Title

Essays in Hospital Organization, Infrastructure, and Productivity

Author

Chu, Bryan Paul

Publication Date

2024

Peer reviewed|Thesis/dissertation

Essays in Hospital Organization, Infrastructure, and Productivity

by

Bryan Chu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin Handel, Chair
Professor Jonathan Kolstad
Professor Nano Barahona

Spring 2024

Essays in Hospital Organization, Infrastructure, and Productivity

Abstract

Essays in Hospital Organization, Infrastructure, and Productivity

by

Bryan Chu

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Benjamin Handel, Chair

Nonprofit teaching hospitals contribute almost half of Health Care and Social Assistance GDP and educate more than 90% of all future physicians. Training physicians involves an important trade-off between the short-term delivery of health services and the long-term benefits of physician training. I leverage unusually detailed electronic health record and audit log data from the emergency department of a large, urban teaching hospital to characterize the static costs of training across a range of granular patient outcomes and process measures. Using panel variation in patient assignment to residents, I find that hospitals must extend length of stay for complex patients by 1% to make a resident 0.053% faster in the future. Over the four-year program, this accrues to a reduction of about 10.3% and implies faster patient throughput.

I develop and estimate a dynamic model of physician training and care quality to understand how the emergency department of an academic hospital trades off costs today with the future benefits of more intense physician training. Results inform the policy discourse aimed at improving healthcare efficiency and extend existing models of nonprofit hospitals to account for the teaching objective. I find that commonly-discussed payment reforms for insurers to reduce costs may increase the shadow cost of training. This could have negative effects on the career outcomes of graduating physicians over four times larger than the savings for the teaching hospital, but feasible remedies such as increasing the staffing of attending physicians by 5% lessens the penalty by 65%.

Medicine has a reputation of being a gender-egalitarian profession, but there is also evidence of persistent differences in hours worked as well as procedures and tasks performed. We investigate gender differences in the intensive margin in detail by leveraging a unique dataset that contains granular information based on the Electronic Medical Records and Audit Log at a large teaching hospital. Our primary analysis sample contains 1,620 physicians, of which about 47% are women. In this highly standardized environment, we find that even after controlling for a detailed set of physician attributes, women spend about 10% more

time on notes per shift than men. Next, we show that patients quasi-randomly assigned to female physicians upon inpatient hospital admission receive 7.6% fewer orders without any declines in quality of care (readmissions or days in the hospital). Analysis of note text reveals that women include 23% more clinical concepts in their notes. Despite meaningful improvements in clinical efficiency caused by additional note writing effort, physician salary and other measures of career advancement are not correlated with this value-adding task.

To Ruby Chen

For your inspiration and unwavering love and support through this entire process and beyond.

Contents

List of Figures

List of Tables

Acknowledgments

I thank Ben Handel, Jon Kolstad, and Nano Barahona for their continued guidance and support. I also thank Robert Thombley, Julia Adler-Milstein, Albert Lee, and the UCSF team for their work in assembling, deriving, and helping me interpret the data, as well as Jaskirat Dhanoa, Jeffrey Hollander, Ryan Lichtarge, and Katrina Stime for discussion of institutional features. I am also grateful to Matt Backus, Savannah Bergquist, Zarek Brot-Goldberg, Kaveh Danesh, Jonathan Holmes, Yuki Ito, Yikun Jiang, Benjamin Lacar, Carolyn Stein, and Audrey Tiew, as well as seminar participants at Simon Business School, Stanford SITE Gender, and UC Berkeley for their valuable comments. For Chapter 3, I am grateful for my coauthors Ben Handel, Jon Kolstad, and Ulrike Malmendier, and we thank Nikki Azerang, Margaret Kallus, and James Yixuan Zhang for providing outstanding research assistance.

Chapter 1

Training in Nonprofit Emergency Departments

1.1 Introduction

Healthcare spending is far higher in the United States than in other developed countries, yet Americans experience worse health outcomes.[1] This is despite a large fraction of care–45% of all Health Care and Social Assistance GDP in 2019–flowing through academic teaching hospitals, widely regarded as the best hospitals in the world.[2] It is not obvious that the institutions that are the best at producing healthcare should also the best at training new physicians. This combination is particularly interesting in academic medicine, where "learning by doing" during residency training is paramount. While inexperienced physicians must see patients in order to learn, they may also achieve worse patient outcomes compared to fully-trained physicians due to their inexperience.

I examine the static costs of allocating patients of varying complexity to residents (physician trainees) of varying experience and attending physicians (teaching faculty who also work independently). I leverage detailed electronic health record and audit log data, I characterize the static costs of training: how narrowly-defined components of patient care improve as residents gain experience. I focus on emergency medicine (EM) residents at the University of California, San Francisco (UCSF). The UCSF EM Residency's day-to-day operations are typical of EM Residency programs. Most patients are seen by a single resident, the trainee, who is supervised by an attending physician, a faculty member. The remaining patients are seen by attendings working independently. Residents choose patients with assistance and guidance from attendings. Via their patient allocation decisions, attending physicians execute the hospital's desired trade-off between training and care quality.

The granularity of my data allow me to examine resident learning in great detail. I observe resident and attending identifiers for each distinct, disaggregated action, which allow me to attribute not only patient-level outcomes and decisions but also the dozens of individual decisions and actions for each patient to specific physicians. Timestamps for each action are unmasked, which allow me to not only correctly order patients and actions during each resident's history of work, but also to examine how time duration to important actions evolves with experience. This combination of granularity in actions, physician identifiers, and unmasked timestamps is rare even in health data, much less data from other industries.[3]

I begin by characterizing the ways in which residents learn by doing over the course of the four-year residency program. I find that residents become much more productive in terms of total patients seen per shift. By managing additional patients simultaneously, they go from seeing three patients per eight-hour shift when they enter the program to seeing almost eight patients per eight-hour shift prior to graduation. Residents also improve significantly for each individual patient. For instance, they become 20% faster at signing the first batch

[1] In 2021, the United States spent 17.8% of GDP on healthcare, compared to the OECD average of 9.6%, but life expectancy was 77.0 years compared to the OECD average of 80.4 (Gunja, et al., 2023).

[2] Academic Hospital GDP: the author's calculations using data from the BEA and AAMC. Global hospital rankings from Newsweek.com: https://www.newsweek.com/rankings/worlds-best-hospitals-2023

[3] Notable exceptions include Levitt, et al. (2013) and Adhvaryu, et al. (2023) in the automobile manufacturing industry.

of medical orders. These speed gains only accrue to complex patients, which I define as those who are ex-ante predicted to require inpatient admission: patient length of stay decreases by 10.3% over four years. Despite meaningful learning, I find no evidence of statistically or economically significant improvements in the 14-day readmission rate, the key measure of ED outcome quality, or in the number of orders signed, a measure of efficiency.[4]

When examining heterogeneity in resident learning by patient type, I find that most of the improvement is driven by or only present in complex patients. Therefore, I conclude that residents learn how to treat complex patients and become faster as a result of their increased skill. Conversely, residents learn relatively little about treating simple patients. Under the assumption that residents cannot learn about complex patients by treating simple patients, this means that the hospital is able to choose the amount of training it provides by changing the allocation of complex patients to residents. Because the number of examination rooms is fixed and almost always at capacity, changes in length of stay directly affect the number of patients seen per day, the definition of patient throughput. Therefore, the hospital trades off resident training and patient throughput when determining the optimal patient assignment.

Through studying how organizations manage within-firm learning via task allocation, I combine the literatures on task allocation and on learning by doing. Although it has been shown that task allocation to heterogeneous workers may have large implications for productivity (Adhvaryu, et al., 2023) and that productivity differences within sector can be large and persistent (Syverson, 2011), the task allocation literature typically does not incorporate worker learning. Instead, workers have fixed and exogenous skill and the firm allocates heterogeneous tasks to determine each worker's comparative advantage, as in Adhvaryu, et al. (2023), Bergeron, et al. (2022), Cheng (2019), Cowgill, et al. (2023), Dahlstrand (2023), and Kasy and Teytelboym (2022). Similarly, the literature on learning by doing typically does not consider task assignment. For instance, in medicine, there is work on resident learning in internal medicine (Chan, 2021), learning about match values of patients to procedures (Gong, 2018) and to medications (Currie and MacLeod, 2020), and learning to work in teams (Chen, 2021 and Reagans, et al., 2005). However, although patients may differ in these settings, their arrival to the physician is exogenous.

I explicitly consider both margins, as the hospital chooses the patients to assign to each resident and task-specific resident skill evolves with the history of patients assigned due to learning by doing. The dynamic framework is similar to that in Minni (2023), but the granularity of my data allow me to be more specific. I characterize how residents belonging to the same department and job title differ in skill and show how the organization's assignment of heterogeneous patients to heterogeneous residents optimally differs. In my setting where the learning margin dominates the comparative advantage margin, considering the impact of learning on future productivity is crucial. If resident skill were fixed, then the empirical

[4]An "order" is any diagnostic or therapeutic procedure that is prescribed for the patient. Diagnostic orders are primarily for gathering information and include procedures such as blood tests, echocardiograms (ECGs), and imaging (CT scans, X-Rays, etc.). Therapeutic orders are primarily for treating and stabilizing the patient, and include pain medication, antibiotics, and surgical procedures.

patient allocation patterns would suggest that the teaching hospital is making grave errors in task assignment. In that case, reallocating patients would lead to large, permanent improvements in productivity. However, this is not possible in practice because such an allocation strategy would reduce teaching, resulting in much lower future average resident skill and productivity.

My findings also add to the literature studying cohort turnover, the planned simultaneous exit of a large number of experienced workers and similarly sized entry of new workers. In American teaching hospitals, cohort turnover occurs every July 1, the date when the most experienced residents graduate and are replaced by a new class of fresh medical school graduates. The fear that patient outcomes will suffer due to the decrease in average experience is known in the United States[5] as the "July Effect." I corroborate Hughes (2017), Wei, et al. (2019), and the recent literature that finds an absence of a significant drop in quality in July. I extend the literature by showing that not only patient outcomes but also many process measures related to productivity and efficiency are similarly unchanged on average across July 1. I also add to the findings of Song, et al. (2016) and Hausknecht and Trevor (2011) and describe another method the teaching hospital uses in order to avoid a disruption in output. Notably, this method, strategic patient allocation, is a choice rather than an investment in infrastructure and supervision.

The rest of the chapter proceeds as follows: Section 1.2 gives more details on residency in general and emergency medicine residency at the teaching hospital from which I obtain data. Section 1.3 describes the electronic health record and audit log data. Section 1.4 presents empirical results that documents the ways in which residents learn by doing. Section 1.5 concludes.

1.2 Medical Residency Background

In the United States, graduates of medical school are required to complete a residency program in order to practice medicine independently. Residency is in a specific predetermined specialty (for example, Radiology, Dermatology, Obstetrics and Gynecology, and Emergency Medicine); medical school students apply to and are accepted to a single program-specialty.[6] Matching residents to program-specialties is done centrally and is a well-known application of the Gale-Shapley algorithm. Programs last between three and seven years, depending on the specialty, and some medical students choose complete a fellowship after their residency ends to further specialize and become for example Cardiologists and Oncologists, or to sub-specialize, for instance in Pediatric Critical Care or Cardiothoracic Surgery. Notably, residency training is not only for learning facts but also for developing "habits, behaviors, attitudes, and values that will last a professional lifetime" (Ludmerer, 2015).

[5]In the United Kingdom, this occurs on the first Wednesday in August and is known as both "Black Wednesday" and the "killing season."

[6]Students apply to multiple programs but typically a single specialty: among students who successfully match, the average number of specialties ranked is 1.2 (AMA, 2019).

The focus of this study is Emergency Medicine Residency at the University of California, San Francisco (UCSF). At UCSF, EM Residency is a four-year program.[7] The setup of the program and the day-to-day routine is typical of EM Residency Programs. The majority of patients are seen by a single resident, the trainee, who is supervised by an attending physician, a faculty member. The remaining patients are seen by attendings working independently. Work is shift-based, meaning that once physicians are off-shift, they are no longer responsible for the patients they cared for during their shift. At UCSF, both residents and attendings work eight-hour shifts. The schedule is determined prior to the beginning of the academic year and determined exogenously. All residents and attendings will work day, night, and weekend shifts; there is no sense that seniority or other factors permit attendings or residents to avoid working less-desirable shifts. Teams–groupings of attendings and residents–are ad-hoc, meaning that they change from shift to shift, and throughout the course of the year, all residents will work with all other residents and all attendings.

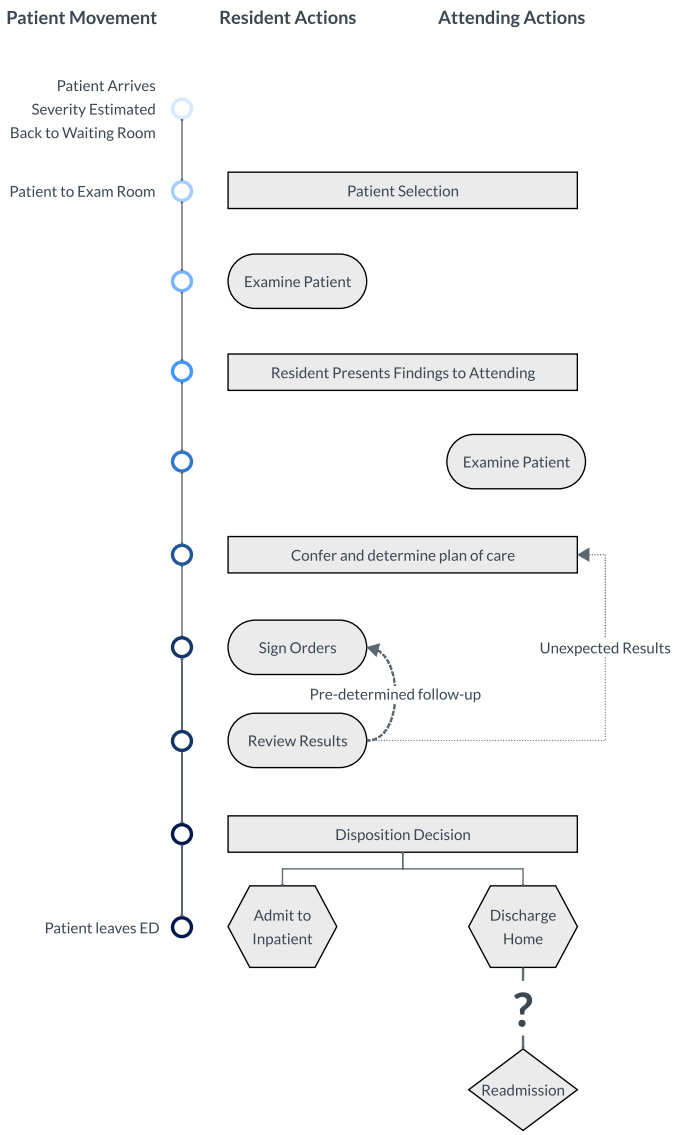
To be clear on terminology, I will use "resident" to refer to the emergency medicine physician trainees who are the focus of this study. At any point of time, EM residents at UCSF must belong to one of four different cohorts–this is defined as the year that they enter the program. Consistent with nationwide averages, I do not observe any attrition or leaves of absence.[8] "Attendings" or attending physicians are faculty members of the medical school, typically on the tenure track, who both supervise residents and see patients independently. I will use the terms "physician" and "provider" interchangeably to refer to residents, attendings, and nurse practitioners, who are also seeing patients independently but do not have supervisory responsibilities. I will use the term "care team" to refer to all providers, nurses, and other medical and non-medical staff (e.g. social workers) who interact with the patient. An "order" is any diagnostic or therapeutic procedure that the care team prescribes for the patient. Diagnostic orders are primarily for gathering information and include procedures such as blood tests, echocardiograms (ECGs), and imaging (CT scans, X-Rays, etc.). Therapeutic orders are primarily for treating and stabilizing the patient and include pain medication, antibiotics, and surgical procedures.

The typical workflow in the ED is depicted in Figure 1.1. When a patient arrives, a triage nurse will interview them, take their vital signs, and estimate their acuity using a five-point scale called the Emergency Severity Index (ESI). This is done independently from the physicians who will subsequently care for the patient. The patient will then return to the waiting room. A resident who is available will select a patient from the waiting room with guidance from the supervising attending. The resident will independently examine the patient and present their findings and plan of care to the attending. The attending will examine the patient, also typically independently, and confer with the resident. An agreement on the plan of care is reached and a set of diagnostic and therapeutic orders are

[7]Most EM Residency programs are three years; four-year programs tend to be located at prestigious and highly-ranked programs such as Johns Hopkins, Massachusetts General Hospital (Harvard Medical School), UCLA, and the University of Washington.

[8]The median EM Residency attrition rate from 2010-2020 is 0.83% (Wang, et al., 2022)

Figure 1.1: Workflow in the Emergency Department



Notes: This flowchart illustrates of the typical workflow in the emergency department. Actions and outcomes are divided into three categories. Left of the timeline are patient movement. To the right of the timeline, actions are classified into those done by residents (left side), attendings (right side), or together (spanning the width of the section). The dotted arrows originating from Review Results indicate that these actions are done only when deemed necessary. Finally, after the disposition decision is made, if and only if the patient is discharged home, they may feel it is necessary to return to the ED within 14 days, which is called an ED Readmission.

signed. Order results are reviewed, typically independently, and if necessary, predetermined follow-up orders are sent and additional examinations and revisions to the plan are agreed upon and executed. The resident and attending will then make a disposition decision: admit the patient to the hospital for additional care or discharge them home. In the event the patient was discharged home, there is a chance they will return to the ED within 14 days. This is called an ED Readmission and is suggestive that the physicians overlooked something important.

Care quality in the ED is multi-dimensional. Broadly, once the patient is stabilized, the goal is to quickly and efficiently assess the patient's condition. The disposition decision is the primary goal of the ED: is the patient healthy enough to send home, or do they need to remain in the hospital for further care? Therefore, the primary measure of ED quality is the accuracy of the disposition decision. A common measure used to evaluate the accuracy of this decision is the 14-day readmission rate (cf. Chan, 2018): among patients who were deemed healthy enough to discharge, at what rate did they return to the ED within 14 days? A second category of quality relates to speed. Doing things faster with no loss in accuracy is also important. Speed is utility-enhancing for patients because they spend less time suffering from their complaint and being in the hospital. It is also efficient for the hospital because it frees up the examination room for the next patient, thereby increasing patient throughput. Important measures of speed I will consider include process measures such as time to first order and patient length of stay in the ED. Finally, I will consider resource utilization as a measure of efficiency. Resources are both costly orders ("materials") as well as labor in the form of supervision and consults by specialists, and being able to achieve the same patient outcomes with fewer orders or consults represents higher efficiency.

1.3 Data and Sample Construction

1.3.1 Data

This research leverages highly granular electronic health record and audit log data from UCSF. The data cover the universe of ED arrivals for patients ages 18-90 over a 24 month period from 2017 to 2019. In total, there are 85,990 patient encounters.[9] In essence, these data record every interaction the physician has with a computer, which is used for gathering information (reading past clinical notes and order results), producing a diagnosis and treating and stabilizing the patient (sending, revising, and canceling orders), and recording information (writing the clinical note summarizing the patient's condition and what was done in the ED).

The data contain an entry for every instance that any provider interacts with an order. For each of these order actions, I observe patient and encounter identifiers, actual, unmasked

[9]The unit of observation is an encounter rather than a patient because the same patient may visit the ED multiple times during the sample period. When this occurs, they are assigned a new encounter_id for each visit but retain the same patient_id.

timestamps for when each order was signed, completed (or canceled), and results became available (when applicable). I also observe identifiers for both the physician who signed the order (typically a resident) and the physician who authorized it (must be an attending). These are both unusual features. For physician identifiers, most medical datasets only contain the data of the attending physician, as they are the entity who is financially and legally responsible. As for timestamps, most datasets either only have the date of the encounter or have detailed but de-identified data that preserves the time between actions but scrambles the start dates. Both of these elements are crucial for this analysis as otherwise I would not be able to attribute residents to patients in the correct order and would not be able to examine the speed and duration of important actions.

I also observe the consumption and production of information. Specifically, I observe the time, duration, and provider for each order result view (e.g. reading the radiologist's report for an MRI; viewing the numerical results of a blood test) and the same information for clinical notes that contain other physicians' impressions of the patient.[10] I also observe the time and duration of edits to the patient's clinical note from the current encounter, as well as the character length of the note. I do not observe any note content.

For patients, in addition to typical covariates such as age, gender, race, and diagnosis codes, I also observe a set of characteristics that I call "ex-ante" characteristics. These are characteristics that are exogenous to the care team who will subsequently care for the patient. Examples include the patient's chief complaint that induced the ED visit, the acuity level assigned to them by the triage nurse, and indicators for abnormal vital signs upon entry to the ED (ex. abnormal pulse). Contrast these with measures such as the final diagnosis, ED disposition, or patient's length of stay in the hospital, which may be endogenous to the composition of the care team and most crucially, resident experience. In the analysis, I use the set of ex-ante and immutable patient characteristics (things that cannot be affected by care, such as the patient's age, gender, and race) to divide patients into those who are ex-ante predicted to require inpatient care and those who are predicted to be safe to discharge home. For simplicity, I refer to these patients as "complex" and "simple." The predictions have high predictive power and fit the observed patterns of inpatient admission well. See Appendix A.1 for more details on the construction and fit of the prediction.

For providers, I observe a set of basic covariates. I observe the role of all providers: resident, attending, nurse practitioner, etc. I observe the specialty for attendings and NPs only and infer the specialty of residents based on the specialties of the attendings who most frequently authorize their orders, which I assume are their most frequent supervisors. Residents use different templates in the system if they are in their first two years compared to years thereafter. I also observe their start and end dates if they occur within the sample period; with these two pieces of information I infer the cohort of each resident.

In a separate dataset, I have the administrative schedules for both providers and residents for calendar year 2018. I use this data to validate my sample construction and to provide some sample statistics on the number of shifts worked by EM and non-EM residents. I am

[10]Both order result and note views can be from "historical" visits outside the sample period.

unable to match the names in the schedule with the provider identity numbers in the EHR data.

1.3.2 Sample Construction

I focus on EM Residents and attendings. These providers make up a minority of physicians who ever work in the ED but work a majority of the shifts and see a majority of the patients, especially among complex patients. The reason I restrict the analysis to EM Residents is because the ED may have other learning objectives for the residents from other specialties who make short rotations through the ED as part of their training. For instance, Internal Medicine residents complete a three-week rotation in the ED. Not only is this time period is too short for the ED to reap the benefits from training them, but the residents also may have a different set of baseline skills compared to EM residents.[11] Because the incentives and constraints for training other residents may differ from those for training ED residents, I choose to exclude them from my analysis.

During the sample period, there were 15 residents in each cohort of EM residents. I am unable to identify them based on names or identifiers so I classify them using the total number and fraction of orders that were signed in the ED context (as opposed to inpatient or outpatient). For residents belonging to each cohort, I define as EM residents those who sign over 80% of their orders in the ED context and are also one of the top 30 residents in terms of number of orders signed. The discontinuities in at least one of these measures are generally quite sharp. I show two examples in Appendix Figure A1.

Table 1.1 shows the breakdown of the residents who work in the ED in terms of the number of individuals, shifts worked, and patients seen. My algorithm slightly under-identifies the true number of EM residents, identifying 83 residents instead of the expected 90 in the six cohorts in my data. In calendar year 2018, where I am able to validate my resident selection by comparing shift summary statistics with administrative shift data, I also under-match slightly, identifying 67 of 75 residents. Perhaps as a result, I find that they work 60% of the shifts rather than the 69% as suggested by the 2018 administrative data. As expected, the majority of patients are seen by EM residents: almost 70% in the two years of EHR data.

Table 1.2 Panel (a) shows sample selection for patient encounters. Over the two years of data, there are a total of 85,990 patient encounters. I first exclude encounters where the patient left early or against medical advice, or passed away in the ED, so that I can be sure that I capture the full extent of the physician's process rather than some interrupted version. These total roughly 7.7% percent of all encounters. Then, I exclude the patients who the triage nurse categorized upon arrival as being the most urgent (Emergency Severity Index category 1) or the least urgent (ESI 5), who together represent about 2.3% of all arrivals. This is because the ESI 1 patients represent "codes" where the entire ED team on staff contributes to the patient's care, so it is an exception to the usual resident-attending

[11] For instance, they likely completed a different set of clinical rotations while in medical school and focused their research on different topics.

Table 1.1: Sample Selection: Residents

	Residents	Shifts (EHR)	Shifts (admin)	Patients (EHR)
All Residents	610	9,340		54,217
EM Residents	83	5,802		37,463
EM Residents (%)	14%	62%		69%
2018 Residents	389	4,512	4,012	26,775
2018 EM Residents	67	2,725	2,765	18,044
2018 EM Residents (%)	17%	60%	69%	67%

Notes: This table shows basic sample statistics on the set of residents who work in the emergency department. I focus on EM Residents, who make up 14% of all residents who work in the ED during the two-year sample based on my classification. They work 62% of the shifts worked by residents and see 69% of all patients seen by residents. I compare the share of shifts with the share of shifts in the administrative data that cover one calendar year and find that EM residents worked 69% of all shifts worked by residents, which compares favorably to the 60% I classify in the data.

pairing and may not represent cases where the resident is directing care. ESI 5 patients are the other extreme: they are cases where the patient does not need urgent medical care, such as patients with a chief complaint of "Medication Refill" and also do not represent resident learning about urgent patients. Next, among the remaining encounters, I am unable to identify the physician in charge ("Primary MD") for 6.6% of the patients. The next step results in our first sample of interest: EM Residents and Attendings see a total of 65.3% of all patients. Finally, EM Residents see 40.4% of all patients, or about 62% of the patients assigned to EM Residents or Attendings. Panel (b) reveals that the 62% of patients are not evenly distributed among patient types: residents see a greater share of complex patients (about 77%) relative to simple patients (about 58%) by two measures of ex-ante patient complexity.

These tables show that EM residents are doing a plurality of work in the ED and a majority of the work for complex patients. It is not the case that they are only being used as low-cost labor by seeing only the low-risk patients they know how to manage and leaving the complex ones for attendings to care for. In the following section I show how patient outcomes, process measures, and the allocation of complex and simple patients vary with resident experience.

1.4 Documenting Resident Learning

I begin by documenting the ways that residents improve during the four year program. I first show measures related to overall productivity. Then, I present results on within-patient improvement both graphically via binned scatterplots and in regression form with additional

Table 1.2: Sample Selection: Encounters

(a) Encounter Selection

	Number or Percent of Patients
All ED Arrivals	85,990
Did not Leave Early	92.4%
Did not Pass Away	92.3%
Triage Nurse ESI 2, 3, or 4	90.0%
Primary MD Identified	83.6%
Seen by Attending or EM Resident	65.3%
Seen by EM Resident	40.4%

(b) Resident Encounters by Complexity

		by Predicted Admission		by Triage Nurse ESI	
	All Patients	Complex	Simple	Complex	Simple
All with Primary MD identified	71,892	17,916	53,976	14,935	56,957
Seen by Attending or EM Resident	78.1%	73.1%	79.7%	74.1%	79.1%
Seen by EM Resident	48.3%	56.0%	45.8%	58.0%	45.8%
Percent EM / Attending or EM	61.8%	76.6%	57.5%	78.3%	57.9%

Notes: This table shows the sample selection of patient encounters. Panel (a) shows the steps of sample selection. Patients who Leave Early are those who have leave without being seen, against medical advice, or pass away in the ED. Triage Nurse ESI 2, 3, or 4 are the three middle categories of the triage nurse's assigned Emergency Severity Index. The two excluded categories are extremely severe cases ("codes") where the entire ED team contributes to the patient's care, or cases where the patient does not need urgent care, such as patients with a chief complaint of "Medication Refill." Primary MD Identified means I was able to identify who the primary provider for the patient was. Panel (b) shows the breakdown of the last three steps of Panel (a) by two ex-ante measures of "complex" and "simple" patients. The first is the primary measure I use in the paper: by a prediction of inpatient admission using only ex-ante and immutable patient characteristics. The second is by the triage nurse's evaluation: ESI category 2 vs. 3 and 4. The bottom row of Panel (b) shows the percent of patients of each patient type seen by EM residents relative to the patients seen by EM Residents and Attendings and reveals that residents see a greater share of complex patients than of simple patients.

specifications and patient heterogeneity and perform robustness checks. Third, I present evidence that when making patient allocation choices, the hospital is aware of the trade-offs between care quality and learning. Finally, I relate the findings to the literature on cohort turnover and summarize the important takeaways.

1.4.1 Overall Productivity: Patients Seen per Shift

The number of patients each resident sees per shift is a basic measure of productivity. Since shifts are always eight hours long regardless of experience, this measure is analogous to each resident's units produced per hour. Figure 1.2, Panel (a) plots the relationship between the average number of patients seen per shift seen by residents and resident experience in months. Patients are grouped by their ex-ante predicted complexity: whether or not they are predicted to be admitted to inpatient care. Panel (b) depicts the growth in the average number of patients managed per hour of each shift for residents of each month in the program. It shows that residents only average about 1.5 patients per hour when they begin residency and improve such that they are managing about 3 patients per hour by the fourth year of residency. Comparing the two panels reveals that the growth in patients seen per shift is mainly in the "simple" category of patients and that it appears to be driven by managing additional patients in parallel rather than large increases in speed per patient, a fact corroborated in Figure 1.3. Overall, residents make significant gains in productivity, going from seeing about three patients per shift in the first month of the program to seeing almost eight patients per shift in the month prior to graduation. However, growth in this productivity measure is mainly in simple patients. Differences between complex and simple patients will be a recurring theme in this section.

