that comes from providing inputs $\mathbf{t}_i$. $C_i(\cdot)$ is physician i's private cost of effort, and $P_i(\cdot)$ is the disutility of peer pressure, which is a function of the worker's effort choice (T_i) and characteristics (a_i), and the effort choices ($\mathbf{T}_{-i}$) and characteristics ($\mathbf{a}_{-i}$) of her peers.

Importantly, this setup presumes that peer pressure is placed on a physician's pace of care, and only affects input choices $\mathbf{t}_i$ by increasing or decreasing the pressure on a physician's time, or in other words the shadow price of time spent on a case.

One way to parameterize the peer pressure function is as an effort norm specific to a worker in a given group:

$$P_i(T_i; g) = p(T_i; \lambda_{ig})$$

where λ_{ig} is the effort norm for worker i working with coworkers g. This formulation allows different members of group g to compare themselves to different effort norms, as might be the case if there are hierarchical relationships between peers or if some workers only respond to a subset of a given peer group, for instance their friends (Bandiera et al. (2005) and Bandiera et al. (2010)). This says nothing about the determinants of any given λ_{ig} but makes the point that a worker feels different pressures from working in different peer groups. The literatures in economics and sociology suggest rich sets of determinants of effort norms λ_{ig}. These norms are influenced by, among other factors, the speed of a worker's peers, her relationships with her peers, and the gender and seniority of her peers (e.g. Gneezy et al. (2003); Niederle and Vesterlund (2007)).

3.2 Peer Pressure and the Shadow Price of Time

When a physician works in a group where her effort norm λ_{ig} is high, she effectively faces a higher shadow price of time spent on a case. To illustrate this point, consider the case of two time-costly inputs and the peer-pressure function $p(T_i; \lambda_{ig}) = T_i\lambda_{ig}$, so that the peer-pressure parameter scales the private cost of taking more time per case. I can rewrite the physician's utility function (suppressing i and g subscripts) as $U = f(t_1, t_2) - C(T) - T\lambda$ and the physician's constraint as $t_1 + t_2 = T$.

The physician's problem can be framed equivalently as a two-stage budgeting problem, where she first decides how much time to spend on a case and then chooses how to optimally distribute the time to maximize $f(t_1, t_2)$. The physician can solve this two-stage problem by backwards induction. Her second-stage indirect utility function is:

$$V(T) = \max_{t_1,t_2} f(t_1, t_2) \text{ s.t. } t_1 + t_2 \leq T$$

The first-order conditions of this problem are $f_1 = f_2 = \mu$, where μ is the Lagrange multiplier on the time constraint, i.e. the shadow price of time. Applying the envelope theorem:

$$V'(T) = \mu \equiv \text{shadow price of time}$$

The physician's first-stage problem boils down to:

$$\max_T V(T) - C(T) - \lambda T$$

with the resulting first-order condition $V' = C' + \lambda$, or equivalently

$$\mu = C' + \lambda$$

This formulation shows that peer pressure (λ) directly influences the shadow price of time μ. When a physician works in an environment with high peer pressure, she intuitively faces a higher shadow price of time. This decreases her time spent per case ($\frac{\partial T^*}{\partial \lambda} < 0$).

In this formulation, a physician in a high peer-pressure environment cuts back on inputs by moving along her expansion path, defined by the condition $f_1 = f_2$. Thus, for a given physician, the time she spends on a case, as well as her input choices are all implicitly functions of peer pressure. When she is induced to speed up by a given peer group, she is forced to cut back on inputs into care. To the extent that these inputs are valuable in the production of patient outcomes, patients may suffer.

Figure 5 provides an illustration of this model of peer pressure, input choices, and patient outcomes in the two-input case. Higher peer pressure has the first-order effect of increasing the shadow price of time. This has the first-order effect of shifting in a physician's time budget (drawn on the floor of this figure). When a physician's input choices move along her expansion path, the effects of these changes on patient health outcomes are found by projection of the expansion path onto the physician's health care production function. In the figure, the surface above the floor represents the physician's inverse health production function (the probability of an error). These production functions may differ if physicians vary in their productivity, in which case the marginal returns to care can vary across physicians even at the same input levels. Some physicians may be on a steeper part of the health production function, while others may be on the "flat of the curve". Physicians may also have different treatment styles, which amounts to different expansion paths in terms of their testing choices. My setting allows me to provide evidence on these important forms of heterogeneity.

If overtesting and overtreatment are rampant, then cutbacks in care should have minimal impact on patient health outcomes. In this world, policies to incentivize physicians to work faster (and thus cut back on testing) would be warranted. If instead patient health outcomes are sensitive to peer-induced within-physician variation in speed and testing, these policies are unlikely to pass cost-benefit calculations.

The logic of my instrumental variables strategy leans on this model of peer-induced speed and inputs. If quasi-randomly assigned peer groups affect physicians only by generating different levels of pressure on

a physician's pace of care, then the identity of a physician's coworker group serves as a valid instrumental variable for estimating the effects of speeding up a physician on other inputs and on patient outcomes. I provide a variety of evidence in support of the predictions of this model in my empirical work.

4 Data and Sample Selection

4.1 New York Hospital Discharge Data

My primary source of data is the New York State Department of Health's Statewide Planning and Research Cooperative System (SPARCS) hospital discharge database. Every short-term nonfederal hospital in New York State is required to submit discharge data through SPARCS, which are then reviewed for quality and completeness by the Department of Health. These data contain rich information on each discharge, including age, race, sex, a list of ICD-9 diagnoses and treatments, service and revenue codes (UB-92 Codes). UB-92 Revenue Codes can be used in tandem with ICD-9 procedure codes to identify tests such as CT scans and EKGs performed on each case. I use the Healthcare Cost and Utilization Project (HCUP) Utilization Flag software to determine whether a given case has received any of 30 types of services or accommodations, e.g. imaging and diagnostic tests (Chest X-rays, Computed Tomography (CT) scans, Ultrasounds).[10]

SPARCS has collected these data on all *inpatient* hospital stays since 1982. In 2005 SPARCS began the collection of all *emergency department* discharges. Many other states have also adopted the collection of emergency department discharges, but New York's database is unique in a few important ways:

1. **Physician state license numbers** provide consistent identifiers of physicians across hospitals in New York. These license numbers make possible the linkage to the state license register[11] and other publicly available physician profiles.[12]

2. **Patient encrypted identifiers** allow for analyses of patient readmissions to any acute care setting in New York State, and also importantly provide the key for record linkage with New York's Vital Statistics death records.

[10] See https://www.hcup-us.ahrq.gov/toolssoftware/util_flags/utilflag.jsp
[11] Available at http://www.op.nysed.gov/opsearches.htm
[12] Notably, New York Physician Profile, the National Plan & Provider Enumeration System (NPPES), and Physician Compare.

3. **New York Vital Statistics death records**:[13] SPARCS performs in-house matching of discharges to death records, which provide date and cause of death. In combination with the date of discharge, this allows me to construct 30-day mortality.

4. **Date and hour of admission and discharge**: My analysis leans heavily on the reporting of the timing of arrival and discharge for each case. I use these elements (a) to measure throughput (hours from arrival to discharge), (b) to measure the stocks and flows of cases in the hours around each visit, (c) to contruct physician shifts based on open cases, and (d) to construct teams of physicians on duty in any given hour.[14] Additionally, these variables enter my analysis directly as controls, allowing me to identify parameters net of the variation that comes from physicians and teams working in different environments, as captured by time of day, day of week, etc.

My analysis focuses primarily on emergency department discharges; I bring in the inpatient records for some robustness checks.

An important limitation is that SPARCS (as with all other state discharge databases) only requires submission of one discharge per hospital visit. For patients arriving at the ED and eventually being admitted, the discharge records will only contain that patient's inpatient record, with a flag indicating that she was admitted via the hospital's emergency department. The inpatient records list physician license numbers from the inpatient stay and not from the ED visit. I can attribute these cases to ED physicians based on their time of arrival to the inpatient setting, but only at a coarsened level, as the exact hour of admission to the inpatient setting does not reflect when the patient first arrived in the ED. My analysis is thus restricted to patients not admitted to the hospital from the ED. Admission from the ED to the hospital in my sample occurs in about 18 percent of cases. I address the potential biases generated by this feature of the data in my robustness checks. Future data collection efforts should focus on reporting both the inpatient record and the ED record for those patients who are eventually admitted to provide a more comprehensive view of hospital treatment.

4.2 Sample Selection and Descriptive Statistics

This section briefly describes how I select my main sample of discharges. I provide descriptive statistics on this sample. For more details on sample selection, refer to Appendix B.

[13] New York State and New York City have separate Departments of Vital Statistics. SPARCS provides linkages to both upon separate IRB approval by each agency.

[14] For more detail on the construction of shifts and teams, see Appendix A

My analysis sample is comprised of discharges from hospitals where typically more than one physician is on duty in a given hour, and where given team configurations are observed frequently enough that I can precisely measure the effects of each of these configurations. I restrict attention to hospital-months with minimal missing data on time stamps, as reporting compliance of time stamps varies across hospitals over time. Only physician licenses with ≥ 1000 associated cases, and only jobs (physician-hospital combinations) with ≥ 500 associated cases are maintained in the sample. I make a few other restrictions on shift lengths and team sizes that result in my final sample, which includes data from 137 hospitals. The sample includes 3,445 full-time physicians working 5,089 jobs, 1.4 million shifts, and 19.3 million cases.

4.2.1 Case-Level Descriptives

The emergency department caseload is quite varied, as demonstrated in Table 1, which provides descriptive statistics of all ED discharges as well as those I end up using in my empirical work.

Females are overrepresented at 54 percent of all cases, as are the young – patients under 5 years old represent over 10% of all cases. Low-income and minority populations are also high utilizers of the ED. Medicaid is the primary payer for over a third of all visits,[15] and less than half of all cases involve a non-Hispanic white patient, despite their constituting roughly 70% of New York's population (US Census Bureau, 2013). EDs are open 24 hours a day but see the majority of cases during waking hours, with only 11% of cases arriving between midnight and 5am.

Table 1 also provides a comparison between case characteristics, inputs, and outcomes of the full sample and those of my selected sample of discharges. Overall, my sample is similar to the full sample, with a few exceptions. Patients are slightly older in my sample, mostly due to my exclusion of pediatric emergency departments. Also, because my sample selection favors smaller hospitals with relatively small team sizes (the case-weighted average number of coworkers of a physician is 1.27 in my sample, as opposed to 2.76 in the full sample), and because minorities are concentrated in large, urban hospitals, my sample skews towards non-Hispanic white patients. Relatedly, patients in my sample are more likely to be privately insured (35% vs. 29% in the full sample), on Medicare (16% vs. 13%), and less likely to be on Medicaid (28% vs. 35%) or to lack insurance (self-pay; 13% vs. 16%).

Inputs into care are slightly higher in my sample; 12.5% of cases receive a CT scan in my sample relative

[15] SPARCS only began systematic collection of insurance variables (primary payer) in 2008. In auxiliary analysis I show that limiting the sample to 2008 and beyond and including payer indicators in my main risk-adjustment models has negligible impact on estimated physician and team effects, despite insurance having important predictive power for length of stay and charges.

to 10.7% across the board. X-rays and EKGs are also more extensively used. However, length of stay measured as a case's duration in hours is slightly lower in my sample, as there is less congestion in smaller EDs than in large, urban settings. Charges are lower as well in my sample, likely reflective of price differences across hospitals, as well as true differences in intensity of care. These input differences reflect combinations of differences in provider and patient characteristics. Likely due to the fact that my patients tend to be older, 30-day mortality rates in my sample are slightly higher (5 per 1000 compared to 4 per 1000 across all cases).

Table 2 lists the top 15 classes of complaints in the ED, as well as throughput, log charges, and mortality rates for each of these complaints in my analysis sample. The first thing to notice is how common these complaints are; these 15 classes constitute roughly 60% of the entire ED caseload. Second, among these complaints, care and outcomes vary substantially. Open wounds tend to take around 2 hours to treat, while patients complaining of abdominal pain have much longer stays, around 4 hours, and receive more intensive treatment by upwards of 80 log points. The differences in throughput reflect both differences in wait times and differences in the time it takes a physician to treat cases of varying complexity or ambiguity. Patients with open wounds have a more straightforward course of treatment than do patients presenting with abdominal pain. These differences make controlling for heterogeneity in case types important in my empirical work. I discuss risk adjustment in more detail in Section 5.

4.2.2 Hospital Descriptives

Table 3 provides further information on the hospitals in my sample. As noted above, the hospitals I make use of in my analysis are smaller than the hospitals in which the average cases are treated, simply because the emergency departments in the largest hospitals have much higher volumes than the typical hospital. In addition, small community hospitals that have volumes low enough for "single-coverage" (one physician on duty at a time) by definition have no variation in a physician's coworkers and are dropped from my sample. These selection criteria are reflected in the annual volumes and modal team sizes of hospitals reported in Table 3. The 90th percentile hospital in terms of annual volume over the full sample discharges around 60,000 cases annually, whereas the same figure in my sample is 52,000. On the lower end, the 10th percentile of hospital annual volumes is only 278 in the full sample, as there are many facilities in the SPARCS outpatient data that report only a small number of ED discharges; in my sample, this figure is 14,000 annual discharges.

4.2.3 Physician Descriptives

The emergency physicians in my sample are 70% male, while about 76% of cases are treated by males, as shown in Table 4. The average physician graduated medical school in 1994, while the typical case is treated by a physician a few years older, having graduated in 1991. I observe physicians working on 5,610 cases on average over my sample period. Mobility is relatively high; from 2005 to 2013, the typical physician works at 1.5 hospitals. There is also considerable movement in and out of my sample – of the potential 108 months of active work a physician could log in 9 years, the average physician is active for 51 months and works 407 shifts, while the typical case is seen by a physician who is active for 77 months and works around 800 shifts.

As described in Section 2, physicians work in many different team configurations – 346 for the average physician. Many of these configurations are very infrequent, but physicians work more than 50 cases in around 10% of configurations. These frequently observed teams, covering roughly 8 million cases in my sample, make it possible to contrast how a physician performs in different configurations with some precision. This feature of the data is at the heart of my empirical methods described in Section 5.

Shifts in my sample average 9.5 hours and 17 cases, as shown in Table 5. Within a shift, changes in a physician's peer group or team are common. In a third of all shifts, the physician experiences one change in her team, while only in 16 percent of shifts do I observe a stable set of coworkers throughout the span of the shift. For the most part, these shifts are worked alone at night, when some hospitals only have one physician on duty, reflecting the low overnight arrival rate of cases.[16]

4.2.4 Length of Stay Measures

In my analysis, my primary measure of pace of care is log length of stay. To mitigate the influence of outliers, I trim length of stay at a maximum of 12 hours, which affects roughly 3% of my analysis sample. In experimenting with other trimming points and with using levels (as opposed to logs) of length of stay as my primary outcome, I find that physician and team match effects are remarkably stable, suggesting that the exact choice of the LOS outcome has little influence on the ordering and magnitudes of these components.

[16] One other interesting feature to note in Table 5 is that most shifts are scheduled more than a full day after the end of a physician's previous shift, but about one in four shifts begin 13 to 18 hours after the end of the previous shift, as would happen when a physician works the same time slot in two consecutive days.

5 Methods for Identifying Peer Effects

With this background, I turn to a discussion of my methods for identifying the importance of peer effects in physician work pace, measured by log length of stay. This section first describes my statistical model of peer effects. I then address identification concerns. Finally, I present my first set of estimation results. The main contribution of this analysis is a variance decomposition that compares the relative magnitudes of worker effects and peer effects. This analysis sets the stage for my analysis of the marginal returns to care in Section 6, where I use estimated peer effects to form instruments for a physician's pace of work.

My methods for identifying contextual peer effects represent a departure from the standard ways in which they are estimated. Typically, researchers specify a characteristic or set of characteristics of a peer group in which they are interested. Given (quasi-)random assignment of individuals to peer groups, the researcher then estimates the relationship between peer group characteristics and individual outcomes.

These methods implicitly treat omitted or unobserved group characteristics as random effects to arrive at causal statements about peer effects along a given dimension.[17] My methodology relaxes the assumptions of previous work in a straightforward way, by including in my statistical model unrestricted individual-by-peer group match effects to capture the net effect of a given peer group on a physician's behavior.

I can then provide a novel variance decomposition of pace of care into a component due to case characteristics, a component due to physician heterogeneity, and finally a component due to peer effects. This estimation strategy provides a useful way for estimating contextual peer effects via a second-step regression of $\hat{\phi}_{d,g}$ on match-level observables, and is applicable in any setting where an individual is repeatedly observed in different peer group configurations. I pursue this strategy in Section 5.4.

5.1 Statistical Model of Peer Effects

I assume the data generating process for case outcomes is linear in a rich vector of case covariates, $\mathbf{X}_c$. A physician $d(c)$ is assigned to each case c.[18] While the physician is treating case c, she is subject to the influence of other physicians on duty, $g(c)$, in terms of how fast she works. My model for the log length of stay of a given case ($\ln LOS_c$) takes the following form:

[17] For example, Sacerdote (2001) is partly interested in the effect of roommate academic abilities on a student's GPA, while Mas and Moretti (2009) are interested in estimating the response of a cashier to the average speed of her coworker group $\frac{1}{N}\sum_{j\neq d}\theta_j$.

[18] In my regressions, $d(c)$ denotes a physician in a particular hospital, i.e. a job.

$$\ln LOS_c = \mathbf{X}_c'\beta + \theta_{d(c)} + \phi_{d(c),g(c)} + \epsilon_c \tag{1}$$

In my baseline specification, $\mathbf{X}_c$ includes day of the week-by-hour of arrival dummies fully interacted with hospital dummies, a dummy for the month-year (e.g. July 2007) in which the case took place also interacted with hospital dummies, patient age-bin dummies interacted with patient gender, indicators for patient race and ethnicity, and finally an exhaustive set of indicators for the 3-digit ICD-9 code of the patient's primary complaint on arrival, i.e. the reason for the patient's visit.[19] [20] Hospital time effects (week, hour, month and interactions) allow for arbitrary sorting of physicians and teams across shift types, while also capturing the hospital-specific profiles of length of stay due to organizational features, typical patient flows, and unobserved time-specific patient characteristics (e.g. intoxication on Friday nights).

Next, θ_d is a hospital-specific physician effect (i.e. a job effect) that captures the part of length of stay due to the physician's typical practice style in a given hospital. Estimated as fixed effects, these physician effects are allowed to be arbitrarily correlated with time and patient characteristics. Most importantly, this approach controls for physician sorting across environments, i.e. shift types with different typical patient flows within the hospital. If truly faster physicians sort to shift types where care is slower in general, then failing to allow for this correlation would make faster physicians appear slower than they are.

The team match effects ($\phi_{d,g}$) capture the influence on physician d's pace of work in coworker group g, and are of primary interest for this section of the paper. These effects are meant to capture the influence of a given peer group on a physician's working pace and arise if workplace spillovers, social or otherwise, are an important determinant of a physician's pace of work.

These effects are only identified relative to one another for each physician, so in my empirical work I make the natural normalization that these effects have a case-weighted average of 0 for each physician. To implement this restriction in practice, I begin by estimating a regression with the full set of physician-by-team effects:

$$\ln LOS_c = \mathbf{X}_c'\beta + \gamma_{d(c),g(c)} + \epsilon_c \tag{2}$$

I then decompose my estimates of $\gamma_{d,g}$ into the physician component θ_d, which equals the case-level average

[19] My use of 3-digit ICD-9 code fixed effects follows the analysis of Doyle et al. (2010).

[20] To speed computation, I restrict the coefficients on these risk adjusters to be common across hospitals. Running regressions at the hospital-level as opposed to the full sample, and thus allowing for the risk-adjustment coefficients to differ across hospitals has no demonstrable effect on my estimated physician or team effects. See Table A.1 for correlations of team effects between full-sample and hospital-specific regressions.

of the $\gamma_{d,g}$ parameters for physician d, and the mean-zero team effect $\phi_{d,g}$, which is simply the residual $\gamma_{d,g} - \theta_d$. A coworker group that tends to speed up a physician relative to her normal pace of work (θ_d) will have a negative $\phi_{d,g}$.

Modeling peer effects as match effects for each physician-by-peer group dyad departs from previous work where, instead of match effects, various peer group characteristics (potentially interacted with individual characteristics) are included as regressors of interest. As such, these studies are only able to capture the component of individual behavior or outcomes due to *observable* peer group characteristics. My approach has the advantage that I capture the full extent to which peer effects matter in determining individual behavior, including effects on *unobservable* dimensions. An additional advantage of my estimation strategy is that I rely on weaker assumptions about sorting of cases to workers and teams, by explicitly including team match fixed effects.

My match effects model contrasts with two other potential parameterizations of how a peer group affects the care of a case handled by an individual physician. First is a model of additive physician and coworker group effects. This alternative would lend itself to settings in which the external group of peers is thought to impact each individual with whom they interact equally. A second alternative model includes the physician herself in the coworker group, which would be appropriate if peer groups affect all members of the team in the same way.[21] Modeling the team effect as a match effect between the set of coworkers and the physician, as I do, nests both of these models and allows for unrestricted forms of coworker interaction.

Findings in previous research on workplace peer effects suggest that match effects between worker and coworkers are important. For instance, Mas and Moretti (2009) document substantial heterogeneity in responsiveness to coworker speed by whether the worker himself is fast or slow. These types of stories impose testable (given the right data elements) structure on the parameters $\phi_{d,g}$. For example, if physicians respond to the average speed of their coworkers, as in Mas and Moretti (2009), then $\phi_{d(c),g(c)} = \beta \frac{1}{N} \sum_{j \neq d} \theta_j$, in which case a regression of $\hat{\phi}_{d,g}$ on $\frac{1}{N} \sum_{j \neq d} \hat{\theta}_j$ will produce an estimate of β.[22] I explore these explanations for the team match effects in Section 5.4.

The final piece of Equation 1 is the term ϵ_c, which may contain both an unobserved contemporaneous

[21] Finkelstein et al. (2014) provide a recent example of a two-way fixed effects model to study the relative contributions of patients and places in Medicare spending. These models also have analogs in wage models with person and firm components, e.g. Abowd et al. (1999) and Card et al. (2013).

[22] In general, this regression produces an attenuated coefficient estimate due to measurement error in the explanatory variable. Attenuation is unlikely to be a major concern in this setting, as the estimated physician effects, $\hat{\theta}_j$, are measured typically with over 1000 observations each and are thus quite precise.

shock common to cases treated by physician $d(c)$ in a given shift (call this $\nu_{s(c)}$) and an idiosyncratic case-level error (η_c): $\epsilon_c = \nu_{s(c)} + \eta_c$. $\nu_{s(c)}$ is similar to the types of classroom-level shocks discussed in the teacher value-added literature, and can complicate the estimation of the team-match effects, if certain teams work on only a handful of shifts, in which case a single shift-level shock can be influential in the estimation of a given match effect. Demand shocks are one type of shift-level shock that I can deal with in my empirical work.[23] If sorting of teams across high- and low-demand periods were driving the estimated team match effects, then controlling for the observed hospital-level demand shock (the deviation from the predicted number of arrivals in the most recent period) would alter the estimated team match effects. That my team match effects are stable regardless of whether I control for this shock, as I show in my empirical analysis below, lends more credibility to the identification of the team effects. Nevertheless, I approach this problem cautiously. It motivates my split-sample construction of team-based instruments for physician speed discussed in Sections 5.3 and 6.