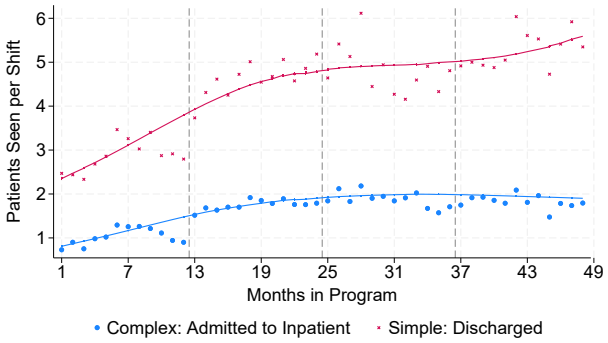
1.4.2 Within-Patient Quality, Efficiency, and Productivity

While they indicate significant improvement, the previous set of results do not capture within-patient differences with experience. Therefore, the results may either understate or overstate the degree of improvments in productivity. For example, if the quality of care also increases with experience even as residents are seeing more patients per shift, then the results understate productivity improvements. On the other hand, if quality suffers with the additional patient load then the results overstate productivity improvements. This subsection investigates the evolution of a variety of patient-level measures of quality and efficiency with resident experience and finds that learning how to treat patients mainly occurs in complex patients.

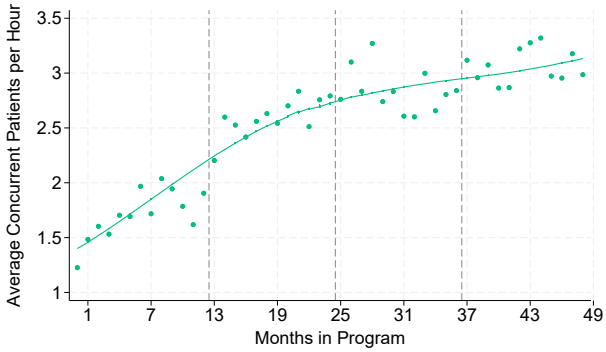
I begin with a graphical depiction of resident improvement via binned scatterplots of various patient outcomes and process measures in Figure 1.3. For each outcome of interest, I regress both the outcome and resident experience on selected patient covariates P_i and show a binned scatterplot of the residuals. I add the overall outcome mean back to the outcome residuals so that the values are more easily interpretable. The slope and standard error of the regression line displayed correspond via Frisch-Waugh-Lovell to the coefficient

Figure 1.2: Patient Load Breakdown

(a) Patients Seen per Shift by Predicted Admission



(b) Simultaneous Patients Managed



Notes: These figures show the evolution of patient load over the 48 months of the EM residency program. Panel (a) shows the breakdown of total number of patients managed during the shift. Patients are grouped by their ex-ante predicted complexity: whether or not they are predicted to be admitted to inpatient care. Panel (b) depicts the average number of patients managed per hour of each shift for residents of each month in the program. Comparing the two panels reveals that the growth in patients seen per shift is mainly in the "simple" category of patients and that it is driven by managing additional patients in parallel rather than large increases in speed per patient, a fact corroborated in Figure 1.3.

on experience β in the regression given by

$$Y_i = \beta \text{Experience}_{j(i)} + P_i'\gamma + \varepsilon_i \tag{1.1}$$

In this regression, i indexes encounters, and $\text{Experience}_{j(i)}$ is the experience of resident j who is in charge of patient i in years. In the binned scatterplots, I select the ex-ante and immutable patient characteristics P_i by hand. The covariates include fixed effects for 10-year bins of patient age, the Charlson comorbidity index, Medicaid status, nonwhite, an interaction of broad chief complaint category and triage nurse assigned emergency severity index, an interaction of indicators for if the encounter began on a weekday and during business hours, and continuous ex-ante predictions of patient complexity and its square from Chu, et al. (2023). The residency program lasts four years, but because my data span two years, I observe each resident for a maximum of two years. Hence, the data are an unbalanced synthetic panel.

The various panels of Figure 1.3 break down resident learning into various components. I first examine improvements in two key measures of care quality and efficiency. I observe in Panel (a) that there does not appear to be a statistically or economically significant change in 14-day ED Readmissions, suggesting that conditional on patient observables, the

accuracy of the disposition decision made by inexperienced and experienced residents is similar. Similarly, in Panel (b), I find that there is also no relationship between experience and the number of costly diagnostic and therapeutic resources signed. Therefore, I conclude that neither patient outcomes nor efficiency in costly resource utilization is a cost of training.

But does that mean that residents do not learn? Panels (c) and (d) refute this. Panel (c) plots the fraction of orders signed by the resident in charge of the patient, rather than the supervising attending, other attendings such as consulting physicians from other specialties, nurses, or other residents. This measure of resident independence increases linearly with experience, and the magnitude over four years is approximately 10% of the mean of 59.5%. This implies that with experience, residents gain independence and are less apt to leave out important orders. Panel (d) plots a measure of speed: how long does it take providers to sign the first order after the patient enters the examination room? Over four years, this decreases approximately 15% of the mean of about 34 minutes. This means that residents become faster at discerning the patient's underlying condition through history-taking and physical examination and determining which set of orders are appropriate for treating and refining the working diagnosis. Becoming faster and more thorough are important clinical skills that affect care quality after the resident graduates and begins practicing independently.

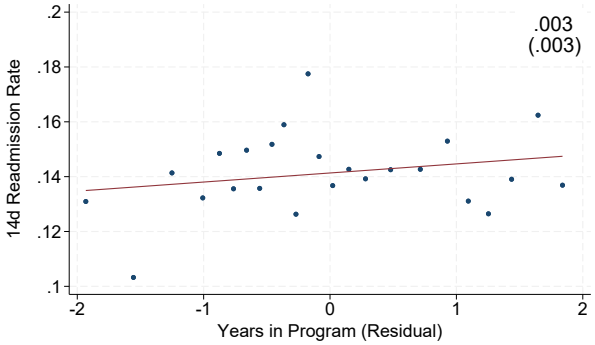
I find that the increases in independence and speed only flow through to the total length of stay for complex patients. Panels (e) and (f) show the evolution of the natural logarithm of the number of hours the patient spends in the ED with resident experience. The patient sample is split by whether I predict that they were admitted to the hospital ("complex") or were discharged home ("simple"). While there is no change for simple patients, there is is a significant and meaningful improvement for complex patients. Under the assumption of linear learning, the four-year improvement of 8.8 log points is almost 25% of the standard deviation of log(length of stay) conditional on patient covariates and is relative to a mean length of stay of about 6.6 hours.[12] Therefore, experience greatly decreases time spent in the ED for complex patients but has limited effects for simple patients. The difference in learning for simple and complex patients motivates examining heterogeneity for the other process measures and the design of the structural model.

The previous findings are supported by the regression results in Table 1.3. The regression specification differs slightly from the binned scatterplots in Figure 1.3. First, the regressor of interest is the natural logarithm of days in the program. I tend to prefer this over the linear specification because the literature generally finds that learning exhibits diminishing returns (cf. Benkard, 2000 and Levitt, et al., 2013). Second, I include resident fixed effects in all regressions to focus on within-resident learning. With these fixed effects, estimates are

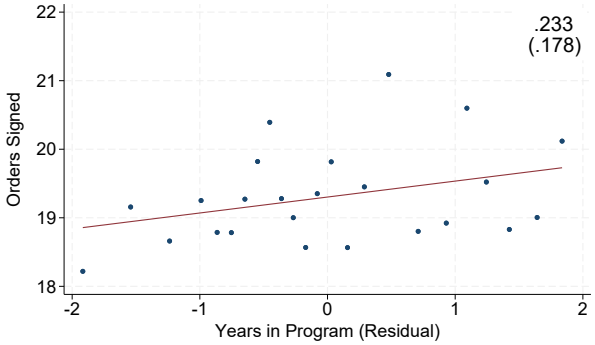
[12]The reason the mean for complex patients is so similar to the mean for simple patients is because for actual admitted patients, I end the the length of stay at the moment the patient is confirmed for inpatient upgrade. At that moment, the patient may not leave the ED, but the ED care team's involvement has concluded and the patient is now the responsibility of the admitting department, whether it be cardiology, surgery, hospital medicine, or something else. Unfortunately, there is no consistent analogous marker for discharged patients (discharge orders are inconsistently signed and disappear entirely midway through the sample period). Both relationships are similar if I instead use total time in the ED for both sets of patients.

Figure 1.3: Learning over Time: Binned Scatterplots

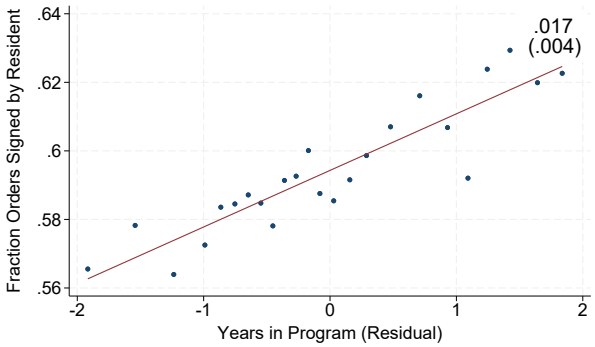
(a) 14-day ED Readmission Rate



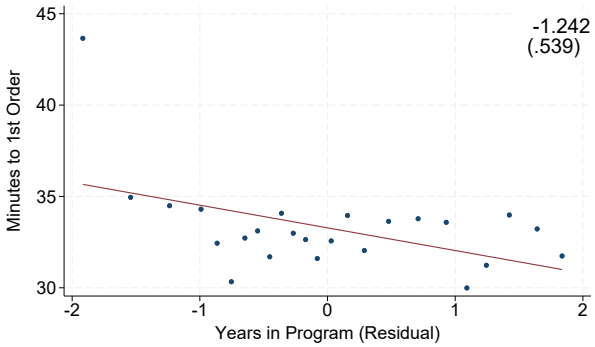
(b) Medical Orders Signed



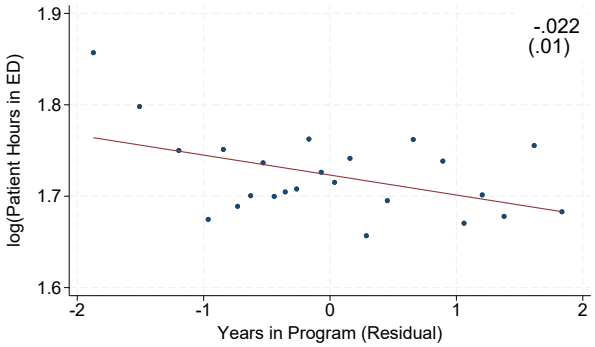
(c) Fraction Orders Signed by Resident



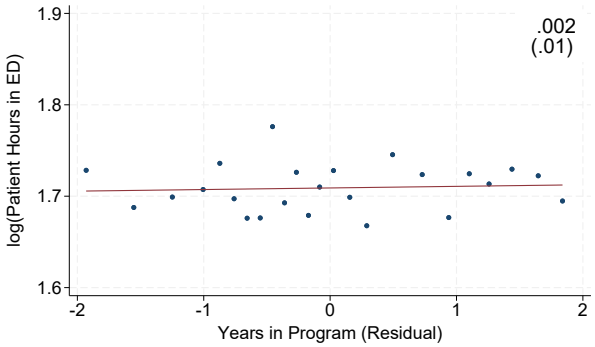
(d) Time to First Order (Minutes)



(e) Log(Length of Stay): Complex Patients



(f) Log(Length of Stay): Simple Patients



Notes: These figures are binned scatterplots with 24 bins of patient outcomes and process measures of interest on the residual of years in the program by the resident in charge of the patient. The sample is all sample patients seen by EM residents unless otherwise specified (Panels (e) and (f), where the sample is split into "complex" and "simple" patients based on an ex-ante prediction of inpatient admission). Residuals are after removing selected patient covariates. The coefficient and standard error, clustered by physician, are displayed. See text for more details.

not subject to bias from individuals in earlier cohorts (e.g. starting residency in 2015) being inherently "better" or "worse" than individuals in later cohorts (e.g. starting residency in 2018). Third, when I include patient covariates, I select them using the post-double-selection LASSO method of Belloni, et al. (2014). Inspection of the covariates chosen by the algorithm reveal that they are more sparse than the set that I manually selected, and tend to include indicators for the number of abnormal vital signs upon entry, which I did not include in the binned scatterplots.

Table 1.3 shows that the results in Figure 1.3 are generally robust to the more sophisticated selection of patient covariates and the inclusion of resident fixed effects. Notably, as in the figures, the log(Length of Stay) relationship is only statistically and economically significant for complex patients. In Table 1.4, I examine heterogeneity by patient complexity for the other process measures.[13] For the natural logarithm of medical orders signed, I find that the small positive effect in Table 1.3 masked offsetting effects for complex patients and simple patients. One potential explanation, supported by the results on diagnostic orders shown in Appendix Table A2, is that with experience, residents obtain less diffuse priors for complex patients, but they substitute effort with costly resources for simple patients in order to save time.[14] Next, we observe that the increase in fraction of orders signed by the resident is also primarily driven by improvements for complex patients, but that decreases in the minutes to the first order are proportionally similar for complex patients compared to for simple patients. Overall, these results suggest that the bulk of the learning that occurs during the residency relate to learning how to treat complex patients and that there is relatively little learning for simple patients.

Robustness

The main threat to the within-patient analyses is that they are biased by selection on unobserved patient characteristics. Specifically, if more experienced residents are assigned patients who are unobservably more complex, my estimates will be biased towards zero. Similarly, if they are assigned unobservably less complex patients because they are seeing additional patients simultaneously, then my estimates will be larger in magnitude than the true improvement with experience. I believe this is unlikely in my setting for two reasons. First, providers observe a limited amount of information when allocating patients, and I am able to control for almost all of these covariates. The main thing I do not observe are the patient's appearance and answers to brief questions, but to the extent that is captured in the triage nurse's estimation of the patient's severity, I do control for it. Second, other than for the first six months of the program, observable patient severity averages per patient are stable across the four years of experience as can be seen in Appendix Figure A2, Panel (b).

[13] By definition, ED Readmissions are only possible for discharged patients, so a breakdown by complexity is not appropriate.

[14] This finding may facilitate the increase in managing additional patients simultaneously with experience shown in Figure 1.2.

Table 1.3: Learning over Time Regressions: Main Results

	Readmissions		log(Medical Orders)		Resident Signed Frac.	
log(Days in Program)	-0.000	0.003	0.040**	0.030***	0.009*	0.012***
	(0.004)	(0.004)	(0.016)	(0.011)	(0.005)	(0.004)
DepVar Mean	0.140		19.321			
Patient Type	Discharged		All		All	
Controls		X		X		X
Obs	22,544	22,544	31,317	31,317	31,317	31,317

	log(Mins to 1st Order)		log(Length of Stay, Hours)			
log(Days in Program)	-0.119***	-0.103***	-0.063***	-0.053***	-0.000	-0.002
	(0.028)	(0.016)	(0.015)	(0.016)	(0.010)	(0.009)
DepVar Mean	39.399		7.514		7.327	
Patient Type	All		Complex		Simple	
Controls		X		X		X
Obs	26,985	26,985	8,828	8,828	22,506	22,506

Notes: Regressions of selected patient outcome and process measures on various measures of resident experience. The sample consists of all patients seen by EM residents. Every regression includes provider fixed effects. Standard errors are clustered by physician. Patient Controls are chosen from the set of immutable and ex-ante patient covariates using the post-double-selection LASSO method of Belloni, et al. (2014) and differ from the covariates used in the binned scatterplots. The 14-day ED readmission rate is the rate at which patients who are discharged home from the ED have a repeat visit within 14 days. By definition, the measure only exists for discharged patients. log(Medical Orders) is the natural logarithm of the sum of diagnostic and therapeutic orders signed in the ED. Frac. Orders Signed by Resident is the fraction of medical orders that are signed by the resident, rather than the attending, nurses, or other residents assisting. log(Mins to 1st Order) is the time between the moment the patient is moved from the waiting room to an exam room and the time that the first medical order is signed. This value is missing if the first order is signed prior to being roomed; see Appendix Table A1 for the extensive margin. log(Length of Stay) is the natural logarithm of the hours the patient spent in the ED under the care of EM providers. It is split into "complex" and "simple" patients based on an ex-ante prediction of inpatient admission. Dependent variable means are listed, always in levels. See text and Appendix A.1 for additional details.

Table 1.4: Learning over Time Regressions: Heterogeneity by Patient Complexity

	log(Medical Orders)				Frac. Orders Signed by Resident			
log(Days in Program)	-0.032**	-0.035**	0.047**	0.049***	0.023***	0.023***	0.008	0.008**
	(0.014)	(0.017)	(0.018)	(0.012)	(0.007)	(0.007)	(0.005)	(0.004)
DepVar Mean	30.072		15.104		0.475		0.643	
Patient Type	Complex		Simple		Complex		Simple	
Controls		X		X		X		X
Obs	8,817	8,817	22,500	22,500	8,817	8,817	22,500	22,500

	log(Medical Orders)			
log(Days in Program)	-0.128***	-0.102***	-0.109***	-0.103***
	(0.042)	(0.034)	(0.029)	(0.018)
DepVar Mean	26.199		44.265	
Patient Type	Complex		Simple	
Controls		X		X
Obs	7,269	7,269	19,716	19,716

Notes: Regressions of selected patient outcome and process measures on various measures of resident experience. The sample consists of all patients seen by EM residents, split by ex-anted predicted complexity. Every regression includes provider fixed effects. Standard errors are clustered by physician. Patient Controls are chosen from the set of immutable and ex-ante patient covariates using the post-double-selection LASSO method of Belloni, et al. (2014) and differ from the covariates used in the binned scatterplots. Dependent variable means are listed, always in levels. See notes to Table 1.3, text, and Appendix A.1 for additional details.

Therefore, in terms of ex-ante patient assignment patterns, I believe I sufficiently control for selection on observables, and that unobservables are of limited importance.

It is not entirely straightforward to confirm this formally. I would like to perform the test proposed in Oster (2019) and Altonji, et al. (2011), but that requires the use of a measure of model fit such as R^2. Because I use LASSO to select covariates, the reported R^2 is not correct because it does not take into account uncertainty in covariate selection. I proceed regardless of this limitation and use the R^2 as if there was no uncertainty. This means that the test results will be biased towards rejecting the null of no treatment effect due to omitted variables bias because I will be overestimating the improvement in model fit from including observable covariates.

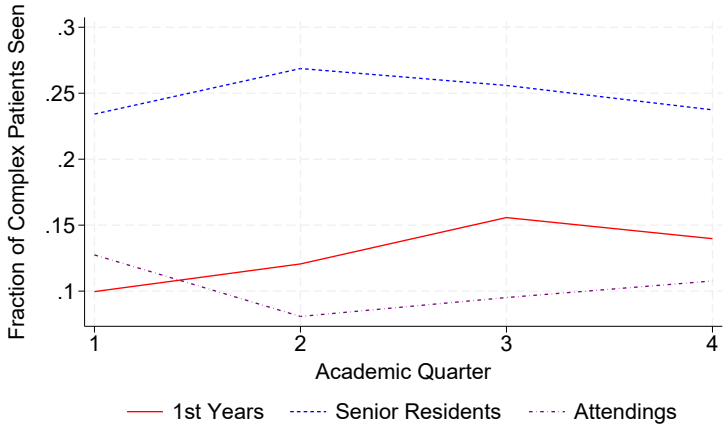
With these caveats in mind, results suggest that the size of omitted variables bias in this context are small. For instance, when considering length of stay for complex patients, the improvement in R^2 from going from a specification with only physician fixed effects to the covariates chosen by post-double-selection is from 0.017 to 0.092, and the coefficient on experience decreases in magnitude from -0.063 to -0.053. If I assume that the maximum R^2 that can be explained by the model is 0.3 (in other words, outside "randomness" such as ED congestion explains the other 0.7), then if the true effect was zero, the omitted variables would need to have 2.18 times the amount of selection as the observable factors to produce the results I obtain. If I assume the maximum R^2 is 0.5, then the omitted variables would need to have 1.15 the amount of proportional selection, whereas if the maximum R^2 is 1, then the omitted variables would need to have 0.53 the amount of proportional selection to obtain the results I have if the true effect is zero. Based on the qualitative arguments based on the context I outlined previously that limit the potential for unobserved selection, I find these magnitudes to be unlikely, especially if the maximum R^2 is limited by factors orthogonal to resident experience such as waiting time for imaging and lab results. Based on the qualitative and quantitative evidence, I conclude that selection on unobservables should not meaningfully affect my results.

1.4.3 Hospital Awareness of Trade Off between Quality and Learning

Because learning is mostly in complex patients, attendings can control the trade-off between care quality and training by changing the allocation of complex and simple patients to residents of varying experience and themselves. But are attendings aware of the trade-off? The answer appears to be yes. Figure 1.4 plots the average fraction of complex patients seen by individual providers during each shift across the four quarters of the academic year. This figure illustrates three interesting facts. First, the fraction of complex patients seen increases from 10% to 15% during the first year (solid red line), corroborating the results of Figure 1.2. Second, again in line with Figure 1.2, the fraction seen by the other three cohorts of residents is relatively stable during the year (dashed blue line). Third, it is attendings who "pick up the slack" in July through September and see the patients that the first year

residents are unable to treat (dotted and dashed purple line).[15]

Figure 1.4: Average Fraction of Complex Patients Seen, by Role



Notes: This figure depicts the average fraction of complex patients seen per shift, by role, for each quarter of the academic year. Complex patients are those with the highest values of ex-ante predicted admission. "Senior Residents" are the average shares of residents in years 2-4. This choice is informed by the results in Figure 1.2, Panel (b) and Appendix Figure A2, Panel (b), where the share of "Most Urgent" and Admitted patients does not continue growing after the first year. The figure shows that first year residents see more patients as they gain experience, but that the patients they are not able to see in academic quarter 1 (July through September) are seen by attending physicians rather than other residents. Not shown is that there are no meaningful differences in provider staffing or arriving patient composition across the academic year.

Next, I provide suggestive evidence that attendings are aware of the trade-off on a more micro level via correlational logit regressions. In these regressions, I regress the probability that a first year resident is assigned a complex patient on the number of complex patients currently being seen in the ED, the number of patients in the waiting room, fixed effects for the patient's chief complaint, and other characteristics of the physicians on staff and the index patient. Results are in Appendix Table A4. I find that first year residents are much less likely to be assigned complex patients when there are many patients in the waiting room. As the number of patients in the waiting room increases from the 25th to 75th percentile, the probability that first year residents are assigned a complex patient decreases by 15%. To conserve space, I do not show the coefficients on patient chief complaint, but these estimates are meaningful. I find that *ceteris paribus*, first year residents are much more likely to be assigned patients from more common chief complaints (e.g. chest pain, abdominal pain, and shortness of breath) compared to the pooled "less common" category.

Taken together, these findings suggest that the hospital is aware of the costs of teaching because they teach less when the costs are higher due to congestion, and that the hospital

[15]Indeed, there is no difference in the composition of arriving patients across the year.

is aware of the benefits because they first train residents in the patients they are most likely to encounter.

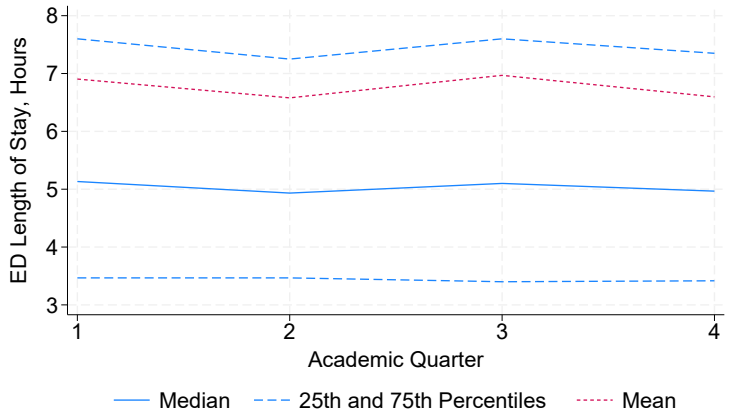
1.4.4 Summary of Resident Learning

Based on the combination of the results on shift-level efficiency and within-patient outcomes, I conclude that EM residents improve in two dimensions. The first is in medical skill for individual patients: the processes of gathering, synthesizing, and interpreting information about each patient's underlying state and signing the correct set of orders given that information. Table 1.4 shows that these improvements are primarily for treating complex patients. The second is improvement in "bandwidth" or cognitive capacity: residents become able to manage additional patients simultaneously. Most of the growth in capacity is for less complex patients, as shown in Figure 1.2. I am primarily interested in the improvement in medical skill for individual patients, which I believe is a more appropriate application of the learning by doing and task allocation frameworks. Because the majority of improvement in within-patient medical skill is for complex patients, attendings can affect the trade-off between care quality and training by changing the allocation of complex and simple patients among residents of varying experience and themselves.

The within-patient results relate to the literature on cohort turnover in residency programs. Hughes (2017), Wei, et al. (2019), and the recent literature finds an *absence* of a significant drop in care quality in July, when the most experienced residents graduate and are replaced by new medical school graduates. In my findings, this follows from the lack of a gradient with respect to ED readmissions and resource utilization: if inexperienced residents achieve the same outcomes as experienced residents, there cannot be a drop-off in quality. But what about complex patient length of stay? I show that there is a notable decrease in length of stay with respect to resident experience. Therefore, outcomes for individual patients can change during the academic year. But recall that attendings see more patients in July through September, when the experience cost is greatest (Figure 1.4). Figure 1.5 shows that average length of stay is unchanged throughout the academic year as a result of the patient allocation strategy. This method of reducing the impact of cohort turnover complements the investments in infrastructure studied by Song, et al. (2016) and Hausknecht and Trevor (2011), such as better nurses, a transition period for continuing residents in June prior to turnover, and better attending supervision. Hospitals can also strategically allocate patients to physicians of varying skill in order to maintain average outcomes. Notably, unlike better infrastructure and training practices, this method is an operational choice that does not require costly investment.

Broadly, resident progress for complex patients can be divided into two categories: patient-relevant and not patient-relevant. The main patient-relevant change is complex patients' length of stay in the ED, which decreases by approximately 10.3% over the four-year program. Crucial patient outcomes, as measured through ED readmissions, are unchanged. The time to first order does decrease, but the average magnitude is only about five minutes, so

Figure 1.5: Average Length of Stay, Complex Patients



Notes: This figure plots summary statistics of patient length of stay over the four quarters of the academic year for patients seen by EM residents and attending physicians. These are raw summary statistics without any patient or physician controls. The median ED length of stay is very stable with respect to academic quarter. The mean is slightly less stable, but that is driven by the top 25 percent of patient encounters.

it is relatively unimportant. Patients are not affected by who signs orders for them, so the fraction of orders signed by the resident is not a patient-relevant outcome. Arguably, they are also relatively insensitive to the number of orders signed, insofar as it does not affect their outcomes and the change in out-of-pocket cost is small due to insurance coverage. It is also ambiguous whether the hospital desires a reduction in orders signed as this depends on how the payer will reimburse them. I return to reimbursements in the first counterfactual.

Therefore, the training environment can be described as follows: residents are more or less capable of treating simple patients when they begin the residency program. However, they need to learn how to diagnose and treat complex patients, but the only way to learn is to learn by doing: by seeing complex patients. Attendings are aware of this, and also of the primary trade-off: inexperienced residents are slower than experienced residents. Therefore, the cost of training an inexperienced resident is greater than the cost of training a more experienced resident. But because learning is concave, inexperienced residents learn more from seeing each patient. Furthermore, there is additional time left in the program for the hospital to benefit from their increased skill compared to more senior residents. Therefore, the benefits of training inexperienced residents may be greater than the benefits of training experienced residents. The hospital must take these trade-offs into account and strategically allocate complex and simple patients to inexperienced residents, experienced residents, and attendings working independently in order to maximize the discounted sum of its stream of payoffs. I formally describe and estimate the nonprofit teaching hospital's dynamic optimization problem in the next chapter.

1.5 Discussion and Conclusion

I examine and quantify the trade-offs that academic emergency departments face when training inexperienced residents. I first investigate at a granular level the costs of training and find that despite substantial increases in independence and the ability to manage additional patients simultaneously, there are no differences in patient outcomes or costly resource utilization. I find notable differences in patient throughput, but only for complex patients who are predicted to require inpatient admission: the median fourth-year is able to arrive at a disposition decision and complete working up these patients 10.3% faster than the median first-year. The improvement in length of stay means that the hospital trades off patient throughput today with patient throughput tomorrow.

Even though my focus is on the emergency department of a single, top-ranked teaching hospital, there are key lessons for the broader healthcare sector. First, time is essential for teaching across specialties and departments: Ludmerer writes, "Time was the irreducible element of good medical education, whatever clinical setting happened to be used." Third, speed is an important quality measure across medical care, even in non-urgent situations. For instance, in the surgical context, it has been shown that longer operative time is associated with increased odds of complications (Jackson, et al., 2011). Next, although there may be some variation in care correlated with teaching hospital rankings, prior work has shown that the basic production function of health services does not differ in outcomes with respect to residency program prestige (Doyle, et al., 2010).

Chapter 2

Task Allocation in Academic Emergency Departments

2.1 Introduction

Recently, both private and public insurers have been turning to financial incentives such as payment reform to reduce costs and improve patient outcomes. However, these policies typically do not consider the dual role of teaching hospitals of treating patients and training the next generation of physicians. While the changes may incentivize teaching hospitals to increase care quality, they may also induce them to reduce teaching, which would have serious consequences for future patients. Understanding how private, nonprofit teaching hospitals trade off the quantity and quality of patient care with resident training is crucial in order to properly assess the impact of such policy changes.

In the previous chapter, I describe the ways in which emergency medicine residents learn with experience. I find that the hospital trades off resident training and patient throughput when determining the optimal patient assignment. The magnitudes reported are the static costs of training: how much patient throughput the hospital must sacrifice today when assigning patients to residents of varying experience. But these are not the costs that the hospital uses to inform its allocation strategy. This is because while the throughput costs of training are paid today, the benefits accrue in the future. Consequently, a model of the hospital's objective function must incorporate dynamics.

Therefore, I develop a discrete-time, infinite-horizon model where the hospital allocates complex patients to residents of different cohorts and attendings working alone to maximize the discounted sum of resident training, subject to a budget constraint written in terms of patient throughput. This builds upon models of nonprofit hospital behavior by Newhouse (1970), Lakdawalla and Philipson (1998), and others. My contributions are to add a teaching objective and the necessary dynamics to the hospital's utility function and to estimate the parameters empirically. The estimates allow me to simulate how training behavior might respond to counterfactual changes in the hospital's payoffs to higher productivity in the present.

I find that an objective function where the hospital maximizes training with respect to a lower bound of patient length of stay can rationalize the observed patient assignment shares during the academic year. That is, the hospital allocates complex patients to maximize the skill of graduating residents, subject to the constraint that average patient length of stay is constant in each quarter of the academic year. I apply the model estimates to two counterfactual exercises and consider the impact of decreased training on physician career outcomes and patient utility. Decreased training means that physicians take longer to see each patient, but because shift lengths are fixed, they will see fewer patients. Career outcomes will suffer because EM physician compensation is often based on the number and complexity of the patients they see.[1] Patients will also suffer because even though they will receive the

[1]Compensation tied to Relative Value Units (RVUs) is increasingly popular for EM physicians (ACEP, 2021). RVUs are a standardized measure of the value of a service or procedure used by the Center for Medicare & Medicaid Services (CMS) and is positively correlated with patient complexity. Therefore, the more patients per shift or complex patients per shift seen, the more RVUs generated and the higher the compensation.

same care and experience the same outcomes, they will have to wait longer if they are seen by a less-trained physician. I compare the impact to career outcomes and patient utility without further adaptations to alternatives where the hospital takes a mitigating action, such as loosening the care quality constraint, increasing the speed of the attendings working independently, and increasing the speed of resident learning.

In the first counterfactual, I quantify the implications on patients of graduating residents of a reduction in training required to achieve a 2% increase in current patient throughput. A desire to increase the number of patients seen could arise from payment reform intended to decrease patient length of stay. This would result in less revenue per patient than the status-quo, which means that the hospital would need to see additional patients to fulfill its budget constraint.[2] Assuming that residents go on to a 30-year career, this would result in costs to future career outcomes and patients over four times larger in present-value than the hospital's gains. However, investing in attending speed–most simply by staffing additional attending physicians so that teaching responsibilities are distributed among additional physicians–such that their aggregate speed increases 5% would greatly reduce the impact on training. With this remedy, training reductions are lowered and future costs decrease by 65%.

In the second counterfactual, I consider the impact and potential responses to a disruption in training. This mirrors the training disruption that affected residents during the Covid-19 pandemic, when both the number and composition of patients seeking emergency care changed.[3] In the counterfactual, I assume that the disruption causes affected residents enter their final year of training with half of the usual steady-state skill. I find that the hospital does not immediately return to the steady-state level of training for the incoming cohort. Furthermore, the skill of the affected cohort is reduced by 1.1%. A one-period 2.5% increase in attending speed allows the hospital to train sufficiently to restore the steady-state for future residents and also allows senior residents and future patients to recover 65% of the costs relative to the no disruption baseline. However, the benefits of further improvements in training capacity accrue to junior residents and future cohorts rather than the affected cohort. Therefore, continuing education for after the senior residents is necessary to fully counteract the effects of the disruption. As illustrated in both counterfactuals, small decreases in current training can have large consequences for residents' career outcomes and hence for future patients. However, straightforward and feasible actions can greatly mediate the reduction in training.

This work contributes to several strands of literature. First, I add teaching considerations to the literature on private, nonprofit hospitals and the literature on payment reform. Research modeling the objectives of private, nonprofit hospitals began with the seminal theoretical contributions of Arrow (1963), Newhouse (1970), Feldstein (1971), and Pauly and Redisch (1973). Since then, the bulk of the theoretical literature has consisted of models

[2]Alternatively, if the reduction in revenue per patient caused the hospital to decrease the number of patients seen, residents would see fewer patients over the course of the program and the impact on training is identical.

[3]Patients delayed both routine and emergency care (Czeisler, et al., 2020).

where the hospital maximizes the weighted sum of profits and quality or quantity of care (cf. Lakdawalla and Philipson (1998); see Gaynor and Town (2012) for an overview). These models have the appealing feature that nonprofit hospitals have similar objective functions to their for-profit counterparts, but with a lower marginal cost for quality or quantity (Gaynor, 2006), and this is consistent with subsequent empirical findings. For instance, nonprofit and for-profit hospitals are very similar in their responses to financial incentives (Duggan, 2000), CEO compensation incentives (Brickley and Van Horn, 2015), pricing behavior with regard to competition (Gaynor and Vogt, 2003), and provision of charitable care (Capps, et al., 2017). Similarly, the literature on payment reform also typically does not consider teaching. This is true both in the theory (cf. McClellan, 2011) as well as the empirical evidence (cf. Clemens and Gottlieb, 2014).

After I add a teaching objective to the nonprofit hospital's utility function, I estimate the parameters of the theoretical model and use it to simulate counterfactuals related to payment reform. Thus, I quantify the extent to which the hospital reduces teaching in response to counterfactual payment policies that reduce its revenue. My findings apply to almost all future physicians and academic medical centers: across specialties, between 83.1% and 96.6% of residency programs were affiliated with nonprofit institutions in 2021 (Lassner, et al., 2022a; Lassner, et al., 2022b). Additionally, Kocher and Wachter (2023) find that academic hospitals tend to do poorly on measures used in value-based payments, which means that many would stand to lose revenue if commonly-discussed payment reforms to decrease costs and increase quality were implemented. Hence, this chapter addresses a shortcoming in the nonprofit hospital literature first raised by Reder (1965): "Still further complications exist: hospitals produce not only current treatment but also train personnel for the production of future treatment. The costs and benefits of this training to the hospitals providing it are not well known." I go further by not only considering the costs and benefits to the teaching hospital itself, but also the costs and benefits to the graduating resident's career and their future patients.

The rest of the chapter proceeds as follows: Section 2.2 introduces the dynamic framework, Section 2.3 discusses estimation, and Section 2.4 provides results. Section 2.5 motivates and presents counterfactual exercises that explore separately the hospital's response to a change in the shadow cost of training and to a one-time disruption to resident training, as well as the effectiveness of mitigating actions it could take. Section 2.6 concludes. For an overview of the setting and data, see Sections 1.2 and 1.3 in Chapter 1.

2.2 Dynamic Framework

In this section, I present a dynamic model of patient allocation. It is necessary to consider dynamics because I am interested in estimating how the hospital allocates patients to trade off current care quantity and quality and future care quantity and quality via training. Unless the hospital acts myopically, a static model cannot capture these trade offs because it does not take into account future benefits of training. In other words, forward-looking hospitals

take opportunity costs and future benefits into account when optimizing patient allocation. The dynamic choice model is a discrete-time, infinite-horizon model, where the state-space, resident skill, evolves akin to overlapping-generations models.

Each shift, attendings first observe the the skill of residents who were assigned to work. An infinitesimally divisible unit mass of complex patients arrives and attendings choose a share of patients to assign to each resident and themselves. Attendings help residents see patients and may also see some patients independently. Patient utility, a function of length of stay and therefore a function of resident skill, is realized. At the end of the shift, resident skill increases by the share of patients they saw. Each July 1, 4th years graduate and are replaced by new 1st years with zero skill. Attending skill is fixed.

This means that skill X is a four-dimensional vector where each element represents the experience of one of the four resident cohorts–first years, second years, third years, and fourth years. I do not need to track attending skill because it is fixed. Within the academic year, skill in the next period is simply skill in the current period plus the share of patients seen in the current period. On July 1, the fourth years graduate, the continuing three cohorts are promoted, and the new first years who join enter with zero skill. This structure for the evolution of skill is similar to the structure in Bloesch and Weber (2023) and Jovanovic (2014).

The hospital's choice of patient allocation share is the path of allocations $\{S_t\}_{t=0}^{\infty}$ that maximizes

$$\begin{aligned} &\sum_{t=0}^{\infty} \beta^t [u(S_t|X_t) + \varepsilon_{St}] \\ \text{subject to } X_{t+1} = &\begin{cases} (0, x_{t1} + s_{t1}, x_{t2} + s_{t2}, x_{t3} + s_{t3}) & \text{if } t \text{ is the end of an academic year} \\ X_t + S_t & \text{otherwise} \end{cases} \end{aligned} \tag{2.1}$$

Flow utility u is a function of the choice of patient allocation shares in period t and is conditional on the state of resident skill X_t and also includes a component that is observable to the hospital staff but not to the econometrician, ε_{St}. This term is different for every allocation choice S and period t and could reflect things such as congestion or features of the patient that make them particularly suitable or costly for training that I do not observe.

This setup leads to the standard Bellman equation describing the value of being in any particular state X given by

$$V(X, AY(t)) = E_{\varepsilon}\left[\max_{S}\{u(S|X) + \varepsilon_{St} + \beta V(X', AY(t+1))\}\right] \tag{2.2}$$

where I now explicitly separate out the resident's knowledge state X from the relative time within the academic year $AY(t)$. I do this to make clear that the value of being in state X differs based on when in the academic year the current period t is. That this should affect the value of being in knowledge state X is intuitive: assume that each period t is a day and consider the state in which all residents have zero skill: $X = [0, 0, 0, 0]$. In this case, it is

far less undesirable for the hospital to be in this state in the first day of the academic year ($AY = 1$), when it can still train the residents, than it would be for the hospital to be in this state on the final day of the academic year ($AY = 365$), where the training utility for the senior residents is about to be realized.

The model is very general and can accommodate any objective function. Motivated by the stable average patient length of stay in the data shown in Chapter 1, Figure 1.5 and the finding that length of stay is the main patient-relevant outcome that improves with experience, I consider a utility function where the hospital maximizes utility from training subject to a lower bound of utility from average patient length of stay.

$$\begin{aligned} \max_{S} \quad & \sum_{t=0}^{\infty} \beta^t [K(S_t; X_t, AY(t)) + \varepsilon_{St}] \\ \text{such that} \quad & L(S_t; X_t) \geq L^* \text{ for all } t \end{aligned} \tag{2.3}$$

L is the hospital's utility from patient length of stay and is the average length of stay utility f for patients, given allocation S and skill X.

$$L(S_t; X_t) = \sum_{c=1}^{4} s_{tc} f(x_{tc}) + \left(1 - \sum_{c=1}^{4} s_{tc}\right) f(x_A) \tag{2.4}$$

This is simply the share of patients allocated to the residents in each cohort $c \in \{1, 2, 3, 4\}$ and attendings in each period t, times the length of stay utility f for providers with each skill. Length of stay utility f is increasing and concave in provider skill x. Given the hospital's choice of L^*, the hospital conditionally maximizes the discounted sum of K, the utility from training. The minimum quality threshold L^* must be satisfied in every period and is to be estimated.

Flow utility from training K is the cumulative skill of the graduating senior residents after they graduate. That is,

$$K(S_t; X_t, AY(t)) = \begin{cases} f(x_{t4} + s_{t4}) & \text{if } t \text{ is the end of an academic year} \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

In this specification, K is only nonzero in the period just before graduation. Note does this does not automatically mean the hospital chooses not to train if it is not in the final period because the value function $V(X, AY(t))$ will generally be increasing in the state X within each quarter: higher levels of skill at any point enable the residents to achieve a higher skill upon graduation. It only means that the hospital does not directly derive utility from resident skill within the academic year–it does not benefit from increased resident skill within the academic year beyond its effect on length of stay utility L.

This model nests many potential theoretical models of hospital behavior. First, patients clearly demand care quality and may view speed of care as an important component of quality. Second, given the time cost of training, teaching hospitals face a trade-off between

patient revenue and training. That said, many hospital administrators view training as the more important goal: "Trustees and administrators of teaching hospitals were charged with making their institutions academic leaders, not financial profit-centers. Fiscal responsibility was required for the institutions to do good work, but ultimately teaching hospitals were measured by their academic and professional accomplishments rather than their balance sheets" (Ludmerer, 2005). The model is able to accommodate the full range of potential weights between revenue and training. Setting $L^* = -\infty$ represents the extreme where hospitals only care about maximizing training. On the other extreme, if they only care about care quality or revenue, then they would choose L^* corresponding to the level of quality that would be provided if attendings provided all care independently such that training is impossible if the quality constraint is to be achieved. Convex combinations of the two objectives are accommodated with intermediate values of L^*.

This objective function is similar to the canonical specifications of Newhouse (1970) and Lakdawalla and Philipson (1998). I include revenue through the hospital utility channel, which is a function of patient length of stay. But length of stay is inextricably linked with revenue: in general, hospitals receive more revenue for each additional patient they see. Because facility size is fixed, the only way to increase revenue is to increase patient throughput via shorter length of stay. I do not separately include revenue to avoid double-counting care quality and revenue. In Appendix A.2, I consider an alternative flow utility function where the hospital maximizes a weighted sum of throughput and training, with the weight to be estimated, but reject it because it is inconsistent with the data: the optimal patient assignment rule does not result in a stable steady-state.

Note that even with constrained maximization, the dynamic problem does not reduce to a static problem. This is because training today and training tomorrow are intertemporal complements. More training today means that residents are faster tomorrow, and if residents are faster tomorrow, then the cost of additional training tomorrow is lower and more training is possible. Therefore, the myopic or static approach of maximizing the senior cohort's training and ignoring the junior cohort in each period is not optimal. With that approach, training for the current seniors will be maximized, but future cohorts will suffer. Instead, the hospital must take into account the intertemporal complementarities and train both cohorts in every period. The optimal division of training across cohorts and time is the result of solving the dynamic problem.

Next, I describe how the model incorporates the key trade-off between care quality and training. For each level of minimum patient care utility L^*, there is a corresponding maximum level of training that can be achieved. Therefore, one way that the hospital can control the trade-off between care quality and training is by tightening or relaxing the care quality threshold. If the hospital relaxes the care quality threshold, then more patients can be allocated to slow, inexperienced residents, who will then become faster in later periods and have higher skill upon graduation. Increasing attending skill x_A has a similar effect as relaxing the threshold L^*. Faster attendings allow for additional patients to be allocated to lower-skilled residents while maintaining mean care quality above L^*. Finally, if the rate of learning,

or the slope of care utility with respect to skill $\frac{df}{dx}$, increases, then the benefits of patient allocation to residents increase. In this case, the hospital will desire to allocate additional patients to residents, but the degree to which it can do so is limited by the quality constraint L^*. In the first counterfactual, I quantify the amount that training decreases when L^* is tightened and consider the effectiveness of mitigating changes such as increasing attending skill and increasing the rate of learning.

Because the hospital maximizes training subject to the quality constraint, it maximizes the sum of discounted outcomes of graduating residents and the patients the residents see in the future. But the objective is one of pure efficiency: if there is a disruption to training that results in one cohort being trained less than the steady-state, the hospital will not take action to "smooth out" resident skill across cohorts beyond what is optimal given the concavity of training utility K. The shape of the optimal training function for each cohort given the skill of continuing cohorts and the corresponding impulse response function govern how many future cohorts are affected by disruptions to a single cohort's training. These are empirical questions where the answers may vary based on the estimated parameters of the model. I quantify the impact in the second counterfactual and consider how one-time changes to the learning environment, such as a temporary relaxation of L^*, increase in attending skill, and increase in the rate of learning mitigate the impact of a training disruption.

2.3 Estimation and Identification

2.3.1 Simplifying Assumptions and Parameterizations

In order to make progress, I make some simplifying assumptions and parameterizations that keep the problem manageable and facilitate estimation. I describe these assumptions and the rationale behind them in this subsection.

The first set of simplifying assumptions I make keeps the state-space and action-space manageable. I assume that the program lasts for two years, so that there are only two cohorts: new residents, or "juniors," who I denote with subscript j, and residents who will graduate at the end of the academic year, or "seniors," who I denote with subscript s. Hence, $c \in \{j, s\}$. I further assume that there is no within-cohort variation in skill. I take the time period t to be a quarter of the academic year, meaning that $AY(t) \in \{1, 2, 3, 4\}$ returns the quarter of the academic year t belongs to. Without loss of generality, I impose that a mass of measure 0.25 complex patients arrives each quarter, so that each academic year a unit mass of patients arrives. Consequently, the maximum steady-state value of resident knowledge is 1.0, achieved when every patient is assigned to one of the cohorts. I discretize both the state space of resident knowledge and the choice variable of the share of patients assigned to each cohort and to attendings working independently.

These simplifications are necessary for the following reasons. First, managing a continuous choice of patient share and experience is intractable when taking first-order conditions is not possible, which applies here because the value function is unknown (V in Equation

(2.2)). Instead, I consider an interval of knowledge space $[0.01, 1.2]$ and discretize it into 200 evenly-spaced values. I choose 0.01 as the starting value because residents begin transitioning into the program in June prior to their first year and they see a share of complex patients that is equivalent to about 0.01 when dividing by the full quarter. This is also necessary because I need a finite value for the natural logarithm of experience. I censor the upper bound of the knowledge space to 1.2 because the maximum steady-state allocation for rising senior residents is 1.0 (achieved if every patient in every quarter is assigned to them). I choose 1.2 rather than 1.0 because this reduces estimation error when rising senior skill is near 1.0: otherwise, the hospital will begin allocating fewer patients to senior residents because it cannot benefit from the increased skill. For example, if the upper bound was 1.0 and rising seniors had 0.9 starting knowledge, the hospital's incentives to allocate more than 0.1 patients to the seniors is drastically reduced because there is no additional training benefit to doing so. In keeping with the discretization of the state-space, I constrain the hospital to dividing the mass of arriving patients into increments that correspond to the grid of valid knowledge values. This means the hospital chooses one of 42 values evenly spaced from 0 to 0.25 to allocate to each cohort and to the attendings such that the sum of allocations to all providers equals 0.25. The full state-space must contain one dimension for each cohort, and must have one such array for each time period.[4] Even with just two cohorts and time in quarters, the state space array with my discretization has dimension $[200, 200, 4]$.

The second set of assumptions concerns the steady-state. I assume that the hospital is in the steady-state and that the steady-state is such that training each year is identical to training every other year. The first assumption facilitates estimation because it does not require me to infer the skill of the rising seniors–this is not data because I do not observe their full patient history. The assumption of a stable steady-state is consistent with the patient assignment data. In Appendix Figure A3, I show that patient assignment shares are similar for each class across the two years of data, suggesting that cohorts are treated similarly. It also is consistent with the intuition that a teaching hospital would treat all cohorts similarly. The assumption rules out models and parameters where in the optimal solution the hospital alternates between training cohorts that enter during even years and ignoring the cohorts that enter on odd years.

Estimation proceeds in two steps in the spirit of Hotz and Miller (1993), Bajari, et al. (2007), and Pakes, et al. (2007). The unknowns and methodology are summarized in Table 2.1 Panel (a). In the first, "offline," step, I estimate the parameters relating to the learning rate $\{\alpha_0, \alpha_1\}$ using OLS in the panel data, and infer attending speed x_A using the estimates $\hat{\alpha}$. Then, for three candidate values of the discount rate[5] β, I find via iteration the lower bound on quality in the utility function given by Equation (2.3) that produces optimal patient assignment shares most similar to the observed shares. I consider three functional

[4]This is because the value to the hospital of any level of resident skill is potentially different depending on when in the academic year it is.

[5]I choose not to estimate the discount rate, which is a common choice in the dynamic model literature. For instance, Pakes, et al. (2007) writes, "We usually think that the prior information we have on δ [the discount rate] is likely to swamp the information on δ available from estimating an entry model."

Table 2.1: Summary of Estimation Parameterizations and Methodology

(a) Parameters to Estimate and Methodology

Category	Unknowns	Estimation Methodology
Learning rate	$\{\alpha_0, \alpha_1\}$	OLS in panel data, "offline"
Attending speed	x_A	Back out from panel data using $\hat{\alpha}$
Discount rate	β	Calibrated; yearly $\beta = \{0.90, 0.95, 0.99\}$
Weight on quality vs. training	ϕ	Dynamic, match patient shares averages
Lower bound on quality	L^*	Dynamic, match patient shares averages

(b) Parameterizations of Length of Stay Utility f

Linear	$f(x) = -\alpha_0 x^{\alpha_1}$
Quadratic	$f(x) = -(\alpha_0 x^{\alpha_1})^2$
Log	$f(x) = \log(C - \alpha_0 x^{\alpha_1})$

Notes: This table enumerates the unknown parameters and the estimation method employed in order to estimate them, as well as the functional forms for the length of stay utility function f. In Panel (a), the first section lists the parameters to be estimated in "offline" in panel data without any dynamics. The middle sections shows that the discount rate is calibrated using various reasonable yearly values, as it is not well-identified in the dynamic model. The final section shows the two parameters that are estimated using the dynamic model and take the offline parameters as fixed; these come from two different utility functions that the hospital may use. Panel (b) lists the three functional forms used for the length of stay utility f. Note that the linear and quadratic parameterizations do differ because the shape parameters α are fixed in the offline estimation. C is a constant chosen to ensure that $C - \alpha_0 x^{\alpha_1}$ is positive for all values of skill x. See text for additional details, as well as Subsection 2.3.2 for more details on offline estimation and Subsection 2.3.3 for more details on the dynamic estimation.

forms for the length of stay flow utility f that vary in concavity: linear, quadratic, and log, as delineated in Table 2.1 Panel (b). Note that the linear and quadratic parameterizations of f differ because the shape parameters α are determined in the offline estimation.

2.3.2 Step 1: Offline Parameter Estimation

I begin by estimating the learning parameters outside of the dynamics, or "offline," via OLS. The goal is to recover how patient length of stay improves as residents gain experience with complex patients and to estimate the skill of attendings working alone. Experience is measured as the cumulative fraction of complex patients seen, which is valid when equal numbers of patients arrive every quarter as in the data. Two factors make this not entirely straightforward. First, I must restrict to the subset of residents who begin the program

during the sample period because I do not observe the resident's full history of patients seen otherwise. Second, the residency schedule is such that the residents work at another hospital in the city that I do not have data from, meaning that I must infer the total fraction of complex patients seen by each resident. I first describe the assumptions I make and then the tests I do in order to test the validity of the assumptions.

The coefficients recovered by OLS are unbiased under the same assumptions on omitted variables as outlined in Chapter 1, Section 1.4. Two additional assumptions are necessary in this setting. First, I assume that the natural logarithm is the correct functional form for resident progress with respect to the cumulative fraction of complex patients seen. Second, I assume that patient assignment inference is in expectation correct. In other words, residents who see "excess" patients relative to their peers at the hospital from which I have data also see similar proportions of "excess" patients at the other location. Both assumptions are fundamentally untestable but I offer arguments in favor of accepting them.

First, to test the validity of the functional form assumption, I compare the quarterly patient share results for the subset of residents with the results using years in the program as the measure for experience. The rationale behind this is that years in the program is a "reduced-form" measure of share of complex patients, as much of the variation is in the time-series rather than in the cross-section. I test the validity of my assumptions in Table 2.2 by seeing if the results for continuous tenure are similar to that of quarterly fraction of patients seen.

Next, because UCSF EM residents work in two locations but I only have data from one, I must infer patient assignment at the other location. The assumption I make is that patient assignment at the other location (Zuckerberg San Francisco General Hospital, ZSFG) mirrors patient assignment at the observed location (UCSF). In other words, if a resident sees 12% of patients at UCSF within a period, I assume they are also seeing 12% of patients at ZSFG. Importantly, this means that attendings are not assigning patients in a mean-reverting manner or that observed differences at UCSF are not magnified or diminished at ZSFG.

I find evidence consistent with this assumption in Appendix Table A6, which shows that the standard deviation of complex patients seen per shift is relatively stable across the academic year. The rationale is that if residents were assigned more patients at UCSF to "make up" for seeing fewer patients at ZSFG for exogenous reasons such as ED congestion, then I would expect to see more dispersion in the number of patients per shift in earlier academic quarters compared to later quarters. This is because of the law of large numbers: in later quarters, variation in patients seen due to exogenous factors should be more similar across residents and any additional variation will have a smaller impact on the total number of patients seen. Furthermore, I believe factors such as the ad-hoc team structure and variation in congestion and patient arrivals make this assumption reasonable, as it is difficult for the rotating attendings to know the resident's history and adjust their assignment instructions accordingly.[6] To lessen the impact of this assumption as well as differences due to exogenous

[6]According to EM residents at UCSF, many decisions regarding progress are made at the cohort level.

factors such as congestion, I use as the measure of patient-specific experience the average of complex patients seen during the calendar quarter. This measure has considerably less variation than experience at the two-week level, but still contains some variation, as can bee seen in Appendix Figure A4.

Now, I describe how I infer attending skill with the learning parameters in hand. First, note that attending physicians are included in the regression samples with their value of log(experience) equal to zero throughout, and with a single, "pooled" fixed effect for all physicians when physician ID fixed effects are included. The method in the specifications without physician fixed effects is simple. Because I observe the length of stay for complex patients seen by attending physicians and I know the functional form of learning, I just take the inverse of that function in the set of patients seen by attendings. For the specifications with both physician fixed effects and patient controls, I first normalize to zero the physician fixed effect of the most prolific first-year resident. Patient controls are relative to the modal patient. Hence, the constant represents the average log(length of stay) for the modal patient seen by this first-year resident when they have log(experience) equal to zero. To get attending skill, I assume that their "individual fixed effect" is identical to that of the most prolific first-year resident. Therefore, the pooled attending physician fixed effect represents α_1 times the natural log of their experience.

2.3.3 Step 2: Dynamic Parameters

The goal of estimation is to find the unknown parameter L^* that gives the best fit between the model-predicted optimal quarterly patient shares and the observed quarterly patient shares. The metric of fit used is RMSE, with each quarter receiving equal weight. In other words, I find the value of L^* that minimizes

$$\sum_{t=1}^{4} \sum_{r \in \{j,s,a\}} \sqrt{(s_{tr} - MSS(L^*;\beta)_{tr})^2} \tag{2.6}$$

where the subscripts j, s, and a represent the shares assigned to each role: the junior resident, senior resident, and attending. s_t are the observed patient allocation shares, and MSS_t is the model-predicted steady-state shares given the parameter L^* and discount rate β in period t.

In order to find the value of L^* that minimizes Equation (2.6), I perform a grid search over values of L^*. For each value of L^* and choice of β, I first perform value function iteration on Equation (2.2) in order to solve for $V(X, q; L^*, \beta)$, the value of being in knowledge state X in academic quarter q given L^* and β. I then use the estimated $V(X, q; L^*, \beta)$ in conjunction with the flow utility K to find the optimal patient allocation choice S for each X and quarter q: $S(X, q; L^*, \beta)$. Finally, I find the steady-state given the allocation choices S. That is, I

For example, at the beginning of second year all residents are expected to take on additional patients and there is limited "personalization" of this directive based on individual progress. This is unsurprising because of the ad-hoc team status and because there are 60 EM residents for the various attendings to keep track of.

search for a value of rising senior knowledge $\sum_{t=t'}^{t'+4} s^*_{tj}$, the cumulative share of patients seen in their first year, such that the optimal training results in the new cohort of junior residents finishes the first year with the same knowledge. In my notation, I search for $\sum_{t=t'}^{t'+4} s^*_{tj}$ that satisfies for all t':

$$\begin{aligned} &\sum_{t=t'}^{t'+4} s^*_{tj} = \sum_{t=t'+5}^{t'+8} S(X^*_t, q(t); L^*, \beta)_j \\ &\text{such that } X^*_{t+1} = X^*_t + S(X^*_t, q(t); L^*, \beta) \\ &\qquad \text{and } t' \text{ is the first quarter of an academic year} \end{aligned} \tag{2.7}$$

The left hand side is the starting knowledge of the rising seniors, which is equal to the cumulative share of patients seen in their first year. The right hand side is the sum of patient shares seen by the new juniors in the next academic year, because S is the function that maps knowledge X and time q to a vector of optimal patient assignment decisions and skill X accumulates in the usual way.