5.2 Threats to Identification of Team Match Effects

Identification of the team match effects in this model requires that the error term ϵ_c be uncorrelated with the identity of a physician's peer group $g(c)$. The primary threat to this identification assumption is that a physician sees different types of cases when she works with different peer groups. This section presents a variety of quantitative tests of this assumption.

5.2.1 Sorting

As discussed earlier, emergency physicians are considered to be generalists, especially in smaller hospitals where only a few physicians are on duty at any given time, so that scope for sorting of patient types to a particular physician within a team is limited. In this section, I provide a series of tests of this claim. I also provide evidence on whether systematic sorting of physicians to teams could bias my results. In line with anecdotal evidence and other recent research on emergency physicians (Chan (2015); Van Parys (2013); Gowrisankaran et al. (2014)), I find limited evidence of sorting.

Balance First, Table 6 presents an assessment of balance on covariates. I ask whether a given physician sees different case types when she works with observably different coworker groups. For simplicity, I show

[23] Other types of shift-level shocks are unobservable, e.g. whether a physician is sick or particularly tired on a given shift.

the difference in (conditional) means of case observables between when a physician works in a younger versus an older coworker group, a male versus a female coworker group, and a faster versus a slower coworker group. The comparisons I make in this table are all *within-physician*; that is, estimates are adjusted for physician fixed effects.[24] I simultaneously adjust my estimates for a set of hospital-specific time effects, which I use throughout my analysis, to capture the fact that physicians may have preferences over shift types (weekend, night, etc). Thus the contrasts in this table represent observable compositional differences in a physician's cases when she works with observably different coworkers, but in the same narrowly defined environment. My tests here attempt to uncover any additional sorting patterns that may cause concern about my identification strategy.

The main covariate I consider is a risk score for a patient's probability of 30-day mortality.[25] This risk score is generated using predictions from a Random Forests machine learning algorithm (detailed in Appendix C) that incorporates information from all of the case-level observables used in my baseline risk adjustment models. Importantly, because it is a tree-based method, Random Forests allow for important interactions and non-linearities in the covariate space that would be difficult for a researcher to model. My risk scores from these models are out-of-sample predictions and so do not suffer from overfitting problems common to situations in which high-dimensional interactions lead to small cell sizes.

The first row of this table presents the difference in risk scores of patients assigned to a physician when that physician is working with a largely male coworker group. The difference (.0006) is very small and statistically insignificant; the p-value of the test that difference is non-zero is .9051. In the second row, I again find no difference in risk scores when a physicians works in a younger coworker group (point estimate = -.0078; p-value = .0977). Finally, I find no evidence that physicians are sorted sicker cases when working with coworker groups who are slower, as measured by the average log length of stay of a physician's coworkers. The difference is very small (-.0011) and insignificant (p-value = .7925). These tests all mimic the tests one would run in standard linear-in-means peer effects models, where the righthand-side variable of interest is the average characteristic of one's peers.

In short, I find considerable degrees of balance in physicians' case types across observably different coworker groups. I now turn to another set of evidence on the potential for sorting biases in my estimated team effects.

[24] Because some physicians move across hospitals, in practice my physician effects are actually physician-by-hospital (job) effects. In the remainder of the paper, I call them physician effects for clarity.

[25] Results for other covariates yield similar results in terms of balance.

Sensitivity In Tables 7 and 8, I perform an exercise to assess the sensitivity of estimates to inclusion or exclusion of important risk-adjusters and time effects. This exercise is similar to tests of student-teacher sorting in the education literature, e.g. Table 6 of Chetty et al. (2014). I ask how sensitive my estimates are to particular choices of the risk-adjustment model. If there were truly random assignment of patients to physicians or of physicians to teams on observables, my estimates of physician effects and team match effects would be insensitive to including or excluding particular elements of the risk adjustment model.

In line with this reasoning, I show that my estimates are highly stable across a series of risk-adjustment models. The correlation of team match effects in my baseline model with models where I *add* important additional covariates is no less than 0.97. Remarkably, when I *eliminate* all the patient-level risk adjusters (age, race, sex, ICD-9 complaint, and interactions) and simply control for detailed hospital-time effects, team match effects have a correlation of .958 with the baseline model with patient risk adjusters. Exclusion of hospital-time effects does lead to considerable changes in my estimates. These result strongly support the anecdote that the primary sorting mechanism of physicians to teams and to patients is through physicians working different shifts. Conditional on my detailed set of hospital-time effects, there is no evidence of sorting patterns on case-level observables that would indicate potential biases.

Throughout all of these specification tests, physician effects are highly stable, with a minimum correlation with the baseline model of .931, as presented in Table 8. I now describe in more detail the particular variables I add and subtract from the risk-adjustment model and their impacts on my estimated team match effects.

I first test the sensitivity of team match effects to *inclusion of omitted observables*. When I include dummies for the number of ED visits a patient has had in the previous 30 days, team match effects have a correlation of .999 with the baseline model estimates. Further including measures of contemporaneous hospital congestion,[26] produces a correlation with baseline match effects estimates of .996 with the baseline estimates. This rules out bias in the match effects stemming from physicians begin paired with specific coworker groups during busy periods.

I next add an exhaustive set of indicators for the vingtile of a patient's risk score, as defined by Random Forest predictions of 30-day mortality risk based on the patient's characteristics (age, race, sex, 3-digit ICD-9 of complaint on arrival, and hospital). These predictions allow for difficult-to-model interactions

[26] In particular, a set of indictors for the size of arrival shocks to the emergency department in the two hours preceding the index case's arrival, defined as deviations from hospital-year-day of week-hour of day average arrivals.

between patient covariates (e.g. age and primary complaint) while limiting overfitting that is common to logistic regression methods for propensity score estimation (Breiman (2001)). Differential sorting of sicker cases to physicians across different teams does not bias my estimates of team match effects: the correlation with the baseline model is once again .996.

Including instead a flexible function of propensity scores from a fully-interacted logistic regression produces similar results. I opt for the Random Forest risk scores in the body of the paper because they produce much better out-of-sample prediction and capture more of the (baseline) unmodeled information in the patient's characteristics. Thus they provide stronger tests for patient sorting than logistic predictions. For more details on the construction of these risk scores, see Appendix C.

Finally, I add to the model a flexible function for how far a physician is through her shift at the time she sees the index case. This addresses the concern that physicians are paired with particular teams only towards the beginning or end of their shifts, when the physician is fresh or tired. My estimates of team match effects are once again virtually unaltered (correlation = .971) by inclusion of this measure.

Next, I assess the degree of bias that would occur if some of my baseline covariates were not observed. Removing *all* risk-adjusters and retaining only hospital-time effects produces correlations in the team effects upwards of 0.95, as mentioned earlier. However, removing some or all of the hospital-time effects leads to much lower correlations with the baseline model. In the extreme, a pure risk-adjustment model with no hospital-time effects produces team match effects that are correlated with the baseline model at around 0.60.

The results in this section bolster the notion that the main lever for sorting is through physicians working in different teams in different times. My findings here are in line with previous work documenting minimal degrees of sorting in ED care.

5.2.2 Selection: Is There Evidence of Differential Admitting Behavior?

An additional concern about sample selection arises in this setting. State-level discharge databases do not report the ED record of patients eventually admitted to the hospital, and instead only report the inpatient records for these patients, with a flag indicating that the patient originated in the ED.

This limitation makes it difficult to study physician admitting behavior, and how a physician's decision to admit a particular case depends on her team environment, as I do not observe the emergency physician in

charge of admitted cases.

To partially address this problem, I consider a hospital-day level analysis that asks whether exposure to different day-level coworker-group environments is associated with admission rates from the ED or inpatient mortality for those admitted from the ED.

To construct this hospital-day level dataset, I extract the set of inpatient records from cases admitted via the emergency department. I then calculate the number of patients admitted from the ED each day, as well as average inpatient inputs (log charges) and outcomes (30-day mortality) for those cases. I perform similar calculations on the emergency department records. I am then able to calculate the fraction of patients arriving to a particular ED on a date who were admitted to the hospital. Finally, I match on to this dataset the estimated case-level average team match effect at the hospital-day level. This setup allows me to use variation across days within a hospital in average team-induced physician slow-downs to test for differential admitting behavior. I also calculate the average physician effect at the hospital-day level to control for any potential correlation between the types of physicians on duty and the types of coworker-induced slow-downs on that day.[27]

I then regress different hospital-day level measures of outpatient and inpatient care and outcomes on these average team and physician exposure variables, controlling for hospital-month and hospital-day of week effects. I cluster standard errors at the hospital level.

The results of this analysis are in Table 9. The top panel reports coefficients on the day-level average team match effects. Column 1 reports a "first-stage" regression, with hospital-day average log LOS of those discharged from the ED (not admitted to the inpatient setting) as the outcome. The point estimate of .604 ($s.e. = .0598$) reassuringly indicates that on days when on-duty physicians are largely paired with teams that are estimated to slow them down, the average pace of care in the emergency department is substantially slower. Column 2 provides the effects on log charges for patients discharged from the ED. The point estimate of .103, divided by the first-stage estimate, suggests that on days when on-duty physicians are working at paces 10% slower than their usual paces, they spend 1.7% more per case. This result is very similar to micro-level estimates later in the paper.

Importantly, I find no evidence that team-induced speed leads to changes in the admitting rate of physicians at the day-level. Column 3 reports the day-level effect of team environment on the fraction of ED

[27] To avoid problems of correlated day-level shocks, I use split-sample versions of the team and physician match effects, as documented in Section 5.3.

cases admitted to the hospital. This coefficient is small (.00308) and statistically insignificant, suggesting that physician slow-downs are not associated with differential admitting probabilities. The remaining columns report reduced-form effects on log charges, length of stay (in days) and 30-day mortality for ED cases admitted to the hospital. None of these tests find any effects of peer-induced speed of ED physicians on duty on the day of admission.

The bottom panel of Table 9 further controls for the average physician type, as measured by case-weighted average physician effects in log LOS, again using split-sample estimates. Coefficients on the day-level average physician effects reflect the effects of being treated in the ED on a day when many slow or many fast physicians are on duty. Coefficients on the main team-based instrument are very similar when this control is included.

Interestingly, coefficients on the physician effects suggest that when the average physician on duty is 2 standard deviations (22.8%) slower in the physician log LOS distribution, admission rates significantly increase by 0.388 percentage points on a base of 19.27%. This indirectly suggests that slower physicians admit more patients.[28] Consistent with the notion that these marginally admitted cases are healthier than inframarginal admissions, log charges, length of stay, and 30-day mortality of admitted patients are all modestly lower on these days.

To summarize, my main results are unlikely driven by differential admitting behavior. The hospital-day level analysis in this section finds no evidence for differential admitting behavior, suggesting that this channel does not lead to important sample selection issues for the analysis of physician behavior across teams.

5.2.3 Dynamic Response to Peers: Within-Shift Event-Study Evidence

Next I present evidence on the timing of the response to the team environment *within* a physician's shift. The evidence here lessens concerns that estimated team effects are driven by physician-specific trends within shifts, which may be the case if physicians only work in specific team configurations during certain unobservably fast or slow periods.

I use the fact that physicians sometimes experience changes in who they are working with over the course

[28] Alternatively, when a group of slow physicians staffs the ED all day, the resulting congestion could lead to admission of marginal patients if, for example, physicians need to free up beds to alleviate waiting times.

of a shift as, for example, the morning shift ends and the afternoon shift begins.[29] I make use of changes in the estimated team effects induced by shift changes to trace out the dynamic adjustments physicians make to working in faster-paced team environments.

I map actual length of stay on a case onto leads and lags of changes in $\hat{\phi}_{d,g}$ within a physician's shift:

$$\ln LOS_c = \sum_{\tau=-4}^{4} \beta_\tau \times \Delta\hat{\phi}_{d(c),g(c+\tau)} + \mathbf{X}_c\delta + \nu_{s(d(c),c)} + \epsilon_c \tag{3}$$

where τ indexes the order of cases seen by physician d around a team change. This specification importantly includes physician-shift effects $\nu_{s(d(c))}$ to absorb any variation within physicians and across teams driven by week, month, or year effects. I use split-sample estimates of $\hat{\phi}_{d,g}$ from OLS estimation of Equation 1 to avoid problems with correlated temporal shocks at the physician or hospital level; see Section 5.3 for details on these split-sample estimates. I also control for the usual case-level observables $\mathbf{X}_c$.

Figure 6 displays the results of this analysis. There is an immediate adjustment of physician length of stay on cases she begins working with the new team, preceded and followed by stable pace of care. The estimated adjustment in this setting is .05 of the difference in the estimated team effects. This estimate differs substantially from a coefficient of 1 (full adjustment) because of the high degree of measurement error in $\hat{\phi}_{d,g}$. The event-study representation also requires substantial sample restrictions that limit the precision of the estimated jump: I throw out any team changes where another team change occurs in the estimation window, I throw out any team changes occurring in the first or last 4 cases of a physician's shift, and I only consider team changes between frequently occurring teams where each team has at least 50 underlying observations. Due to these limitations, I only present the speed effects of these team changes, and to increase power I rely on the full variation in the data when I estimate the effects of speed-ups on other inputs and on patient outcomes.

Despite these limitations, the event study provides simple evidence that dynamic adjustments to team changes are minimal, and that when a physician begins working in a new team configuration, she immediately adjusts her speed of care. This quells fears that team effects are driven by congestion effects that build up over time, or by some unmodeled physician-specific trend.

[29] This analysis is similar to the event studies presented in Mas and Moretti (2009).

5.3 Quantifying the Importance of Teams

How much do teams matter for physician performance in the emergency department? Here I document the importance of team match effects in my data. I estimate various forms of Equation 1, and decompose the variance in log length of stay into parts attributable to the physician and to the physician-team matches. The comparison between the variation in speed across physicians and the variation in speed across the teams with which a physician works provides an intuitive measure of the importance of teams.

In Table 10, I present estimates of the variance of the physician and team-match components of length of stay, as well as variance of other components. For my basic estimates, I report the sample variances of the estimated components. The physician effects θ are normalized to be mean zero in each hospital, so the variance component I estimate ($Var(\theta)$) is the within-hospital variance of physician effects. Similarly, the team match effects ϕ are normalized to be mean zero for each physician in a given hospital (i.e. for each job), so that $Var(\phi)$ is the within-physician variance in peer effects.

My basic estimates in Column 4 show that physician effects have a case-weighted variance of .013 for log length of stay. This implies that on average, a one-standard deviation faster physician in a hospital works through each case in 11.4% less time.[30]

Team match effects have a case-weighted variance of .008. Taken at face value, these estimates would suggest large peer effects in pace of care; moving a physician into a one-standard deviation faster team environment would result in the physician working 8.9% faster. However, this variance estimate is upwardly biased by two sources of additional variation: (a) measurement error resulting from the fact that these team match effects are estimated off of relatively small samples,[31] and (b) correlated shocks at the physician-day (i.e. physician-shift) level. Correlated shocks exacerbate the small-sample problem, in that the nominal number of cases associated with a physician-team match overcounts the number of independent observations for that match.

As a way to address both of these concerns directly, I employ a split-sample technique and estimate Equation 1 separately on two partitions of the data. I randomly sample the physician shifts associated with each physician-team pair, so that for each physician-team pair, I end up with roughly half of the cases for that

[30] The estimated variance of worker effects is on the same order as found in other studies of worker speed, notably Mas and Moretti (2009), who find that the variance of worker effects among supermarket cashiers is .0081 ($s.d. = .0901$), and Chan (2015), who documents very similar results among emergency physicians in a single hospital.

[31] To partially address this, I limit attention to physician-team pairs that I observe on 50 or more cases.

pair in the first partition.[32] I estimate Equation 1 separately on each partition to yield two noisy estimates of each physician effect ($\hat{\theta}_{d,1}$ and $\hat{\theta}_{d,2}$) and each team-match effect ($\hat{\phi}_{d,g,1}$ and $\hat{\phi}_{d,g,2}$).

Let us focus on the vector of estimated team-match effects:

$$\hat{\phi}_i = \phi + e_i, i \in 1, 2$$

where e_i represents estimation error. The stratified sampling of shifts for each physician-team match makes plausible the assumption that the samples and estimation errors are independent: $Cov(e_1, e_2) = 0$. This allows me to estimate the variance of the physician-team match component as the covariance of the two estimates, since

$$\begin{aligned} Cov(\hat{\phi}_1, \hat{\phi}_2) &= Cov(\phi + e_1, \phi + e_2) \\ &= Cov(\phi, \phi) + \underbrace{Cov(e_1, \phi)}_{=0} + \underbrace{Cov(\phi, e_2)}_{=0} + \underbrace{Cov(e_1, e_2)}_{=0} \\ &= Var(\phi) \end{aligned}$$

Similar approaches to estimating variance components are found in the value-added literature in education (e.g. Kane et al. (2002), Kane et al. (2013), and Chetty et al. (2014)), where either test-score residuals from consecutive years or pairs of classrooms within a year are used to construct estimates of the variance of teacher effects.

Split-sample variance component estimates for job effects and team match effects are found in the bottom rows of Table 10. The split-sample method yields an estimate of $Var(\theta)$ that is nearly identical to the basic estimates. For log length of stay trimmed at 12 hours (Column (4)), I again estimate a variance of job effects equal to .013. This is reassuring, as I observe the physicians in my sample working thousands of cases and hundreds of shifts, so that neither small samples nor correlated shocks should be expected to have biased their estimated variance.[33]

Turning to the variance of team match effects, the split-sample estimator yields variance estimates that are

[32] I lump together teams with fewer than 50 associated discharges and cases in which the physician works alone into a single team category for each physician, and I randomly sample shifts for this composite category in the same way as for all the other physician-team cells. This strategy ensures that roughly half of the data for each physician is in each partition.

[33] I have also calculated observable patient and hospital-time variance components using split-sample estimates of β. Similar to the job effects variance estimates, these split-sample variance component estimates are virtually identical to the basic estimates.

half the size. The variance of team match effects is estimated to be .004, which is slightly less than one third of the estimated variance of physician effects. This variance estimate is still quite large. Putting a physician in a team in which she is estimated to be one standard deviation faster speeds her up by 6.3% on each case. One interesting interpretation of this result is that a given physician can be induced to work as fast as her one-standard deviation faster colleagues if she is put in a team environment that is approximately 2 standard deviations faster for her. This exercise makes it clear that an individual physician's speed varies substantially across coworker environments. To the best of my knowledge, mine are the first estimates in the literature of how important coworkers are in determining individual work behavior.

5.4 Determinants of Peer Effects

How do peer effects arise in the workplace? This has been the focus of most of the prior research in the personnel economics literature on peer effects (e.g., Falk and Ichino (2006); Mas and Moretti (2009); Bandiera et al. (2005); Chan (2015); Cornelissen et al. (2013); Mas and Herbst (2015)). My estimation strategy in the previous section lends itself to a simple second-step analysis that I pursue briefly here.

My match effects estimates $\phi_{d,g}$ represent a convenient reduction of the data to the level of the physician-team pair. With these moments in hand, one can estimate extremely general models of peer interactions, including non-linear models and models that impose different assumptions on the cross-physician within-team differences, such as those discussed in Section 5.1. A full exploration of these mechanisms and models is beyond the scope of this paper but represents an interesting path for future research.

As a first pass, I provide the results of a simple least-squares regression of $\hat{\phi}_{d,g}$ (measuring log LOS of a physician-team match) on a vector $\mathbf{Q}_{d,g}$ of physician-team characteristics. The regression takes the following form:

$$\hat{\phi}_{d,g} = a_d + \mathbf{Q}'_{d,g} b + e_{d,g} \tag{4}$$

For comparison to the literature on productivity spillovers (Falk and Ichino (2006); Mas and Moretti (2009)), I include in $\mathbf{Q}_{d,g}$ peers' average pace of care, constructed as the leave-out mean:

$$\bar{\hat{\theta}}_{d,g} = \frac{1}{N_g - 1} \sum_{j \in g; j \neq d} \hat{\theta}_j$$

I also include the average spending proclivities of peers and the gender and age composition (measured by average medical school graduation year) of the peer group, all similarly calculated as leave-out means.

5.4.1 Results

The results of this exercise are in Table 11. I include physician-by-hospital (job) fixed effects in each of these regressions, and I cluster standard errors at the hospital level. I keep only physician-team cells with greater than or equal to 50 underlying cases, and I weight by the number of underlying cases.

Spillovers of pace of care in my setting are similar in magnitude to those documented in the lower-skilled or less complex settings in Mas and Moretti (2009) and Falk and Ichino (2006).[34] In Column 1, I find that a 10% increase in peer average speed is associated with a .871% increase in a physician's pace of work on a given case. Further controlling for peer spending tendencies, peer gender, and peer experience in Column (5) boosts the spillover point estimate to 0.156 ($s.e. = .0221$), indicating spillovers of 1.56% for a 10% increase in peer average speed. Peer spending tendencies, while unconditionally uncorrelated with a physician's speed, have large estimated effects on speed of care conditional on the peer group's typical individual speed. A team whose average member performs more tests and racks up more charges in a given amount of time induces a physician to speed up (i.e. decrease LOS). One interpretation of this result is that teams that appear busier create an atmosphere in which physicians are more time-pressured and respond by speeding up care.

These types of time pressures could come from other workplace dynamics, as well. I find that working with more male colleagues induces physicians to work faster, both conditional on other peer group characteristics and unconditionally. Working with an all-male group of physicians as opposed to an all-female group, the typical physician would work at a pace that is 1% faster. Interestingly, this responsiveness does not seem to vary with the reference physician's gender, as shown in Table A.6.

In a set of Appendix Tables, I provide a series of heterogeneity tests to assess whether observably different types of physicians (based on speed, spending, gender, and cohort) respond differentially to their peers' characteristics. In short, I find slighlty larger spillovers for women and for more experienced physicians. There is also suggestive evidence that slower physicians are more responsive than faster physicians to working with faster peers, similar to the findings of Mas and Moretti (2009). These interaction effects, which are common in the literature on peer effects, are evidence of important non-separabilities between

[34] There is some evidence of workplace spillovers in the setting of the emergency department. In related work, Atal & Silver (2015) estimate a 1 to 2% increase in own speed when working with colleagues who are 10% faster using detailed electronic health records from 2 hospital-based emergency departments. In a related study of a single hospital's electronic medical records, Chan (2015) finds evidence that physicians engage in "foot-dragging" when other peers are present. The author also estimates a spillover parameter of 0.1 of working with a faster coworker. This is remarkably similar to my estimates across the 137 hospitals in my study.

physician and team that necessitate a match effects model as opposed to a model with additive physician and team effects.

Understanding workplace spillovers and complementarities could help managers in allocating shifts to improve patient flows. Given the size of workplace peer effects as estimated in my variance decompositions, there are likely substantial gains in patient flows from optimal matching of workers to peer groups. Digging deeper into the determinants of team effects is a goal of future research.

I now turn to a description of my instrumental variables strategy using peer-induced speed-ups to estimate the within-physician marginal returns to care.

6 Marginal Returns to Emergency Care

The results in the previous section show that the group of coworkers with which a physician is working has a substantial influence on her speed of care for a case. The remainder of the paper is concerned with tracing out how physicians speed up, and how physician speed-ups affect patient health. To preview my findings, coworker groups that induce physicians to speed up also induce physicians to cut back substantially on other dimensions of care, namely time-consuming diagnostic tests. This finding is consistent with the model laid out in Section 3 and contrasts with other models of physician speed-ups, which I discuss shortly. Finally, patient outcomes suffer when physicians work in higher-speed environments. In particular, physicians who are already fast (and tend to be low-spending) at baseline have the largest estimated marginal returns to care. Patient outcomes do not vary with physician speed for physicians who are typically slow. This heterogeneity is suggestive of diminishing marginal returns to care. I further explore the physician-level heterogeneity in the subsequent section.

This section proceeds in the following steps. I first present an analysis of how physicians respond along other dimensions of care when they work with coworker groups who induce them to work faster. I then discuss my instrumental variables (IV) strategy for estimating within-physician marginal returns to care using peer-induced speed-ups. Finally I provide results of this IV analysis and discuss robustness.

6.1 How Do Physicians Speed Up?

My model in Section 3 predicts that increased peer pressure (λ) on a physician's pace of care translates into speed-ups and cutbacks on other time-consuming inputs into care. In practice, this implies that when a physician works in a peer group that induces her to speed up, her measured inputs (e.g. total spending and particular diagnostic tests) into patient care should decrease. This mechanism implies a *positive* correlation between peer-induced changes in length of stay and changes in spending. In this subsection, I descriptively establish that coworker groups that speed up a physician also induce her to order fewer time-costly tests and thus rack up fewer charges. This finding is very robust. My findings in this section support the primary implications of my model, which I build on to develop my instrumental variables strategy.[35]

To shed light on how physicians hasten care, I estimate another version of the match effects model in Equation 1 with log charges as the dependent variable rather than log LOS. I then examine the bivariate relationship between the original estimated match effects for log LOS (ϕ_{LOS}) and the estimated match effects for log charges (ϕ_{charge}).

The results of this analysis, presented in Figure 7, are in strong support of my model. Teams that induce physicians to speed up also induce physicians to cut back on care (thus a positive relationship between log LOS and log charges). These two estimated match effects are highly correlated ($Corr = .333$), despite begin estimated with substantial noise. Similar results hold when looking at the relationship between match effects in CT scan or X-ray utilization and log charges or log LOS, as shown in the top panels of Figure 8. Finally there is no evidence that CT scans and X-rays are substituted for one another in different team configurations, as would be the case if speed-ups led physicians to cut out X-rays in favor CT scans or vise versa. This is evidenced in the bottom panel of Figure 8, which plots estimated team match effects for CT scan utilization against those for X-ray utilization, again finding a strong positive relationship. This set of results suggests that teams are not inducing physicians to reorganize their testing behavior, as might be the case if physicians consulted with coworkers and tended to change the *mixture* of tests they perform in some team configurations. This evidence is largely consistent with my Conceptual Framework and suggests that coworker groups affect physicians in ways that are largely captured by their effects on physician pace of care.

[35] Alternatively, physicians may cut out slack time during which the patient is idly waiting in bed or in the waiting room while no care is being delivered.In this case, physicians should be able to provide faster care without cutting back on any other inputs. Additionally, physicians could speed up care by immediately ordering any test that might prove useful without first spending any mental effort to assess the value of each test. If this behavior is pervasive, peer-induced changes in length of stay should be *negatively* correlated with changes in spending.

The fact that physicians cut back on testing as they speed up, in line with the peer-pressure model from Section 3, provides a lens for interpreting my instrumental variables results below. The relationship between physician speed-ups and patient outcomes thus reflects both the direct effect of physician haste and the indirect effects of induced cutbacks on other inputs. My instrumental variables approach to estimating the effects of physician speed-ups on patient outcomes incorporates both of these channels, providing the relevant policy parameter.

6.2 Using Peer-Induced Speed as an Instrument

In this section, I describe how I use peer-induced speed-ups to estimate the within-physician marginal returns to care in an instrumental variables framework.

My model for estimating the impact of the log time a physician spends on a case ($\ln LOS$) on a variety of outcomes (y) is:

$$y_c = \beta \ln LOS_c + \mathbf{X}_c'\gamma + \omega_{d(c)} + \epsilon_c \tag{5}$$

I condition on the same rich set of case characteristics $\mathbf{X}_c$ as in the match effects analysis in Section 5.1. This model includes physician effects ω_d, so that β is interpretable as the marginal return to a given physician's time spent caring for a case. OLS estimation of this model is unlikely to provide a causal estimate of β because unobserved case characteristics (e.g. severity, acuity) influence both the time it takes to care for the patient and the expected outcome of the patient. When a physician is observed working at different speeds through different cases, it does not necessarily represent variation in work pace that is orthogonal to the error terms ϵ.

To address this endogeneity concern, I use a first-stage equation for predicting case c's LOS based on the quasi-randomly assigned peer group of physician $d(c)$. This first-stage equation takes the form:

$$\ln LOS_c = \phi_{d(c),g(c)} + \mathbf{X}_c'\pi + \theta_{d(c)} + \nu_c \tag{6}$$

where $\phi_{d,g}$ is the team match effect in $\ln LOS$ from Section 5. The team match effect measures the team-induced $\ln LOS$ of physician d working in the team setting g and is the excluded instrument in the system of equations defined by Equations 5 and 6. Equation 6 is identical to Equation 1.

Use of the full set of team match fixed effects as instruments is likely to suffer from finite-sample bias if the number of observations per team match does not go to infinity with sample size (Bound et al. (1995)). I opt

for a split-sample strategy that avoids overfitting in the first-stage by using parameter estimates from an independent partition of the data, in the spirit of Angrist and Krueger (1995). I use the same split-sample technique (block-sampling of physician-team-shift cells) as described in Section 5.3 to partition the data into two subsamples, A and B. By estimating Equation 6 on each partition, I recover two independent estimates of ϕ_g: $\hat{\phi}_g^A$ and $\hat{\phi}_g^B$. I construct from these estimates a single split-sample instrumental variable:

$$\tilde{\phi}_c = \begin{cases} \hat{\phi}^A_{d(c),g(c)} \text{ if } c \in B \\ \hat{\phi}^B_{d(c),g(c)} \text{ if } c \in A \end{cases}$$

The instrument $\tilde{\phi}_{d,g,c}$ can be thought of as the adjusted leave-out average $\ln LOS$ of physician d working with coworker group g. My block-sampling of physician-team-shift cells is meant to acknowledge the fact that my data are not independent. With cross-sectional iid data, split-sample IV amounts to simply splitting the data into two random partitions, fitting the model on one partition, and using the estimates to generate first-stage predictions in the complement partition. In my setting, correlations between the error terms induced by contemporaneous physician-shift or hospital-shift shocks make it necessary to split the data more systematically. I use block-sampling of physician-team-shift cells to eliminate the influence of correlated shocks within a physician's shift. In principle, there are a number of ways one can construct instruments from sample-splitting or jackknife techniques. I opt for this simple partitioning method for transparency and ease of computation.[36]

After constructing the instrument, I run a series of reduced-form (intention-to-treat [ITT]) regressions

$$y_c = \delta \tilde{\phi}_c + \mathbf{X}_c' \tilde{\gamma} + \tilde{\theta}_{d(c)} + \tilde{\epsilon}_c \tag{7}$$

with the typical set of time controls and risk adjusters (x_c) and physician fixed effects $\tilde{\theta}_d$. These estimates contrast a given physician working in groups with which she is estimated (on the complementary partition of the data) to work quickly or slowly. Since teams are as good as randomly assigned, OLS estimation of this reduced-form regression provides clean estimates of δ. I also run the analog of the first-stage regression:

$$\ln LOS_c = \psi \tilde{\phi}_{d,g,c} + \mathbf{X}_c' \tilde{\pi} + \tilde{\omega}_{d(c)} + \tilde{\nu}_c \tag{8}$$

[36] For example, one could use leave-shift-out estimates (which are computationally expensive if one has to run regressions for every leave-out-shift sample), or rely on team effects using other months' or years' estimates. I have experimented with some of these alternatives, which provide very similar results.

where ψ is expected to be considerably less than 1 due to measurement error in $\tilde{\phi}_c$. The indirect least squares estimate of β in Equation 5 is just the ratio δ/ψ. Finally, I estimate the model using two-stage least squares (2SLS) with $\tilde{\phi}_c$ as the instrument for $\ln LOS_c$ in Equation 5 , which produces the same estimate of β as the indirect least squares estimate. I report standard errors clustered at the hospital level. This choice of clustering is conservative relative to clustering at the physician level.[37]

6.3 Identifying Assumptions

My instrumental variables strategy for identifying marginal returns to care relies on a relevance condition, a monotonicity assumption, and an exclusion restriction. Relevance requires that coworker groups be influential in a physician's pace of care ($Var(\phi_{d,g}) \neq 0$, so that $\psi > 0$ in Equation 8). I have already presented a battery of evidence that coworker groups do indeed affect a physician's speed on a given case, thus satisfying this condition. I present evidence in support of monotonicity and exclusion below.

6.3.1 Monotonicity

As discussed in Imbens and Angrist (1994), identification of local average treatment effects (LATEs) in an IV framework also requires a monotonicity assumption. This amounts to assuming that a coworker group that speeds up a physician on one case also speeds her up on all other cases. Although monotonicity is fundamentally unprovable, I provide a few quantitative tests here that lend support to this assumption. Namely, I re-estimate my match effects model Equation 1 on mutually exclusive samples of case types and work environments and test whether the effect of a team on a physicians pace of care is in the same direction across these sampels. Figure 9 plots the estimates of $\phi_{d,g}$ from these samples against each other, along with coefficients and standard errors from WLS (case-weighted) regressions relating each measure of team match effects. Estimated team match effects are highly positively correlated across cases with different observable characteristics: gender, age, severity, day vs. night, and weekdays vs. weekends. For example, a WLS regression of male-patient on female-patient team match effects produces a coefficient of .6097 ($s.e. = .0064$). These results lend some credence to the monotonicity assumption.[38]

[37] My use of clustered standard errors is in line with other work using jackknife or split-sample instrumental variables estimation strategies (e.g., Doyle et al. (2015)). Correcting standard errors for the fact that the instrument is a generated regressor is a goal of future work (Murphy and Topel (2002)).

[38] Figure A.1 displays the same plots for physician effects, also documenting that physicians do not seem to be fast on some types of cases and slow on others.

6.3.2 Exclusion Restriction

The exclusion restriction in this setting requires that the error term in Equation 5 is independent of the identity of a physician's coworker group, i.e. $\mathrm{E}[\epsilon_c \mid g(c)] = \mathrm{E}[\epsilon_c] = 0$, for each physician. This condition could be violated in a few important ways. First, sorting would lead to an exclusion-restriction violation if patients are allocated to on-duty physicians according to those physicians' comparative advantages. Second, sample selection from differential admitting behavior across coworker groups may cause important compositional changes in a physician's caseload. I have presented a number of tests of these two types of sorting earlier in the paper, finding little evidence of either. Throughout my analysis, I provide placebo tests for the hypothesis that either sorting or selection change the composition of a physician's workload in a way correlated with the peer-induced speed-ups. My placebo tests again use the Random Forests risk scores, an omitted variable in the risk-adjustment model, as a dependent variable to test whether the instrument predicts the type of cases a physician treats. Despite these risk scores having high correlations with patient mortality conditional on the baseline risk-adjustment model, I once again fail to reject the null of no sorting or selection.

A final way in which the exclusion restriction may be violated concerns whether coworker groups have direct, independent effects on patient outcomes not captured by their induced speed-up. If the true data generating process involves the coworker group's direct effect on patient outcomes, then ϵ_c in Equation 5 includes a component that is a function of a physician's peer group, so that $\mathrm{E}[\epsilon_c \mid g(c)] \neq 0$.[39] This condition could be violated if peers directly provided care for the patients of a given physician. This is not the case in general; physicians autonomously decide on the course of care for each of their assigned cases. As laid out in Sections 2 and 3, the primary way that peer groups influence an emergency physician's care is via their effect on the physician's time budget for each case. Evidence in Section 6.1 supports this notion, by showing that peer-induced speed-ups line up very tightly with cutbacks in charges generated by time-costly diagnostic tests.

In short, while the exclusion restriction is fundamentally untestable, I find no evidence of important violations to the exclusion restriction. Systematic sorting of physicians to teams is minimal; there is no evidence that physicians differentially admit sicker patients when working with different teams; and direct effects of coworkers on patient care are unlikely.

[39] Kolesár et al. (2014) lays out milder conditions (namely an orthogonality condition rather than an independence assumption) under which leave-out, grouped IV estimators like the one used in this paper are consistent.

6.3.3 Outcomes: Other Inputs and Patient Health Outcomes

I use my instrumental variables strategy to examine a broad range of intermediate inputs to provide a comprehensive picture of how physicians speed up by rationing other types of care, similar to Section 6.1. I estimate regressions with other inputs into care as dependent variables, notably log charges and indicators for specific tests and services provided (CT scans, chest X-rays, EKGs, diagnostic ultrasounds, and respiratory services). As is common in the literature, I use log charges as a summary index of all the inputs into care, less the physician's time.[40] The response of charges to a physician working in a fast or slow team environment almost surely reflect true differences in inputs, rather than mere price differences, as it is unclear how a physician would use differently priced resources within a hospital when working at different paces.

Finally, I examine patient health outcomes. I focus the majority of my analysis on 30-day mortality, an unambiguous marker of the quality of healthcare delivered by the ED physician. In order to facilitate this analysis, I focus on the subsample of patients who are at risk of 30-day mortality. I define this high-risk group using Random Forest prediction models, as discussed in more detail in Appendix C.[41] My approach to estimating quality of care is similar to previous work focusing on groups of patients with high *ex ante* risk of mortality (e.g. Card et al. (2009); Doyle et al. (2015)).

Importantly, when I limit analysis to these cases, I do not recalculate the instrument for this group, but instead rely on the constructed instrument from the full sample. This is desirable for two reasons. First, recalculating the instrument on this $\approx 10\%$ sample amounts to dropping 90% of the data in the construction of the instrument, which severely reduces predictive power in the first-stage. Second, it is intuitively appealing to use the team-induced behavior of the physician on the full set of patients, who are majority low-risk, for prediction of the physician's behavior on high-risk cases. Her peer-induced behavior on low-risk patients is much more likely to reflect responses to typical social pressures, since these cases are fairly commonplace and the physician is expected to be independent in her care for those cases.

[40] I could instead compute costs using a deflator cost-to-charge ratio (CCR), but to my knowledge hospital-level CCRs only exist for inpatient care. To the extent that my analysis uses only within-hospital variation with a multitude of time effects, any inflation-related changes to charges that are not reflective of true costs of care are subsumed. Using CCRs to deflate may be problematic, since they are well known to introduce noise (Almond et al. (2010)).

[41] These prediction models, fit within each hospital, take a case's predetermined characteristics as inputs into a group of bootstrapped, de-correlated tree models (see Breiman (2001)). I define a patient's out-of-sample risk score (i.e. propensity score) as the vote share of all trees that do not include the observation itself (called the "out-of-bag" vote share in the statistics literature). Tree-based methods allow rich, difficult-to-model interactions between the inputs in determining the risk score, and generally far out-perform other methods such as logistic regression in out-of-sample prediction.

6.4 Results of Within-Physician Design

Table 12 reports results of my instrumental variables analysis for cases at risk of mortality. Column 1 contains reduced-form estimates, while Column 2 contains estimates from 2SLS estimation using the split-sample instrumental variables approach laid out above. Note that this table restricts attention to the top decile of mortality risk, as reflected in the sample's 30-day mortality rate of 4.2% (see Column 3 for averages of outcomes in the estimation sample). This compares to a mortality rate of 5 per 1000 in the full sample. These patients are also more intensively treated and receive more tests than the typical ED patient. 22% of high-risk patients receive a CT scan, while 30% receive a chest X-ray, and 39% receive an EKG.

The first-stage using the split-sample instrument is strong. The coefficient from the regression of log length of stay on the split-sample instrument controlling for the baseline risk adjusters, time effects, and physician effects is .241 ($s.e. = .0217$ clustered at the hospital-level). That the coefficient is substantially less than 1 reflects the noise in the team match effects estimates. By splitting the sample, the average physician-team cells, which contain 60-70 cases in the full sample, only contain $\approx$ 30-35 cases in each of the partitions. This induces more noise is the estimated match effects, which attenuates the reduced-form but does not attenuate the 2SLS coefficients. 2SLS accounts for this attenuation by dividing the reduced-form effect sizes by the first-stage coefficient.

The top panel of Table 12 confirms the analysis of team-induced charges from Section 6.1[42] Focusing on the 2SLS results in Column 2, a physician working in a coworker group that induces her to slow down by 10% per case also spends 2.3% ($s.e. = .463\%$) more per case. In terms of the team match effects, working in a 2-standard deviation (SD) slower peer environment slows a physician down by 12.6% (see Table 10) and increases her total per-case charges by 2.9%. On a base of $\approx \$2000$ in typical charges for this group, this amounts to an increase of $\approx \$58$ per case. This increase reflects in part an increased odds of ordering extra diagnostic tests, as shown in the subsequent rows of the table.

CT scans, X-rays, and EKGs all increase in use as a physician is induced to slow down care. Working in a 2SD slower peer environment, a physician's use of CT scans increases by 0.94 percentage points ($s.e. = 0.22$ percentage points) on a base rate of 22.35%. Given the same perturbation, X-ray utilization increases by 0.64 percentage points ($s.e. = 0.28$), and EKG use increases by 0.88 percentage points ($s.e. = 0.29$). There is also suggestive evidence that physicians do more diagnostic ultrasounds in the high-risk population

[42]The difference between the analysis in Section 6.1 and the present analysis is the use of split-sample adjusted group means as opposed to single-sample adjusted group means. The split-sample approach, as usual, has the benefit of removing finite-sample bias and bias from correlated shocks.

when working at a slower pace, although this result is statistically insignificant.

The effects of physician speed on patient health are substantial. When physicians work in 2SD slower team environments, 30-day mortality in high-risk cases decreases by 0.21 percentage points ($s.e. = .108$ percentage points). This represents a reduction of 5% given baseline 30-day mortality. This result suggests that incentivizing physicians to work faster through cases may have deleterious effects on patient outcomes, at least for the 10% of cases at high risk of mortality.

Table 12 also presents the result of a placebo test to check whether mortality effects are driven by physicians treating and discharging lower-risk cases (as measured by Random Forest risk scores) when working in coworker groups estimated to slow them down. Given the lack of evidence of patient sorting in the previous sections of this paper, it would be surprising to find any sorting of this type here. Reassuringly, there is no evidence that the risk of a physician's caseload is related to the instrument.[43]

Figure 10 displays graphically the reduced-form effects on 30-day mortality of team-induced speed on patient outcomes. The solid blue line shows the relationship between team-induced speed and 30-day mortality, while the dashed red line helps evaluate the degree of selection on observables by replacing 30-day mortality with the random-forest mortality risk score described in Appendix C. In summary, for these at-risk cases, a physician's team-induced speed-up leads to an increase in 30-day mortality. This increase in mortality is not driven by observably sicker patients being sorted to physicians when they work in high-speed environments, evidenced by the mortality risk scores being flat across the distribution of the instrument. These mortality increases likely stem from not only the decreased time the physician spends evaluating the case, but also decreases in orders for potentially revelatory diagnostic tests. Decreased investigation of high-risk cases is likely to cause oversights and missed diagnoses that negatively impact health outcomes.

Given the modest magnitudes of spending in the ED, the mortality effects are quite large. To add some context, I next provide a very simple back-of-the-envelope calculation, which ignores a number of factors such as the social value of wait times, the implicit costs of a physician's time, and the costs associated with other labor inputs. Nonetheless, if we take my estimates at face value, a 10% increase in peer-induced *spending* coincides with a $\frac{-.0040}{.0556} \times 0.1 = -0.72$ percentage-point reduction in 30-day mortality in the high-risk population. In this population of 1.539 million cases, this amounts to, at a minimum, $0.0072 \times 1,539,000 \approx 11,000$ life-months, or 922.66 life-years saved. Given average charges of $\approx$ \$2000 in this

[43] Placebo tests using other omitted variables, such as Charlson Comorbidity Indices (CCI) and the number of visits to the ED in the past 30 days also fail to detect any compositional differences related to the instrument.

group, the cost of this 10% across-the-board increase in spending is $\$200 \times 1,539,000 = \$307,800,000$. This implies a cost-per-life-year saved of $\$307,800,000/922 \approx \$334,000$. This importantly assumes all mortality effects evaporate after 30 days, and it assumes that listed charges reflect costs, whereas in reality they are probably inflated over cost. This estimate is around the typical value of a statistical life year (Aldy and Viscusi (2008)).

6.4.1 Results on the Full Sample

Table 13 presents my reduced-form and IV estimates on the full sample, where I do not limit to cases at risk of mortality. I find a slightly stronger first-stage relationship between the constructed instrument and LOS. In the 2SLS analysis, I find very similar effects of physician slow-downs on charges and specific diagnostic tests. The placebo test on case risk scores once again fails to detect any substantial compositional differences in a physician's caseload related to the split-sample instrument. Finally, in the full sample, physician slow-downs have no effect on 30-day mortality.

As discussed previously, most patients arriving to the ED have conditions or complaints that are not acutely life-threatening. These cases' mortality risks are likely unaffected by speed of care or testing. Heterogeneous treatment effects by patient risk group make it so that the average treatment effect across the full sample is weighted towards the majority of cases with negligible treatment effects. For this reason I focus the remainder of the analysis on at-risk cases.

6.4.2 Non-Specific Complaints

Many of the top complaints of patients arriving to the ED are vague or non-specific in nature. Patients complaining of abdominal pain make up 8.2% of discharges, while another 4.1% of patients arrive with nonspecific chest pain. These cases and others like them involve a patient with a symptom that could indicate many potential underlying causes. Some of these underlying causes may be dangerous, such as acute headaches, which are largely unproblematic, but in rare cases may indicate a stroke or a transient ischemic attack, a precursor to stroke. In these vague cases, emergency physicians rely on diagnostic tools, from formal tests to more nuanced techniques including eliciting a patient's history in order to come to the right conclusion.[44] Since missed diagnoses are likely more prevalent in this population, non-specific

[44] See Sanders (2010), for example, on the popular press opinion of the importance and nuances of patient histories.

complaints represent an interesting subgroup to assess whether missed diagnosis is a likely mechanism through which physician speed-ups lead to increased mortality.