I iterate until the average L2-norm (Euclidean distance) between successive elements of the value function is less than 10^{-6}. Although the grid search over possible values of L^* is slightly cumbersome, this method has the advantage that I in theory do not risk finding a local minimum rather than the global minimum. In practice, I begin with a relatively coarse grid and perform a finer grid search around the minimum given by the coarse grid.

2.4 Results

In this section, I first present and discuss estimates of the offline estimation and provide evidence in support of my assumptions. Then, I discuss the results for the dynamic model.

Results of the offline OLS estimation of the learning parameters are in Table 2.2. My preferred estimates are the bolded set in the rightmost column, which are the results using the natural logarithm of quarterly patient share with physician fixed effects and patient controls. Starting from the bottom row, the results suggest that the average attending physician has skill similar to a resident with a cumulative experience share of 2.0 patients (recall that a mass of 1 patient arrives each year). While this may be lower than expected, this measure includes interruptions to attending speed due to supervisory duties so it is not a measure of pure attending skill. Next, for learning speed, the results suggest that for each 1% increase in cumulative quarterly patient share, patient length of stay will decrease by about 0.048%.

The remaining columns of Table 2.2 are tests of the assumptions necessary for the patient share results to be valid. I compare those estimates to estimates from the full sample of residents and for continuous measures of experience. The first pair of columns presents the results using years in the program as a continuous measure and for the full sample of residents, which represent the baseline. Next, I restrict to full history residents and find minimal changes in the estimated coefficients. Coarsening the years of experience measure to quarterly snapshots results in coefficient estimates of increased magnitude, but the standard

errors are large enough that I cannot reject that they are equal to the coefficients from the continuous measures. Similarly, changing the measure of experience to patient share does not create a large difference in the estimates. The estimated attending skill is similar for all specifications other than the full resident sample, which I believe is due to some imprecisely estimated resident fixed effects that have outsize influence on the grand mean of inferred attending skill. Values across the other specifications with patient controls are all similar.

Results for the dynamic model are similar across specifications and discount rates. Regardless of the specification or discount rate, the estimated lower bound, converted from utils into hours, is always around 6.6 hours per patient. The model that fits the data best[7] is the specification with quadratic utility from length of stay and a yearly discount rate of $\beta = 0.95$. Full results can be seen in Appendix Table A5. The model fits the qualitative patterns of increasing allocation to first year residents and decreasing allocation to attendings well, but the gradient for assignment to attending physicians is steeper than observed in the data. This can be seen in Appendix Figure A5.

I next evaluate model fit by examining how well it fits non-targeted moments. Specifically, I examine how the length of stay predictions compare to length of stay averages in the data. First, I examine the average length of stay across the academic year. The estimated quality bound of 6.61 hours is greater than the raw median length of stay but slightly less than the raw mean in the data. The reason it differs is related to the fact that the inferred constant is calculated from the modal patient and the most prolific first year resident, so the levels may differ from raw average values in the data. Therefore, when assessing model fit, I will focus on matches with changes over calendar time. The model predicts that length of stay is stable over the academic year, as the quality bound binds in every quarter. Figure 2.1 Panel (a) shows that the median ED length of stay is very stable with respect to academic quarter, just as the model predicts. The mean shows more movement, but that is driven by the top 25 percent of encounters and potentially related to encounters where patients were in worse condition than expected or the affected by the arrival of a code patient in critical condition who demanded the attention of the entire ED. Next, I examine how average length of stay varies across quarters of experience, and compare it to the model predicted values. Figure 2.1 Panel (b) shows that see that average length of stay by resident experience predicted by the model has similar shape as the median length of stay in the data. The similarities between average length of stay in the data and predicted length of stay in the model were not a moment that was targeted in the estimation–only patient share assignment was–and the comparisons give me more confidence in the estimates.

[7] If I strictly choose by model fit, then the quadratic model with $\beta = 0.99$ fits better. However, examination of the fit parameters in the neighborhood of the best fit threshold suggests that this is likely an artifact of choosing a finite grid. Estimates for $\beta = 0.95$ and $\beta = 0.99$ using a coarser grid are similar than with the grid sized used in the main estimation. Furthermore, $\beta = 0.95$ is likely a more realistic discount rate given that hospital management and residency directors have finite terms.

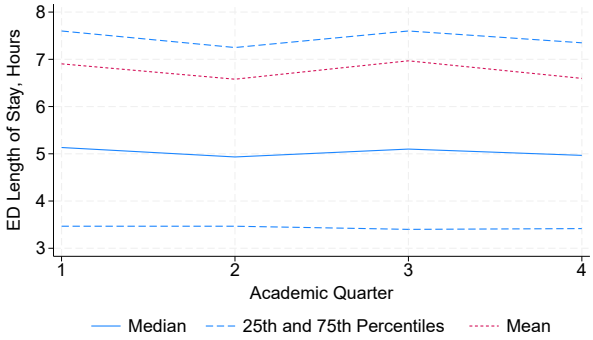
Table 2.2: Offline Parameter Estimates of Learning Speed and Attending Skill

	log(Patient Length of Stay, Hours)							
Experience Type	Tenure (continuous)				Tenure (quarterly)		Patient Share (quarterly)	
α_1: log("Experience")	-0.044	-0.049	-0.067	-0.062	-0.1272	-0.112	-0.036	**-0.048**
	(0.028)	(0.036)	(0.014)	(0.019)	(0.029)	(0.036)	(0.010)	(0.015)
α_0: Constant	1.674	1.879	1.668	1.869	1.673	1.882	1.663	**1.837**
	(0.012)	(0.012)	(0.017)	(0.064)	(0.018)	(0.064)	(0.019)	(0.065)
Sample	All		Full History Residents					
MD ID and Patient Ctrls		Y		Y		Y		Y
Observations	10,255	10,255	2,527	2,527	2,527	2,527	2,377	2,377
Inferred x_A	0.88	5.34	0.84	2.65	0.95	1.91	0.40	**2.00**

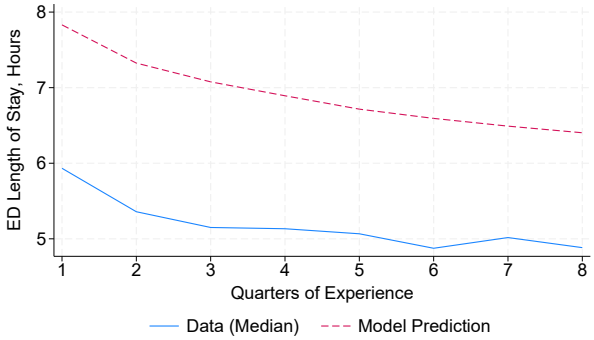
Notes: This table shows the results of the offline estimates for the parameters governing learning speed as well as the value of attending skill inferred using these estimates. The bolded estimates in the final column are what are used in the dynamic estimation as they use the desired measure of experience: share of complex patient seen. This measure is only available for residents who begin the program during the sample (the 2018 and 2019 cohorts). In order to test that this sample of residents is not significantly different from the full sample and that discretizing experience to quarterly intervals similarly does not result in different estimates, I begin with the full sample of residents and a continuous measure related to patient share: years in the program. This first pair of regressions is similar to the regressions in Chapter 1, Table 1.3. I then restrict to the 2018 and 2019 cohorts and find that the coefficients do not change much in the specification with physician fixed effects and patient controls. Similarly, I cannot reject equality of coefficients when I only allow tenure to change each quarter, and again when using the quarterly patient share definition of experience. Standard errors are clustered by physician in the first two columns but are the greater of the clustered and robust otherwise because there are fewer than 40 residents who identify α_1.

Figure 2.1: Model Fit: Non-Targeted Moments

(a) Average Length of Stay



(b) Average Length of Stay by Experience



Notes: These figures show how well the model fits the non-targeted moments relating to patient length of stay. In Panel (a), a reproduction of Chapter 1, Figure 1.4, we see that the median ED length of stay is very stable with respect to academic quarter, just as the model predicts. The mean is slightly less stable, but that is driven by the top 25 percent of encounters. In Panel (b), we see that average length of stay by experience has approximately the same shape as the median length of stay for residents with each level of experience. The levels differ because the model fits patterns for the modal patient, who are sicker than the empirical average patient median shown.

2.5 Counterfactuals

I use the model to assess the consequences of a policy change and of a training disruption on both patient care quality and resident skill. In addition to quantifying the impact of the changes, I consider the effectiveness of various remedies that the hospital may enact in order to counteract the effects of the counterfactual changes.

The first remedy I consider is an increase in the speed at which attendings see patients independently, which in the model is represented by x_A. This is a feasible action because it does not necessarily require that the hospital hire higher-skilled attendings. Instead, they can simply staff more attending physicians on each shift. This works because x_A includes the supervision portion of the attending's duties. If there are additional attendings working on each shift, then supervisory duties will be split among more physicians, thereby reducing the number of disruptions each attending faces when caring for patients individually. This will reduce length of stay for patients assigned to attendings and increase their effective speed x_A. This remedy is also realistic: a similar change was proposed by the Institute of Medicine in 2009: they estimated that $1.7 billion was required to improve residency, with the bulk of the spending for more providers to assume some of the patient load currently seen by residents, thereby allowing them more time to reflect, study, and learn (Ulmer, et al., 2009).

Second, I consider an increase in the learning rate of residents. This is potentially more

difficult to implement because it would likely involve redesigning the curriculum or partnering with additional hospitals so that residents see additional patients.[8] On the other hand, it could also have significant returns because residency shapes the habits and approaches that physicians will continue to use throughout their careers (Ludmerer, 2015).

For the first counterfactual, I consider permanent changes, but in the second counterfactual, the changes are for one period only. In the second counterfactual, I also explore the effectiveness of a temporary relaxing of the care quality constraint L^*. To simulate the counterfactuals, I change the relevant parameters and apply the model with other parameters held fixed to find the new optimal patient assignment function. In the first counterfactual, changes are permanent so I re-solve the model following the same procedure as outlined in Subsection 2.3.3. In the second counterfactual, changes are for one period only, and the hospital knows that they are temporary. Hence, the optimal allocation of patients is one that maximizes current cohort utility plus the discounted value function corresponding to training for the junior cohort, where the value function is the one from the steady-state.

The two teaching outcomes I consider are average patient length of stay and the total number of patients seen over the resident's career. For now, I make the extreme assumption that no further learning occurs after the resident graduates from the program.[9] Under this assumption, calculating average patient length of stay is straightforward: it is simply the average length of stay given by the resident's skill upon graduation. This is equivalent to the intensive margin of patient utility: for each patient the resident sees, what is the difference in their length of stay? Estimating the total number of patients seen requires making additional assumptions. I assume that graduates see $\frac{8}{\alpha_0 X^{\alpha_1}}$ patients per shift, where X is the skill they leave residency with, and that they go on to work 18 8-hour shifts per month (AMA, 2017) for 30 years. Differences in total patients seen represent the extensive margin of the change.

2.5.1 Increasing the Bound on Quality

In the first counterfactual, the hospital decides to increase the lower bound of care quality. In the model, this is governed by an increase in L^*. There are real reasons for why hospital administrators may choose to make this change. The first is that length of stay is an important part of Medicare's Hospital Report Cards.[10] Hospital administrators may care about these ratings both because higher ratings help attract more patients and for intrinsic or reputational concerns (Kolstad, 2013). The second may be due to payment reform, which is a heavily-discussed policy lever to reduce healthcare costs (see McClellan, 2011). I next

[8] EM Residents typically are not constrained by the ACGME's hours limit so this change would be legal, but it ignores general equilibrium effects, such as the possibility of slower learning due to increased fatigue or a change in selection into specialties (cf. Wasserman, 2023).

[9] In progress is a version where graduating residents learn at half the speed as they did during residency. This approximation takes into account the facts that attendings work fewer shifts per month than residents and they no longer have formal supervision for every patient.

[10] See the "Timely and Effective Care" subsection of Medicare's Care Compare website (accessed October 25, 2023): https://www.medicare.gov/care-compare/

explain how payment reform may interact with the hospital's choice of care quality bound L^*.

Generally, payments from both private and public insurers have been trending away from the traditional fee-for-service (FFS) system to alternatives such as value-based payments and capitated, prospective payment systems (PPS). Under FFS, providers are paid for every procedure, order, and service they provide to the patient. One of the issues with this system is that providers are not incentivized to reduce utilization or cost and have financial incentives to provide marginally necessary care (cf. Marmor and Gordon, 2021). Two leading alternatives are value-based care and PPS. In value-based care, providers are paid more if they realize better quality outcomes regardless of utilization, for example, for lower rates of complications from surgery or shorter ED length of stay. Because payment is independent of utilization, value-based care incentivizes physicians to reduce cost and improve quality. PPS works similarly in that providers are paid the same amount for every patient type regardless of utilization, so again providers have incentives to reduce costs.

Issues may arise for teaching hospitals if the reimbursement rates for value-based care and PPS are set uniformly across hospital types. An example would be if under the two systems, the average hospital's revenue is identical. Teaching hospitals would lose revenue due to a change like this because they tend to do poorly on many typical quality and efficiency metrics (Kocher and Wachter, 2023). Furthermore, teaching hospitals currently have very high FFS reimbursement rates, estimated at 10-20% above FFS payments at non-teaching hospitals, although quality of care is higher for some patient types, which offsets the additional cost somewhat (Sloan, 2021). In any case, if the switch from FFS to PPS or value-based care occurs without sufficient accommodations for teaching hospitals, they would stand to lose significant revenue from patient care. This is the spirit motivating the first counterfactual.

Consider a very simple payment model where instead of being paid for each hour with patients (similar to FFS), hospitals are instead paid a fixed amount for each patient (similar to PPS). Further assume that all hospitals began with the same FFS rates and that the PPS rate is set so that the average hospital in the nation does not experience a change in revenue. My results suggest that UCSF could see at least 8% more patients each day if they did not train at all and instead had attendings see all of the patients.[11] Therefore, under this simple payment structure, they would lose 8% of revenue.[12] The hospital could recover some of the lost revenue by increasing patient throughput via increasing the quality constraint L^*. Particularly in the short run, the hospital must reduce training since residents are slower

[11]This is calculated from the model predictions for average length of stay during the academic year given optimal patient assignment under the current parameters, and the inferred value of attending skill. It is a lower bound because the current attending skill measure assumes that attendings also have teaching and supervisory duties, which would be reduced if the hospital reduced teaching.

[12]If the teaching hospital had higher FFS reimbursement rates than non-teaching hospitals prior to the policy change, then it would stand to lose even more revenue. Additionally, while it is true that even in the current world, the hospital could increase revenue by training less, it has chosen not to. This is because the hospital has chosen L^* at the current level from maximizing preferences over care quality, quantity, revenue, and teaching. I take this choice given and do not model it. As long as revenue is a normal good, changes that decrease revenue will cause the hospital to seek ways to increase it.

than attendings in order to increase throughput, and this is precisely what occurred in the 1980s when the first PPS reforms were implemented (Ludmerer, 2015). However, there are two mitigating actions the hospital can take. First, they could increase the rate of learning α_1 so that residents gain more skill with each patient seen. Second, they could increase the speed of attendings seeing patients independently. In the counterfactual, I assume that the hospital chooses to become 2% faster at caring for complex patients.

Table 2.3 shows the impact of the increase in the patient quality constraint alone and in combination with mitigating actions. The figures presented compare the new steady-state with the current steady-state, and ignores contributions to revenue, patients seen, and minutes per patient during the transition period. The first row reports current outcomes. Under the assumptions on speed and shifts worked described above, residents see 8,165 complex patients over the course of their career and spend 381 minutes per patient. If the hospital adjusts training to increase the length of stay by 2% and makes no further changes, then in the new steady-state graduating residents see 186 fewer patients during their career and are almost 9 minutes slower for each individual patient. The 2% gain in teaching hospital revenue from seeing patients faster is paid by the hospital that employs the resident after graduation because its new physicians are slower, and this cost is over 4 times larger than the revenue gain. This future cost is an externality from the teaching hospital's point of view because it undervalues the future productivity of their graduating residents when making the decision to reduce training.

However, the social planner can induce the teaching hospital to take mitigating actions and reduce the impact on training required by the 2% stricter length of stay requirement. For instance, teaching hospitals can increase the speed that attendings see patients independently by 5%. This is very effective in reducing the difference in counterfactual training from the current training level, making up 65% of the loss relative to when no other actions are taken. The intuition behind this is that the hospital desires to maximize training given a constraint, and increasing the speed at which attendings work in effect makes the constraint less binding. This allows them to increase training while still meeting the quality constraint. On the other hand, increasing the rate of learning by 5% less effective, only allowing the hospital to make up 53% of the loss. This is again due to the constrained maximization problem faced by the hospital. Although the benefits of training are increased, the hospital has difficulty taking advantage of this and increasing the fraction of patients allocated to residents because it still must meet the same care quality constraint in every quarter. In other words, the hospital is not permitted to intertemporally substitute decreased speed in earlier quarters due to increased allocation of patients to residents with increased speed in later quarters because residents have gained more skill as it would in the absence of the constraint. Only in the later quarters of the academic year, when residents who learn faster are more skilled can the hospital actually increase patient allocation relative to current levels, and even then it is not by much. Finally, taking both actions actually has the effect of slightly improving future outcomes, as now the hospital is able to take advantage of the increased benefits of faster learning and actually allocate additional patients to residents.

Table 2.3: Quality Bound Changes: Steady-State Counterfactual Resident Training and Mitigating Factors

Plan	Revenue "Externality"	Lifetime Patients	Minutes per Patient
Current Outcomes		8,165	381
Decrease length of stay 2% and...			
No other changes	-4.1:1	-186	+8.9
Attending Speed +5%	-1.4:1	-65	+3.0
Learning Rate +5%	-1.9:1	-87	+4.1
Learning Rate +5% & Att +5%	+1.1:1	+50	-2.3

Notes: This table shows the loss for future patients of senior residents if UCSF decides to decrease patient length of stay by 2%, adjust the care quality utility constraint by the corresponding amount, and take the listed mitigating action. The revenue externality captures only the financial cost and is the ratio of the change in the present value of future patient revenue to the current patient revenue increase due to the policy change, assuming no changes in reimbursements. For example, if no other changes are taken, the present value of the cost to future employers of the resident is 4.1 times the additional revenue generated by UCSF by speeding up, because the graduating residents have less skill. This is in essence the discounted difference in lifetime patients seen and assumes that residents go on to work 18 shifts per month, as is typical for EM attendings, for 30 years. Lifetime patients is the total difference in patients seen (extensive margin), and minutes per patient is the length of stay difference for each patient (intensive margin) given graduating resident skill. For now, I make the extreme assumption that no further learning occurs post-residency.

These results show that small changes in training by the teaching hospital can have outsize effects for future patients and future employers of residents. It is important for policymakers to consider these externalities when designing payment systems so that future patients do not end up paying orders of magnitude greater in costs in order to save a little today. Fortunately, there are feasible and straightforward remedies available that can mitigate these costs. The counterfactual shows that increasing the speed at which attendings see patients individually by just 5% can recover 65% of the loss in training resulting from a desire to increase patient throughput by 2%. This can be satisfied by staffing additional attending physicians, which admittedly may be difficult if general equilibrium effects are considered, but is likely far easier and more effective than finding ways to redesign the resident curriculum or having them work additional hours in order to learn faster. Another possibility that I have not discussed is to increase Medicare's Indirect Medical Education Payment to cover non-Medicare patients. These payments are a lump-sum bonus paid by Medicare to Medicare PPS academic hospitals, which CMS recognizes have higher costs than non-teaching hospitals due to the teaching responsibility. Increasing the IME Payments would also be effective as it would decrease the hospital's need to gain more revenue from patients due to payment

reform, thereby reducing the need to increase patient throughput by decreasing training.

2.5.2 Unexpected One-Time Training Disruption

In the second counterfactual, I consider the consequences and effectiveness of policy responses to a one-time, unexpected disruption in training. This scenario resembles disruptions during the Covid-19 Pandemic, which affected residents in at least two general ways. First, the composition of patients who went to the hospital changed as patients delayed and avoided both routine and urgent or emergency care (Czeisler, et al., 2020). This, in addition to the influx of Covid patients, changed the pool of patients that residents could see and learn from, decreasing effective patient share in every period. Second, medical workers were under extreme stress during this period (HHS, 2022), which also likely reduced residents' ability to learn.[13]

I ask three questions. First, how many incoming classes of residents will the one-time disruption affect through spillovers? Second, for affected residents, what are the long-run effects of the disruption on their future careers? Third, what temporary mitigating factors could reduce the impact on residents affected by the training disruption? For simplicity, I assume that future incoming cohorts are equally skilled as their historical counterparts even though this may be contrary to evidence (Jhajj, et al., 2022).

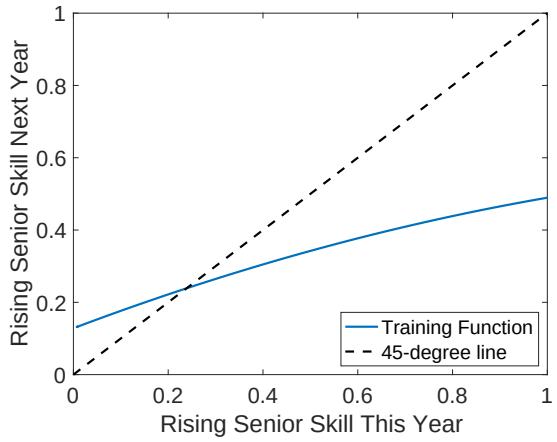
Under no additional changes, how long does the hospital take to return to the steady-state of training? The answer to this question reveals how many incoming classes of residents will be affected and is simply the impulse response function of the system. It can be inferred from the Optimal Training Function presented in Figure 2.2 Panel (a). The figure illustrates the utility-maximizing choice of patient allocation during the incoming cohort's first year (vertical axis) as a function of the skill of the rising senior (horizontal axis). The steady-state of the system is when rising senior skill today is equal to the resulting rising senior skill tomorrow, which occurs at the point which the optimal training function intersects the 45-degree line (the dashed line in the figure). In the figure, we observe that for decreases in rising senior skill today, the hospital trains the incoming cohort less than usual, but that the decrease is relatively small since the slope of the training function is relatively flat. Nevertheless, the number of affected cohorts depends on the initial size of the disruption as the hospital transitions back to the steady-state.

Figure 2.2 Panel (b) shows the impact on the rising seniors. This figure plots the graduating skill of the senior residents as a function of their skill when they become seniors. Below the steady-state, the decrease in training is greater than the gains above the steady-state. In combination with Panel (a), we see that under a training disruption, both cohorts are trained less than in the steady-state, but the decrease in training for the senior cohort is much greater than for the new incoming cohort.

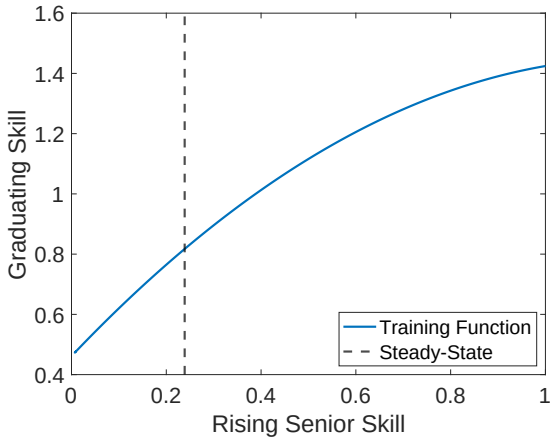
[13] As with K-12 education, medical student education during this time also suffered (Jhajj, et al., 2022) so it is reasonable to infer that residents were also affected.

Figure 2.2: Optimal Training Function and Outcomes

(a) Optimal Training Function



(b) Training Outcomes for Senior Residents



Notes: This figure depicts the hospital's optimal choice of total training for the two cohorts given the training the current senior residents received in their first year. Panel (a) depicts the total first-year training for tomorrow's junior residents in their first year given the training the current senior residents received in their first year. The steady-state is where the optimal training choice intersects with the dotted 45-degree line. Panel (a) illustrates that for values of current rising senior resident skill that differ from the steady-state, the hospital takes a few periods to "zig-zag" back to the steady-state. Panel (b) depicts the hospital's optimal choice of total training for senior residents conditional on their skill acquired during their first year. The steady-state is depicted by the vertical line. This figure illustrates that below the steady-state, the gradient in training is steeper than above the steady-state. When combined with Panel (a), one notices that below the steady-state the hospital prioritizes training the junior residents to return to the steady-state and trains the seniors much less. This is due to the value of training the junior resident in the following years and because the hospital is relative patient ($\beta = 0.95$): the higher skilled the junior resident, the more training both cohorts can receive and still meet the quality constraint.

The intuition behind these findings is as follows. Because the rising seniors are less skilled than usual, the hospital is limited in the total fraction of patients it can assign to residents and maintain the care quality constraint. Therefore, it is very costly for the hospital to train the disrupted cohort more than the steady-state amount because then not only would the incoming cohort receives less training, but also future incoming cohorts would also receive less training. Hence, the hospital trades off training for the affected cohort vs. training for all future cohorts. The combination of the yearly discount factor of $\beta = 0.95$ and the concavity of the training utility with respect to patient share means that the hospital is patient enough to sacrifice training in the current period in order to return to the steady-

state sooner, rather than spreading the cost of the disruption over additional future periods. This result means that one-period policies to alleviate the training disruption are unlikely to induce the hospital to fully restore training to the senior cohort. Hence, as I will show, policies that allow it to increase total training will benefit the incoming cohort more so that it can both minimize the impact to future cohorts and benefit from additional periods of increased training and patient throughput.

For the counterfactual outcomes, I assume that the result of the disruption is that the affected rising seniors begin the academic year with half of the steady-state skill, but that the incoming cohort is identical to all other incoming cohorts. For the temporary changes, I find the optimal patient allocation with the different model parameters and the lower-than-usual starting value of rising senior skill but knowing that there will be a return to the status-quo in the following academic year. I consider combinations of eight one-time, temporary policy changes: relaxing the care quality bound by 2.5% and 5.0% in length of stay, increasing attending speed by 2.5% and by 5.0%, increasing the rate of learning by 5.0%, and combining a 5% increase in learning speed with either a relaxation of the lower bound by 2.5%, a 5% increase in attending speed, or both.

Results are in Table 2.4. As before, I compare outcomes with the steady-state outcomes in lifetime patients seen and minutes per patient. Those outcomes are represented in the first row of the table. The remaining rows (under the single dividing line) display the outcomes under various one-time changes when the rising senior class begins with half the knowledge as in the steady-state. As we saw in Figure 2.2, there is a significant training cost in the Status Quo–if no other actions are taken. Under the same assumptions on resident careers as the first counterfactual, residents see 90 fewer complex patients during their career and spend 4.2 additional minutes on each patient they see. The hospital decreases training by allocating 10% of complex patients away from residents to attending physicians (column 3) and incoming residents continue to be affected negatively (column 4). Temporarily relaxing care quality by 2.5% is able to make up about 1/3 of the disruption by permitting about the same amount of training as in the status quo. However, because the senior residents begin their senior year with half of the typical skill, they require more training than typical to make up lost training during the disruption. It turns out that the hospital prefers to train the incoming cohort the same amount as in the status quo so that it returns to the steady-state in the next academic year over training the affected senior cohort more.

Temporarily relaxing care quality by 5.0% illustrates the trade-off between the affected senior cohort and future cohorts. In this scenario, the hospital trains more than in the steady-state, allocating about 9.3% additional complex patients to residents instead of attendings. However, the additional training goes to the junior residents, who are trained by more than in the steady-state, and the senior residents actually receive less training than in the case where care quality is relaxed only 2.5%! This is because the hospital prefers to collect gains from many future years of additional training instead of training the affected cohort up to the usual steady-state.

This counterfactual illustrates two important but slightly counter-intuitive findings. First,

Table 2.4: Training Disruption: Counterfactual Resident Training and Mitigating Factors

Counterfactual	Lifetime Patients	Minutes per Patient	Total Training	Future Periods
Steady-State	8,165	381	0.790	=
Status Quo	-90	+4.2	-0.106	↓
Relax L^* 2.5%	-58	+2.7	+0.009	=
Relax L^* 5.0%	-79	+3.7	+0.093	↑
Att Speed +2.5%	-109	+5.2	-0.075	=
Att Speed +5.0%	-113	+5.3	-0.039	↑
Learning Rate +5.0%	-79	+3.7	-0.063	↓
L^* −2.5% & Learning	-65	+3.0	0.057	↑
Att Speed +5.0% & Learning	-83	+3.9	+0.009	↑
L^* 2.5% & A.S. +5.0% & Learning	-75	+3.5	+0.100	↑

Notes: This table shows the future implications when senior residents experience disruptions in their first-year training. For illustration, I assume they enter their second year with half of the knowledge they otherwise would have, approximately a 0.118-unit decrease in skill. In the status quo, the hospital does not adjust anything and behaves as depicted in Figure 2.2. The other entries show the outcome if the hospital makes a one-year change in behavior as indicated. The outcomes are relative to the steady-state outcome and compare the total number of patients the physician can see after the graduate (extensive margin) and the extra time per patient seen (intensive margin). The third column is the total amount of training, in patient mass, that are allocated to residents in the year of the policy change. Future Periods indicates whether or not the hospital experiences higher or lower utility from training in future periods based on if the policy change causes the hospital to train the junior residents more than they would have in the steady-state. See text and notes to Table 2.3 for more details.

it is relatively straightforward to restore training to the incoming cohort of residents: simple, feasible policies such as relaxing the care quality constraint by 2.5% or staffing additional attendings such that the speed with which they care for patients increases by 2.5% is sufficient. However, the second finding is that it is much more difficult to endogenously induce the hospital to restore training to the senior cohort. Even polices that permit the hospital to provide sufficient training to make up the forgone 0.118 units of training such as temporarily relaxing the length of stay constraint by 5.0% or the combination of relaxing the length of stay constraint by 2.5%, increasing attending speed by 5%, and increasing the rate of learning by 5% see the increased training accrue to the junior cohort rather than the senior cohort.

Therefore, while it is important to provide such temporary policies to mitigate the impact to future cohorts, other policies, such as continuing education after the senior residents graduate and become attending physicians , will be necessary to make up for the disruption that they have experienced. Such policies have the additional benefit of helping the cohort who were senior residents during the disruption, who are not considered in this counterfactual were also harmed.

2.6 Discussion and Conclusion

When profit maximization is not the primary goal, how multi-product nonprofit firms adjust the production of their products due to changes in revenue from one product is ambiguous. I study nonprofit teaching hospitals, which have the dual role of providing health services and training the next generation of physicians. Because the teaching component in this environment requires learning by doing, the hospital faces a trade-off between care quality and teaching. I study how the hospital allocates complex patients to residents and attendings to make this trade-off. I find that short-run increases in quality achieved with reducing training are dwarfed by long-run quality decreases because residents see many patients over the rest of their career. Policies that use revenue to incentivize quality improvements to current patients such as value-based care and prospective payment systems are an increasingly popular tool among both public and private insurers (Sokol, 2020). I show that when designing such policies, policymakers should be aware of potential unintended reductions in teaching: in respose to the decrease in revenue, academic hospitals may reduce teaching and the resulting reductions in physician skill may result in costs for future patients orders of magnitude larger than the savings for current patients.

I examine and quantify these trade-offs in the emergency department of a large, urban teaching hospital. I develop and estimate a dynamic model of training and find that the hospital acts as if it maximizes training conditional on a minimum average patient throughput level in each quarter. In counterfactuals, I find that if hospital administrators increase throughput by 2%, the required reduction in training will result in lower future throughput losses over 4 times larger than the current gains. However, there exist simple and feasible

changes that can reduce the externality. For instance, I find that a 5% increase in attending speed would mitigate the training reduction and reduce the future costs by 65%.

Even though my focus is on the emergency department of a single, top-ranked teaching hospital, there are key lessons for the broader healthcare sector. First, the link between throughput and revenue applies to all departments. Second, although there may be some variation in care correlated with teaching hospital rankings, prior work has shown that the basic production function of health services does not differ in outcomes with respect to residency program prestige (Doyle, et al., 2010).

CMS is aware of the increased costs faced by teaching hospitals. The Medicare Prospective Payment System (PPS) includes a bonus paid to academic hospitals, known as the Indirect Medical Education Payment (IME). Although my findings show that patient throughput costs may be significant, they should not be used as the sole basis for determining the size of these payments. I believe that a large part of the reason that patient outcomes and resource utilization do not change with experience is due to the success of attending supervision. Staffing high-quality attendings can be very expensive, especially when they are also spending significant time conducting valuable research, and the IME Payments should account for this cost. My hypothesis is supported by the workflow, in which attendings and residents confer to determine the plan of care for each patient, as well as the empirical results, which show that resident independence increases most notably and significantly in the first hour of the patient's encounter when the plan of care is developed. Further research exploring the ways in which variation in supervision affects both patient outcomes and teaching quality could be valuable in improving both care quality and training outcomes. After all, the degree to which policy can improve patient outcomes and reduce costs is reliant on improvements in physician habits and practice, much of which is taught and absorbed during residency (Ludmerer, 2015).

Finally, while teaching hospitals are a single example of a specialized organization, they constitute an outsize share of both the economy and individual utility. The United States spent 17.8% of GDP on healthcare in 2021, and in 2019, teaching hospitals contributed 45% to Health Care and Social Assistance GDP.[14] Preserving life and increasing the quality of life, the main functions of hospitals, are arguably the most important components of individual utility, well-being, and happiness, and the continued production of high-quality health services requires continued investment in teaching. That said, the model and empirical strategy can be applied to related settings as many nonprofit institutions are in essence multiproduct firms. Most notably, this group includes research universities, which through their research are principal drivers of innovation in the modern economy (Lerner, et al., 2023) yet are also responsible for educating undergraduate, professional, and graduate students. Studying how they make this trade-off and respond to changes in government funding could be a fruitful area for future research.

[14]Gunja, et al. (2023) and the author's calculations using statistics from the BEA and AAMC

Chapter 3

Gender Differences in Non-Promotable Tasks: The Case of Clinical Note-Taking (with Benjamin Handel, Jonathan Kolstad, and Ulrike Malmendier)

3.1 Introduction

Much work has cited medicine as one of the most gender-egalitarian fields (cf. Goldin, 2014). However, it has been documented that female physicians still lag their male counterparts in career advancement and salary (Sasser, 2005). For example, the gender gap in Obstetrics and Gynecology disappears after controlling for specialty, private vs. group practice, and procedures performed (Reyes, 2007). Therefore, it appears that differences remain along the intensive margin in medicine, not only by hours worked but also by procedures and tasks performed.

In some respects, these facts should not be surprising. A large literature has documented differences between men and women in domains that are potentially important for the practice of medicine such as competitiveness (Niederle and Vesterlund, 2011), task selection (Gneezy, et al., 2003), speaking up (Coffman, 2014; Thomas-Hunt and Phillips, 2004), altruism, cooperation, and risk tolerance (Buser, et al., 2004).

Detailed understanding of gender differences in specific tasks performed broadens our understanding of gender inequities in medicine and how they relate to both patient outcomes via variation in care and physician outcomes via compensation and career advancement. An example for differences in patient outcomes is Currie, et al. (2016), who study clinical decision making in treating heart attack patients. They find significant differences by gender, with male cardiologists systematically making lower quality diagnosis and providing more intensive treatment to less clinically appropriate patients. An example for differences in physician outcomes is Sarsons (2019), who shows that surgeon gender impacts the way in which referring physicians interpret surgeon skill, which leads in disparities in surgeon careers.

At a broader level, flexibility, particularly with respect to child care, has been shown to play an important role for women in high skill professions (see Goldin and Mitchell (2017) for a review). The degree to which this flexibility plays a role in medicine, particularly in academic settings, is less well understood. Furthermore, little is known about how such flexibility plays out in day-to-day work.

We focus on clinical documentation: the production and consumption of clinical notes. Note writing and reading is a significant task performed by physicians, measured both in terms of time spent and by impact on both patients and physicians. The rapid pace of Electronic Health Record (EHR) adaptation means that the nature of work has shifted dramatically for physicians. Physicians average 3.8 hours per day working on EHR (Verma, et al., 2020), and there is evidence that EHR use is associated with increased clinical efficiency (cf. Holmgren, et al., 2022), the original motivation for adaptation. On the other hand, there is mounting evidence of a link between EHR usage and burnout. For example, Gardner, et al. (2019) find that physicians who self-report that EHRs add to their daily frustration have 2.4 times the odds of burnout as measured by the Physician Work Life Study compared to those who disagree.

In this chapter, we investigate how physician gender interacts with the changing nature of medicine. First, we document granular differences in clinical documentation habits between

men and women. Second, we ask whether increased note-taking benefits patients. Third, we ask whether there are costs or benefits for the physicians themselves in terms of compensation and career advancement.

We investigate these margins in a highly standardized environment at a top teaching hospital, the University of California, San Francisco (UCSF). Our analysis leverages unique Electronic Health Record (EHR) and Audit Log Data, in which we observe the exact time and duration of every instance of note activity as well as the time of every medical procedure and medication ordered for the patient. These data follow each of the 85,990 patients who enter the hospital through the emergency department during a two-year period from 2017-2019. The combination of rich patient-level controls and quasi-random assignment to physicians given the context allow us to not only measure differences in documentation habits, but also link medical decision-making and patient outcomes to documentation. To shed light on differences in note content, we rely on another dataset from UCSF that contains de-identified note text.[1] Finally, we link to external sources for physician salary, grants, and publications.

First, we find as in previous studies that women spend more time than men on note-taking (cf. Gupta, et al., 2019). Even after controlling for a set of physician characteristics such as specialty, medical school graduation decade, and medical school rank, we find that women spend 10% more time on notes per shift compared to their male counterparts. These differences are mainly due to note writing rather than note reading, and we do not find gender differences in "words per minute," note length divided time spent writing. Furthermore, we document hourly patterns in the differences: women spend more time taking notes during their shift, especially between 10am and 4pm, but spend about the same amount of time outside of typical scheduled work hours.

Next, we examine the value of note-taking for patient care. We ask whether patients who have longer notes written about them receive higher quality or more efficient care. In this section, we focus on hospital medicine notes for patients who were admitted to the hospital by the emergency department. Leveraging the quasi-random assignment of patients to physicians in this setting, we use a Wald Estimator to estimate the effect on resource utilization of an increase in time spent writing notes due to physician gender. We find that a change from an all-male to all-female care time on day $t-1$ leads to a 5.4% reduction in the number of orders signed on day t. Then, we collapse measures to the encounter level and find that this day-to-day reduction of orders is a pure reduction in orders signed, rather than a "shifting forward" of the same set of orders. Notably, the proportional effects are larger for overnight orders, where the care team always changes, than for day orders, where care teams change less frequently from one day shift to the next due to the institutional shift schedule. Finally, we show that there are no statistically or economically significant differences in either care quality (as measured by the 30-day readmission rate) or in the total length of the patient's stay in the hospital. Therefore, we conclude that the reduction

[1] UCSF has run algorithms that remove any identifying information from the note, such as names, and also scrambles the dates. Unfortunately, we are not able to link the note text data to our main EHR and Audit Log dataset.

in orders does not lead to lower quality care, but is unlikely to be correlated with faster diagnosis or patient recovery. Still, these results suggest that additional note writing may lead to increases in clinical efficiency: the same patient outcomes are achieved with fewer costly resources.

We attempt to shed light on the mechanisms behind these reductions in two ways. First, we analyze gender differences in note content. Second, we ask if there are differences in note utilization by others correlated with author gender. We find in our examination of note content that female physicians include about 23% additional clinical concepts in their notes. The increase is similar across categories of clinical concepts relating to patient condition and those relating to procedures and medications. Because female physicians use fewer resources, that means that they write about a greater fraction of the procedures, lab tests, and medications that they do sign for each patient. In terms of tone, female physicians include about 15% more negated concepts ("The patient does not have a history of..."), but are no more likely to use conditional concepts that may indicate clinical hypotheses or differential diagnoses. We also find no differences in the clinical use of notes written by men and by women, conditional on length and note type. In combination with the finding that there are also reductions in orders overnight, these results suggest that at least some of the improvement in clinical efficiency we find does come from the note content and not just from differences in physician skill (cf. Currie and MacLeod, 2020) or practice style (cf. Cutler, et al., 2019).

However, despite the productivity benefits of increased effort spent on notes, we do find evidence that physicians benefit, and if anything, they are harmed. We compare 2018 note-taking intensity with 2018 salary, 2019-2020 grant receipt, and 2019 publications.[2] Confidence intervals are wide, and model fit as measured by R^2 does not improve meaningfully with the inclusion of note intensity terms. Point estimates suggest that increased note-taking intensity is correlated with lower salary, less grant receipt, and fewer publications. In terms of salary, point estimates suggest that an additional standard deviation in time spent writing notes decreases male salaries by about 8% and female salaries by about 2%. Similarly, an additional standard deviation in time spent writing notes decreases the likelihood of grant receipt in the next two years by about 6.6 percentage points (about 31% of the mean) for men and increases it by about 1.3 percentage points (about 6% of the mean) for women, and decreases publications in the following year by about 5.9%, or 1.5 publications, for men and by about 7.2%, or 1.8 publications, for women. These results suggest that note-taking is costly in terms of salary and crowds out efforts to produce academic research, which are crucial elements of career advancement in academic medicine.

This project relates primarily to three strands of literature. The first is on gender gaps in wages and promotions, as well as differences in behavior, as detailed above. We show that even within the same hospital, there are notable differences in how men and women approach their narrowly-defined jobs. Furthermore, we show that these differences meaningfully affect

[2]Note that physicians publish primarily in medical and health services journals, which have a far shorter time from submission to publication compared to economics journals.

productivity, yet physicians are not rewarded for their increased effort. The second is on practice variation across physicians, such as Chandra and Staiger (2007), Molitor (2018), and Finkelstein, et al. (2022). A notable difference is that while these papers tend to study differences across locations, we focus on differences across physicians within the same hospital. This means that the differences we find are in spite of the fact that the physicians in our sample have already self-selected into a specific "firm" and are subject to the same work environment and culture. Finally, we contribute to the growing literature, primarily in the health services literature, that uses EHR and audit log data to show differences in physician behavior and differences in patient care, such as Patel et al. (2018) and Huigol et al. (2022).

The remainder of this chapter proceeds as follows: Section 3.2 describes the data used and some descriptive statistics, Section 3.3 presents methodology and results on note activity, Section 3.4 presents methodology and results on the clinical value of longer notes as well as mechanisms including note content and utilization by others, Section 3.5 presents methodology and results on physician outcomes, and Section 3.6 concludes.

3.2 Data

This chapter uses three main sources of data. The first are 104 weeks of EHR and Audit Log Data within 2017-2019. The second are one synthetic calendar year of de-identified clinical notes[3] written by hospital medicine physicians. The third set of data are from non-UCSF sources that we use for physician education history, salary, publications, grants, and attrition. We expand upon each of them and provide summary statistics in the following subsections.

3.2.1 Electronic Health Record and Audit Log Data

Our primary data consists of 104 continuous weeks from 2017-2019 of Electronic Health Record (EHR) and Audit Log data from the University of California, San Francisco. The data covers all patients who enter the hospital via the emergency department (ED) and follows them until they are discharged from the hospital.[4] Observations are organized at the encounter level, which is one discrete hospital visit; the same patient can present for multiple encounters and we also observe patient identifiers. For each encounter, we observe detailed information on note activity by their physicians, including exactly when and for how long each note was created, viewed, edited, and signed. This includes views of "historical" notes from prior patient encounters in the hospital system even if they fall outside our sample period. We also observe the note length in characters for each version of each note. In addition to documentation, we observe each Medical Order (medications, procedures) that

[3]As part of the de-identification process, dates are randomly scrambled. Our sample uses notes with de-identified year 2018, which in expectation comprise notes from one actual calendar year.

[4]Discharge is by far the most common outcome. Alternatives are leaving without being seen, leaving against medical advice, and (rarely) death.

is prescribed, including orders that are executed as well as orders that are subsequently canceled. For labs and imaging orders, we observe when order results are available, when they are viewed, by who, and for how long. For most labs we also observe a flag for whether the results are abnormal. We observe provider identifiers and exact timestamps for every note and order action.

We also observe a detailed set of patient characteristics. These include both relatively "typical" measures such as age, sex, and race, but also a set of unique "ex-ante" measures taken prior to physician intervention by the ED's triage nurse. These measures include things such as the patient's chief complaint that brought them to the ED that day, a set of indicators for abnormal vital signs, and the triage nurse's estimation of how urgent the patient's condition is (Emergency Severity Index, or ESI). We leverage these ex-ante measures along with a set of immutable patient conditions that cannot be affected by care decisions within an encounter (ex. age at arrival, means of arrival, patient sex) and use them to calculate a continuous index of ex-ante patient complexity. This measure is therefore exogenous to the team that will subsequently care for the patient and is the measure developed and used in Chu, et al. (2023).

Throughout, we focus on attending physicians. The data contain a limited set of physician covariates, including gender, role (resident, attending, nurse practitioner, etc.), and specialty. In addition to internal user identification codes, we also observe each physician's National Provider Identifier (NPI) and their full name.

As seen in Table 3.1, the full sample contains 2,559 physicians, which we narrow down to 1,620 for our Note Taking (Section 3.3) analyses with the indicated sample restrictions. There are large gender differences unconditional on physician specialty, shifts worked, and patient seen in the median number of hours spent on notes.

Table 3.1: Physician Sample Selection for the Main Analysis

Restriction	Total Physicians	Fraction female	Median Note Hours, F	Median Note Hours, M
All attendings	2,559	0.509		
...with note actions	2,531	0.509	22.4	19.2
...match Physician Compare	1,820	0.478	34.2	25.5
...internal vs. PC gender match	1,806	0.476	34.5	25.5
...specialties with 2+ of each gender	1,620	0.473	36.8	26.1

Notes: This table shows the results of each physician sample selection restriction for the main analysis. Physician Compare (abbreviated PC) is a dataset of physician practice locations and characteristics distributed by the Centers for Medicare & Medicaid Services (CMS). Median Note Hours is the median of the sum of total hours spent editing and viewing clinical notes in the data.

For the Clinical Outcomes analysis in Section 3.4, we use a different set of encounters and

physicians. Because we want to investigate downstream clinical utilization of notes, we focus on the hospital medicine setting. After patients are admitted to the hospital from the ED for inpatient care, they fall under the care and supervision of a hospital medicine team. Out of the full sample of 85,990 encounters, 18,320 encounters (21%) have an inpatient admission.

We limit our analysis to physicians who specialize in internal medicine who as hospitalists are ultimately responsible for managing and coordinating each patient's care. This excludes physicians who belong to other specialties that may be involved with patient care as consulting physicians and leaves us with 230 attending physicians, with physician-level averages seen in Table 3.2.

Table 3.2: Internal Medicine Physicians for the Clinical Outcomes Analysis

	All	Men	Women
Number of Attending Hospitalists	230	0.42	0.58
Mean Patient-Days Worked	202	220	190
Edits Action Count	548	523	566
Edits Hours Spent	48	42	52
...per Patient-Day: Action Count	2.37	2.16	2.53
...per Patient-Day: Minutes Spent	12.72	11.81	13.38
Orders Authorized	6,031	6,027	6,033
...per Patient-Day: Orders Auth.	22.89	23.23	22.62
Orders Signed	1,734	1,665	1,789
...per Patient-Day: Orders Signed	7.10	6.84	7.32

Notes: This table shows sample statistics for the 14,707 patient encounters and hospital medicine attending physicians in the Clinical Outcomes analysis. Mean patient-days worked is mean of the number of patient-shifts the hospitalist works on notes or authorizes an order. This measure increases if hospitalists care for the same patient for additional shifts or cares for additional patients during the same shift. Edits Action Count is the number of edit actions, and Edits Hours Spent is the duration of those actions. Orders Authorized are the orders where the index hospitalist was clinically responsible for. This includes both orders signed themselves and orders signed by residents under their supervision. Orders Signed are the orders that the hospitalist signs themselves. Per Patient-Day normalizes these measures by the number of patient-days worked to get at an average workload per patient, per shift.

3.2.2 De-Identified Note Text

We obtain de-identified note text from a separate dataset within UCSF for the note content analysis in Subsection 3.4.2. These data cover the universe of encounters within the UCSF hospital system, including the facility from which the EHR and Audit Log data is derived from. However, these data are not linkable to the EHR data either by encounter, patient, or physician. To be consistent with the quantitative analysis in Section 3.4, we select hospital medicine notes from the same qualitative sample of patients: those who are admitted to inpatient care from the emergency department. We use all notes from qualifying encounters in de-identified calendar year 2018. Note that as part of the de-identifying process, the dates of encounters within this data have been scrambled, so these notes are not all necessarily from 2018 encounters, but they should in expectation consist of one calendar year's worth of notes.

As can be seen in Table 3.3, this sample consists of 46,395 notes, of which about 53% have female authors. About 77% of the notes are daily Progress Notes, and there are also longer History & Physical (about 12%) and Assessment & Plan (about 10%) notes.

Table 3.3: Summary Statistics for De-Identified Notes

Note Type	Count	Fraction Female Author
Progress Notes	35,273	0.523
History & Physical	5,717	0.512
Assessment & Plan	4,753	0.600
All Notes	45,743	0.530

Notes: This table shows the count of hospitalist notes of each type and the fraction female author for the note text analyzed for gender differences in note content.

In addition to the de-identified text, each note is also associated with a set of clinical "Note Concepts." These are a standardized set of clinical procedures and terms that have been extracted using Natural Language Processing (NLP) specifically designed for clinical text and customized by UCSF.[5] Clinical concepts are categorized into seven categories: Signs & Symptoms, Diseases, Patient History, Family HIstory, Procedures, Lab Tests, and Medications. Furthermore, each concept has a binary "Negated" label if the concept was negated (ex. "Not indicative of [concept]" or "Patient denies [concept]"), and a binary "Conditional" label if conditional modifiers are used on the concept (ex. "potentially," "suggestive of"). Finally, each concept has a Model Confidence score in the interval $[0, 1]$ which corresponds to the algorithm's uncertainty in categorizing the text into a specific concept.

[5]The algorithm is a customized version of Apache cTAKES. Clinical concepts directly map to SNOMED codes.

3.2.3 External Data Sources

We complement the internal data with Physician Compare, a dataset provided by CMS, that lists each clinician-group in the country with Medicare enrollments. We merge our sample of physicians with the physicians in Physician Compare using NPI in order to obtain four items. First, we obtain a second elicitation of gender, which we use to validate our internal data. Our sample selection depicted in Table 3.1 only keeps physicians for which the Physician Compare gender matches our internal data. Second, we obtain the physician's name, which we use to match physicians to administrative shift schedule data to verify the shifts inferred in the EHR. Third, we obtain the medical school from which the physician graduated. Fourth, we obtain their medical school graduation year. We also merge to Sacramento Bee's California State Worker Salary Database for salary information, and to the National Institutes of Health (NIH)'s PubMed for publications, and to the NIH RePORTER for grants. These three items are the Physician Outcomes we consider in Section 3.5.

3.3 Note Activity

In this section, we describe the differences in note writing and viewing behavior between male and female physicians. We begin by investigating overall differences. To do this, we collapse all data to the physician level and regress the natural logarithm of total minutes spent viewing and editing notes during the two-year sample on an indicator for physician gender that takes on value 1 if the physician is female and 0 if they are male. We progressively add more controls for physician characteristics including specialty, medical school graduation decade, and medical school rank. That is, we estimate regressions of the form

$$\log\left(\sum_j \text{Note Time}_{jn}\right) = \beta_0 + \beta_1 \text{Female}_j + \beta_2 \log(\text{Shifts Worked}_j) + X_j'\beta_3 + \varepsilon_j \qquad (3.1)$$

where j indexes physicians and n specific actions, such that Note Time_{jn} is the time elapsed of a specific action and X_j are a vector of physician characteristics. β_1 is the coefficient of interest and represents the percent difference in time spent on notes by women vs. men.

Results can be seen in Table 3.4. Our preferred specification is the rightmost one, and the coefficient on Female of 0.1 indicates that female physicians spend about 10 log points, or approximate 10% more time viewing and editing notes per shift than their male counterparts. This is somewhat surprising since we are comparing physicians who work at the same hospital and in the same division (specialty) with approximately the same amount of experience (graduation decade).

Table 3.4: Time Spent Viewing and Editing Notes Per Shift

	log(minutes spent, all actions)			
Female	0.150***	0.144***	0.097**	0.100**
	(0.054)	(0.042)	(0.042)	(0.042)
log(shifts worked)	1.043***	0.989***	0.996***	0.997***
	(0.024)	(0.023)	(0.023)	(0.023)
Specialty FE		X	X	X
Grad Decade FE			X	X
Med School Rank FE				X
Obs	1,553	1,553	1,551	1,551
R-squared	0.663	0.798	0.808	0.808
Adj R-squared	0.663	0.790	0.800	0.799

Notes: This table shows regressions at the physician level of the natural logarithm of total minutes spent on clinical notes on an indicator for Female, the natural logarithm of shifts worked, and indicated physician controls. Our preferred specification is the rightmost one, which includes fixed effects for physician specialty, medical school graduation decade, and medical school rank. The coefficient on female of 0.100 indicates that conditional on the natural logarithm of shifts worked and physician characteristics, women spend a total of 10% longer on documentation than their male counterparts. See text for additional details.

We next examine differences in when within the day that note actions are performed. We estimate regressions of the form

$$y_{js} = \alpha + \beta_0 \text{female}_j + \sum_{h=1}^{23} \left[\delta_h \text{hour}_{h(s)} + \beta_k \text{female}_j \cdot \text{hour}_{h(s)}\right] + X'_{js}\gamma + \varepsilon_{js} \tag{3.2}$$

Each observation is an individual physician j times hour s: for instance, physician 1 on January 1, 2018 at 10am is a separate observation from physician 1 on January 2, 2018 at 10am. y_{js} are outcomes of interest, and $h(s)$ is a function that maps the specific hour s to one of the 24 hours of the day h. We consider two outcomes: (1) a binary indicator for any note action performed within the hour, which represents the extensive margin of note activity, and (2) the natural logarithm of total minutes spent on notes within that hour, which represents the intensive margin of note activity conditional on having any activity. X_{js} contain both physician level controls (specialty, graduation decade, and medical school rank) and calendar month fixed effects (January, February) to capture seasonality. The coefficients of interest are β_h: how much more women work on notes than men during each hour of the day (h subscript).

In Figures 3.1 and 3.2, we plot $\beta_0 + \beta_h$ for each hour h as well as β_0, which represents the difference at midnight. This is done to preserve the overall gender difference, as we have

established in Table 3.4 that women spend more time overall than men. The shaded area represents the modal shift for attendings derived from our shift data. In Figure 3.1, we see that women are about 2% more likely to perform any note activity during the middle of the modal shift, from 11am to 4pm, whereas men are about 2-3% more likely to perform any note activity in the three hours prior to the modal shift, from 5am to 7am. Moving to Figure 3.2, we see that during the entire modal shift, women who do any note activity spend about 1.5% more time on notes on average compared to men who do any note activity. Women also spend about 1.5% more time on notes than men in the evening, between 8pm and 11pm. There are no times during the day where men spend more time on notes than women. Together, these figures show that most of the differences in note activity are during the workday, and that differences are a combination of women being more likely than men to perform any note activity and spending more time than men conditional on doing so.

These findings relate to Goldin and Mitchell (2017), who find that flexibility of when and where to do work is important for gender equality. In principle, physicians could spend less time on notes during the workday and complete their note tasks in the evening at home via VPN. Although we find that women spend about 1.5% more time on notes in the evening hours compared to men, they are no more likely to engage in any note activity–there are no differences along the extensive margin. This suggests that either the flexibility benefits of EHR are overstated (for instance, if it is not possible to leave a shift early because of other duties such as supervising trainees), or that they are less important compared to other sources of flexibility such as "non-standard" shifts that overlap better with the school day.[6]

We next examine what types of note activity drive these differences. There are two main actions that physicians can take on notes: they can either produce information (editing and writing notes) or consume information (viewing preexisting notes). As seen in Table 3.5, differences are driven by time spent editing notes rather than referring to notes written by others. Recall that these coefficients are proportional differences. It turns out that all physicians spend more time editing notes compared to viewing them, so the effect in levels is even greater than the proportional effects shown in Table 3.5: the median physician spends 70% of all note activity time editing notes and 30% of time viewing notes.

3.3.1 Heterogeneity by Patient Complexity

In this subsection, we examine whether results are results driven by more or less complex patients. To do this, we split the sample of encounters into two equally sized groups based on their predicted ex-ante complexity measure. Then, we sum all note activity for physician-complexity group pairs and regress a version of Equation (3.1) with an additional indicator for high complexity and the interaction of the Female indicator variable and the high complexity indicator.

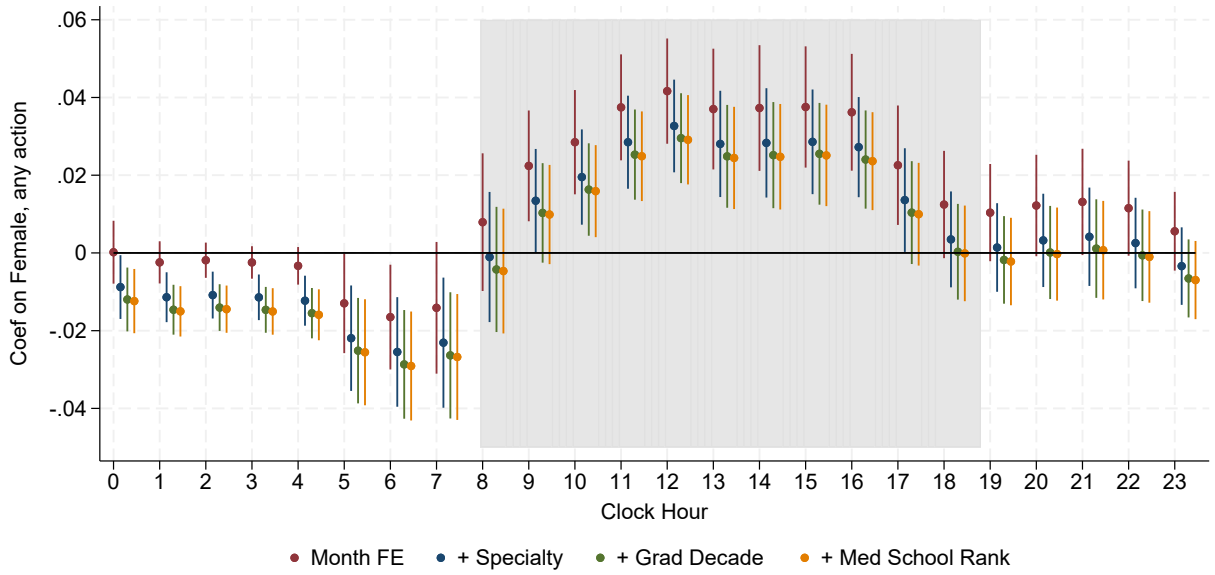
[6]Based on administrative shift data, these are uncommon in our setting.