Table 14 presents IV estimates limiting attention to high-risk cases with non-specific or vague complaints. I identify these cases using the arriving ICD-9 complaint of the case; see Appendix D for a listing of the diagnosis classes underlying my categorization. The first-stage relationship between team environment and own speed has very similar magnitude as in the full sample and the high-risk sample. Physicians working in faster-paced coworker environments cut care for this subpopulation at rates very similar to the full sample. CT scans and EKGs see large cutbacks among this population when physicians speed up. Working in a 2SD faster coworker environment and thus speeding a physician up by 12.6% decreases charges by 2.5%, which is very similar to the 2.9% cutback across all high-risk cases.

Turning to 30-day mortality, the IV estimate is large and significant. Speeding up a physician by 10% leads to an increase in mortality of 0.1 percentage points. Baseline mortality for the high-risk non-specific population is significantly lower than for the high-risk population at large (1.13% versus 4.21%), so the IV estimate implies an increase in 30-day mortality of 9% for a 10% physician speed-up. Put another way, moving a physician to a 2SD faster team environment increases 30-day mortality for this subpopulation by 19.9 percent ($\frac{-0.126 \times -0.021}{.0133}$). ED care for these sick, non-specific populations is highly valuable on the margin, such that cutting back on health care inputs causes substantial harm.

6.5 Heterogeneity in Marginal Returns Across Physicians

My design lends itself to considering whether different physicians in the same hospital operate at different margins of the returns to care. Here I present one particular classification of physicians: their average speed. A natural inquiry is whether in the same hospitals, slower physicians have lower marginal returns, as would be the case in a world of diminishing returns to treatment. As shown in Tables 15 and 16, there does seem to be substantial heterogeneity across physicians of different average paces. Among at-risk patients treated by fast physicians (Table 15; where "fast" is defined as below the median physician effect for LOS in a given hospital), 30-day mortality is very sensitive to team-induced LOS. The IV estimate of the marginal increases in 30-day mortality due to speeding up by 10% is $-0.0286 \times .1 = -0.00286$ percentage points, a $-.00286/.0411 = 7$ percent reduction in mortality among this group. Spending and testing are all sensitive in similar magnitudes to the full at-risk sample.

For slow physicians treating at-risk cases, marginal returns are estimated to be near 0. The IV estimate

for these physicians of slowing down 10% is a little more than one-tenth of the size of the effect for fast physicians ($-0.0032 \times .1 = -0.00032$). Log charges are still estimated to be sensitive to physician speed-ups, suggesting that it is *not* the case that slow physicians are able to speed up without cutting back, as they might be able to do if they had more slack time than fast physicians. There is suggestive evidence that slower physicians' cutbacks in response to the same increase in speed are somewhat smaller.

This finding is quite striking and holds up to finer binning of physicians into speed groups. I plot the ITT estimates from 8 regressions that stratify physicians by their hospital-specific octile of speed in Figure 11. This graphical illustration suggests that marginal returns to time diminish to near 0 by the fourth octile and are relatively flat thereafter. Given the magnitudes, one conclusion is that fast physicians are working *too* fast. On the other hand, estimated mortality effects near zero for the slower physicians may imply that policies aimed at speeding this group up would not harm at-risk patients.

These results beg the question: are slow physicians, who appear to have reached a point of small marginal returns, providing better average care? We might be tempted to assume all physicians in a hospital face the same production function, in which case, slower, more cautious care should be met with better outcomes on average. I address this question in the next section, where I compare my within-physician results to the results of a cross-physician, within-hospital design.

7 Physician Heterogeneity: Do Slower Physicians Save More Lives?

In the previous section, I found that marginal returns to care are highly positive for faster physicians and near zero for slower physicians in the same hospital. This section seeks to explain these heterogeneous marginal returns.

I first provide an auxiliary cross-sectional analysis relating a physician's *typical* speed to her *average* outcomes to answer the question: Do slower physicians save more lives? I find that slower physicians' adjusted mortality rates are very similar to their faster colleagues', despite slower physicians taking more time and spending substantially more on testing and other inputs.

I then discuss a simple stylized model of physician productivity and endogenous input choices that can explain the two facts about physician heterogeneity: (a) that faster physicians have higher marginal returns to care, and (b) that faster physicians provide no worse care on average.

7.1 Cross-Sectional Analysis

Here I use an auxiliary research design based on quasi-random assignment of cases to physicians of different average speeds to estimate the cross-sectional relationship between physician speed and patient outcomes.

The following two-equation system provides the foundation for this cross-sectional analysis:

$$y_c = \beta \ln LOS_c + \mathbf{X}_c'\gamma + \epsilon_c \tag{9}$$

$$\ln LOS_c = \theta_{d(c)} + \mathbf{X}_c'\pi + \nu_c \tag{10}$$

$\mathbf{X}_c$ contains the usual hospital-time effects and case risk adjusters, and θ_d represents an effect on log length of stay of the case's quasi-randomly assigned physician $d(c)$. In practice, I construct the physician effects using the same split-sample methodology as in the teams analysis.[45] Outcomes (y) are the same as in the within-physician analysis, including measures of case-level inputs and 30-day mortality.

Tables 17 and 18 provide the results of this exercise for the high-risk sample and for the full sample, respectively. Slower physicians are found to have slightly *worse* outcomes in terms of patient mortality on average. A physician who is 10% slower spends $.5086/.91 = 5.59\%$ more, in part by using more diagnostic tests, and has a 30-day mortality rate that is .033 percentage points *higher*, representing a 0.78% difference in mortality. There is no evidence of patient sorting across physicians, as evidenced by the placebo tests using constructed mortality risk scores. This result is similar to findings in the Dartmouth Atlas, that vast differences in levels of care across places or across hospitals appear unrelated to patient outcomes. It is also similar to a recent paper (Doyle et al. (2010)) that studies patients randomly assigned to physicians with different training backgrounds, documenting that costs are much lower among physicians with elite training, but that patients fare similarly in terms of mortality across the two settings. Importantly, though, my analysis uses only within-hospital variation – it cannot be the case that the excess costs incurred by some physicians are due to organizational differences across firms.

The across-physician result in this section stands in contrast to the positive within-physician marginal returns to treatment documented earlier. Taken at face value, researchers or policy makers might conclude from these across-physician estimates that emergency care is on the flat of the curve, and recommend

[45] Physician effects do not suffer from the same finite-sample bias problems as team match effects, so the split-sample and single-sample estimates of the physician effects are nearly identical.

instituting incentives for physicians to speed up. These policies would, however, be misguided. On the margin, the average physician is providing valuable care. Previous research using across-provider designs should be interpreted cautiously when policy is aimed at incentivizing providers to cut back, as the relevant parameters for these policy interventions are within-provider marginal returns.

In the next section, I provide a simple framework based on differences in physician productivity that rationalizes these across-physician results in light of my primary within-physician estimates of returns to care.

7.2 The Role of Productivity Differences Across Physicians

Physicians who generally spend less time on each case have higher marginal returns to time than their colleagues who spend more time on each case. This finding alone could be consistent with a few underlying stylized models of physician behavior. First, physicians could simply have different preferences. Some physicians may care more about wait times than others, for example, leading them to choose to work faster on average to keep wait times down. Under the null that these physicians are otherwise exchangeable – in particular that they have the same concave production function – physicians who choose to work faster will have a higher return to care on the margin. However, so long as the production function is monotonically increasing, faster physicians will provide *worse* outcomes on average. In contrast, I find that faster physicians fare slightly better on average than their slower counterparts.

An alternative model of physician behavior relies on heterogeneity in physician productivity, paired with physicians' preferences to "solve" each case. To be more precise, imagine that physicians work on each case until they come to a point where their probability of having missed a crucial diagnosis (i.e. a type II error) is near some (potentially physician-specific) threshold. One can think of this threshold as a standard of care, or as the physician having done due diligence in working through a case. Physicians with similar thresholds for having "solved" a case will have similar average outcomes, but the more productive physicians will reach the threshold while using fewer inputs and taking less time. On the margin, these endogenously faster physicians will have higher returns to treatment; cutting back on their time may force them to omit an important diagnostic test.

A graphical illustration of this simple model is presented in Figure 13. In this figure, physicians differ in their productivity. More productive physicians require fewer inputs to reach a given probability of an error (i.e. the inverse of quality). If all physicians try to meet a similar threshold for the probability of

an error on a case, then lower-productivity physicians will use more inputs on average. However, on the margin, higher productivity physicians will have higher returns to care, as reflected by the slope of the physician's production function at the threshold.

This model, while very simple, provides an intuitive explanation of the main results of my analysis. More work investigating physician behavior in similar contexts could provide valuable insights into how much of the large variation in healthcare spending is due to differences in physician practice and productivity (Chandra et al. (2012); Chandra and Staiger (2007); Chandra and Staiger (2011); Finkelstein et al. (2014); Abaluck et al. (2014); Chandra et al. (2015)).

8 Conclusion

Policymakers are increasingly interested in ways to curtail health care spending, with the belief that much of the time and money spent on healthcare in the US is of little to no value. As the population of the US ages, demands on healthcare providers to be efficient in their use of resources and time are increasing. Physicians, the central arbiters of treatment choices, have become the focus of much of the discussion on cutting back care. Hospitals worried exclusively about queueing and wait times may provide piece-rate style incentives to their physicians. Even CMS has recently adopted measures of wait times and throughput for reimbursement purposes.

This paper presents some of the first evidence on the within-physician marginal returns to care. Workplace peer effects generate an exogenous within-physician source of input variation. I leverage this source of variation using detailed administrative data. I provide a novel approach to estimating peer effects in a workplace that features physicians autonomously working through cases. I find that a physician's pace of care on a particular case is highly sensitive to the identity of her coworker group. A physician's pace of care across peer groups has a variance that is one-quarter of the variance of the pace of care across physicians in a hospital.

Contrary to opinions that health care spending has reached the flat of the curve, I document large positive marginal returns to health care inputs. When the average physician works in peer groups that induce her to speed up and cut back on other inputs, patient mortality among at-risk cases increases. Marginal returns to care are larger among patients whose symptoms are vague, where time for careful scrutiny may help a physician make an accurate diagnosis. Importantly, I find that marginal returns to care are highest

for the fastest physicians in a hospital, and that patient outcomes for slower physicians are unaffected by cuts in care.

More nuanced incentives, such as a two-sided incentive that incentivizes fast physicians to slow down and vise versa, may provide improvements in patient health (via decreased mortality for cases treated by fast physicians) while maintaining or even decreasing total input use (via potentially large reductions in spending among the slower physicians).

This study has a few important limitations. First, a problem that is common to the literature: because objective health outcomes are hard to come by, this paper limits analysis of marginal returns to care to patients at risk of an objective health outcome: 30-day mortality. My results do not shed light on the marginal benefits of increased care in terms of patient morbidity or patient satisfaction.

Secondly, the result that physicians cut back nearly equally on care for the sickest subgroups and for the full population may speak to the interpretation of my findings. We might expect physicians to be more or less targeted in their cutbacks under different policies. If I interpret speed-ups in my setting as being generated by contemporaneous social pressure, physicians may be less patient-centered in their cutbacks than under alternative incentives such as an annual volume bonus, where the rewards to the physician are spread out over time.

Next, because of data limitations, I am only able to provide indirect evidence of returns to speed and spending on an important margin: the decision to admit the patient to the hospital on the current visit. Current databases discard information from the ED visit for admitted patients, and only keep the inpatient discharge record for those cases. These inpatient records do not include time stamps from the ED visit, nor do they include the license number of the emergency physician of record. Including ED records for these admitted patients in data collection is likely quite easy for hospitals in the age of electronic medical records. The costs are likely administrative, but the benefits to research and to public health efforts of having these additional records would likely exceed those costs.

Finally, future work linking state hospital discharge records with patient identifiers to other administrative claims-based records, for example Medicare claims data or All-Payer Claims Databases, would provide a more complete picture of care for a subset of ED visits (those in the claims records), in which designs similar to the one I have presented here could be fruitfully applied.

References

Abaluck, Jason, Leila Agha, Christopher Kabrhel, Ali Raja, and Arjun Venkatesh (2014) "Negative Tests and the Efficiency of Medical Care: What Determines Heterogeneity in Imaging Behavior?" Working Paper 19956, National Bureau of Economic Research.

Abowd, John M, Francis Kramarz, and David N Margolis (1999) "High Wage Workers and High Wage Firms," *Econometrica*, Vol. 67, pp. 251–333.

Aldy, Joseph E and W Kip Viscusi (2008) "Adjusting the value of a statistical life for age and cohort effects," *The Review of Economics and Statistics*, Vol. 90, pp. 573–581.

Almond, Douglas, Joseph J Doyle, Amanda Ellen Kowalski, and Heidi L Williams (2010) "Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns," *The Quarterly Journal of Economics*, Vol. 125, pp. 591–634.

Angrist, Joshua D and Alan B Krueger (1995) "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, Vol. 13, pp. 225–235.

American Medical Association (2003) *Physician Socioeconomic Statistics*, Chicago, IL.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005) "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *The Quarterly Journal of Economics*, Vol. 120, pp. 917–962.

——— (2006) "The Evolution of Cooperative Norms: Evidence From a Natural Field Experiment," *Advances in Economic Analysis & Policy*, Vol. 5.

——— (2010) "Social Incentives in the Workplace," *Review of Economic Studies*, Vol. 77, pp. 417–458.

Berwick, Donald M and Andrew D Hackbarth (2012) "Eliminating Waste in US Health Care," *JAMA*, Vol. 307, pp. 1513–1516.

Bound, John, David A Jaeger, and Regina M Baker (1995) "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, Vol. 90, pp. 443–450.

Breiman, Leo (2001) "Random Forests," *Machine Learning*, Vol. 45, pp. 5–32.

Card, David, Carlos Dobkin, and Nicole Maestas (2009) "Does Medicare Save Lives?" *Quarterly Journal of Economics*.

Card, David, Jörg Heining, and Patrick Kline (2013) "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, Vol. 128, pp. 967–1015.

Chan, David (2015) "Teamwork and Moral Hazard: Evidence from the Emergency Department," *Journal of Political Economy*.

Chandra, Amitabh, David Cutler, and Zirui Song (2012) "Who Ordered That? The Economics of Treatment Choices in Medical Care," *Handbook of Health Economics*, Vol. 2, pp. 397–432.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson (2015) "Healthcare Exceptionalism? Performance and Allocation in the U.S. Healthcare Sector," Working Paper 21603, National Bureau of Economic Research.

Chandra, Amitabh and Doug Staiger (2007) "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, Vol. 115, pp. 103–140.

Chandra, Amitabh and Douglas O Staiger (2011) "Expertise, Underuse, and Overuse in Healthcare."

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

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg (2013) "Peer Effects in the Workplace."

Cutler, David M. (2007) "The Lifetime Costs and Benefits of Medical Technology," *Journal of Health Economics*, Vol. 26, pp. 1081–1100.

Doyle, Joseph J (2011) "Returns to Local-Area Health Care Spending: Evidence from Health Shocks to Patients Far From Home," *American Economic Journal: Applied Economics*, pp. 221–243.

Doyle, Joseph J, Steven M Ewer, and Todd H Wagner (2010) "Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams," *Journal of Health Economics*, Vol. 29, pp. 866–882.

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

Epstein, Andrew J and Sean Nicholson (2009) "The Formation and Evolution of Physician Treatment Styles: An Application to Cesarean Sections," *Journal of Health Economics*, Vol. 28, pp. 1126–1140.

Falk, Armin and Andrea Ichino (2006) "Clean Evidence on Peer Effects," *Journal of Labor Economics*, Vol. 24, pp. 39–58.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams (2014) "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," Working Paper 20789, National Bureau of Economic Research.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini (2003) "Performance in Competitive Environments: Gender Differences," *The Quarterly Journal of Economics*, Vol. 118, pp. 1049–1074.

Gowrisankaran, Gautam, Keith A Joiner, and Pierre Thomas Léger (2014) "Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments.."

Imbens, Guido W and Joshua D Angrist (1994) "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 62, pp. 467–75.

IOM (2006) "Hospital-Based Emergency Care: At the Breaking Point,"Technical report, Institute of Medicine.

Jena, Anupam B., Seth Seabury, Darius Lakdawalla, and Amitabh Chandra (2011) "Malpractice Risk According to Physician Specialty," *New England Journal of Medicine*, Vol. 365, pp. 629–636.

Kachalia, Allen, Tejal K Gandhi, Ann Louise Puopolo, Catherine Yoon, Eric J Thomas, Richard Griffey, Troyen A Brennan, and David M Studdert (2007) "Missed and Delayed Diagnoses in the Emergency Department: A Study of Closed Malpractice Claims from 4 Liability Insurers," *Annals of Emergency Medicine*, Vol. 49, pp. 196–205.

Kandel, Eugene and Edward P Lazear (1992) "Peer Pressure and Partnerships," *Journal of Political Economy*, Vol. 100, pp. 801–17.

Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger (2013) "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," *Seattle, WA: Bill and Melinda Gates Foundation*.

Kane, Thomas J, Douglas O Staiger, David Grissmer, and Helen F Ladd (2002) "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," *Brookings papers on education policy*, pp. 235–283.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens (2014) "Identification and Inference With Many Invalid Instruments," *Journal of Business & Economic Statistics*, pp. 00–00.

MaCurdy, Thomas, Jason Shafrin, Diana Zheng, and Frederick Thomas (2011) "Optimal Pay-for-Performance Scores: How to Incentivize Physicians to Behave Efficiently."

Marschak, Jacob and William H Andrews (1944) "Random Simultaneous Equations and the Theory of Production," *Econometrica, Journal of the Econometric Society*, pp. 143–205.

Mas, Alexandre and Daniel Herbst (2015) "Peer Spillovers in the Workplace: A Meta-Analysis."

Mas, Alexandre and Enrico Moretti (2009) "Peers at Work," *American Economic Review*, Vol. 99, pp. 112–45.

McClellan, Mark, Barbara J McNeil, and Joseph P Newhouse (1994) "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables," *JAMA*, Vol. 272, pp. 859–866.

McClellan, Mark and Douglas Staiger (2000) "Comparing the Quality of Health Care Providers," in *Forum for Health Economics & Policy*, Vol. 3.

Molitor, David (2011) "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration."

Mundlak, Yair (1961) "Empirical Production Function Free of Management Bias," *Journal of Farm Economics*, Vol. 43, pp. 44–56.

Murphy, Kevin M and Robert H Topel (2002) "Estimation and Inference in Two-Step Econometric Models," *Journal of Business & Economic Statistics*, Vol. 20, pp. 88–97.

Newman-Toker, David E, Ernest Moy, Ernest Valente, Rosanna Coffey, and Anika L Hines (2014) "Missed Diagnosis of Stroke in the Emergency Department: A Cross-Sectional Analysis of a Large Population-Based Sample," *Diagnosis*, Vol. 1, pp. 155–166.

Niederle, Muriel and Lise Vesterlund (2007) "Do Women Shy Away From Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics*, Vol. 122, pp. 1067–1101.

Phelps, Charles E. and Cathleen Mooney (1993) "Geographic Variation in Health Care: The Role of Private Markets," *Competitive Approaches to Health Care Reform*, pp. 139–175.

Pope, J Hector, Tom P Aufderheide, Robin Ruthazer, Robert H Woolard, James A Feldman, Joni R Beshan-

sky, John L Griffith, and Harry P Selker (2000) "Missed Diagnoses of Acute Cardiac Ischemia in the Emergency Department," *New England Journal of Medicine*, Vol. 342, pp. 1163–1170.

Rothstein, Jesse (2014) "Revisiting the Impacts of Teachers."

Sacerdote, Bruce (2001) "Peer Effects With Random Assignment: Results For Dartmouth Roommates," *The Quarterly Journal of Economics*, Vol. 116, pp. 681–704.

Sanders, L. (2010) *Every Patient Tells a Story: Medical Mysteries and the Art of Diagnosis*: Broadway Books.

Skinner, Jonathan and Elliott Fisher (2010) "Reflections on Geographic Variations in US Health Care."

Thomas, Eric J, David M Studdert, Helen R Burstin, E John Orav, Timothy Zeena, Elliott J Williams, K Mason Howard, Paul C Weiler, and Troyen A Brennan (2000) "Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado," *Medical Care*, Vol. 38, pp. 261–271.

Thompson, David A, Paul R Yarnold, Diana R Williams, and Stephen L Adams (1996) "Effects of Actual Waiting Time, Perceived Waiting Time, Information Delivery, and Expressive Quality on Patient Satisfaction in the Emergency Department," *Annals of Emergency Medicine*, Vol. 28, pp. 657–665.

Van Parys, Jessica (2013) "What Makes an Efficient Physician? Evidence from Florida Emergency Room Visits."

Wilson, Michael, Jonathan Welch, Jeremiah Schuur, Kelli O'Laughlin, and David Cutler (2014) "Hospital and Emergency Department factors Associated With Variations in Missed Diagnosis and Costs for Patients Age 65 Years and Older With Acute Myocardial Infarction Who Present to Emergency Departments," *Academic Emergency Medicine*, Vol. 21, pp. 1101–1108.

9 Tables

Table 1: Case descriptives

Variable	All discharges Mean (1)	All discharges SD (2)	Analysis sample Mean (3)	Analysis sample SD (4)	Variable	All discharges Mean (5)	All discharges SD (6)	Analysis sample Mean (7)	Analysis sample SD (8)
Patient characteristics					Case characteristics				
Female	0.540		0.546		Hour of arrival: 12am to 5am	0.110		0.126	
Age < 1	0.029		0.016		Hour of arrival: 6am to 11am	0.248		0.269	
Age $\in [1, 5)$	0.079		0.049		Hour of arrival: 12pm to 5pm	0.348		0.324	
Age $\in [5, 15)$	0.101		0.074		Hour of arrival: 6pm to 11pm	0.294		0.280	
Age $\in [15, 35)$	0.337		0.352						
Age $\in [35, 50)$	0.207		0.225		Weekend	0.279		0.303	
Age $\in [50, 65)$	0.142		0.156		Winter (Jan-Mar)	0.245		0.240	
Age ≥ 65	0.106		0.128		Spring (Apr-Jun)	0.257		0.257	
White	0.499		0.637		Summer (Jul-Sep)	0.251		0.256	
Black	0.255		0.209		Fall (Oct-Dec)	0.247		0.248	
Asian	0.025		0.014						
Other race	0.204		0.132		Personnel assignment				
Missing race	0.018		0.008		Assigned to active physician	0.822		1.000	
Hispanic	0.182		0.135		Number of coworkers of attending physician	2.772	2.599	1.260	1.081
Missing ethnicity	0.085		0.064						
Medicaid	0.349		0.283		Inputs				
Medicare	0.132		0.162		Length of stay (hrs), trimmed at 12	3.065	2.433	2.871	2.293
Other govt. payer	0.010		0.010		Total charges (dollars)	1459.750	2424.070	1477.136	2999.668
Workers' compensation	0.023		0.030		CT scan	0.107		0.129	
Private payer	0.292		0.349		Chest X-ray	0.119		0.127	
Self pay	0.164		0.127		Electrocardiogram	0.119		0.140	
Other payer	0.030		0.039		Ultrasound	0.029		0.031	
Charlson weighted comorbidity index	0.108	0.427	0.119	0.451	Respiratory services	0.025		0.023	
ED visits in past 30 days	0.290	0.954	0.290	0.962					
					Outcomes				
Hospitals		211		137	30-day mortality	0.004		0.005	
Active physicians		4,260		2,969	Left against medical advice	0.018		0.019	
Teams w/ ≥ 100 cases		40,294		20,937	Readmitted to ED or hospital within 30 days	0.203		0.200	
Physician-team cells w/ ≥ 50 cases		64,849		38,670	30-day total follow-up charges	2,095	14,873	2,126	14,971
Cases (discharges)		52,305,056		17,969,850					

Source: SPARCS

Notes: Hospitals restricted to those treating ≥ 1000 cases in sample. Physicians restricted to licenses with at least 1000 associated cases, and at least 500 cases in a single hospital. Teams are defined as groups of physicians on duty in a given hour.