Table 3.5: Breakdown of Time Spent Viewing and Editing Notes Per Shift

	log(minutes editing)				log(minutes viewing)			
Female	0.329***	0.215***	0.139**	0.149**	0.122	0.216***	0.092	0.104*
	(0.081)	(0.066)	(0.065)	(0.065)	(0.076)	(0.064)	(0.059)	(0.060)
log(shifts worked)	0.914***	0.860***	0.886***	0.887***	0.924***	0.867***	0.890***	0.890***
	(0.034)	(0.033)	(0.032)	(0.032)	(0.032)	(0.031)	(0.029)	(0.029)
Specialty FE		X	X	X		X	X	X
Grad Decade FE			X	X			X	X
Med School Rank FE				X				X
Obs	1,463	1,463	1,461	1,461	1,534	1,534	1,532	1,532
R-squared	0.404	0.655	0.677	0.679	0.431	0.628	0.680	0.682
Adj R-squared	0.403	0.642	0.663	0.664	0.431	0.614	0.667	0.667

Notes: This table is similar to Table 3.4, except that it separates time spent editing notes (first four columns) and time spent viewing notes (last four columns). See notes to Table 3.4 and text for additional details.

Figure 3.1: Breakdown of Note-Taking Differences Throughout the Day: Extensive Margin



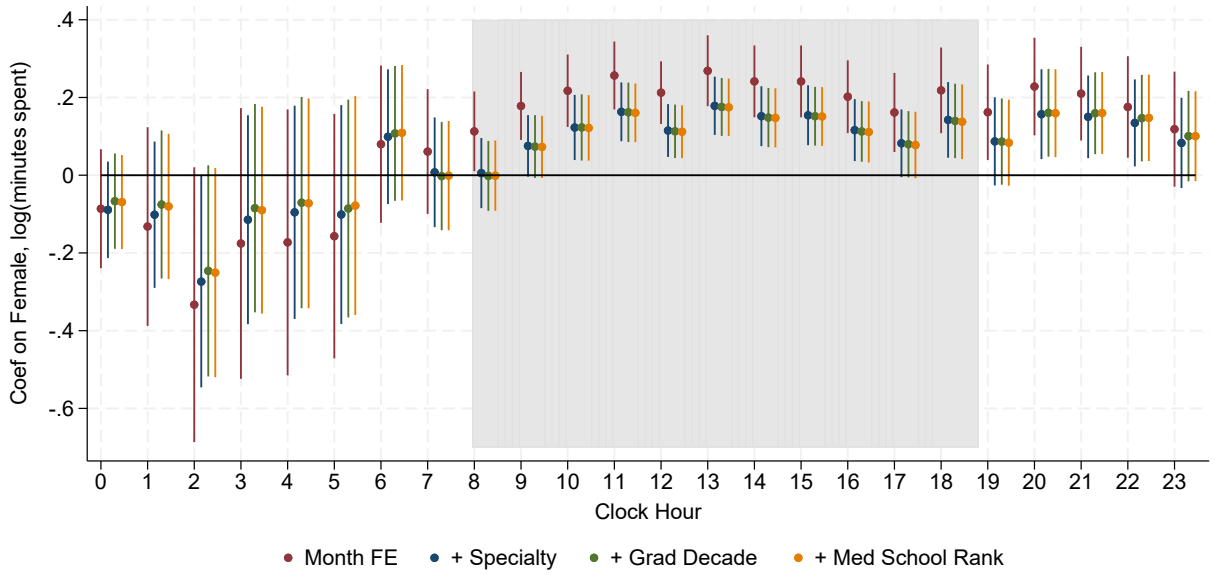
Notes: This figure shows the coefficients and standard errors for the 23 hourly-times-female fixed effects (β_k) and the main effect on female (β_0) estimated in Equation (3.2). Each hour contains four coefficient estimates: the first only has fixed effects for month, the second adds physician specialty fixed effects, the third adds medical school graduation decade fixed effects, and the fourth adds medical school rank fixed effects. The shaded region (8am to 7pm) indicates the modal shift worked by physicians in the sample. The dependent variable is a binary indicator for any note activity and thus represents the extensive margin of note activity. At 12pm, the fourth estimate with all fixed effects (labeled "+Med School Rank") is about 0.03, indicating that controlling for the fixed effects, women are about 3% more likely to do any note activity at noon compared to men. See text for additional details.

These regressions take the form

$$\log\left(\sum_{jk} \text{Note Time}_{jkn}\right) = \alpha_0 + \alpha_1 \text{HiComplexity}_k + \alpha_2 \text{Female}_j + \alpha_3 \text{HiComplexity}_k \cdot \text{Female}_j + \alpha_4 \log(\text{Shifts Worked}_j) + \alpha_5 \log(\text{Encounter-Days}_{jk}) + X_j' \alpha_6 + \varepsilon_{jk} \quad (3.3)$$

where $k \in \{1, 2\}$ indexes complexity group. In our preferred specification, we also add the natural logarithm of encounter-days that each physician is involved in, which we call $\log(\text{Encounter-Days}_{jk})$. This is important because even with random assignment of patients to physicians, so physicians see drastically different proportions of low-and high-complexity

Figure 3.2: Breakdown of Note-Taking Differences Throughout the Day: Intensive Margin



Notes: This figure show the coefficients and standard errors for the 23 hourly-times-female fixed effects (β_k) and the main effect on female (β_0) estimated in Equation (3.2). Each hour contains four coefficient estimates: the first only has fixed effects for month, the second adds physician specialty fixed effects, the third adds medical school graduation decade fixed effects, and the fourth adds medical school rank fixed effects. The shaded region (8am to 7pm) indicates the modal shift worked by physicians in the sample. The dependent variable is the natural logarithm of minutes spent on notes and thus represents the intensive margin of note activity, conditional on any activity. At 12pm, the fourth estimate with all fixed effects is about 0.1, indicating that among men and women who do any note activity at noon, women spend about 10% longer than men. See text for additional details.

patients during their shifts. Results are in Table 3.6. In this table, all patient encounters are included. As expected, the coefficient on High Complexity is large and positive, indicating that physicians spend longer on notes for high complexity patients than for less complexity patients. Next, the coefficient on female remains positive and significant. In our preferred specification, this is about 0.225, meaning that women spend about 22.5% longer on notes for low complexity patients than men. Next, the coefficient on the interaction of Female and High Complexity patients is negative, suggesting that the gender difference in note time is smaller for high complexity patients than for low complexity patients.

Because the majority of low-complexity patients are not admitted to the hospital, they typically only have a single note written about them, the Emergency Department Provider

Table 3.6: Note Activity Heterogeneity by Patient Complexity: All Patients

	log(minutes spent, all patients)				
High Complexity	1.638***	1.632***	1.630***	1.629***	0.902***
	(0.064)	(0.055)	(0.055)	(0.055)	(0.061)
Female	0.293***	0.245***	0.207***	0.209***	0.225***
	(0.069)	(0.062)	(0.062)	(0.063)	(0.058)
Female X High Complexity	-0.185**	-0.177**	-0.177**	-0.177**	-0.132*
	(0.091)	(0.079)	(0.078)	(0.079)	(0.070)
log(shifts worked)	1.015***	0.967***	0.972***	0.973***	0.617***
	(0.021)	(0.020)	(0.021)	(0.021)	(0.024)
log(patient encounter-days)					0.377***
					(0.017)
Specialty FE		X	X	X	X
Grad Decade FE			X	X	X
Med School Rank FE				X	X
Obs	3,026	3,026	3,022	3,022	2,731
R-squared	0.616	0.714	0.719	0.719	0.768
Adj R-squared	0.615	0.709	0.713	0.713	0.762

Notes: This table is similar to Table 3.4, but encounters are divided into low and high complexity based on whether their predicted ex-ante complexity measure is greater than or less than the median. Each observation is a physician times complexity group and the sample is all patient encounters.The rightmost column additionally controls for the natural logarithm of patient encounter-days that each physician is involved in; this is because even with random assignment of patients to physicians, some physicians see drastically different proportions of low- and high-complexity patients during their shifts. See notes to Table 3.4 and text for additional details.

Note. Therefore, this analysis compares drastically different patient types. To address this concern, we restrict to the 21% of patients who are admitted to the hospital as inpatients, regardless of the admitted department (for instance, including admissions for surgery, to the general "medicine" floor, and to other services). We redefine the groups based on the median ex-ante complexity for these patients and repeat the analysis. Results are in Table 3.7. Our preferred specification is the rightmost column. As before, we find that the coefficient on high complexity patients is positive and statistically significant, but is of a much lower magnitude than before: now we find that on average, male physicians spend about 9% longer on notes for more complex patients than for less complex patients. The coefficient on female remains positive, but is imprecisely estimated. The coefficient on the interaction is small in magnitude and also imprecisely estimated.

Taken at face value, these results suggest that there may be limited differences in note

Table 3.7: Note Activity Heterogeneity by Patient Complexity: Admitted Patients

	log(minutes spent, admitted patients)				
High Complexity	0.422***	0.401***	0.398***	0.398***	0.091**
	(0.072)	(0.058)	(0.056)	(0.056)	(0.036)
Female	0.039	0.045	-0.029	-0.024	0.057
	(0.070)	(0.058)	(0.057)	(0.058)	(0.039)
Female X High Complexity	-0.029	-0.019	-0.020	-0.021	0.015
	(0.105)	(0.085)	(0.083)	(0.083)	(0.051)
log(shifts worked)	1.060***	0.968***	0.981***	0.980***	0.352***
	(0.023)	(0.022)	(0.022)	(0.022)	(0.020)
log(patient encounter-days)					0.653***
					(0.014)
Specialty FE		X	X	X	X
Grad Decade FE			X	X	X
Med School Rank FE				X	X
Obs	2,881	2,881	2,877	2,877	2,644
R-squared	0.477	0.665	0.684	0.685	0.868
Adj R-squared	0.477	0.659	0.677	0.677	0.864

Notes: This table is similar to Table 3.4, but encounters are divided into low and high complexity based on whether their predicted ex-ante complexity measure is greater than or less than the median. Each observation is a physician times complexity group and the sample is restricted to patients who are admitted to inpatient services. The complexity groups are recalculated from Table 3.6 so there are equal numbers of encounters in the low- and high-complexity groups. The rightmost column additionally controls for the natural logarithm of patient encounter-days that each physician is involved in; this is because even with random assignment of patients to physicians, some physicians see drastically different proportions of low- and high-complexity patients during their shifts. See notes to Table 3.4 and text for additional details.

activity by gender for patients who are ex-post unhealthy enough to require emergency surgery or inpatient care. However, note that the standard errors are large enough such that we cannot rule out a positive coefficient on Female. Additionally, the patient split may be misspecified. It may be that splitting by pre-emergency department complexity is a noisy way of splitting patients because we do not include any new information discovered by the emergency department care team that is exogenous to the team of physicians who will subsequently care for the patient. More problematically, we may not be correctly controlling for encounter-days because these are constructed from note activity rather than orders activity. We may be understating gender differences if men who supervise patients and other physicians (chiefly, residents) do not do any note activity for their patients instead of doing

a very small amount. If this is the pattern of behavior, then we will not count this as an encounter-day. In progress is a specification that instead infers encounter-days by authorized orders, which are a more robust method of inferring patient involvement. We have reason to believe that this may be a representative pattern of behavior for a meaningful sample of physicians based on subsequent analysis presented in patient heterogeneity analysis in Section 3.4, which show that women spend even more time on notes for more-complex patients relative to less-complex patients: see Appendix Tables A9, A10, and A11.

3.3.2 Differences in the Production of Notes

We have established that women spend more time on notes than men, and most of this is due to additional time spent writing notes. But do they write longer notes as a result? In other words, are there differences in the note-writing production function, for instance if women require additional time to write the same length note as men? We investigate this by comparing the relationship between the length of notes, measured with character count, and the time spent writing them for notes written by single authors.

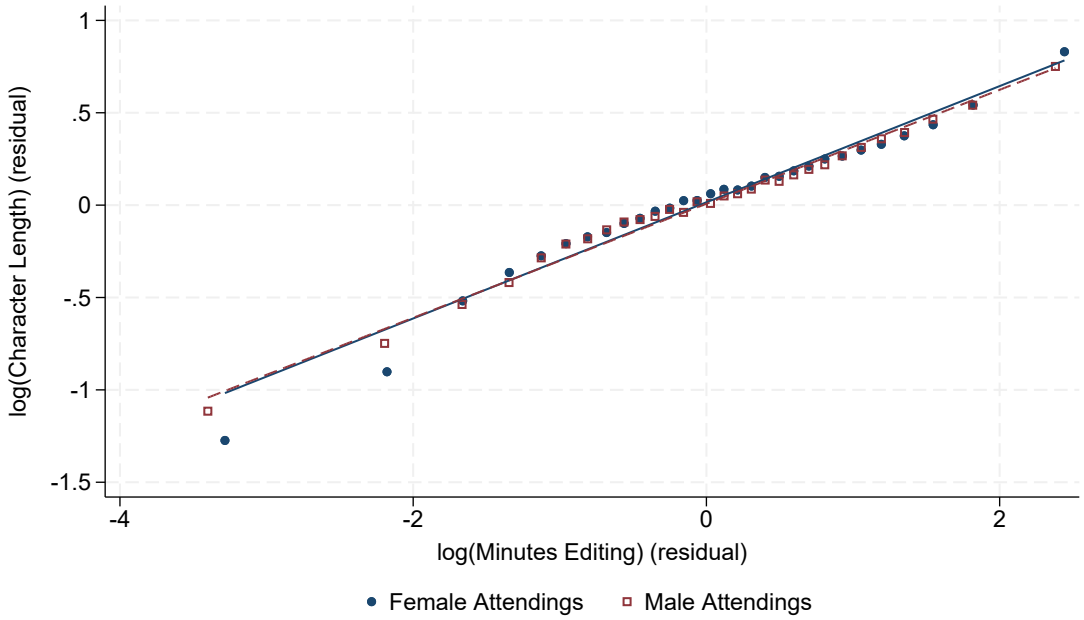
Figure 3.3 is a binned scatterplot of the relationship between the natural logarithm of minutes spent editing notes and the natural logarithm of the final character length of the note. Each point in the figure is one of 30 bins of notes written by a single attending. Both the log(Minutes Editing) and log(Character Length) are the residuals after removing note type fixed effects. We observe that the intercepts and slopes of the line of best fit for men and women are virtually identical, as are the shape of the relationship for all but the bottom two quantiles of notes where physicians spend the least time writing. This suggests that the production function of notes for men and women are identical, meaning that for a given time spent writing a note, there are no gender differences in the final length of the note. Therefore, because women spend longer writing notes than men, their notes are on average longer.

We supplement the visual analysis in Figure 3.3 with a set of regressions. The regressions take the form

$$\begin{aligned}\log(\text{CharsAdded}_i) = \beta_0 + \beta_1 \log(\text{MinsEditing}_i) + \beta_2 \text{Female}_{j(i)} \\ + \beta_3 \text{Female}_{j(i)} \cdot \log(\text{MinsEditing}_i) \\ + \text{NoteType}_i' \beta_4 + \beta_5 \text{Complexity}_{e(i)} + X_{j(i)}' \beta_6 + \varepsilon_i \quad (3.4)\end{aligned}$$

where each observation remains a note i, $j(i)$ is a function that returns the identity of physician j who edited note i, and $e(i)$ returns the encounter e that note i is associated with. The coefficients of interest are β_2 and β_3, which represent differences in "characters per minute" between men and women. As can be seen in Appendix Table A7, we find (imprecise) zeros on these coefficients across specifications, supporting our visual analysis suggesting that conditional on the amount of time spent, women and men write similar length notes on average. Note that the specification corresponding to Figure 3.3 is the one in the second column.

Figure 3.3: Note Length vs. Time Spent Editing for Men and Women



Notes: This figure is a binned scatterplot showing the relationship between the natural logarithm of minutes spent editing notes and the natural logarithm of the final character length of the note. Each point is one of 30 bins of notes written by a single attending. Both the log(Minutes Editing) and log(Character Length) are the residuals after removing note type fixed effects and thus the slopes correspond to those in Column 2 of Appendix Table A7. See text for additional details.

3.3.3 Summary of Findings on Note Activity

In this section, we have established that on average, women physicians spend about 10% more time on notes compared to their male counterparts. The additional time is both on the extensive and the intensive margin, and is concentrated during the workday rather than after hours or before the shift begins. The additional time is mostly in time spent editing notes, and leads to women writing longer notes on average. We have suggestive evidence that differences may be larger in less-complex patients, but gender differences in note activity may still be important for more-complex patients. Next, in Section 3.4, we investigate the clinical value of longer notes.

3.4 The Clinical Value of Longer Notes

In Section 3.3, we established that there are statistically and economically significant differences in the note-taking behavior of male and female physicians, that this difference is driven by women spending more time writing notes, and that the additional time does lead to longer notes. We now examine whether their patients benefit from longer notes. First, we describe our main approach and discuss findings. Then, we investigate two potential mechanisms: differences in the content of the notes and differences in time that other clinical staff spend reading notes.

Our ideal setting for this investigation is one where patients are sick enough where clinical note content is an important input into care decisions that we can observe. We choose to focus on hospital medicine ("hospitalist") notes for patients who were admitted to the medicine floor by the emergency department. This means that the patients were deemed not healthy enough to send home directly from the emergency department, but with ailments that did not require immediate surgery or attention from a specialized service (for instance, cardiology for heart attack patients). Regardless, because work is shift-based, care is provided by a team of physicians. At minimum, the team consists of the hospitalists and other providers who work the day shift, where the bulk of diagnosis and treatment occurs, and the hospitalists and providers who work the night shift, where the focus is on maintaining a stable condition. Meanwhile, the day-shift hospitalist also coordinates care from other providers, such as consulting specialists.

Therefore, the notes that hospitalists write upon patient intake describing patient condition and the plan for their care (the History and Physical note and Assessment and Plan note, respectively), as well as daily Progress Notes may meaningfully affect the path of care and patient outcomes. In this setting, a major purpose of the note is to put past clinical decisions and order results into context: "why a certain medication was prescribed or a specific test was ordered" (Schrager, 2022). Notably, because patient stays last multiple days, more detailed notes on any day of the stay may lead to better care either overnight or during subsequent days of the encounter. This setting has an additional attractive feature: patients are quasi-randomly assigned to physicians based on bed availability. This institutional detail allows us to control using a Wald Estimator for an important source of omitted variables bias: unobservable patient condition correlated with both note length and care patterns.

We consider three process measures and outcomes that we believe are good proxies for various aspects of clinical value: (1) the number and timing of costly medical orders signed for the patient, (2) the patient's total length of stay in the hospital, and (3) their 30-day readmission rate. The number of orders signed directly impacts medical spending and "overuse" of resources relates to the inefficiencies documented in Chandra and Staiger (2007). The timing of orders relates to diagnostic skill and quality of care: even with no reduction in the total number of orders, clinical notes are valuable if additional notes cause the same orders to be signed earlier. In other words, a finding that orders from certain physicians occur earlier in the patient stay would suggest that notes increase quality of care by increasing the speed of diagnosis, which directly increases patient welfare. A related measure is the

patient's total length of stay in the hospital, which we consider a reduced-form measure related to both quality of care and medical spending. Finally, the 30-day readmission rate is the outcome that is the most important proxy for care quality: ultimately, the hospital's goal is to treat the patient, and the readmission rate is a common (albeit imperfect) measure of the hospital's success rate (see for instance Benbassat and Taragin (2000) and the Hospital Readmissions Reduction Program used by CMS in Medicare Value-Based Purchasing). If we find that additional notes causes physicians to use fewer resources but is associated with a higher 30-day admission rate, then there likely is not an increase in efficiency because readmissions are both costly both financially and to patient well-being.

3.4.1 Main Approach: Effect of Additional Notes on Medical Orders

We investigate if additional notes lead to more efficient care. First, we investigate whether additional notes yesterday affect ordering behavior today. Then, we investigate whether changes are due to a pure reduction in orders signed or if the same total quantity of orders are signed, but are signed earlier.

Day-to-Day Ordering

Conceptually, we leverage our findings from Section 3.3 that female physicians spend more time writing notes compared to male physicians, and use a two-stage Wald Estimator to estimate the difference in orders the following day that can be explained by physician gender. In the first stage, we regress edit time on day $t-1$ on the fraction of note edit actions by female hospitalists on day $t-1$ and patient controls and recover a coefficient similar to the overall gender difference from Table 3.4. In the second stage, we regress orders on day t on the predicted edit time on day $t-1$ from the first stage and the same set of controls. The coefficient on predicted edit time is the Wald Estimate of interest: the reduction in orders on day t that can be explained by physician gender on day $t-1$. In this analysis, we focus on the active diagnosis and treatment portion of patient care, which occurs during the day. Hence, we limit both note activity and orders to those that occur during the hospitalist day shift, from 7am to 7pm.

Our first stage is

$$\text{IHS}(\text{MinsEditing}_{i,t-1}) = \alpha_0 + \alpha_1 \text{FracFemale}_{i,t-1} + \alpha_2 \text{MaleAttending}_{i,t} + X'_{i,t}\alpha_3 + v_{i,t-1} \quad (3.5)$$

and the corresponding Wald Estimate in the second stage is

$$\text{IHS}(\text{orders}_{i,t}) = \beta_0 + \beta_1 \text{IHS}\,\widehat{\text{MinsEditing}}_{i,t-1} + \text{MaleAttending}_{i,t} + X'_{it}\beta_3 + u_{i,t} \quad (3.6)$$

IHS is the inverse hyperbolic sine. It behaves similarly to the natural logarithm except that it is defined at zero, which is important because there are encounter-days with zero note actions by attendings as well as encounter-days with zero orders signed. Each observation

is an encounter-day (i,t). X_{it} contains controls such as fixed effects for the encounter's final diagnosis group, fixed effects for the current day of stay t, and consults split by specialty on both day $t-1$ and day t. The coefficient of interest is β_1: how much orders on day t change by given differences in edits on day $t-1$. Because we instrument, this Wald Statistic gives the change in orders caused by additional notes written about the patient solely because the care team was more female.

We report both OLS and Wald Estimates in Table 3.8 for all orders as well as four important categories of orders individually: diagnostic, therapeutic, medication, and monitoring orders. Diagnostic orders are orders that return a result that we categorize as primarily diagnostic in nature, such as a CT Scan. All Therapeutic includes all medications and curative procedures such as surgery. Medications are a subset of All Therapeutic orders. Monitoring orders are laboratory tests that we categorize as primarily used to monitor the patient's condition rather than diagnose their principal diagnoses. They consist of the orders that return a result that are in the empirical top 10% in quantity ordered across all inpatient encounters. Examples of monitoring orders include Complete Blood Count, Comprehensive Metabolic Panel, Measure Weight, Vital Signs, and ECG (electrocardiogram).

The first stage is positive and statistically significant: the coefficient on Fraction Female is 0.123 with a standard error of 0.0194. The interpretation is that if the composition of attendings in the care team on day $t-1$ goes from all male (Fraction Female = 0) to all female (Fraction Female = 1), then all providers will spend about 12.3% longer editing notes for the patient on day $t-1$. This is comparable to our 0.100 estimate for the full sample of physicians who belong to all specialties in Table 3.4. The standard error is a bit larger than what we would prefer for a two-step approach, but we proceed regardless.

Across the board, we find positive and statistically significant OLS coefficients. This is expected due to omitted variables bias: conditional on the controls, patients who are in worse condition require more orders and also have more important information to document.[7] Once we use the two-step Wald approach, the sign changes and coefficients become much larger in magnitude. We prefer to scale the coefficients by the first stage in discussions about their magnitude because the maximum change in the first stage is exactly the 0.123 coefficient: changing from an entirely male to an entirely female care team. Under this scaling, we find that changing from an entirely male to an entirely female care team reduces orders signed the following day by about 5.4%. The reduction in diagnostic orders is proportionally smaller, at 3.3%, but larger for therapeutic orders at 7.3%. The bulk of therapeutic orders are medications rather than procedures (e.g. surgery), and there is a similar proportional decrease of 7.5% for medications. Monitoring orders primarily used to verify that the patient's condition is stable show a smaller proportional decrease of about 2.0%. In terms of levels, this means that the increase in notes associated with changing

[7]This is apparent both from intuition and from comparing coefficient estimates from OLS estimates with no patient controls and our preferred specification with the full set of included controls (available upon request). For example, for All Orders, the OLS coefficient decreases from 0.109 to 0.055 and the R^2 increases from 0.016 to 0.222 after we add patient controls.

Table 3.8: Daily Clinical Impact of Longer Notes: Summary

	All Orders	Diagnostic	Therapeutic	Meds	Monitoring
OLS (IHS)	0.055	0.056	0.039	0.038	0.048
	(.006)	(0.007)	(0.007)	(0.007)	(0.006)
Wald Coef (IHS)	-0.436	-0.266	-0.593	-0.607	-0.162
	(0.145)	(0.160)	(0.169)	(0.174)	(0.138)
Wald Effect Size (IHS)	-0.054	-0.033	-0.073	-0.075	-0.020
Daily Mean (levels)	10.3	2.1	4.8	3.7	1.5
Mean Daily Effect (levels)	-0.55	-0.07	-0.35	-0.27	-0.03

Notes: This table shows the relationship between notes yesterday and medical orders today. It is a summary of coefficients from OLS and Wald estimates of the relationship between the inverse hyperbolic sine (IHS) of note edit time on day $t-1$ and the IHS of orders of the indicated type on day t corresponding to equations (3.5) and (3.6). The dependent variable is the IHS of the indicated order type. All Orders are any order signed. Diagnostic orders are those categorized as primarily diagnostic in nature, such as a CT Scan. All Therapeutic includes all medications and curative procedures such as surgery. Medications are a subset of All Therapeutic orders. Monitoring orders are laboratory tests that are primarily used to monitor the patient's condition rather than diagnose their principal diagnoses and consist of the orders that return a result that are in the empirical top 10% in quantity across all inpatient encounters. The Wald Effect Size scales the Wald Coefficient by the first stage estimate, which is 0.123 (0.0194), and each regression has 22,616 encounter-day observations. The daily mean across encounter-days of each order type is also reported, and the Daily Change is simply the Daily Mean times the Wald Effect Size. All regressions control continuously for ex-ante patient complexity and via fixed effects for final diagnosis category, day of stay, and specialty of any consults on day t and on day $t-1$. As expected, the sign changes from the OLS to the Wald Estimator: in the OLS, (unobservably) more complex patients tend to have more notes written about them and also have additional orders signed, but the Wald Estimator isolates the difference in notes to that correlated with physician gender, which due to random assignment of patients to physicians is uncorrelated with patient unobservables. Standard errors are clustered by encounter. See text for additional details.

from an entirely male to an entirely female care team reduces all orders the following day by about 0.55, of which the bulk are therapeutic orders (0.35 decrease). Diagnostic orders show a much smaller decrease of 0.07 orders per day.

We argue that the finding that most of the reduction is due to fewer medication orders is not inconsistent with the role of clinical notes in the patient care process. Recall that the purpose of the note is to add context to why a certain medication was prescribed or a specific test was ordered (Schrager, 2022). In this light, patients with relatively "straight-

forward" diagnoses require shorter notes than those with ambiguous conditions. However, many underlying conditions do not have a clear diagnostic. We hypothesize that a reasonable response under uncertainty is to give antibiotics and hydration, and to observe the patient's response. In that scenario, changes in patient condition under this treatment regimen informs both next steps and the final diagnosis, and the cases that require the most clinical skill and intuition. Because care teams change, recording non-quantitative information is crucial and therefore these are precisely the cases where notes may have the most value. The data align with this hypothesis: the primary medication classes driving differences in medication orders are exactly antibiotics and electrolytes.