Table 2: Most common reasons for ED visit

Chief complaint	Percent of cases	Patient age	Length of stay (hrs)	Total charges ($)	30-day mortality (%)
Other injuries and conditions due to external causes	9.9	35.2	2.3	1,418	0.26
Abdominal pain	8.2	37.9	4.2	2,828	0.23
Other lower respiratory disease	5.5	36.1	2.8	1,393	0.84
Spondylosis; intervertebral disc disorders; other back problems	5.3	42.2	2.7	1,329	0.20
Other connective tissue disease	4.7	42.6	2.6	1,233	0.24
Nonspecific chest pain	4.1	44.8	3.6	2,282	0.38
Other non-traumatic joint disorders	3.8	42.5	2.5	1,228	0.21
Open wounds of extremities	3.0	36.6	2.0	995	0.08
Headache; including migraine	3.0	38.2	3.3	1,816	0.14
Superficial injury; contusion	2.8	35.6	2.2	1,096	0.17
Other upper respiratory infections	2.7	25.4	2.0	731	0.03
Fever of unknown origin	2.5	15.1	2.7	1,130	0.17
Nausea and vomiting	2.3	31.1	3.7	1,789	0.30
Other skin disorders	2.0	31.1	2.0	693	0.06
Open wounds of head; neck; and trunk	2.0	29.0	2.2	1,279	0.30

Source: SPARCS; tabulation of analysis sample.

Table 3: Hospital descriptive statistics

	All discharges	**Analysis sample**
Annual volume: Average hospital	30,167	30,012
Annual volume: 10th percentile hospital	7,998	13,769
Annual volume: 90th percentile hospital	65,297	51,942
Modal team size: Average hospital	1.91	1.87
Modal team size: 10th percentile hospital	1	1
Modal team size: 90th percentile hospital	3	3
Ownership: Government	.113	.0438
Ownership: Private non-profit	.425	.467
Ownership: Church non-profit	.127	.153
Ownership: Other non-profit	.283	.277
Ownership: Proprietary	.0519	.0584
Type: Critical access	.066	0
Number of hospitals	212	137

Source: SPARCS; tabulation of analysis sample.

Table 4: Physician descriptive statistics

	Physician-weighted		**Case-weighted**	
	Mean	SD	Mean	SD
Male	0.677		0.752	
Female	0.300		0.238	
Missing gender	0.023		0.010	
Medical school cohort	1993.8	10.7	1991.3	10.1
Missing med school info	0.006		0.001	
Cases per physician	6,052	7,151	14,499	10,312
Hospitals worked at in sample	1.545	0.948	1.940	1.241
Months active in sample	53.5	29.4	79.1	29.4
Shifts in sample	432.3	387.1	822.8	438.7
Teams worked with	351.3	422.4	441.6	514.8
Fraction of teams w/ ≥ 50 obs	0.096		0.167	
Cases per physician-team w/ ≥ 50 obs	68.5	7.8	69.5	80.1
Physicians		2,969		

Source: SPARCS
Notes: Limited to physicians with at least 1000 associated discharges in total and at least 500 discharges in the hospital of record.

Table 5: Shift-level descriptive statistics

Variable	Mean	SD
Shift length (hours)	9.522	2.478
Cases discharged	16.923	10.327
No team changes in shift	0.163	
One team change in shift	0.329	
Two team changes in shift	0.286	
Three+ team changes in shift	0.222	
Start hour of shift: midnight to 5am	0.069	
Start hour of shift: 6am to 11am	0.462	
Start hour of shift: noon to 5pm	0.247	
Start hour of shift: 6pm to 11pm	0.222	
9 to 12 hours since last shift ended	0.091	
13 to 18 hours since last shift ended	0.261	
19 to 30 hours since last shift ended	0.116	
31+ hours since last shift ended	0.533	
Number of shifts	1,400,240	

Source: SPARCS
Notes: Shifts are physician-specific. See Appendix A for details on how shifts are constructed from discharge data.

Table 6: Balance on patient risk scores

	(1) Difference (high-low)	(2) s.e.	(3) p-value
Dep var: mortality risk score (in percent, mean = .5411)			
Peer fraction male	.0006	[.0053]	.9051
Peer graduation year	-.0078	[.0047]	.0977
Peer LOS	-.0011	[.0041]	.7925

Do a physician's assigned case characteristics depend on coworker observables? This table present within-physician differences in means of Random Forest risk scores (on a 0-100 scale) of patients the physician treats when working in different classes of peer groups as defined by peer average characteristics. Each row's categorization indicates whether a peer group of a physician on a given case is above or below the average for that characteristic for the physician. Standard errors clustered at the hospital level. Reported p-values from test that difference in means is 0.

Table 7: Sensitivity of team match effects to risk- and time-adjustments

	LOS 8	Log LOS 8	LOS 12	Log LOS 12
Corr w/ baseline				
Baseline			1.000	
Add dummies for # ED visits in past 30 days	0.999	0.999	0.999	0.999
Add arrival shock vingtiles	0.997	0.996	0.998	0.996
Add 30-day mortality risk vingtiles	0.997	0.996	0.998	0.996
Add 5 bins for time through shift	0.970	0.972	0.971	0.971
Baseline			1.000	
Only time effects, no risk-adjustment	0.955	0.957	0.962	0.958
Only month-year-by-hospital and DOW-by-hospital dummies	0.896	0.914	0.855	0.899
Only month-year-by-hospital dummies	0.894	0.911	0.852	0.896
Take away all time effects, only risk-adjust	0.682	0.573	0.635	0.565

Notes: This table presents correlations between baseline estimated team match effects and team match effects estimated using alternative sets of covariates. Each row of the top panel builds on the model of the previous row. The bottom panel selectively omits the indicated variables in each row from the match effects model. In each alternative model, team match effects are normalized to be mean 0 at the case-level for each physician. All correlations are case-weighted.

Table 8: Sensitivity of physician effects to risk- and time-adjustments

	LOS 8	Log LOS 8	LOS 12	Log LOS 12
Corr w/ baseline				
Baseline			1.000	
Add dummies for # ED visits in past 30 days	0.999	0.999	0.999	0.999
Add arrival shock vingtiles	0.999	0.999	0.999	0.999
Add 30-day mortality risk vingtiles	0.999	0.999	0.999	0.999
Add 5 bins for time through shift	0.998	0.999	0.999	0.999
Baseline			1.000	
Only time effects, no risk-adjustment	0.985	0.993	0.982	0.991
Only month-year-by-hospital and DOW-by-hospital dummies	0.985	0.992	0.982	0.991
Only month-year-by-hospital dummies	0.950	0.951	0.931	0.946
Take away all time effects, only risk-adjust	0.981	0.990	0.984	0.989

Notes: This table presents correlations between baseline estimated physician effects and physician effects estimated using alternative sets of covariates. Each row of the top panel builds on the model of the previous row. The bottom panel selectively omits the indicated variables in each row from the match effects model. In each alternative model, team match effects are normalized to be mean 0 at the case-level for each physician. All correlations are case-weighted.

Table 9: Hospital-day level split-sample IV analysis of admitting behavior

	(1) Avg log LOS Discharged cases	(2) Avg log charges Discharged cases	(3) Fraction admitted	(4) Avg log charges Admitted cases	(5) Avg LOS (days) Admitted cases	(6) Avg 30D mortality Admitted cases
Day-level avg LOS match effect	0.604***	0.103***	0.00308	-0.00378	-0.249	-0.000519
	(0.0598)	(0.0181)	(0.00369)	(0.0150)	(0.178)	(0.00410)
Hospital-days	377,585	377,585	377,585	377,585	377,585	377,585

	(1)	(2)	(3)	(4)	(5)	(6)
Day-level avg LOS match effect	0.533***	0.0723***	0.00175	-0.00290	-0.233	-0.0000421
	(0.0484)	(0.0163)	(0.00350)	(0.0148)	(0.176)	(0.00408)
Day-level avg LOS physician effect	0.951***	0.406***	0.0177***	-0.0116	-0.212*	-0.00633*
	(0.0212)	(0.0347)	(0.00312)	(0.00989)	(0.0947)	(0.00287)
Hospital-days	377,585	377,585	377,585	377,585	377,585	377,585

Table 10: Estimated variance components of length of stay

	(1) LOS trim 8hrs	(2) Log LOS trim 8hrs	(3) LOS trim 12hrs	(4) Log LOS trim 12hrs
Basic estimates				
Total variance	3.572	0.516	5.225	0.563
R-squared, full model	0.375	0.464	0.398	0.473
Variance of patient Xb	0.446	0.050	0.579	0.056
Variance of hospital-time Xb	0.262	0.053	0.534	0.062
Variance of job effects	0.095	0.012	0.130	0.013
Variance of team match effects	0.056	0.007	0.088	0.008
Split-sample estimates				
Variance of job effects	0.093	0.012	0.127	0.013
Variance of team match effects	0.028	0.003	0.048	0.004

Notes: Basic estimates created from full-sample OLS estimation of 1. Split-sample estimates created from covariances in split-sample OLS estimation of 1, as described in text. Split-sample estimates weighted by combined number of observations in physician-team cell in full sample. Job effects calculated as case-weighted averages of team match effects for each physician-hospital pair. Variance of job effects calculated using deviations from hospital case-weighted average to remove the hospital-level component. As such, the variance of job effects measures how variable physician work paces are within hospitals. Variance of team match effects calculated using deviations from physician-by-hospital case-weighted averages. Because team match effects are nested within jobs, there is no covariance term between job and team match effects. Team match effects limited to physician-team cells with at least 50 associated observations.

Table 11: Team characteristics correlated with estimated LOS team match effects

	(1)	(2)	(3)	(4)	(5)
Peer avg log LOS	0.0871				0.156
	[0.0180]***				[0.0221]***
Peer avg log charges		0.00368			-0.101
		[0.0116]			[0.0209]***
Peer avg grad year			-0.000263		-0.000228
			[0.000156]		[0.000143]
Peer fraction male				-0.01000	-0.00913
				[0.00276]***	[0.00241]***
Physician-team pairs	39,089	39,089	39,089	39,089	39,089

Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Marginal returns estimates: top decile of mortality risk

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	-0.0040	-0.0166	0.0421
	[0.0021]**	[0.0086]**	
Other inputs			
Log charges	0.0556	0.2308	7.2718
	[0.0120]***	[0.0463]***	
CT scan	0.0180	0.0748	0.2235
	[0.0044]***	[0.0179]***	
Chest X-ray	0.0123	0.0509	0.2994
	[0.0056]**	[0.0222]**	
EKG	0.0168	0.0697	0.3890
	[0.0058]***	[0.0234]***	
Ultrasound	0.0014	0.0057	0.0144
	[0.0014]	[0.0057]	
Placebo			
Predicted 30D mort	0.0006	0.0027	0.0547
	[0.0010]	[0.0040]	
First stage			
Log LOS trim 12	0.2410		1.1800
	[0.0217]***		
Cases	1,539,112		

This table presents reduced form (ITT) estimates of the effects of team-induced speed. Each row contains estimates from a separate regression denoted by the outcome in the first column. The instrument is derived from the split-sample estimates of ϕ_g as discussed in the text. All regressions include the baseline set of controls and physician dummies. Standard errors clustered at the hospital level. Predicted mortality and mortality risk comes from Random Forest out-of-bag vote shares (see Appendix C). High mortality risk defined as top decile of out-of-bag vote shares.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Marginal returns estimates: full sample

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	0.0001	0.0003	0.0054
	[0.0002]	[0.0008]	
Other inputs			
Log charges	0.0643	0.2248	6.8850
	[0.0108]***	[0.0386]***	
CT scan	0.0177	0.0620	0.1305
	[0.0028]***	[0.0099]***	
Chest X-ray	0.0117	0.0411	0.1286
	[0.0021]***	[0.0074]***	
EKG	0.0110	0.0385	0.1412
	[0.0024]***	[0.0085]***	
Ultrasound	0.0040	0.0141	0.0315
	[0.0012]***	[0.0040]***	
Placebo			
Predicted 30D mort	0.0001	0.0005	0.0054
	[0.0001]	[0.0004]	
First stage			
Log LOS trim 12	0.2858		0.9771
	[0.0233]***		
Cases	16,783,148		

This table presents reduced form (ITT) estimates of the effects of team-induced speed. Each row contains estimates from a separate regression denoted by the outcome in the first column. The instrument is derived from the split-sample estimates of ϕ_g as discussed in the text. All regressions include the baseline set of controls and physician dummies. Standard errors clustered at the hospital level. Predicted mortality and mortality risk comes from Random Forest out-of-bag vote shares (see Appendix C).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Marginal returns estimates: non-specific complaints in top decile of mortality risk

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	-0.0049	-0.0201	0.0113
	[0.0023]**	[0.0096]**	
Other inputs			
Log charges	0.0475	0.1949	7.5212
	[0.0140]***	[0.0538]***	
CT scan	0.0221	0.0906	0.3058
	[0.0094]***	[0.0371]***	
Chest X-ray	0.0089	0.0366	0.3834
	[0.0096]	[0.0385]	
EKG	0.0288	0.1183	0.5363
	[0.0089]***	[0.0398]***	
Ultrasound	0.0021	0.0086	0.0297
	[0.0038]	[0.0155]	
Placebo			
Predicted 30D mort	-0.0005	-0.0020	0.0305
	[0.0008]	[0.0034]	
First stage			
Log LOS	0.2436		1.3215
	[0.0287]***		
Cases	432,698		

This table presents reduced form (ITT) estimates of the effects of team-induced speed. Each row contains estimates from a separate regression denoted by the outcome in the first column. The instrument is derived from the split-sample estimates of ϕ_g as discussed in the text. All regressions include the baseline set of controls and physician dummies. Standard errors clustered at the hospital level. Predicted mortality and mortality risk comes from Random Forest out-of-bag vote shares (see Appendix C). High mortality risk defined as top decile of out-of-bag vote shares. Non-specific complaints constitute admitting diagnosis classes as listed in Appendix D.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Marginal returns estimates: Fast physicians, high-risk patients

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	-0.0067	-0.0286	0.0411
	[0.0028]***	[0.0125]**	
Other inputs			
Log charges	0.0549	0.2326	7.2368
	[0.0154]***	[0.0680]***	
CT scan	0.0196	0.0832	0.2112
	[0.0061]***	[0.0264]***	
Chest X-ray	0.0081	0.0344	0.2916
	[0.0069]	[0.0278]	
EKG	0.0274	0.1163	0.3810
	[0.0078]***	[0.0349]***	
Ultrasound	0.0020	0.0085	0.0138
	[0.0017]	[0.0077]	
Placebo			
Predicted 30D mort	0.0016	0.0066	0.0543
	[0.0014]	[0.0056]	
First stage			
Log LOS trim 12	0.2359		1.1097
	[0.0307]***		
Cases	757,114		

See notes to previous tables. This table restricts the sample to faster than median physicians in a hospital. Sample limited to high-risk cases.

Table 16: Marginal returns estimates: Slow physicians, high-risk patients

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	-0.0006	-0.0032	0.0431
	[0.0027]	[0.0135]	
Other inputs			
Log charges	0.0422	0.2097	7.3059
	[0.0159]***	[0.0719]***	
CT scan	0.0129	0.0642	0.2354
	[0.0063]**	[0.0316]**	
Chest X-ray	0.0105	0.0519	0.3071
	[0.0082]	[0.0403]*	
EKG	0.0036	0.0178	0.3968
	[0.0081]	[0.0414]	
Ultrasound	-0.0000	-0.0001	0.0150
	[0.0022]	[0.0109]	
Placebo			
Predicted 30D mort	0.0000	0.0002	0.0551
	[0.0014]	[0.0078]	
First stage			
Log LOS trim 12	0.2014		1.2482
	[0.0219]***		
Cases	781,062		

See notes to previous tables. This table restricts the sample to slower than median physicians in a hospital. Sample limited to high-risk cases.

Table 17: Estimates from *cross-physician* design, top decile of mortality risk

Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	0.0030	0.0033	0.0421
	[0.0019]*	[0.0021]*	
Other inputs			
Log charges	0.5086	0.5590	7.2718
	[0.0542]***	[0.0535]***	
CT scan	0.1465	0.1610	0.2235
	[0.0133]***	[0.0137]***	
Chest X-ray	0.1036	0.1138	0.2994
	[0.0132]***	[0.0136]***	
EKG	0.1151	0.1265	0.3889
	[0.0169]***	[0.0175]***	
Ultrasound	0.0099	0.0108	0.0144
	[0.0014]***	[0.0015]***	
Placebo			
Predicted 30D mort	0.0002	0.0002	0.0547
	[0.0009]	[0.0010]	
First stage			
Log LOS trim 12	0.9100		1.1800
	[0.0154]***		
Cases	1,539,693		

This table presents reduced form (ITT) estimates of the effects of physician-induced speed. Each row contains estimates from a separate regression denoted by the outcome in the first column. The instrument is derived from the split-sample estimates of ω_g as discussed in the text. All regressions include the baseline set of controls. Standard errors clustered at the hospital level. Predicted mortality and mortality risk comes from Random Forest out-of-bag vote shares (see Appendix C). High mortality risk defined as top decile of out-of-bag vote shares.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Estimates from *cross-physician* design, full sample

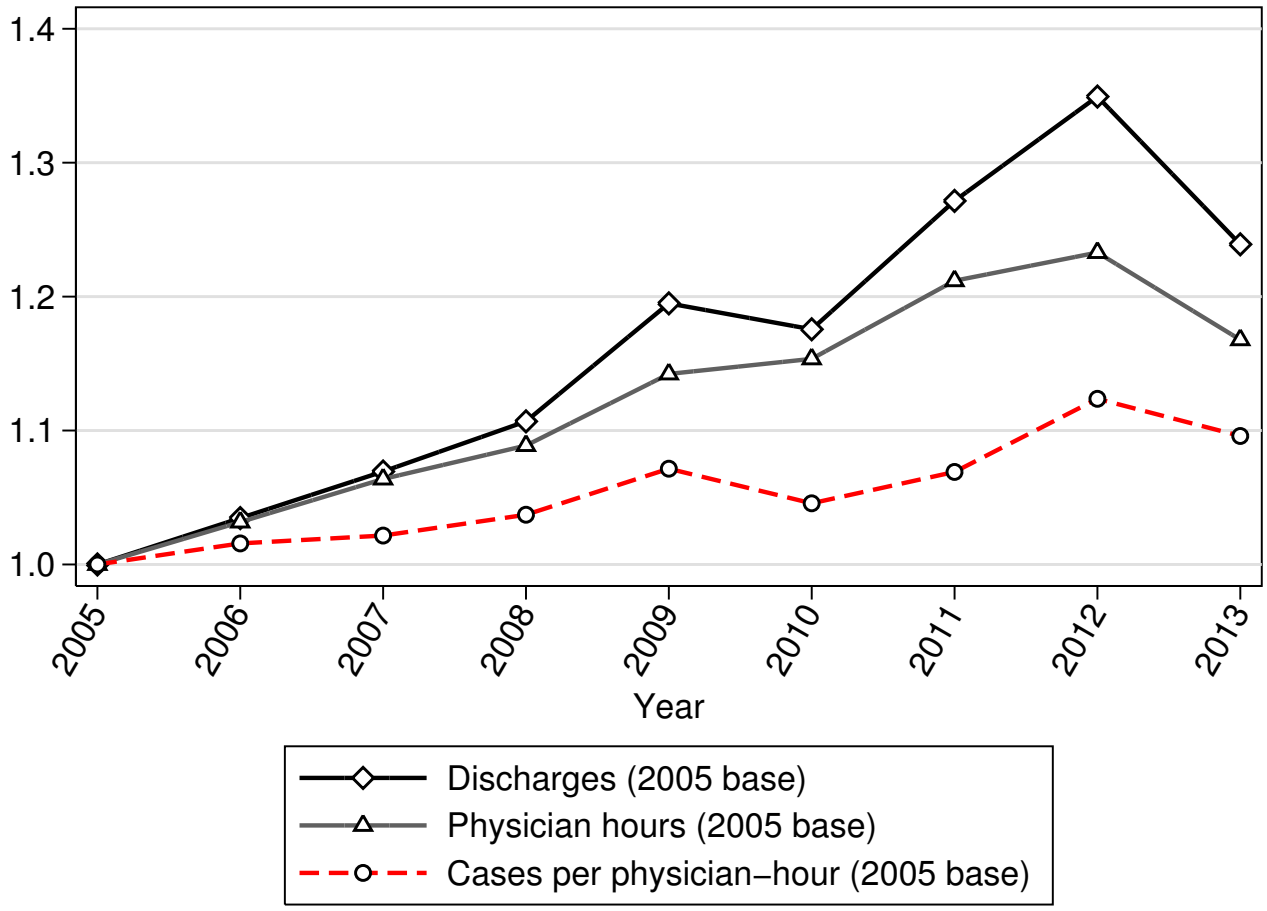
Outcome	(1) Reduced form	(2) 2SLS	(3) Avg outcome
30D mortality	0.0002	0.0002	0.0054
	[0.0003]	[0.0003]	
Other inputs			
Log charges	0.4803	0.5068	6.8849
	[0.0559]***	[0.0573]***	
CT scan	0.1015	0.1071	0.1305
	[0.0094]***	[0.0096]***	
Chest X-ray	0.0770	0.0812	0.1286
	[0.0095]***	[0.0098]***	
EKG	0.0675	0.0712	0.1412
	[0.0092]***	[0.0095]***	
Ultrasound	0.0310	0.0327	0.0315
	[0.0052]***	[0.0054]***	
Placebo			
Predicted 30D mort	0.0002	0.0002	0.0054
	[0.0002]	[0.0002]	
First stage			
Log LOS trim 12	0.9477		0.9771
	[0.0059]***		
Cases	16,790,632		

This table presents reduced form (ITT) estimates of the effects of physician-induced speed. Each row contains estimates from a separate regression denoted by the outcome in the first column. The instrument is derived from the split-sample estimates of ω_g as discussed in the text. All regressions include the baseline set of controls. Standard errors clustered at the hospital level. Predicted mortality and mortality risk comes from Random Forest out-of-bag vote shares (see Appendix C).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

10 Figures

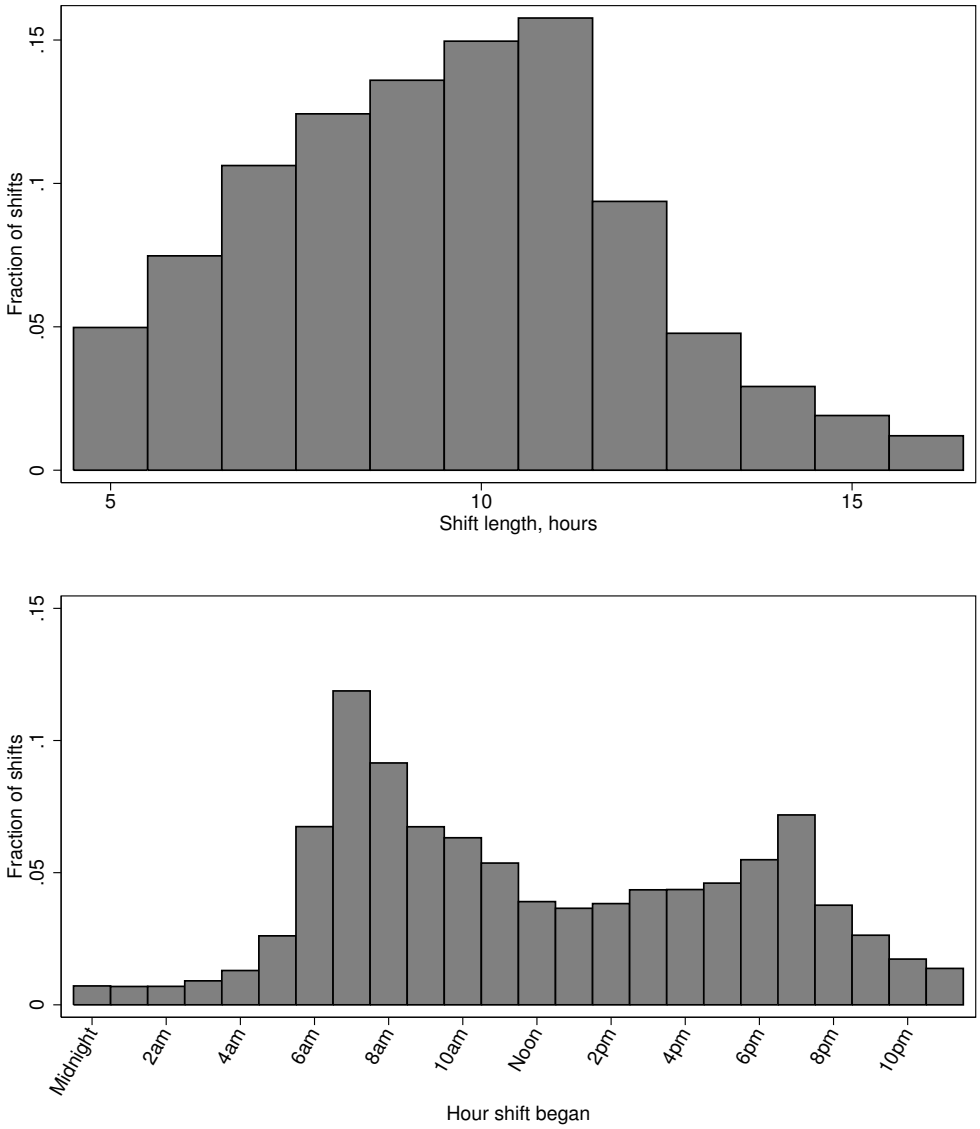
Figure 1: Annual ED volume and average physician cases per hour



Sources: SPARCS ED discharge records and author's calcultions.

Notes: This figure shows the number of emergency department discharges per year alongside the average number of cases per hour physicians discharge in a shift. Only physicians working 100 cases or more in a given year are included in the cases-per-hour calculations.

Figure 2: Shift start times and durations

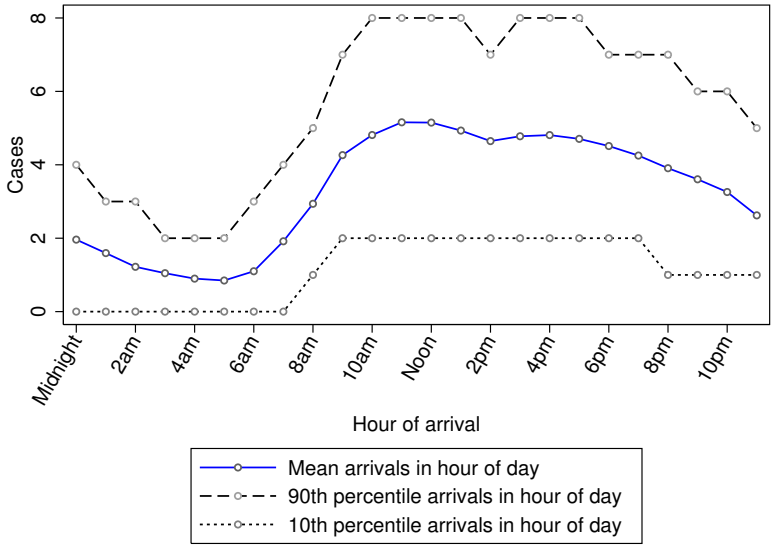


shifts = 1,400,240

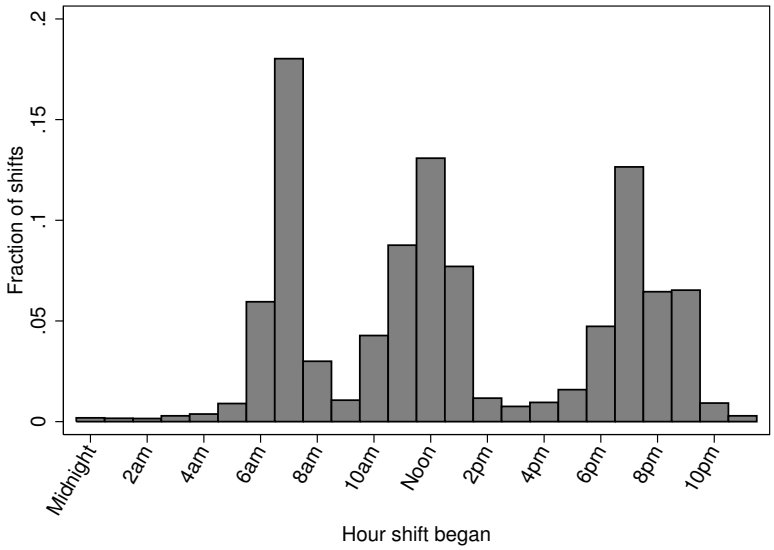
Notes: The top panel of this figure plots the distribution of constructed shift lengths in my sample. The bottom panel plots the distribution across the hours of the day in the starting hour of constructed shifts. Shifts constructed from SPARCS discharge data as detailed in Appendix A.

Figure 3: Arrival flows and shift starting hours at a medium-sized hospital in NY

(a) Arrivals per hour, by hour of day

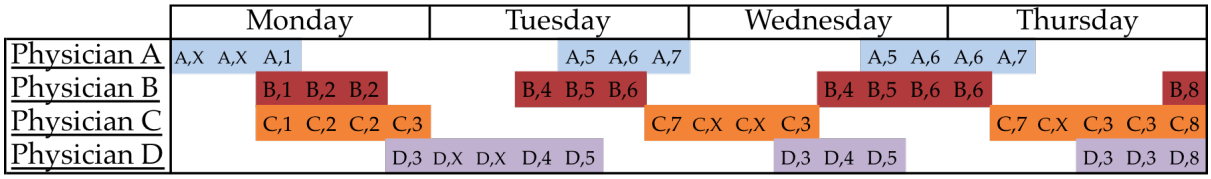


(b) Fraction of shifts beginning, by hour of day



Notes: These figures plot the typical hourly arrival flows (top panel) and the distribution of shift starting hours (bottom panel) for one of the emergency departments in my sample.

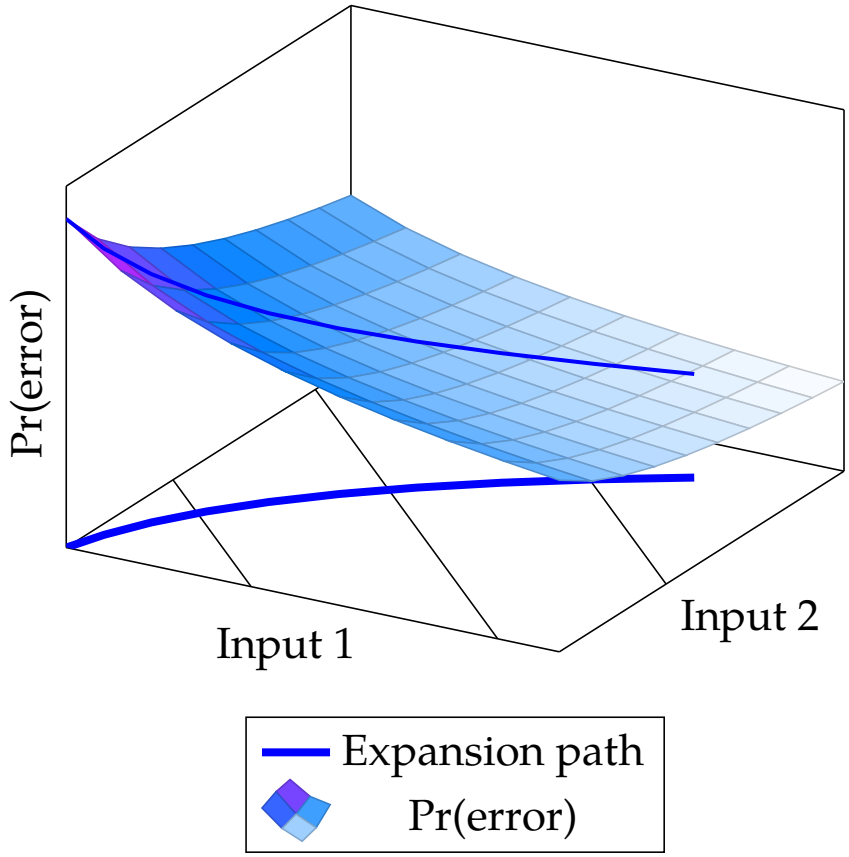
Figure 4: Example scheduling



Group	{A}	{A}	{A,B,C}	{B,C}	{B,C}	{C,D}	{D}	{D}	{B,D}	{A,B,D}	{A,B}	{A,C}	{C}	{C}	{C,D}	{B,D}	{A,B,D}	{A,B}	{A,B}	{A,C}	{C}	{C,D}	{C,D}	{B,C,D}
Team	X	X	1	2	2	3	X	X	4	5	6	7	X	X	3	4	5	6	6	7	X	3	3	8

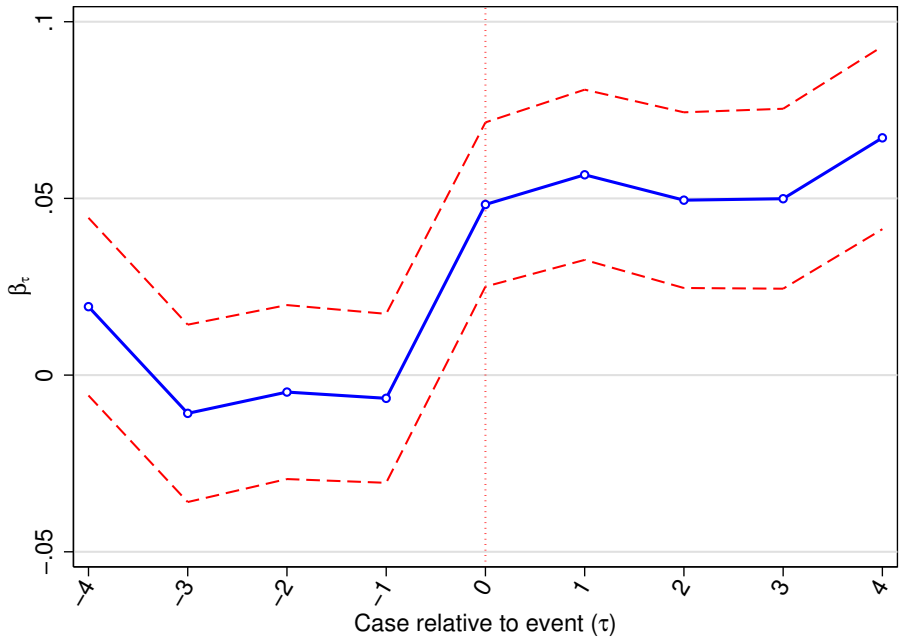
Notes: This figure presents a hypothetical emergency department schedule. Teams are labeled according to their members and the sequence in which they first appear. Teams marked "X" are so labeled because they are singleton teams consisting only of the physician herself. In the data, I observe physicians working with different coworkers groups repeatedly across the sample period. I can thus observe a physician's speed, other inputs, and outcomes across these different team settings. This variation in peer groups for a given physician is the foundation of my empirical analysis of team effects.

Figure 5: Physician expansion paths and patient outcomes



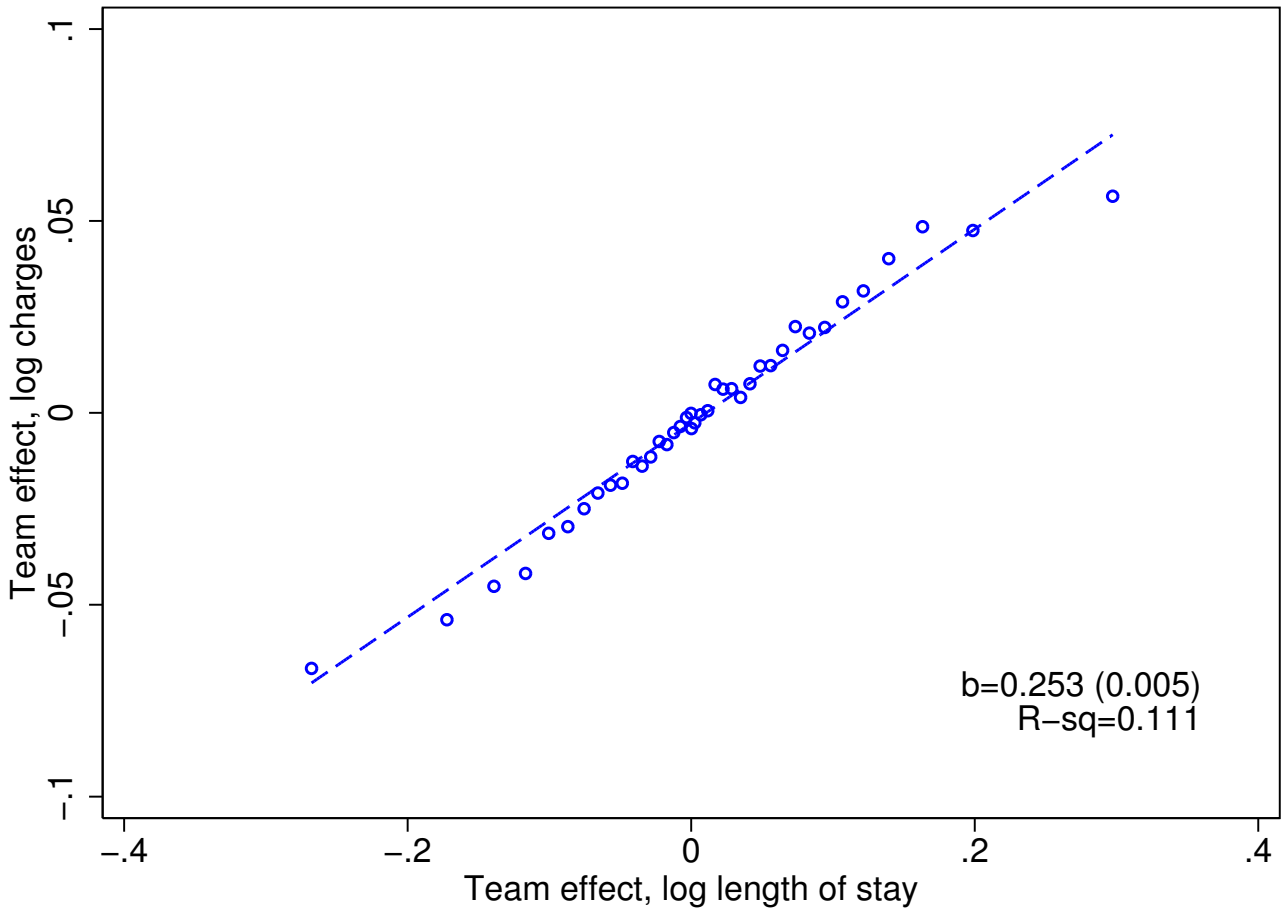
Notes: This figure illustrates the logic of my conceptual framework. Different degrees of peer pressure shift in and out a physician's time budget for each case. Under these differnt budget sets, the physician allocates inputs into care according to her expansion path. Peer-induced differences in case-level inputs are then reflected in the patient's outcome, which in this example is the probability of an error.

Figure 6: Event study of change in team environment: $\Delta\hat{\phi}_g$
Outcome: $\ln LOS$



This figure presents coefficient estimates from OLS estimation of Equation 3: $y_c = \mathbf{X}_c\delta + \sum_{\tau=-4}^{4} \beta_\tau \times \Delta\hat{\phi}_{g(d(c),c+\tau)} + \nu_{s(d(c),c)} + \epsilon_c$. See text for details.

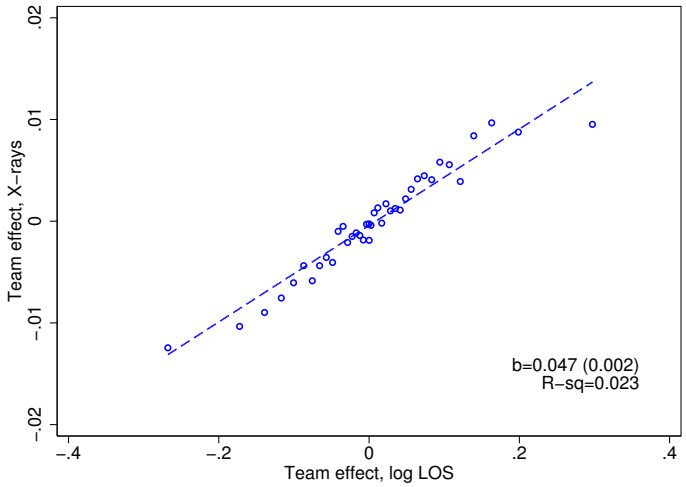
Figure 7: Teams that speed up a physician also cause cutbacks in spending



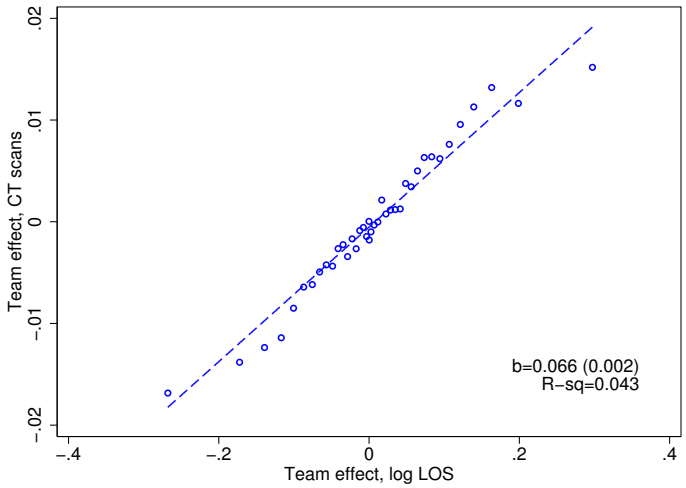
Notes: This binned scatterplot illustrates the relationship between team-induced log length of stay on the x-axis and team-induced log charges on the y-axis. To construct the quantities on each axis, I estimate Equation 1 – $\ln LOS_c = \mathbf{X}_c'\beta^{LOS} + \theta^{LOS}_{d(c)} + \phi^{LOS}_{d(c),g(c)} + \epsilon^{LOS}_c$ – and an auxiliary version of Equation 1 where I replace log LOS with log charges as the dependent variable – $\ln charges_c = \mathbf{X}_c'\beta^{charge} + \theta^{charge}_{d(c)} + \phi^{charge}_{d(c),g(c)} + \epsilon^{charge}_c$. I recover the two team-match components ϕ^{LOS} and ϕ^{charge} and plot them against one another here. Team match effects for each dependent variable are normalized to sum to zero for each physician at the case level, so that coworker groups with whom a physician works at her average pace or spending level (i.e. her physician effect θ_d have $\phi_{d,g} = 0$. The displayed regression coefficient and R^2 are from the bivariate regression at the physician-team level of $\phi^{charge}_{d,g}$ on $\phi^{LOS}_{d,g}$.