Total Orders

Having established that more notes yesterday cause fewer orders today, we now investigate whether these results are due to a pure reduction of orders or a change in the timing of orders. We do this in a relatively simple way: we investigate how the total number of orders during the entire encounter is affected by the gender composition of the care team. In this way, it is like the reduced-form of the Wald Estimate from the previous analysis. We regress

$$\log\left(\sum_{e}\text{orders}_{et}\right) = \delta_0 + \delta_1 \frac{1}{N_j}\sum_{j}\mathbf{1}_{\left\{\text{FemaleEdit}_{j(e)t}=1\right\}} + X_e'\delta_2 + \varepsilon_e \tag{3.7}$$

$$= \delta_0 + \delta_1 \text{ProportionFemaleEdit}_e + X_e'\delta_2 + \varepsilon_e \tag{3.8}$$

Each observation in the regression is an encounter e. The dependent variable is the natural logarithm of the sum of orders signed during all days t of the encounter e. The coefficient of interest is δ_1, which is the coefficient on the proportion of edit actions performed by female hospitalists over the entire encounter, the simple unweighted mean of edits by female hospitalists during the encounter. X_e are a vector of encounter-level covariates and are selected using the LASSO technique of Belloni, et al. (2014). We impose that the algorithm selects the natural logarithm of the number of days the patient spends in the hospital as well as final diagnosis category fixed effects, as those two elements explain a great deal of variation in orders signed.[8]

Results are in Table 3.9. Panel (a) shows the effect on All Orders, Diagnostic Orders, and Therapeutic Orders caused by changes on the fraction of note edits by women in three different periods. The first column shows results for the grand total of orders signed during the encounter. The second column restricts to orders signed during the Day shift (7am to 7pm), where most of the diagnosis and treatment is attempted and achieved. The third column restricts to orders signed during the Night shift, where the primary focus is maintaining a stable condition and continuing the treatments ordered during the day. Throughout, we find

[8]One may object to the inclusion of the length of stay control as it may be endogenous to the care team randomly assigned to care for the patient and therefore a "bad control" in the language of Angrist and Pischke (2008). Results excluding this variable are qualitatively similar and the comparison for orders signed is in Appendix Table A8.

negative and statistically significant effects with point estimates similar to what we found in the day-by-day Wald Estimate in the previous analysis. For example, the reduction in all orders signed during the day is -0.050, corresponding to a 5% decrease over the entire encounter. This is very similar to the scaled daily decrease of 5.4% we found in Table 3.8. These results suggest that the daily decreases are primarily driven by a pure reduction in orders signed rather than a "shifting forward" of the same set of orders: if instead the effect was a shifting forward, then we would find no reduction in this analysis.

Next, in Panel (b) we observe that there is no statistically significant increase in the total length of the inpatient stay. The point estimate corresponds to a 2.8% increase, which would be an increase of about 5 hours relative to the mean length of stay of 7.4 days. Next, we find no change in the 30 day readmission rate, suggesting that the decrease in orders is not associated with lower-quality care. If anything, because the point estimate is negative, the quality of care increases (fewer readmission). The -0.009 point estimate is a 5 percent decrease in the readmission rate relative to the mean of 17.7%. Overall, the point estimates are relatively small and the standard errors allow us to rule out large negative (for the patient) effects. Therefore, we conclude that the decrease in orders does not affect either patient length of stay or readmissions, so additional notes allow physicians to achieve the same patient outcomes with fewer costly resources, an increase in efficiency of care.

Heterogeneity by Patient Type

Now we examine heterogeneity in note-taking and ordering behavior by patient complexity. As we did for Note Activity heterogeneity, we split the sample of patients into two subsamples based on their predicted ex-ante complexity and repeat our analysis. Results for the day-to-day Wald Estimate analysis are in Appendix Table A9. The first stage is slightly larger for the above-median complexity sample of patients (coefficient of 0.138 vs. 0.0925), although standard errors are such that we cannot reject equality. This suggests that the gender difference in note-writing is larger for more complex patients, the finding we alluded to earlier in our Patient Heterogeneity Analysis for Note Activity in Subsection 3.3.1. Although the proportional scaled Wald Effect sizes are similar for the two patient types,[9] the daily mean quantity of orders is larger for complex patients. Therefore, the mean effect in levels is larger for complex patients.

Results for the total number of orders signed paint a more nuanced picture. Results for below-median complexity patients are in Appendix Table A10. We continue to find no impact on length of stay or readmissions. Coefficients for Total and Day orders are larger in magnitude than for the full sample, but coefficients for Night orders are smaller. This suggests that there is still a pure reduction of orders for less-complex patients, but that the contribution from diagnostic and therapeutic orders during the "active" diagnosis and

[9]Standard errors are much larger for the less complex patients such that the estimated coefficients are no longer statistically significant. We do not know whether that is due to additional unexplained variation in ordering behavior for less complex patients or due to the fact that the first stage has less power for less complex patients.

Table 3.9: Encounter-Level Clinical Impact of Longer Notes

(a) Orders Signed

	All Orders, Total		All Orders, Day		All Orders, Night	
Frac. Edits by Women	-0.044	-0.076***	-0.025	-0.050***	-0.064*	-0.081***
	(0.029)	(0.015)	(0.031)	(0.016)	(0.039)	(0.028)
	Diagnostic, Total		Diagnostic, Day		Diagnostic, Night	
Frac. Edits by Women	-0.048	-0.085***	-0.028	-0.067**	-0.081**	-0.093***
	(0.039)	(0.025)	(0.040)	(0.026)	(0.036)	(0.028)
	Therapeutic, Total		Therapeutic, Day		Therapeutic, Night	
Frac. Edits by Women	-0.048	-0.068***	-0.036	-0.048***	-0.062*	-0.068**
	(0.029)	(0.016)	(0.031)	(0.017)	(0.038)	(0.028)
log(Inpatient Days)		X		X		X
Diagnosis Category FE		X		X		X
LASSO controls		X		X		X
Obs	7,375	7,373	7,375	7,375	7,375	7,373

(b) Patient Outcomes

	log(Inpatient Days)		30-day Readmissions		
Frac. Edits by Women	0.055**	0.028	0.006	-0.007	-0.009
	(0.024)	(0.021)	(0.010)	(0.010)	(0.010)
log(Inpatient Days)					X
Diagnosis Category FE		X		X	X
LASSO controls		X		X	X
Obs	7,375	7,375	7,375	7,375	7,375

Notes: This table shows the relationship between cumulative notes in an encounter and cumulative medical orders and clinical outcomes corresponding to Equation (3.7). It can be thought of as the sum of the results in Table 3.8, and the coefficients are the "reduced form" of the Wald Estimator. Each of All Orders, Diagnostic Orders, and Therapeutic Orders are separated into the Total orders signed during the encounter, the orders signed during the Day shift, when most of the diagnoses and treatment is performed, and orders signed during the Night shift, when orders are primarily continuing what was begun during the day. Each setting shows two estimates on the total Fraction of Note Edits by Women: the left estimate is one with no other controls, and the right column contains a continuous control for the natural logarithm of the total number of days of the stay, fixed effects for final diagnosis category, and a set of controls selected using LASSO using the technique of Belloni, et al. (2014). Patient Outcomes are the total length of the stay (where of course there is no control for length of stay) as well as a binary indicator for whether the patient returned to the hospital system within 30-days, a readmission suggestive of incomplete care. Results for orders signed without the Inpatient Days control are in Appendix Table A8. See text for additional details.

treatment period during the Day shift is relatively more important. Results for above-median complexity patients are in Appendix Table A11. As before, there is no impact on length of stay or on readmissions. Coefficients on orders are smaller almost across the board, and are notably smaller during the Day. This suggests that although there is still a reduction in total orders signed, a greater proportion of the day-to-day reduction in orders signed may be due to a "shifting forward" of the same set of orders, although not by enough to permit patients to be discharged from the hospital sooner.

In this subsection, we established that patients who are quasi-randomly assigned a female hospitalist benefit from additional time spent on their clinical notes in the form of fewer orders and more efficient care, with no offsetting sacrifice in quality of care. A meaningful portion of the reduction in orders is overnight, where the care team always changes, suggesting a context in which the additional documentation is useful.[10] Still, there is a potential concern that the differences we find are due to the fact that some of the female physicians in our sample practice medicine more efficiently than the overall average physician and particularly the men.[11]

First, we note that this may not be a well-defined concern. Broadly, recall that the purpose of the clinical note is to describe the patient's condition (especially changes in condition) and to put orders into context. If women are "better" physicians because they "pay more attention" or "do more things," then they will mechanically have more observations and procedures to document. If instead the effect must be transmitted solely through note content, then one possible way this could occur is that men and women do the exact same things, but women record a greater proportion of those things or greater detail about those things than men. Our finding that women sign fewer orders despite writing more directly contradicts this simple potential explanation.

Either way, we provide some circumstantial evidence supporting our interpretation that the notes themselves have value in two ways. First, we examine differences in the content of notes. Second, we examine differences in time spent reading notes by others for notes written by men and women. The next two subsections document our findings.

3.4.2 Note Text Analysis

We first examine gender differences in note content in order to investigate the mechanisms through which additional notes increase care efficiency. Specific differences in the content may inform us whether or not differences in note length are due to additional physician effort or to additional detail about the same set of actions. For instance, an effort story

[10]There is a in-person "checkout" meeting when the shifts change between the outgoing and incoming physicians, but it is typically brief. Anecdotally it is usually just a list of patients to keep an eye on and general, high-level information rather than the specific details that are contained in the daily Progress Note.

[11]Formally, if one were to instead interpret our Wald Estimator as an Instrumental Variables approach, this concern would be that the exclusion restriction fails. The required exclusion restriction is that all of the impact on orders is transmitted through the note rather than for instance differences in practice style, which may be unlikely.

could be that notes from women are longer because they ask more questions or search for more details about patient condition. A related hypothesis is that although women sign fewer orders, they do more things, chiefly physical exams, that are not entered into the EHR but are nevertheless documented in the note. On the other hand, additional detail about the same set of actions could be from additional details elicited from the patient for the same questions, or additional context provided about the same set of orders.

Because the note text data are from a separate dataset from our main EHR and Audit Log data, we are unable to control for the same set of detailed patient covariates. To reduce bias from unobservable patient condition, we must leverage a context in which patients are quasi-randomly assigned to physicians. We choose the same set of patients and physicians that we analyze in the other parts of this section: hospitalist notes for patients who are admitted to inpatient care from the emergency department. We investigate gender differences in the total number of clinical concepts, as previously extracted by UCSF in these notes, unconditional on note length.[12] We estimate regressions at the note level of the form

$$y_i = \beta_0 + \beta_1 \text{Female}_{j(i)} + X_i' \beta_2 + \varepsilon_i \tag{3.9}$$

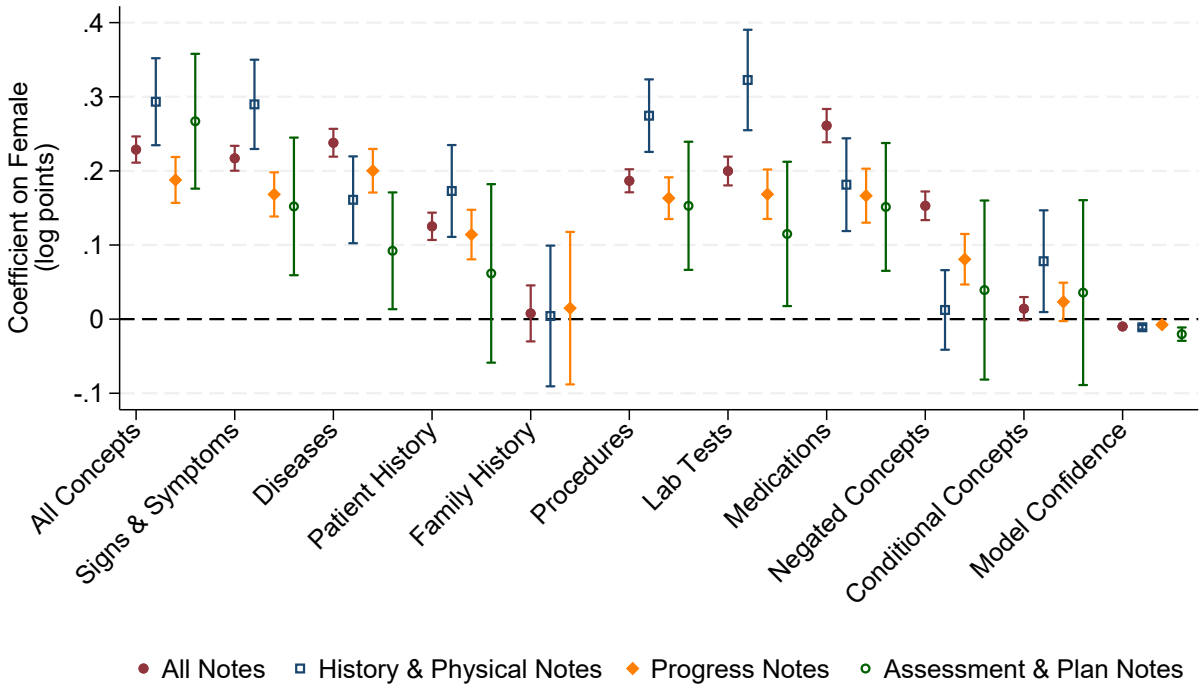
where y_i are outcomes of interest, typically the natural logarithm of the number of concepts of each type contained in note i, $\text{Female}_{j(i)}$ is the gender of the notewriter, X_i are a vector of note types (Progress Note, History and Physical Note, Assessment and Plan Note) and final diagnoses. We also estimate Equation (3.9) separately by note type.

Results are depicted in Figure 3.4, which shows the estimated coefficients and standard errors for the coefficient on female from the regression of Equation (3.9). For each dependent variable, the regressions are run four times: once pooled with note type fixed effects, and three times for the indicated note type only. The sample consists of three types of notes written by hospital medicine attending physicians for patients admitted to the hospital via the emergency department. The coefficient for All Concepts and All Notes (leftmost maroon solid circle) is about 0.23, indicating that women include about 23% additional clinical concepts in their notes compared to men, unconditional on note length.

These results allow us to test some hypotheses related to physician effort vs. documentation. First, coefficients are positive across almost all categories, meaning that female physicians include more note concepts in almost all categories in their notes. This finding suggests that in their longer notes, women are writing about more varied topics rather than providing additional details on the same set of topics. Second, coefficients for Procedures, Lab Tests, and Medications are all positive despite the fact that female physicians sign fewer orders of all three types (Table 3.9). This means that they are writing about a larger share of the orders that they do sign. Third, we find limited differences in tone. Although female physicians are more likely to include negated concepts ("Not [observation]") than men, they are not differentially likely to use conditional concepts ("Perhaps [observation]"). The former

[12] There is no data on note length in the note metadata, but in principle, we can use the length of the de-identified text.

Figure 3.4: Note Text Analysis: Note Concepts



Notes: This figure shows the estimated coefficients and standard errors for the coefficient on female from separate regressions of the natural logarithm of the sum of the note concept category indicated (or the mean Model Confidence score), final diagnosis fixed effects, and note type fixed effects (Equation (3.9)). For each dependent variable, the regressions are run four times: once pooled with note type fixed effects, and three times for the indicated note type only. The sample consists of three types of notes written by hospital medicine attending physicians for patients admitted to the hospital via the emergency department. The note types are the intake History and Physical Note and Assessment and Plan Note, as well as shorter daily Progress Notes. The coefficient for Family History in Assessment and Plan notes is missing because zero Assessment and Plan notes contain clinical concepts categorized as Family History. The coefficient for All Concepts and All Notes (leftmost maroon solid circle) is about 0.23, indicating that women include about 23% additional clinical concepts in their notes compared to men. See text for additional details.

is in line with our expectations, as part of putting things into context is ruling out alternative hypotheses and differential diagnoses. However, we expected to find a similar pattern for conditional concepts as they could be used to indicate the consideration of exactly those alternatives. Overall, these three findings suggest that at least part of the mechanism is via

the additional documentation channel, rather than solely from increased physician effort.

Next, we observe that the coefficients for Patient History are smaller than those for Signs and Symptoms and for Diseases, meaning that female physicians write proportionally more about the latter two categories than for Patient History. A possible interpretation is that while female physicians do ask the additional questions necessary to elicit more details on Patient History, they are also doing more observations and other investigations necessary to report on Signs and Symptoms and Diseases. The point estimate on Family History is zero, but standard errors are large so we cannot rule out positive effects, but family history may be less important than usual for patients admitted via the emergency department so additional detail may be clinically unnecessary.

Finally, coefficients for procedures are similar to those for lab tests and medications. This suggests that women are not performing additional procedures that do not require EHR entry (such as physical examinations) because the additional fraction of those is similar to the additional lab tests and medications, which always involve the EHR and are therefore fully observed by us.

These facts suggest that while female physicians may be exerting additional effort in the spaces of asking questions and observing and recording patient condition, they are likely not performing additional diagnostic or therapeutic procedures that are not recorded in the EHR. Therefore, the additional diagnostic and therapeutic effort mechanism may be limited.

As a robustness check, we observe that the average Model Confidence score for concepts that the algorithm extracts is very similar for men and women. The coefficient of -0.0079 is small relative to the overall standard deviation of about 0.070 and mean of 0.81. It is also likely to be too small to explain a meaningful portion of the gender difference: the mean note contains about 196 concepts, so the gender difference is about 45 concepts. We would need the NLP model to report an additional erroneous concept every time the model confidence is only 0.038 less than the mean for differences in model confidence to fully explain the difference in note concept count, which seems unrealistic relative to the standard deviation of the measure.[13]

3.4.3 Note Views by Others

As a final test of the usefulness of note content, we investigate differences in the time that that other clinical users spend reading notes written by men and by women. If the notes were not useful for clinical purposes, then others would not spend time reading them.[14] If we find no differences in reading by others conditional on note length, it is another piece of evidence supporting the documentation mechanism that it is the note content itself that contributes to increases in clinical efficiency. Specifically, we examine the time that other

[13]This assumes that the remaining concepts for women are identified with equal confidence to men. If instead we assume that the other concepts that women write about are identified with mean confidence of 0.1 standard deviations greater than men, then errors would need to be 0.066 less than the mean to explain the difference. We think this remains unlikely.

[14]Anecdotally, physicians learn quickly which note authors write useful notes.

providers spend reading the note in the seven days after the note is completed. We restrict to providers in order to exclude non-medical staff such as billers and coders who may also read the note, but for non-clinical purposes. The temporal restriction within the next week is also to restrict to clinical purposes, as some notes are read by other providers much later as a teaching example.

The results are summarized in Figure 3.5, which plots the log of minutes reading notes by others vs. log of minutes spent writing the note. Both are residualized for note type. The relationship is surprisingly linear, and notably there appear to be little difference between notes written by women (solid blue line and filled circles) vs. notes written by men (dashed maroon line and hollow squares). As with the relationship between Note Length and Time Spent Editing documented in Figure 3.3, we estimate the same regressions in Equation (3.4) and report the results in Appendix Table A12. As expected, the relevant coefficients are small in magnitude. Therefore, because the longer notes written by women are not differentially less read than longer notes written by men, we consider this additional evidence in support of the documentation mechanism: that the content of the note has clinical value.

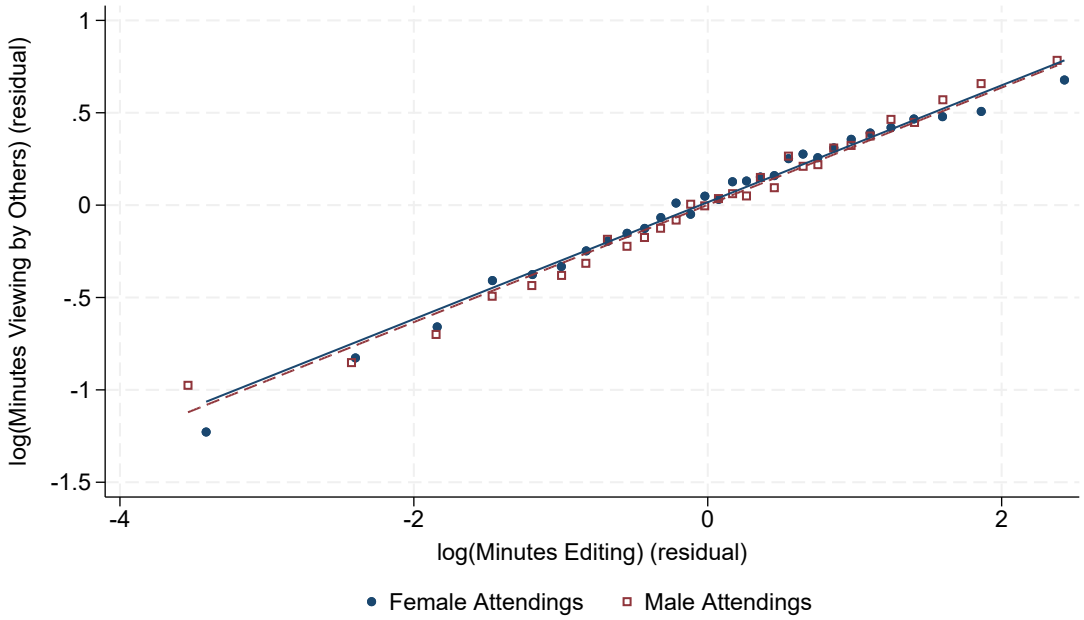
3.5 Physician Outcomes

We have showed above that women spend more time on documentation than men, and that this behavior leads to efficiency benefits for patient treatment. We next ask how these differences affect the physicians themselves. In addition to potentially leading to physician burnout (cf. Gardner, et al., 2019), additional time and mental effort spent on clinical notes may crowd out time for grant-writing and academic research, two important components for how medical school faculty are evaluated and for how they advance on the academic job ladder.[15] We investigate the correlation between 2018 note activity and physician salary in 2018, grant receipt in 2019 and 2020, and journal publications in 2019.

We consider salary a useful "reduced-form" measure of the costs and benefits of additional documentation. In classic labor economics with perfect competition, workers are paid their marginal product. Because female physicians, through note-writing, are more efficient than their male counterparts, they have higher marginal product and should be paid more. However, treating patients is not the only goal of the academic hospital and medical school: research, fund-raising through grants, and teaching are all also important. Therefore, the presence or absence of a relationship between note-taking and salary is insufficient to fully capture the costs and benefits of note-writing for physicians. Therefore, we directly examine the impact of notes on two of these other value-adding activities: grant receipt and research output. We test if additional effort on documentation "crowds out" these other productive endeavors. Although we would like to investigate burnout, as proxied by attrition, the COVID-19 pandemic is a large confounding factor that prevents us from doing so

[15]UCSF attending physicians are also permitted to "buy out" a portion of their clinical teaching responsibilities with grant funding.

Figure 3.5: Minutes Reading Notes vs. Minutes Writing Notes for Men and Women



Notes: This figure is a binned scatterplot showing the natural logarithm of minutes viewing by other clinical staff within one week of note completion vs. the natural logarithm of minutes editing by the attending physician. This measure is a direct proxy of the "utilization" of the note. Each point is one of 30 bins of notes written by a single attending. Both the log(Minutes Editing) and log(Minutes Viewing by Others) are the residuals after removing note type fixed effects and thus the slopes correspond to those in Column 2 of Appendix Table A12. This figure is an analog of Figure 3.3, just with a different dependent variable. Other clinical staff include other physicians, residents, nurse practitioners, physician assistants, and registered nurses. Excluded note readers include researchers and billing staff. See text for additional details.

without making heroic assumptions on note-taking, gender, and the immense personal and professional challenges posed by the pandemic.

For the grants and publications analysis, we restrict to the physicians who are empirically "research professors." Operationally, we keep only physicians who have at least one publication in the years 2016-2018, who we deem as "actively publishing." This excludes physicians whose primary job function is teaching, similar to "teaching faculty" in a research university who typically do not publish or apply for grants.

3.5.1 Constructing the Note Intensity Measure

Our first step is to construct a measure of physician note-taking intensity to facilitate comparisons of note activity relative to the average physician. We focus on the production of notes, and for each note, we sum the time spent performing active production actions[16] and define these as measures of Note Intensity. Conceptually, we predict the average effort that physicians with certain characteristics (excluding gender) would spend on a note of each type, controlling for patient severity. The residuals from these regression ars therefore a measure of "excess" note intensity for the physician-note. We define the simple mean of the residuals for each physician as their measure of Note Intensity, relative to the average physician in the sample. Because these estimates are estimated with error, we perform Empirical Bayes shrinkage to the grand mean prior to using them as regressors, and we will bootstrap so that the subsequent regressions have standard errors that account for uncertainty in their construction.

Formally, our raw measures of physician-level note intensity are the mean residuals $\sum_{j(i)=j} \varepsilon_i$ at the physician j level from

$$\log(\text{EditTime}_{ij}) = \beta_0 \text{NoteType}_i + \beta_1 \text{PatientChars}_{e(i)} + \beta_2 \text{AttendingChars}_j + \varepsilon_{ij} \quad (3.10)$$

Each observation is a note i written in 2018, and we consider as the main outcome the natural logarithm of total time spend producing note i by physician j.[17] We include three sets of control variables. First, we include fixed effects for note type. Second, we include ex-ante complexity and complexity-squared and a broad category of the chief complaint for each encounter e that note i is associated with, $e(i)$. Third, we include a set of fixed effects for the characteristics of physician j, excluding gender, along with the natural logarithms of total shifts they worked in 2018. Shifts worked are included because they are highly predictive of salary, and because we will include them there we also wanted to include them in this "first stage."

After we estimate this regression, we collect the residuals ε_{ij} and take the mean for each physician j. We then perform Empirical Bayes shrinkage to the grand mean over all physicians j using the standard error of each physician mean, defined as the standard deviation of residuals for physician j divided by the square root of the number of notes written by physician j minus one: $SE_j = \frac{\sigma_j}{\sqrt{N_j - 1}}$.

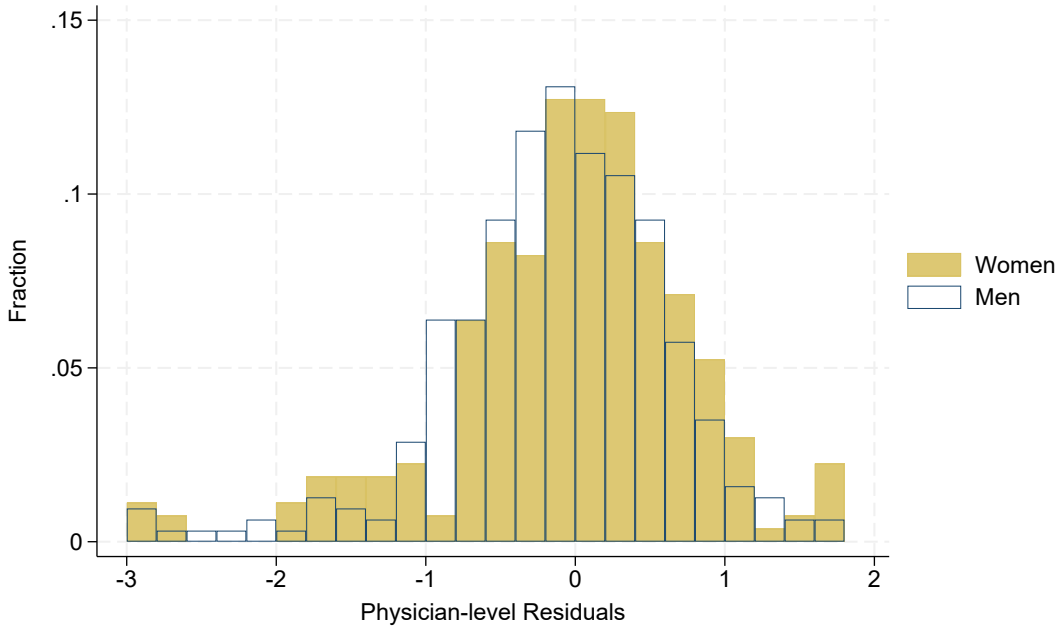
This measure has a natural interpretation. It represents the proportional difference in time spent, in log points, that each physician spends relative to the average physician in the hospital. For example, a Note Intensity measure of 0.10 would represent that the physician spends on average 10% more time on notes compared to their peers with the same specialty, seniority, and frequency of clinical work.

[16] These include create note, edit text, append, sign, cosign, pend, and correct error.

[17] We also consider other outcomes, such as the total number of production actions, and indicator for producing more than one action (the modal action is "sign" the note, which happens instantaneously in our data), and the inverse hyperbolic sine of total time. The measures are highly correlated and produce similar results when used as regressors for salary, grants, and publications.

The distribution, split by gender, of the shrunk physician note intensity is displayed in Figure 3.6. The distribution for women is in the solid gold bars, while the distribution for men is in the hollow blue bars. We can see that the female distribution is shifted slightly to the right; the difference in means is about 0.086 log points. This difference is smaller than our prior estimates, namely the 0.149 result in Table 3.5, for two reasons. First, this sample is much smaller, as it only includes physicians who have more than one edit action (otherwise shrinkage is undefined). Second, and more importantly, it only includes actions on the notes that physicians have any interaction with, whereas the estimate in Table 3.5 implicitly includes the time that physicians could potentially spend on any notes. The overall standard deviation of this measure is about 0.77.

Figure 3.6: Distribution of Physician Note Intensity



Notes: This figure plots the distribution of the note intensity measure derived from the mean residual from regressions of the natural logarithm of time spent editing each note on ex-ante patient complexity and complexity squared, chief complaint category, fixed effects for physician specialty, medical school graduation decade, and medical school rank, and the natural logarithm of total shifts worked (Equation (3.10)). Distributions are plotted separately for men (hollow blue bars) and women (solid gold bars). The difference in means is about 0.086 log points. It is smaller than our prior estimates (e.g. Table 3.5) because it only takes into account notes that the physicians have observed involvement with rather than all notes that they potentially could have edited. See text for additional details.