Figure 8: Relationships between team match effects in X-ray use, CT scan use, and ln LOS

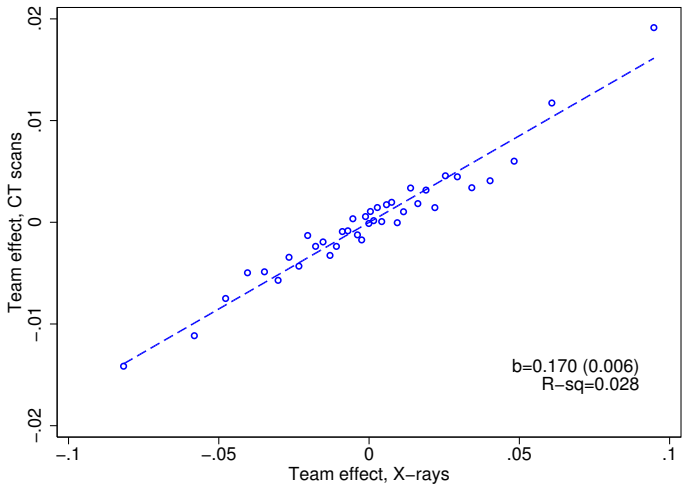
(a) X-rays vs ln LOS



(b) CT scans vs ln LOS

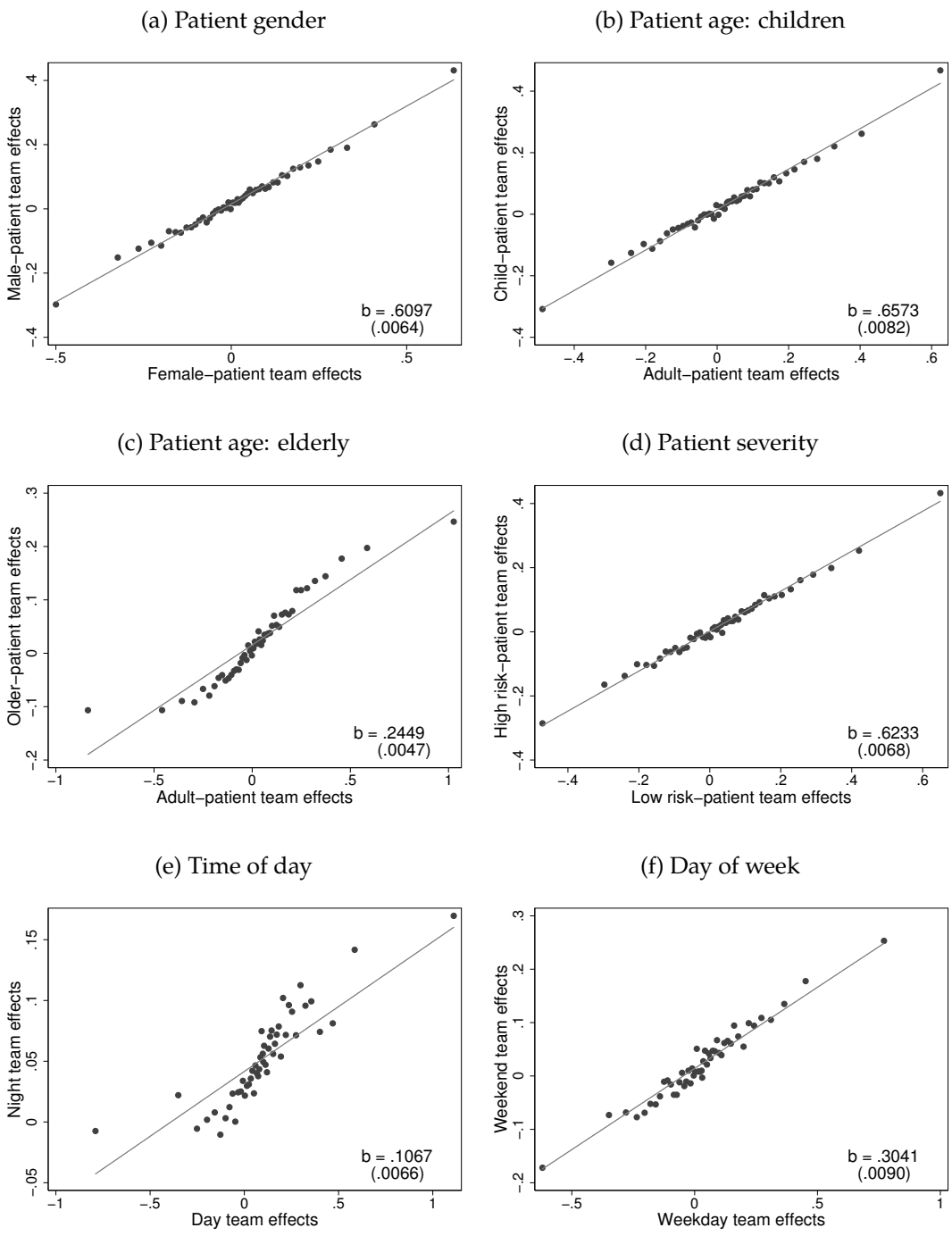


(c) CT scans vs X-rays



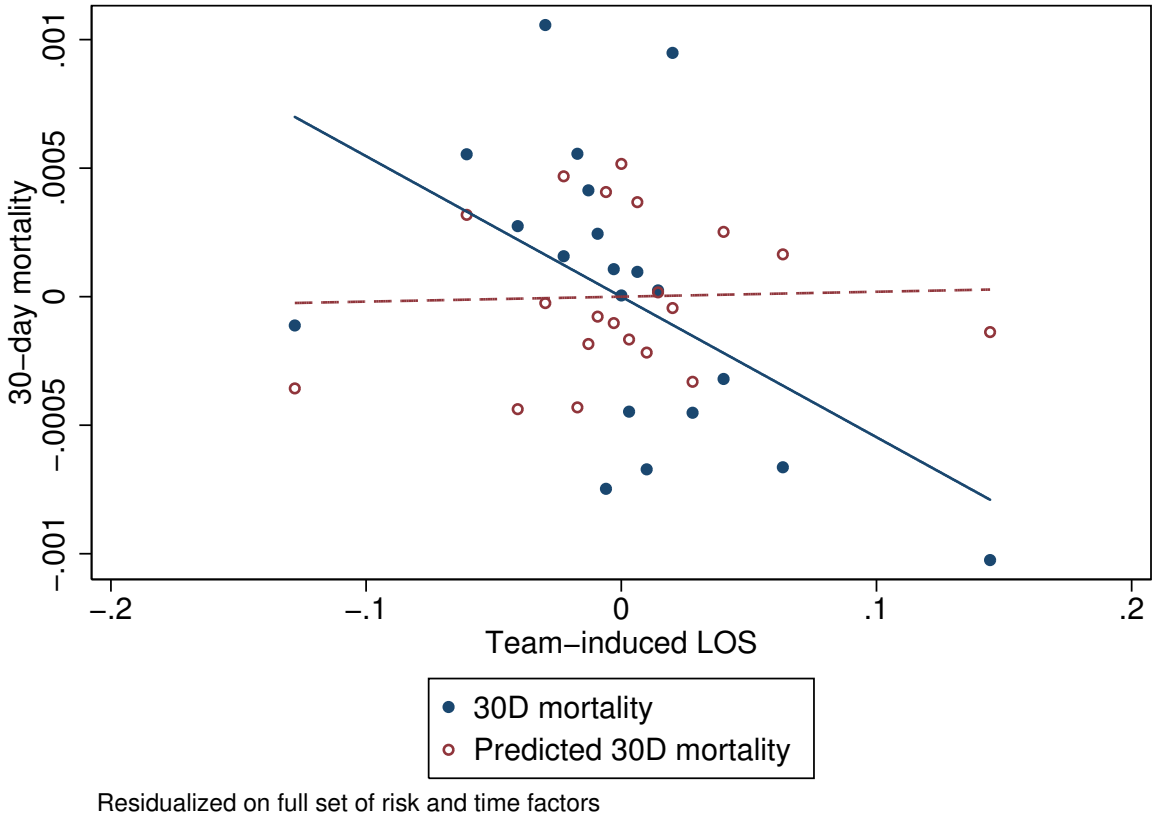
Notes: See notes to previous figure.

Figure 9: Team match effects across subgroups of cases



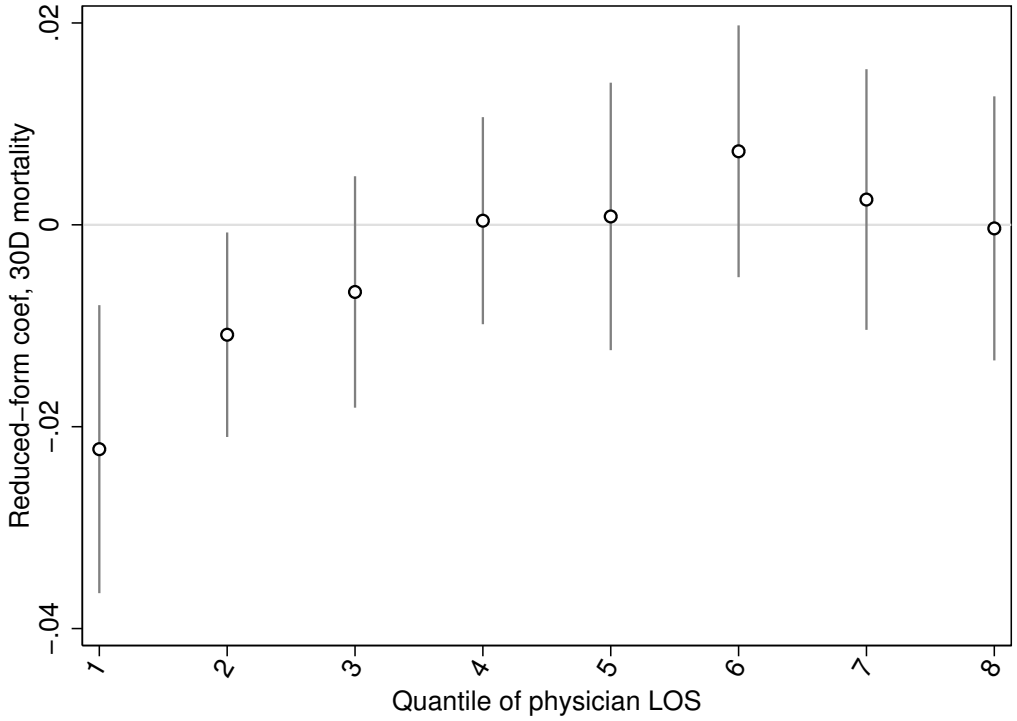
Notes: These figures show the relationships between team match effects in length of stay estimated in separate regressions over mutually exclusive subgroups of discharges. I estimate the match effects model $\ln LOS_c = \mathbf{X}_c'\beta^{LOS} + \theta^{LOS}_{d(c)} + \phi^{LOS}_{d(c),g(c)} + \epsilon^{LOS}_c$ for each subgroup of cases and plot the coinciding match effects $\hat{\phi}_{d,g}$ against one another. Limited to non-singleton teams with at least 100 underlying cases. Figures are weighted by the number of cases used in estimation of the y-axis quantities.

Figure 10: Reduced-form effect of team-induced speed on 30-day mortality



Notes: This figure plots the reduced-form relationship between team-induced log LOS (constructed using the split-sample method described in the text) and 30-day mortality, as well the placebo-test relationship between team-induced log LOS and *predicted* 30-day mortality (predictions generated using Random Forests prediction algorithms, as described in Appendix C). Variables are residualized on the full set of time effects, risk adjusters, and physician effects, as detailed in the text.

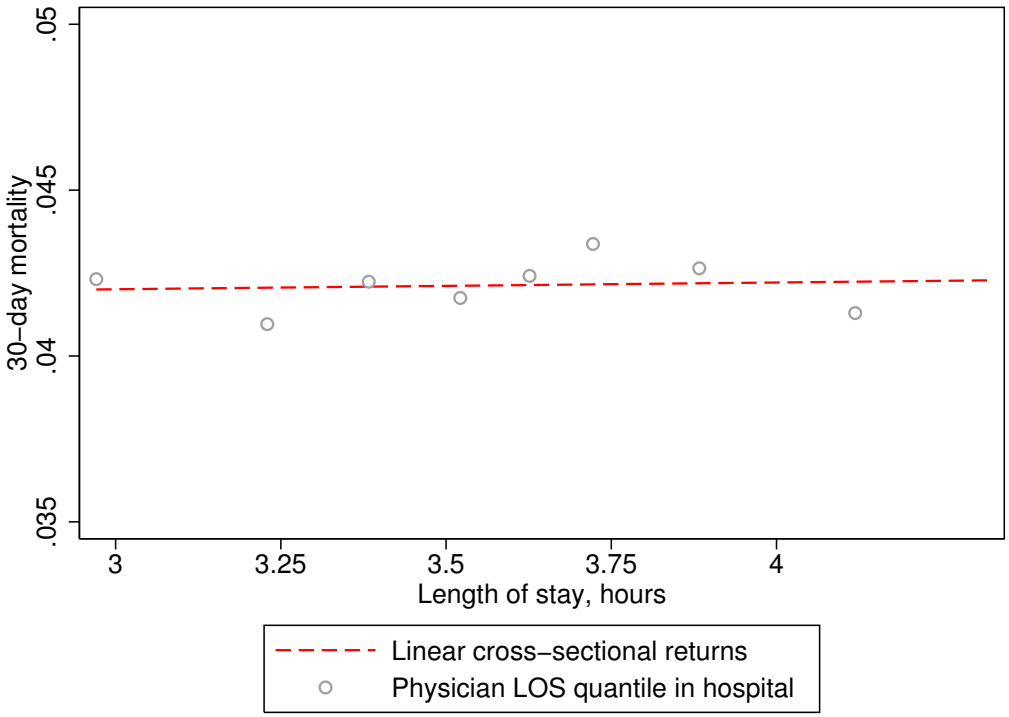
Figure 11: Reduced-form estimates by within-hospital physician speed octile



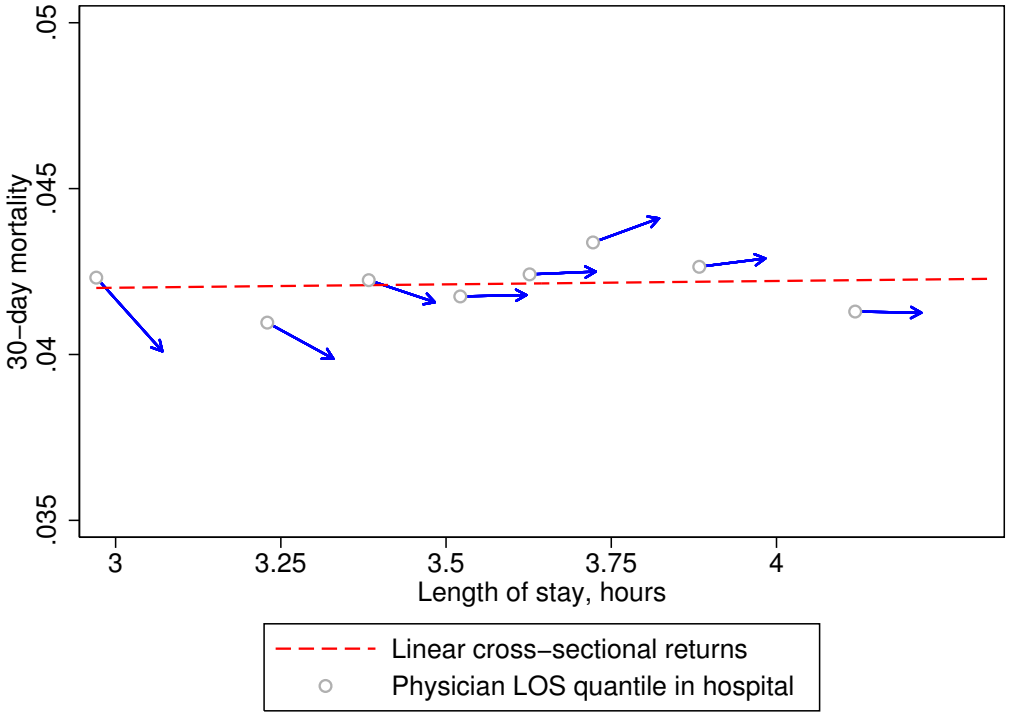
Notes: This figure displays reduced-form coefficients $\hat{\delta}$ from estimation of Equation 7 on separate groups of physicians. I split physicians into within-hospital octiles of the physician-effect distribution for log length of stay. I use the split-sample version of the team match effects, as detailed in the text, as the instrument to which these coefficients correspond. Sample limited to cases in the top decile of the mortality risk distribution for each hospital, as detailed in the text. Vertical lines represent 95% confidence intervals, with standard errors clustered by hospital.

Figure 12: Adjusted 30-day mortality rates across octiles of *physician-level* LOS

(a) Estimated physician mortality by physician speed, at-risk cases

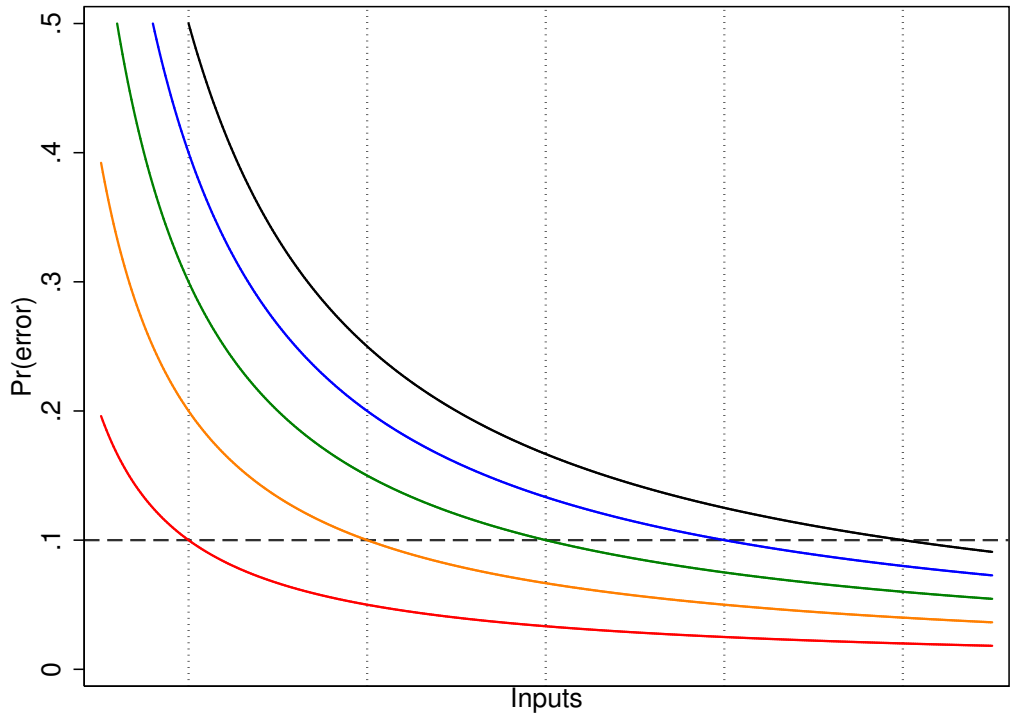


(b) Impose estimated slopes from within-physician designs on each octile



Notes: The top panel of this figure plots adjusted 30-day mortality rates for physicians in each within-hospital octile of physician LOS effects (θ_d), along with the reduced-form relationship between physician-level length of stay and 30-day mortality. The bottom panel imposes the octile-specific *within-physician* reduced-form estimates from Figure 11.

Figure 13: Heterogeneity in physician efficiency
Model of physician input choices



Notes: This figure visualizes a stylized model of physician efficiency differences. Each solid line represents a different physician production function $f(inputs, \alpha_d)$, where the output is the probability of a diagnostic error and α_d is a physician's productivity parameter. If physicians act as problem solvers with similar goals (e.g. represented by the dashed line), they endogenously choose input levels according to their α. The cross-sectional relationship between inputs and outcomes would be flat if all physicians had the same target, while within-physician returns would be high for the most efficient physicians and near 0 for the least efficient.

A Appendix tables and figures

Table A.1: Sensitivity of team match effects to outcome specification and estimation method

	Corr w/ baseline
Full-sample regressions, restricting risk-adjustment	
Team effects, log LOS trimmed at 12 hrs	1.000
Team effects, LOS trimmed at 12 hrs	0.919
Team effects, log LOS trimmed at 8 hrs	0.959
Team effects, LOS trimmed at 8 hrs	0.916
Hospital-level regressions	
Team effects, log LOS trimmed at 12 hrs	0.889
Team effects, LOS trimmed at 12 hrs	0.905
Team effects, log LOS trimmed at 8 hrs	0.848
Team effects, LOS trimmed at 8 hrs	0.901

Source: SPARCS
Notes: This table presents case-weighted correlations between team-match effects estimated with different versions of the LOS outcome.

Table A.2: Variance components of length of stay, using full-sample estimates, with additional controls in X

	(1) LOS trim 8hrs	(2) Log LOS trim 8hrs	(3) LOS trim 12hrs	(4) Log LOS trim 12hrs
Basic estimates				
Total variance	3.572	0.516	5.225	0.563
R-squared, full model	0.375	0.464	0.398	0.473
Variance of patient Xb	0.447	0.050	0.581	0.056
Variance of hospital-time Xb	0.272	0.055	0.545	0.063
Variance of job effects	0.096	0.012	0.131	0.014
Variance of team match effects	0.054	0.007	0.087	0.008
Split-sample estimates				
Variance of job effects	0.094	0.008	0.128	0.009
Variance of team match effects	0.027	0.002	0.046	0.003

Notes: Basic estimates created from full-sample OLS estimation of 1. Split-sample estimates created from covariances in split-sample OLS estimation of 1, as described in text. Split-sample estimates weighted by combined number of observations in physician-team cell in full sample. Job effects calculated as case-weighted averages of team match effects for each physician-hospital pair. Variance of job effects calculated using deviations from hospital case-weighted average to remove the hospital-level component. As such, the variance of job effects measures how variable physician work paces are within hospitals. Variance of team match effects calculated using deviations from physician-by-hospital case-weighted averages. Because team match effects are nested within jobs, there is no covariance term between job and team match effects. Team match effects limited to physician-team cells with at least 50 associated observations.

Table A.3: Correlates of physician LOS and spending

Physician log LOS trimmed at 12hrs	(1)	(2)	(3)	(4)
Male	-0.0398			-0.0385
	[0.00518]***			[0.00562]***
Med school grad year		0.000659		0.000322
		[0.000358]		[0.000372]
Doctor of Osteopathy			0.00313	-0.00181
			[0.00655]	[0.00626]
Physician log charges				
Male	-0.0305			-0.0242
	[0.00511]***			[0.00571]***
Med school grad year		0.00165		0.00137
		[0.000438]***		[0.000463]**
Doctor of Osteopathy			0.0197	0.00959
			[0.00766]*	[0.00752]
Physician-hospitals (jobs)	4,519	4,519	4,519	4,519

This table presents the relationship between observable physician characteristics and estimated physician effects $\hat{\theta}_d$ from Equation 1 for both log length of stay and log charges. Standard errors clustered by hospital. All regressions include hospital dummies. Regressions are weighted by the number of observed cases in each job.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4: Drift in physician effects, unbalanced job sample

(a) LOS trim 8

	2005	2006	2007	2008	2009	2010	2011	2012	2013
2005	1.000								
2006	.801	1.000							
2007	.678	.763	1.000						
2008	.605	.640	.724	1.000					
2009	.607	.644	.705	.756	1.000				
2010	.608	.614	.640	.668	.771	1.000			
2011	.557	.581	.618	.655	.671	.751	1.000		
2012	.521	.541	.584	.601	.636	.700	.750	1.000	
2013	.470	.509	.500	.501	.561	.619	.633	.736	1.000

This table presents the correlations of physician-hospital (job) effects across years for my baseline specification. Job effects are standardized (unweighted) to mean 0, standard deviation 1 for each hospital-year. Jobs weighted by the sum of the number of associated discharges in the years under consideration. Sample uses all job observations regardless of duration (N=5077).

(b) Log charges

	2005	2006	2007	2008	2009	2010	2011	2012	2013
2005	1.000								
2006	.770	1.000							
2007	.679	.768	1.000						
2008	.634	.695	.748	1.000					
2009	.608	.658	.684	.763	1.000				
2010	.553	.599	.614	.714	.806	1.000			
2011	.534	.571	.590	.654	.708	.752	1.000		
2012	.518	.535	.573	.633	.677	.688	.777	1.000	
2013	.448	.518	.535	.600	.652	.650	.698	.791	1.000

This table presents the correlations of physician-hospital (job) effects across years for my baseline specification. Job effects are standardized (unweighted) to mean 0, standard deviation 1 for each hospital-year. Jobs weighted by the sum of the number of associated discharges in the years under consideration. Sample uses all job observations regardless of duration (N=5089).

Table A.5: Autocorrelations and autocovariances of team match effects: log LOS

(a) Autocorrelation matrix

	2005	2006	2007	2008	2009	2010	2011	2012	2013
2005	1.000								
2006	0.149	1.000							
2007	0.102	0.129	1.000						
2008	0.101	0.114	0.146	1.000					
2009	0.080	0.093	0.118	0.156	1.000				
2010	0.091	0.074	0.078	0.092	0.170	1.000			
2011	0.058	0.082	0.089	0.098	0.145	0.133	1.000		
2012	0.073	0.056	0.067	0.082	0.107	0.095	0.151	1.000	
2013	0.059	0.068	0.095	0.067	0.085	0.073	0.113	0.149	1.000

This table presents the correlations of team match effects across years for my baseline specification. Physician-team pairs are limited to those with at least 50 associated cases in the analysis sample. Teams weighted by the sum of the number of associated discharges in the years under consideration. Sample uses all team observations regardless of duration (N=39,805).

(b) Autocovariance matrix

	2005	2006	2007	2008	2009	2010	2011	2012	2013
2005	0.0129								
2006	0.0026	0.0141							
2007	0.0021	0.0024	0.0149						
2008	0.0022	0.0024	0.0028	0.0156					
2009	0.0018	0.0021	0.0025	0.0031	0.0175				
2010	0.0022	0.0018	0.0018	0.0021	0.0038	0.0182			
2011	0.0013	0.0019	0.0021	0.0024	0.0035	0.0028	0.0172		
2012	0.0017	0.0013	0.0016	0.0019	0.0026	0.0021	0.0029	0.0143	
2013	0.0013	0.0016	0.0023	0.0015	0.0021	0.0017	0.0024	0.0026	0.0133

This table presents the covariances of team match effects across years for my baseline specification. Physician-team pairs are limited to those with at least 50 associated cases in the analysis sample. Teams weighted by the sum of the number of associated discharges in the years under consideration. Sample uses all team observations regardless of duration (N=39,805).

Table A.6: Interactions between team characteristics and own characteristics: physician gender

	(1)	(2)	(3)	(4)	(5)
Peer avg LOS $\times\mathbb{I}$(male phys)	0.0799				0.143
	[0.0213]***				[0.0206]***
Peer avg LOS $\times\mathbb{I}$(female phys)	0.0904				0.183
	[0.0255]***				[0.0361]***
Peer avg log charges $\times\mathbb{I}$(male phys)		0.000279			-0.0950
		[0.0131]			[0.0186]***
Peer avg log charges $\times\mathbb{I}$(female phys)		0.00197			-0.126
		[0.0166]			[0.0424]**
Peer avg grad yr $\times\mathbb{I}$(male phys)			-0.000282		-0.000253
			[0.000145]		[0.000135]
Peer avg grad yr $\times\mathbb{I}$(female phys)			-0.000122		-0.0000857
			[0.000264]		[0.000239]
Peer frac male $\times\mathbb{I}$(male phys)				-0.0107	-0.00960
				[0.00260]***	[0.00241]***
Peer frac male $\times\mathbb{I}$(female phys)				-0.00990	-0.0100
				[0.00475]*	[0.00359]**
Physician-team pairs	39,089	39,089	39,089	39,089	39,089

Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.7: Interactions between team characteristics and own characteristics: physician age

	(1)	(2)	(3)	(4)	(5)
Peer avg LOS $\times\mathbb{I}$(young phys)	0.0714				0.128
	[0.0268]**				[0.0410]**
Peer avg LOS $\times\mathbb{I}$(older phys)	0.0930				0.165
	[0.0191]***				[0.0206]***
Peer avg log charges $\times\mathbb{I}$(young phys)		0.0107			-0.0760
		[0.0170]			[0.0295]*
Peer avg log charges $\times\mathbb{I}$(older phys)		0.000824			-0.109
		[0.0136]			[0.0225]***
Peer avg grad yr $\times\mathbb{I}$(young phys)			-0.000359		-0.000379
			[0.000195]		[0.000183]*
Peer avg grad yr $\times\mathbb{I}$(older phys)			-0.000234		-0.000182
			[0.000171]		[0.000152]
Peer frac male $\times\mathbb{I}$(young phys)				-0.00902	-0.00844
				[0.00393]*	[0.00396]*
Peer frac male $\times\mathbb{I}$(older phys)				-0.0104	-0.00951
				[0.00309]**	[0.00273]***
Physician-team pairs	39,089	39,089	39,089	39,089	39,089

Young physicians defined as those graduating medical school in 2000 and beyond Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.8: Interactions between team characteristics and own characteristics: young male physicians

	(1)	(2)	(3)	(4)	(5)	(6)
Peer avg LOS $\times\mathbb{I}$(young male phys)	0.0547					0.0819
	[0.0274]*					[0.0389]*
Peer avg LOS $\times\mathbb{I}$(not young male phys)	0.0931					0.168
	[0.0173]***					[0.0206]***
Peer avg log charges $\times\mathbb{I}$(young male phys)		0.0153				-0.0408
		[0.0175]				[0.0277]
Peer avg log charges $\times\mathbb{I}$(not young male phys)		0.00127				-0.111
		[0.0123]				[0.0223]***
Peer avg grad yr $\times\mathbb{I}$(young male phys)			-0.000197			-0.000265
			[0.000216]			[0.000210]
Peer avg grad yr $\times\mathbb{I}$(not young male phys)			-0.000275			-0.000222
			[0.000165]			[0.000149]
Peer frac male $\times\mathbb{I}$(young male phys)				-0.0112		-0.0101
				[0.00495]*		[0.00487]*
Peer frac male $\times\mathbb{I}$(not young male phys)				-0.00977		-0.00914
				[0.00295]**		[0.00255]***
$\mathbb{I}$(young male in team) $\times\mathbb{I}$(young male phys)					-0.00339	
					[0.00373]	
$\mathbb{I}$(young male in team) $\times\mathbb{I}$(not young male phys)					-0.00498	
					[0.00291]	
Physician-team pairs	39,089	39,089	39,089	39,089	39,089	39,089

Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.9: Interactions between team characteristics and own characteristics: physician speed

	(1)	(2)	(3)	(4)	(5)
Peer avg LOS $\times\mathbb{I}$(slow phys)	0.0952				0.159
	[0.0200]***				[0.0230]***
Peer avg LOS $\times\mathbb{I}$(fast phys)	0.0797				0.153
	[0.0205]***				[0.0255]***
Peer avg log charges $\times\mathbb{I}$(slow phys)		0.00611			-0.0982
		[0.0119]			[0.0190]***
Peer avg log charges $\times\mathbb{I}$(fast phys)		0.00145			-0.104
		[0.0146]			[0.0259]***
Peer avg grad yr $\times\mathbb{I}$(slow phys)			-0.000197		-0.000150
			[0.000179]		[0.000164]
Peer avg grad yr $\times\mathbb{I}$(fast phys)			-0.000327		-0.000303
			[0.000178]		[0.000171]
Peer frac male $\times\mathbb{I}$(slow phys)				-0.0114	-0.0103
				[0.00330]***	[0.00315]**
Peer frac male $\times\mathbb{I}$(fast phys)				-0.00875	-0.00817
				[0.00299]**	[0.00254]**
Physician-team pairs	39,089	39,089	39,089	39,089	39,089

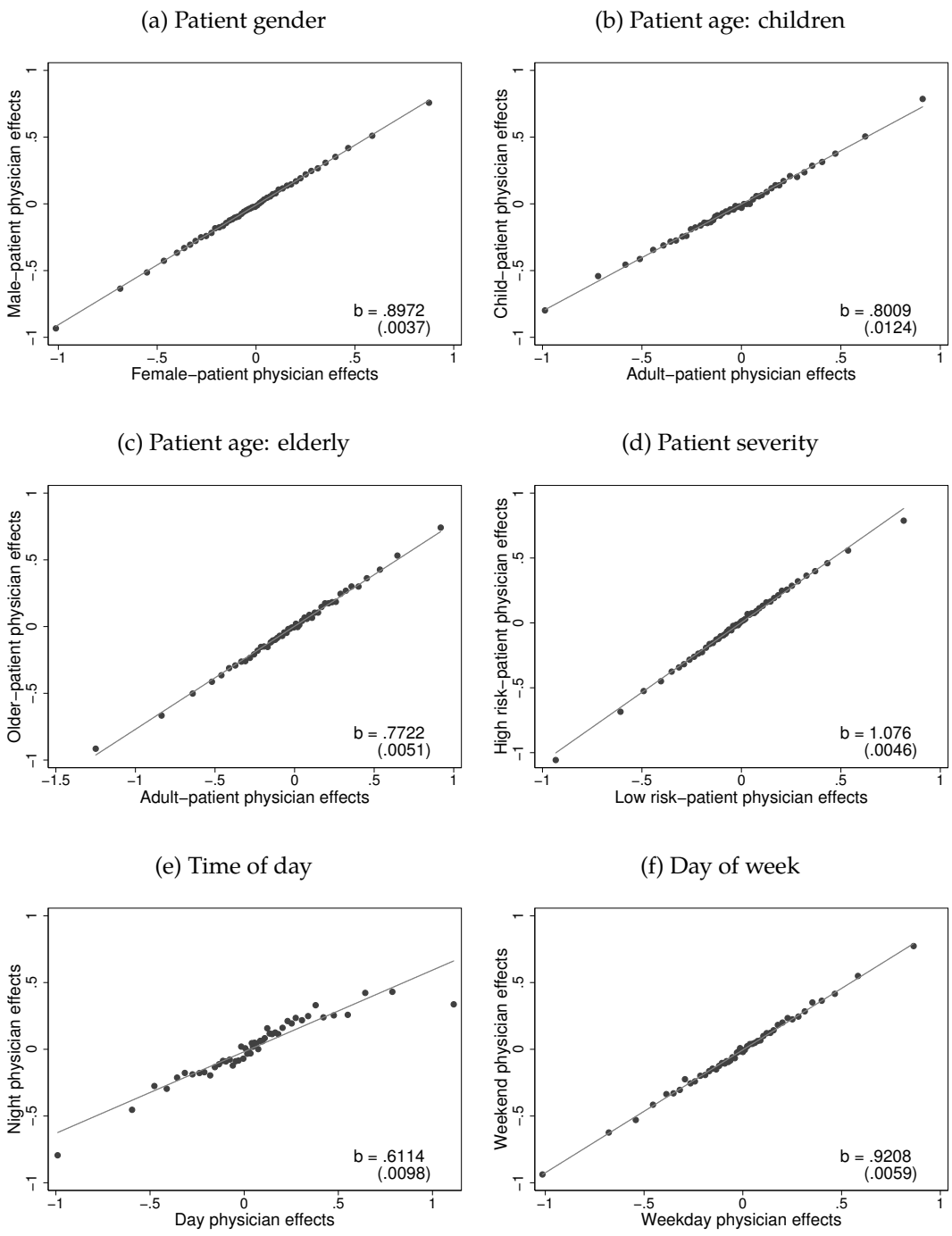
Slow (fast) physicians are those above (below) median in the fixed effects distribution for LOS within a hospital. Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.10: Interactions between team characteristics and own characteristics: physician spending

	(1)	(2)	(3)	(4)	(5)
Peer avg LOS $\times\mathbb{I}$(high-spending phys)	0.0790				0.144
	[0.0179]***				[0.0247]***
Peer avg LOS $\times\mathbb{I}$(low-spending phys)	0.0992				0.173
	[0.0249]***				[0.0270]***
Peer avg log charges $\times\mathbb{I}$(high-spending phys)		0.00339			-0.0924
		[0.0111]			[0.0229]***
Peer avg log charges $\times\mathbb{I}$(low-spending phys)		0.00417			-0.113
		[0.0167]			[0.0247]***
Peer avg grad yr $\times\mathbb{I}$(high-spending phys)			-0.000297		-0.000252
			[0.000175]		[0.000160]
Peer avg grad yr $\times\mathbb{I}$(low-spending phys)			-0.000218		-0.000198
			[0.000163]		[0.000152]
Peer frac male $\times\mathbb{I}$(high-spending phys)				-0.0108	-0.0101
				[0.00300]***	[0.00268]***
Peer frac male $\times\mathbb{I}$(low-spending phys)				-0.00898	-0.00788
				[0.00309]**	[0.00281]**
Physician-team pairs	39,089	39,089	39,089	39,089	39,089

High-spending (low-spending) physicians are those above (below) median in the fixed effects distribution for log charges within a hospital. Standard errors clustered by hospital. All regressions include physician-hospital (job) dummies. Dependent variable is ordered team effect for Log LOS trimmed at 12hrs from estimation of Equation 1 on full sample. All regressions limited to physician-team pairs with no fewer than 50 underlying cases. Regressions are weighted by underlying cell size of physician-team pair.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1: Physician effects across subgroups of cases



Notes: These figures show the relationships between physician effects in length of stay estimated in separate regressions over mutually exclusive subgroups of discharges. I estimate the match effects model $\ln LOS_c = \mathbf{X}_c'\beta^{LOS} + \theta^{LOS}_{d(c)} + \phi^{LOS}_{d(c),g(c)} + \epsilon^{LOS}_c$ for each subgroup of cases and plot the coinciding physician effects $\hat{\theta}_d$ against one another. Figures are weighted by the number of cases used in estimation of the y-axis quantities.

Appendices

A Creating physician schedules and teams data elements

In this appendix I describe how I create physician schedules and define teams of physicians as used in my analysis. The primary inputs for the creation of schedules and teams are the dates and hours of arrival and discharge, the physician license number, and the hospital identifier.

I first create an auxiliary dataset that contains two rows for each discharge – one for the admission event and one for the discharge event. Each row has a case identifier, a physician license number, a hospital identifier, an indicator for which type of event it is (arrival or discharge), and the date and hour of the event. I sort this dataset by hospital, physician, a continuous hour measure generated from the combination of the date and hour variables, and finally by event type. I then create a variable that counts the number of active cases for each hospital-physician pair in a given hour by sequentially summing up the number of arrivals minus the number of discharges for that hospital-physician pair. This procedure leaves me with a count of the number of active cases a physician has in a hospital at the beginning of each hour in the sample.[46]

After calculating caseloads at the hospital-physician-hour level, I turn to coding up whether a physician is on duty in that hospital-hour. This boils down to whether a physician has any active cases in the hospital-hour, with one small exception. During lower-volume days or hours, physicians occasionally may be on duty without any active cases. I avoid coding them as off duty during these hours by asking whether they have had a non-zero caseload in either of the two hours preceding or following the reference hour in this hospital. If a physician has been active in the preceding and the following hours, then I interpolate that the physician is also on duty in the reference hour.

Finally, to generate teams of physicians on duty in a hospital-hour, I merge together all the individual physician schedules at the hospital-hour level to generate the list of on-duty physicians for every hospital-hour. In the analysis, I limit the teams to include only physicians who see sufficiently many cases in a given hospital-month – whom I call *roster physicians*, as there are some infrequent license numbers that come up

[46] Two issues come up in this procedure. First, hour of discharge is not reported in a small fraction of my in-sample cases. In this case, I assume the patient is discharged 3 hours after they are admitted for the purpose of getting caseload counts for each physician. I do not use these cases in my throughput analysis. Second, there are a handful of cases for which throughput is greater than the typical shift length. To avoid counting a physician as active for the full time of these patients' stays, I cap the time a physician can be with a given patient by recoding the hour of discharge for these cases to be the hour of arrival plus 10.

that are likely miscoded. I do not assign the cases attached to these infrequent license numbers to any of the roster physicians. Eliminating these licenses from the teams helps to reduce the dimensionality of the set of teams, aiding the identification of team effects.

B Sample selection

Here I describe my primary criteria for sample selection. The most stringent criterion pertains to missing timestamps. I drop any hospital months where discharge hour was unreported. This was common in the early years of data collection, and reporting remains a problem for some hospitals. This restriction drops all the discharges from 28 hospitals in the sample. I further drop all hospitals in my sample where modal team sizes across discharges are between 2 and 4. For example, I drop all of the very small emergency departments that primarily use "single coverage" – one physician on duty at a time – since in these hospitals, I cannot identify peer effects. On the other hand, as modal team sizes grow, the number of observations per team decreases rapidly, so that direct estimation of team effects is hopeless.

Within hospitals, I further restrict attention to cases arriving in hours when the number of physicians on duty is within one of the modal team size in each hospital to focus on "normal" times.

I keep only physicians with licenses who are observed working over 1000 cases in the original sample. There are a multitude of infrequent or unverifiable license numbers that appear, but these physicians account for a small portion of all cases and are concentrated in a small number of hospitals. I also drop any physician-hospital combination (jobs) with fewer than 500 associated cases.

I also restrict attention to cases cared for during physician-shifts meeting some basic criteria, namely that the inferred shift length falls between 5 and 16 hours. This helps in the construction of stable teams of physicians.

In combination these restrictions leave me with 137 hospitals, with on average of 96 months (of a possible 108 months) of discharge data for each hospital. The final sample includes 3,445 physicians working on 5,089 jobs over 1.4 million shifts and 19.3 million total discharges.

Table B.1: Sample selection

	Hospitals	Hosp-months	Physicians	Jobs	Shifts	Cases
Full	248	22,104	51,382	77,272	5,416,599	55,937,031
Restrict dates	247	21,991	51,194	76,959	5,366,536	55,374,040
Restrict hospitals	247	21,991	51,194	76,959	5,366,536	55,374,040
Hosp-months reporting hour discharged	219	18,469	43,644	62,622	3,785,946	41,862,864
Restrict jobs	192	17,963	4,055	6,376	2,514,324	35,154,533
Restrict shifts	192	17,827	4,050	6,363	1,882,099	29,646,176
Hospital team-size criteria	137	13,144	3,448	5,092	1,532,544	25,212,492
Normal hosp team size in hour	137	13,139	3,445	5,089	1,400,240	19,328,124

C Random Forest prediction models

In this appendix I describe the construction of the readmission and mortality predictions I use in the paper. I assess the risk of a given case for 30-day mortality based on a set of predetermined characteristics of the case, X_c. These characteristics include patient age, gender, race/ethnicity, and complaint on arrival (3-digit ICD-9).

I draw on methods common to the statistical/machine learning literature to make predictions about each case's likelihood of mortality. Logistic regression methods provide the standard alternative and are useful in many settings. However, these methods leave many degrees of freedom for the researcher to choose which interaction terms or nonlinearities to include. Including too many interactions and saturating the model exacerbates overfitting and leads to poor out-of-sample predictions. On the other hand, manually searching for a parsimonious model is expensive, unless the set of possible interactions is sufficiently small. Previous work (Doyle et al. (2010)) has relied on stratifying patients' risk based solely on conditional probabilities of mortality by 3-digit ICD-9 code. For some types of care, knowledge of the ICD-9 is likely sufficient, but there is substantial gain in out-of-sample predictive power in my setting from allowing for rich interactions between ICD-9 and other patient characteristics.

Given the diverse nature of cases treated in the emergency department, the covariates that map into mortality outcomes likely have complex, difficult-to-specify nonlinearities and interactions. This makes the emergency department a prime candidate environment for the use of statistical learning techniques for propensity score estimation.

In this paper, I use tools from the statistical learning literature to predict propensity scores of cases based on a limited set of predetermined observables. Treating age as a categorical binned variable with 20 bins, the number of possible linear terms and single-interaction terms simply between age and a categorical

3-digit complaint variable would lead to a model with over 10,000 parameters. Many of these categories provide little signal and induce estimation noise. Choosing the correct subset of these interaction terms to consider is no simple task. The random forest algorithm is a tree-based ensemble method with desirable properties for these kinds of model selection problems. For a full description of the algorithm, see Breiman (2001).

In short, I fit a random forest of 800 trees separately to each hospital's discharges, using the out-of-bag vote share as the predicted risk score for a given case.[47] The out-of-bag vote share provides out-of-sample predictions by using the set of trees that do not contain a given observation – for each tree, the random forest algorithm selects a bootstrapped subsample of the observations, so that about one-third of the observations are "out-of-bag" for a given tree. The out-of-bag vote share is constructed by running each observation through all of the constructed trees for which it is out-of-bag, collecting the predictions (0 or 1 for a binary classification problem), and taking the average prediction for each observation.

[47] This is a conservative choice for the number of trees, based on a number of cross-validation exercises. It is worth noting that the only cost of adding more trees is computational. Random forests do not overfit as the number of trees per forest increases, so in practice it is recommended to use sufficiently many trees per forest so that the error rate has stabilized.

D Non-specific complaints

I classify cases as having non-specific complaints in the body of the text if their complaint on arrival is in one of the following ICD-9 diagnosis categories as defined by HCUP's Clinical Classifications Software: Abdominal pain (CCS category 251), non-specific chest pain (102), headache/migraine (84), fever of unknown origin (246), nausea and vomiting (250), conditions associated with dizziness or vertigo (93), other ear and sense disorders (94). 84, 250, 93, and 94, while very common, tend to have been present for patients readmitting with strokes and mini-strokes. 251 and 102 are also very common and typically benign, but on rare occasion are precursors of heart attacks or other cardiac events (Pope et al. (2000); Kachalia et al. (2007); Newman-Toker et al. (2014); Wilson et al. (2014)).

E Stability of physician and team effects over time

E.1 Drift in physician effects

To what extent are physicians' practice styles and quality of care fixed over time? In this section, I estimate the degree of drift in my physician-level measures over the sample period. Table A.4provides estimated autocorrelations of physician effects for log length of stay and log charges.

There are a few important features to point out. First, these effects are quite stable over time, displaying first-order autocorrelations upwards of 0.7, up to 0.8. Similar first-order autocorrelation measures in the teacher value added literature (Chetty et al. (2014); Rothstein (2014)) range from 0.2 to 0.56, depending on the subject and grade level.[48] Second, there is evidence of drift in the physician effects for both throughput and log charges, as the autocorrelations grow weaker at longer lags. This suggests that physician practice style evolves over time. In line with this finding, I document faster evolution (greater declines in the autocorrelation vector at longer lags) for younger physicians than for older physicians, consistent with the idea that older physicians are more set in their ways. The degree of practice-style evolution has received theoretical interest, but has been met with little empirical evidence (see Phelps and Mooney (1993); Epstein and Nicholson (2009); Molitor (2011)).

[48] See, e.g., Table 2 of Chetty et al. (2014) and Appendix Table 1 of Rothstein (2014).

E.2 Drift in team effects

Teamwork is likely to evolve over time, as physicians develop relationships (friendship, animosity, cooperation) with one another and learn to work together. Teams could be subject to both slow evolution and sharp changes in their functioning.[49] For these reasons, we may expect that the component of productivity or practice style due to the team would drift more than the physician component.

Team effects, especially when estimated separately across multiple periods, are more prone to measurement error than are physician effects, simply because I do not observe the same team nearly as frequently as I observe the same physician. Due to this higher degree of measurement error, period-to-period team effects are less highly correlated than physician effects.

Nonetheless, team effects do exhibit substantial autocorrelation, as documented in Table A.5, suggesting that relationships between coworkers are somewhat stable, and that my measures of team effects are not merely picking up contemporaneous correlated shocks to the team members.

[49] See **?** and Bandiera et al. (2006) for examples in the fruit-picking industry of the evolution and sharp changes in teamwork among socially connected coworkers.