3.5.2 Effect of Note Intensity on Salary, Grants, and Publications

Now, we investigate the relationship between note intensity on physician salary, grants, and publications. We regress each of the outcomes of interest on the natural logarithm of shifts worked, a binary indicator for female, fixed effects for other physician characteristics, and finally on both the note intensity measure and an interaction of that intensity measure with the binary female indicator. The unit of observation is the physician j. That is, we estimate

$$y_j = \beta_0 + \beta_1 \log \text{ShiftsWorked}_j + \beta_2 \text{Female}_j + \beta_3 \text{NoteIntensity}_j + \beta_4 \text{Female}_j \cdot \text{NoteIntensity}_j + X_j' \beta_5 + \varepsilon_j \quad (3.11)$$

We add controls one-by-one. We cluster standard errors by specialty in the specifications without the Note Intensity measure, and compare these clustered standard errors to a set that are bootstrapped by specialty and patient for the specifications with Note Intensity.

The bootstrap proceeds in two nested loops. In the outer loop, we block bootstrap by specialty to preserve the clustering structure within physician specialty. In the inner loop, for each set of physicians belonging to specialties selected in the outer loop, we block bootstrap notes by the patient that the note was written for. We choose to sample at the patient level rather than the encounter level in the event that patient condition is correlated across encounters, which would lead to notes written for specific patients differ from "average" systematically across encounters for that patient.

In the inner loop, we construct and shrink a new set of physician note intensity measures for each bootstrap replication according to the steps outlined in Subsection 3.5.1. Then, we estimate Equation (3.11) and record the coefficient. This two-step bootstrap accounts for both estimation error in the "first stage" where we estimate note intensity, as well as correlation in standard errors between physicians belonging to the same specialty in the outcomes regressions. Operationally, we bootstrap 300 replications of patients for each of 300 replications of specialty blocks. Overall, bootstrapped standard errors are very similar to standard errors clustered by specialty that do not take into account the noisy first stage.[18]

We first examine the correlation of gender and note intensity with physician salary. Regression results are shown in Table 3.10. The sample of physicians is smaller than for our main analysis sample because we are not able to match all physicians in the Sacramento Bee salary data. We find a very large unconditional gender gap that is drastically reduced with the inclusion of specialty and graduation decade fixed effects. This makes sense because men are overrepresented in both higher-paying specialties (ex. surgery vs. hospital medicine) and in more senior roles (ex. full professor vs. assistant professor), which the graduation decade fixed effects proxy for. Adding the note intensity measures does not meaningfully improve model fit: R^2 increases only from 0.482 to 0.490. Furthermore, the coefficients are noisily estimated.

[18]We do not understand why the standard error on shifts worked is so much smaller in bootstrapped version vs. the clustered version in the Salary regressions of Table 3.10.

Table 3.10: The Effect of Note Intensity on Salary

	log(2018 salary)				
log(2018 shifts worked)	0.770***	0.656***	0.673***	0.674***	0.674***
	(0.200)	(0.212)	(0.202)	(0.202)	(0.061)
Female	-0.274***	-0.154***	-0.084**	-0.066*	-0.066*
	(0.052)	(0.046)	(0.038)	(0.037)	(0.040)
Note Intensity (log minutes)				-0.107*	-0.107*
				(0.059)	(0.052)
Female X Note Intensity				0.079	0.079
				(0.074)	(0.068)
Specialty FE		X	X	X	X
Med School Rank and Grad Decade FE			X	X	X
Bootstrapped Standard Errors					X
Obs	494	494	494	494	494
R-squared	0.113	0.342	0.482	0.490	0.490

Notes: This figure shows results from a regression of the natural logarithm of 2018 salary on the natural logarithm of total shifts worked in 2018 and the indicated physician fixed effects. Note Intensity is the residual, in log minutes, calculated in Subsection 3.5.1, for calendar year 2018. It represents the mean percent time the physician spends on notes relative to the average physician in the sample. The coefficient on Note Intensity can be interpreted as an elasticity: on average, a male physician's salary decreases by 1.1% for every 10% more than average they spend on notes. In the first four columns, standard errors are clustered by specialty. In the rightmost column, they are block bootstrapped by patient (which may include multiple encounters) to construct the Note Intensity measure and then by specialty to cluster the standard errors by specialty. See text for additional details.

We can interpret the coefficient estimates if we ignore the imprecision of the estimates. In our preferred specification controlling for the full set of physician characteristics and for note intensity, we find a gender gap of about 6.6%. That means that female physicians belonging to the same specialty and doing the average amount of notes compared to their peers make about 6.6% less than their male counterparts. Next, we examine the Note Intensity measure, which can be interpreted like an elasticity. A 10% increase in Note Intensity decreases male salary by about 1.1% and decreases female salary by about 0.2%. The effect of a 0.77 log point increase in Note Intensity, corresponding to one standard deviation of the measure, decreases male salary by about 8% and female salary by about 2%.

Next, we use the same method to examine the propensity to receive any grant in 2019-2020. The dependent variable is a binary indicator equal to one if the physician received any grants during 2019-2020 and zero otherwise. The sample of physicians is smaller than in the main analysis sample for two reasons. The first reason is because we restrict to physicians who we think are actively doing research, which we proxy with having at least one publication between 2016 and 2018. The second is because our grant data from NIH RePORTER are incomplete.[19] Results are in Table 3.11.

Estimates for both gender and Note Intensity are imprecise. However, point estimates suggest that there is a 6.4 percentage point gender gap in the probability of receiving grants in 2019-2020. Relative to a mean of about 21.4%, this represents a 30% decrease in the probability of receiving a grant. Additional note intensity also decreases the probability of receiving a grant for men but is near zero for women. For men, the effect of a 0.77 log point increase in Note Intensity decreases the probability of future grant receipt by about 6.6 percentage points. Therefore, if the point estimates are correct, note activity appears to crowd out grant receipt for men, but not for women.

Finally, we apply our analysis to 2019 publications. We consider all publications in all journals, regardless of author position, and we consider as the dependent variable the natural logarithm of 2019 publications. Our data are limited to the same subset of physicians as in the grants analysis. All physicians who have at least one publication in 2016-2018 and have Note Intensity measures have positive publications in 2019. Results are in Table 3.12.

Again, estimates for both gender and Note Intensity are imprecise. Point estimates suggest about a 15.5% gender gap in publications, which relative to a mean of 24.5 publications about four publications. There is also a penalty for Note Intensity for both genders, suggesting that Note Intensity crowds out research and publications. For men, a 0.77 log point increase in Note Activity decreases the number of 2019 publications by 5.9%, or 1.5 publications. The effect is slightly larger for women, about 7.2%, or 1.8 publications.

Overall, we do not find evidence that increased Note Intensity has benefits for physicians in terms of current salary, future grants, or future publications. That said, our estimates are noisy and will benefit from ongoing work to expand the sample of physicians for which we have salary, grants, and publications data. If the point estimates are accurate, then if

[19] We are working on expanding this.

Table 3.11: The Effect of Note Intensity on Future Grant Receipt

	Any Grant: 2019-2020				
log(2018 shifts worked)	-0.107**	-0.101**	-0.100**	-0.097**	-0.097**
	(0.047)	(0.049)	(0.048)	(0.048)	(0.047)
Female	-0.085	-0.087	-0.079	-0.064	-0.064
	(0.053)	(0.070)	(0.069)	(0.071)	(0.063)
Note Intensity (log minutes)				-0.086	-0.086
				(0.064)	(0.064)
Female X Note Intensity				0.103	0.103
				(0.089)	(0.084)
Specialty FE		X	X	X	X
Med School Rank and Grad Decade FE			X	X	X
Bootstrapped Standard Errors					X
Obs	299	299	299	299	299
R-squared	0.053	0.228	0.242	0.250	0.250

Notes: This table is the analog of Tables 3.10 and 3.12, except here the dependent variable is a binary indicator for receiving any grant in calendar years 2019 and 2020. Note Intensity is for calendar year 2018. The dependent variable mean (probability of any grant receipt in 2019 or 2020) is about 21.4% with a median of 0. See notes to Table 3.10 and text for additional details.

Table 3.12: The Effect of Note Intensity on Future Publications

	log(2019 Publications)				
log(2018 shifts worked)	-0.014	-0.087	-0.069	-0.070	-0.070
	(0.097)	(0.093)	(0.092)	(0.095)	(0.100)
Female	-0.256*	-0.150	-0.166	-0.155	-0.155
	(0.147)	(0.148)	(0.129)	(0.132)	(0.126)
Note Intensity (log minutes)				-0.077	-0.077
				(0.128)	(0.139)
Female X Note Intensity				-0.016	-0.016
				(0.207)	(0.190)
Specialty FE		X	X	X	X
Med School Rank and Grad Decade FE			X	X	X
Bootstrapped Standard Errors					X
Obs	299	299	299	299	299
R-squared	0.018	0.305	0.388	0.391	0.391

Notes: This table is the analog of Tables 3.10 and 3.11, except here the dependent variable is the natural logarithm of the number of publications by the physician in 2019. Note Intensity is for calendar year 2018. All physicians in this limited sample have positive 2019 publications. The dependent variable mean (in levels) is 24.5, with a median of 16. See notes to Table 3.10 and text for additional details.

anything, increased Note Intensity harms physician outcomes and career advancement, but does not do so disproportionately for women compared to men.

3.6 Discussion and Conclusion

We have leveraged detailed data to show that even within a single, highly standardized environment, male and female attending physicians exhibit meaningful differences in how they carry out their jobs. Women spend about 10% more time on notes per shift than men with the same specific job description. The additional time spent is concentrated during the work day, rather than remotely from home before or after the shift. We find that the additional notes are correlated with more efficient patient care: female physicians, who write longer notes for their patients, are able to achieve the same patient outcomes with fewer costly resources. In principle, these differences could be due to comparative advantage (Chandra and Staiger, 2007) or to practice style (Cutler, et al., 2019), but we argue using the timing of resource savings, analysis of note content, and analysis of note utilization by others that a portion of the benefit is likely due to additional details recorded in the notes. Finally, we

examine whether the value-adding task of additional note writing leads to pecuniary benefits to physicians or instead crowds out other activity necessary for career advancement. We find suggestive evidence that note taking is costly to physicians employed in academic medicine: physicians of both genders who take more notes earn less money, are less likely to receive a research grant, and publish fewer articles in the future.

Our findings complement existing work showing differences along the intensive margin in medicine (cf. Reyes, 2007) and have implications not only for understanding gender inequities among physicians, but also for understanding variation in both patient outcomes and in treatment styles. Although we study academic medicine, meaningful differences in tasks performed may be important in jobs across the economy. This is especially important if even specific job titles at the same firm are insufficient to fully describe heterogeneity in tasks performed by men and women, as we found in the academic hospital we study. These differences may be important for both the gender gap in earnings as well as the gap in promotions (cf. Blau and Khan, 2017; Haegele, 2024) across the economy, and may also relate to productivity differences across otherwise observationally similar firms.

Bibliography

Adhvaryu, Nyshadham, and Tamayo (2023), "Managerial Quality and Productivity Dynamics," *Review of Economic Studies*.

Altonji, Elder, and Taber (2005), "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*.

American College of Emergency Physicians (2023), "Gauging Emergency Physician Productivity: Are RVUs the Answer?" https://www.acep.org/imports/clinical-and-practice-management/resources/reimbursement-imported/gauging-emergency-physician-productivity-are-rvus-the-answer, accessed November 7, 2023.

American Medical Association (2017), "What it's like to specialize in emergency medicine: Shadowing Dr. Clem," https://www.ama-assn.org/medical-students/specialty-profiles/what-it-s-specialize-emergency-medicine-shadowing-dr-clem, accessed October 25, 2023.

American Medical Association (2019), "Applying to more than 1 medical specialty? What you should know," https://www.ama-assn.org/medical-students/specialty-profiles/applying-more-1-medical-specialty-what-you-should-know, accessed November 3, 2023.

Angrist JD and JS Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*.

Arrow, Kenneth (1963), "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*.

Association of American Medical Colleges (2022),"Economic Impact of AAMC Medical Schools and Teaching Hospitals."

Bajari, Patrick, C. Lanier Benkard, and Jonathan Levin (2007), "Estimating Dynamic Modelsof Imperfect Competition," *Econometrica*.

Belloni, Chernozhukov, and Hansen (2014), "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*.

Benbassat J, and Taragin M (2000), "Hospital Readmissions as a Measure of Quality of Health Care: Advantages and Limitations," *Arch Intern Med.*

Benkard, C. Lanier (2000), "Learning and Forgetting: The Dynamics of Aircraft Production," *The American Economic Review.*

Bergeron, Augustin, Pedro Bessone, John Kabeya Kabeya, Gabriel Tourke, and Jonathan L. Weigel (2022), "Optimal Assignment of Bureaucrats: Evidence from Randomly Assigned Tax Collectors in the DRC," *NBER Working Paper 30413.*

Blau, Francine and Lawrence Kahn (2017), "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature.*

Bloesch, Justin and Jacob P. Weber (2023), "Congestion in Onboarding Workers and Sticky R&D," *Working Paper.*

Brickley, James and R. Lawrence Van Horn (2015), "Managerial Incentives in nonprofit Organizations: Evidence from Hospitals," *Journal of Law and Economics.*

Buser, Niederle, and Oosterbeek (2004), "Gender, Competitiveness, and Career Choices," *The Quarterly Journal of Economics.*

Capps, Carlton, and David (2017), "Antitrust Treatment of Nonprofits: Should Hospitals Receive Special Care?" *NBER Working Paper 23131.*

Chan, David (2018), "The Efficiency of Slacking Off: Evidence from the Emergency Department," *Econometrica.*

Chan, David (2021), "Influence and Information in Team Decisions: Evidence from Medical Residency," *American Economic Journal: Economic Policy.*

Chandra and Staiger (2007), "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy.*

Chen, Yiqun (2021), "Team-Specific Human Capital and Team Performance: Evidence from Doctors," *The American Economic Review.*

Cheng, Yi (2019), "The Unexpected Costs of Expertise: Evidence from Highly Specialized Physicians," *Working Paper.*

Chu, Bryan, Ben Handel, Jonas Knecht, Jon Kolstad, Filip Matějka, and Ulrike Malmendier (2023), "The Effect of Fatigue and Cognitive Load on Medical Provider Decision-Making and Patient Health Outcomes," *Working Paper.*

Clemens, Jeffrey and Joshua D. Gottlieb (2014), "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review.*

Coffman, Katherine (2014), "Evidence on Self-Stereotyping and the Contribution of Ideas," *The Quarterly Journal of Economics*.

Cowgill, Bo, Jonathan M.V. Davis, B. Pablo Montagnes, and Patryk Perkowski (2023), "Matchmaking Principles: Theory and Evidence from Internal Talent Markets," *Working Paper*.

Currie, MacLeod, and Van Parys (2016), "Provider Practice Style and Patient Health Outcomes: The Case of Heart Attacks," *Journal of Health Economics*.

Currie, Janet and Bentley MacLeod (2020), "Understanding Doctor Decision Making: The Case of Depression Treatment," *Econometrica*.

Cutler, D., Skinner, J. S., Stern, A. D., and Wennberg, D. (2019), "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," *American Economic Journal: Economic Policy*.

Czeisler MÉ, Marynak K, Clarke KE, et al. (2020), "Delay or Avoidance of Medical Care Because of COVID-19–Related Concerns — United States," *MMWR Morbidity and Mortality Weekly Report*, June 2020.

Dahlstrand, Amanda (2023), "Defying Distance? The Provision of Services in the Digital Age," *Working Paper*.

Doyle Jr., 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*.

Duggan, Mark (2000), "Hospital Ownership and Public Medical Spending," *The Quarterly Journal of Economics*.

Feldstein, Martin (1971), "Hospital Cost Inflation: A Study of nonprofit Price Dynamics," *The American Economic Review*.

Finkelstein, Gentzkow, Li, and Williams (2022), "What Drives Prescription Opioid Abuse? Evidence from Migration," *NBER Working Paper*.

Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, Linzer M. (2019), "Physician stress and burnout: the impact of health information technology," *J Am Med Inform Assoc.* 2019 Feb 1;26(2):106-114.

Gaynor, Martin (2006), "What Do We Know About Competition and Quality in Health Care Markets?" *Foundations and Trends in Microeconomics*.

Gaynor, Martin and William Vogt (2003), "Competition among Hospitals," *The RAND Journal of Economics*.

Gaynor, Martin and Robert J. Town (2012), "Competition in Health Care Markets," *The Handbook of Health Economics*.

Gneezy, Niederle, and Rustichini (2003), "Performance in Competitive Environments: Gender Differences," *The Quarterly Journal of Economics*.

Goldin, Claudia (2014), "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*.

Goldin and Mitchell (2017), "The New Life Cycle of Women's Employment: Disappearing Humps, Sagging Middles, Expanding Tops," *Journal of Economic Perspectives*.

Gong, Qing (2018), "Physician Learning and Treatment Choices: Evidence from Brain Aneurysms," *Working Paper*.

Gunja, Munira Z., Evan D. Gumas, and Reginald D. Williams II (2023), "U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes," *Commonwealth Fund*, Jan. 2023.

Gupta, Murray, Sarkar, Mourad, and Adler-Milstein (2019), "Differences in Ambulatory EHR Use Patterns for Male vs. Female Physicians," *NEJM Catalyst*.

Haegele, Ingrid (2024), "The Broken Rung: Gender and the Leadership Gap," *Working Paper*.

Hausknecht, John P. and Charlie O. Trevor (2011), "Collective Turnover at the Group, Unit, and Organizational Levels: Evidence, Issues, and Implications," *Journal of Management*.

Holmgren AJ, Phelan J, Jha AK, Adler-Milstein J (2022), "Hospital organizational strategies associated with advanced EHR adoption," *Health Serv Res.* 2022; 57: 259–269.

Hotz, V. Joseph and Robert A. Miller (1993), "Conditional Choice Probabilities and the Estimation of Dynamic Models," *The Review of Economic Studies*.

Hughes, Emily (2017), *Canadian Medical Association Journal* 2017 August 14;189:E1050-1.

Huilgol YS, Adler-Milstein J, Ivey SL, Hong JC (2022), "Opportunities to use electronic health record audit logs to improve cancer care," *Cancer Med.* 2022 Mar 29.

Jackson TD, Wannares JJ, Lancaster RT, Rattner DW, Hutter MM (2011), "Does speed matter? The impact of operative time on outcome in laparoscopic surgery," *Surgical Endoscopy*, 2011 Jul;25(7):2288-95. doi: 10.1007/s00464-010-1550-8. Epub 2011 Feb 7. PMID: 21298533; PMCID: PMC3676304.

Jhajj S, Kaur P, Jhajj P, Ramadan A, Jain P, Upadhyay S, Jain R (2022), "Impact of Covid-19 on Medical Students around the Globe," *Journal of Community Hospital Internal Medicine Perspectives*.

Jovanovic, Boyan (2014), "Misallocation and Growth," *American Economic Review*.

Kasy, Maximilian and Alexander Teytelboym (2022), "Matching with semi-bandits," *The Econometrics Journal*.

Kocher, Bob and Robert M. Wachter (2023), "Why is it so Hard for Academic Medical Centers to Succeed in Value-Based Care?" *Health Affairs Scholar*.

Kolstad, Jonathan (2013), "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *The American Economic Review*.

Lassner, Jared W., et al. (2022a), "Growth of For-Profit Involvement in Emergency Medicine Graduate Medical Education and Association Between For-Profit Affiliation and Resident Salary," *AEM Education and Training*.

Lassner, Jared W., et al. (2022b), "Quantifying For-Profit Outcomes in GME: A Multi-specialty Analysis of Board Certifying Examination Pass Rates in For-Profit Affiliated Residency Programs," *Journal of Graduate Medical Education*.

Lakdawalla, Darius, and Tomas Philipson (1998), "Nonprofit Production and Competition," *NBER Working Paper 6377*.

Lerner, Josh, Carolyn Stein, and Heidi Williams (2023), "The Wandering Scholars: Understanding the Heterogeneity of University Commercialization," *Working Paper*.

Levitt, List, and Syverson (2013), "Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant," *Journal of Political Economy*.

Ludmerer, Kenneth M. (2005), *Time to Heal: American Medical Education from the Turn of the Century to the Era of Managed Care*, New York: Oxford University Press.

Ludmerer, Kenneth M. (2015), *Let Me Heal: The Opportunity to Preserve Excellence in American Medicine*, New York: Oxford University Press.

Marmor, Theodore R. and Robert W. Gordon (2021), "Commercial Pressures on Professionalism in American Medical Care: From Medicare to the Affordable Care Act," *Journal of Law, Medicine, and Ethics*.

McClellan, Mark (2011), "Reforming Payments to Healthcare Providers: The Key to Slowing Healthcare Cost Growth While Improving Quality?" *Journal of Economic Perspectives*.

Minni, Virginia (2023), "Making the Invisible Hand Visible: Managers and the Allocation of Workers to Jobs," *Working Paper*.

Molitor, David (2018), "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration," *American Economic Journal: Economic Policy*.

Newhouse, Joseph (1971), "Toward a Theory of Nonprofit Institutions: An Economic Model of a Hospital," *The American Economic Review*.

Niederle and Vesterlund (2011), "Gender and Competition," *The Annual Review of Economics*.

Reagans Ray, Linda Argote, and Daria Brooks (2005), "Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together," *Management Science*.

Reder, M.W. (1965), "Some Problems in the Economics of Hospitals," *The American Economic Review*.

Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services (2022), "Impact of the COVID-19 pandemic on the hospital and outpatient clinician workforce: challenges and policy responses (Issue Brief No. HP-2022-13)," May 2022.

Oster, Emily (2019), "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business and Economic Statistics*.

Pakes, Ariel, Michael Ostrovsky, and Steven Berry (2007), "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *RAND Journal of Economics*.

Patel, R.S., Bachu, R., Adikey, A., Malik, M., and Shah, M. (2018), "Factors Related to Physician Burnout and Its Consequences: A Review," *Behav. Sci.*

Pauly, Mark and Michael Redisch (1973), "The Not-For-Profit Hospital as a Physicians' Cooperative," *The American Economic Review*.

Reyes, Jessica Wolpaw (2007), "Gender Gaps in Income and Productivity of Obstetricians and Gynecologists," *Obstetrics & Gynecology*.

Sarsons, Heather (2019), "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," *Working Paper*.

Sasser, Alicia (2005), "Gender Differences in Physician Pay: Tradeoffs between Career and Family," *The Journal of Human Resources*.

Schrager, Sarina (2022), "Writing Clinical Notes: Have We Made Progress?" *Fam Pract Manag*.

Sloan, Frank A. (2021), "Quality and Cost of Care by Hospital Teaching Status: What are the Differences?" *The Milbank Quarterly*.

Sokol, Emily (2020), "Private Payers Outpace Public Insurance in Value-Based Care Push," https://revcycleintelligence.com/news/private-payers-outpace-public-insurance-in-value-based-care-push, accessed October 28, 2023.

Song, Hummy, Robert S. Huckman, and Jason R. Barro (2016), "Cohort Turnover and Operational Performance: The July Phenomenon in Teaching Hospitals," *Working Paper*.

Syverson, Chad (2011), "What Determines Productivity?" *Journal of Economic Literature*.

Thomas-Hunt and Phillips (2004), "When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups" *Personality and Social Psychology Bulletin*.

Ulmer, Cheryl, Diane Miller, and Michael M.E. Johns, eds. (2009), *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety*, Washington, D.C.: National Academies Press.

U.S. Bureau of Economic Analysis, "Value Added by Industry" (accessed Thursday, October 19, 2023)

Verma G, Ivanov A, Benn F, Rathi A, Tran N, et al. (2020), "Analyses of electronic health records utilization in a large community hospital," *PLOS ONE* 15(7): e0233004.

Wang, Xixi, et al. (2022), "Residency Attrition and Associated Characteristics, a 10-Year Cross Specialty Comparative Study," *Journal of Brain and Neurological Disorders*.

Wasserman, Melanie (2023), "Hours Constraints, Occupational Choice, and Gender: Evidence from Medical Residents," *Review of Economic Studies*.

Wei, Eric, Laura Sarff, and Brad Spellberg (2019), "Debunking the July Effect Myth." *Journal of Patient Safety*.

Appendix A

Appendix

A.1 Prediction of Inpatient Admission

In this section I provide a brief overview of the ex-ante prediction of inpatient admission. The key feature of this prediction is that only factors that are immutable (e.g. patient age) and determined prior to the physician's involvement (e.g. abnormal vital signs upon entry; chief complaint as recorded by the triage nurse) are included in the prediction. Therefore, it is by construction exogenous to the providers who will subsequently care for the patient.

I use LASSO to select among the large set of ex-ante and immutable patient covariates, with a logit functional form because inpatient admission is a binary outcome. The predictions fit observed patterns of inpatient admission well: the area under the receiver operating characteristic curve (AUC) is around 0.97. One way to interpret AUC is that it is the probability that the model ranks a random positive example more highly than a random negative example. The maximum value is 1, so a value of 0.97 indicates that the model is very successful at predicting the observed outcome.

Similar results are obtained whether the functional form is a linear probability model or a probit.

A.2 Alternative Hospital Objective Function

An alternative flow utility function I consider is one where the hospital maximizes a weighted sum of patient care quality and resident training.

$$\max_{S} \sum_{t=0}^{\infty} \beta^t [\phi L(S_t; X_t) + (1-\phi) K(S_t; X_t, AY(t)) + \varepsilon_{St}] \tag{A.1}$$

L is the hospital's utility from length of stay (care quality). This is a function of the allocation of patients S_t and the state variable of resident knowledge X_t. K is the hospital's utility from training ("knowledge"), which depends on the allocation of patients S_t, the state variable X_t, and the result of a function AY that maps time t to relative time within the academic year. The weight on care quality ϕ is to be estimated.

It turns out that the weighted average specification of Equation page 97 actually does not result in a stable steady state of training. Instead, the steady-state utility-maximizing patient allocation is a two-year cycle where every other cohort is trained. The result holds for all chosen values of the discount rate, as well as all three concavities of utility from length of stay and making the utility from learning more concave than the utility from patient care. This is because for each cohort, training today and training tomorrow are intertemporal complements. A larger amount of training today means that the cost of training tomorrow is decreased because the residents have higher skill. The presence of attending physicians amplifies this feature: instead of distributing training among the two cohorts, it is better to focus it on one cohort and give the remaining patients to the attending to maximize patient utility.

To build intuition for this result, consider a model where there is only one period per academic year. In the first year, if the senior cohort received training in year zero, then it is cheaper to train them further than to train the new cohort, and it is better for patient outcomes if the hospital allocates patients only between the senior cohort and the attendings. In the second year, the senior residents have zero skill because they received no training when they were junior residents. The hospital chooses to ignore them, trains only the new cohort, and gives the remaining patients to the attendings to maximize the patient utility portion of utility. Then, in year three, it trains the senior cohort, because it has already trained them in the previous year, and again ignores the new cohort. Consequently, every other year, both training utility and patient utility are high and we have a two-year cycle.

A.3 Additional Tables

Table A1: Learning Over Time: Immediate Orders Upon ED Admission

	First Order Upon Admission					
Years in Program	0.010**	0.010***	0.017*	0.017**	0.006	0.008**
	(0.005)	(0.004)	(0.010)	(0.008)	(0.005)	(0.004)
log(Days in Program)	0.006	0.006	0.015	0.012	0.001	0.004
	(0.005)	(0.004)	(0.011)	(0.008)	(0.005)	(0.004)
DepVar Mean	0.138		0.175		0.123	
Patient Type	All Patients		Complex		Simple	
Controls		X		X		X
Obs	31,334	31,334	8,828	8,828	22,506	22,506

Notes: Separate regressions of a binary indicator for first medical order upon admission to the ED on measures of resident experience, split by ex-ante predicted patient complexity. This process measure is a complement to the log(Minutes to First Order) outcome in Tables 1.3 and 1.4, which is undefined when orders are immediate. The dependent variable is equal to one if the first medical order is signed at or before the patient is moved from the waiting room to an examination room and zero otherwise. Standard errors are clustered by physician. See text and corresponding Table Notes for additional details.

Table A2: Learning Over Time: Diagnostic Orders

	Any Diagnostic Order					
Days in Program	-0.001	-0.000	0.000	-0.000	-0.002	0.001
	(0.003)	(0.003)	(0.002)	(0.002)	(0.004)	(0.004)
log(Days in Program)	0.005	0.005	-0.001	-0.001	0.006	0.007
	(0.004)	(0.003)	(0.002)	(0.002)	(0.006)	(0.004)
DepVar Mean	0.939		0.992		0.919	
Patient Type	All Patients		Complex		Simple	
Obs	31,334	31,334	8,828	8,828	22,506	22,506
	log(Diagnostic Orders)					
Days in Program	0.012	0.012	-0.020	-0.017	0.020	0.035***
	(0.014)	(0.010)	(0.019)	(0.015)	(0.015)	(0.012)
log(Days in Program)	0.024*	0.009	-0.020	-0.026	0.025	0.031**
	(0.013)	(0.011)	(0.017)	(0.017)	(0.016)	(0.013)
DepVar Mean	12.079		19.509		9.165	
Patient Type	All Patients		Complex		Simple	
Obs	29,434	29,434	8,755	8,755	20,679	20,679

Notes: This table presents regressions of outcomes related to diagnostic orders signed on two parameterizations of resident experience, split by ex-ante predicted patient complexity. It is similar to Tables 1.3 and 1.4. Dependent variable means are listed, always in levels. Diagnostic Orders are medical orders primarily for gathering information about the patient, such as lab tests and imaging, rather than for treating or stabilizing the patient. Any Diagnostic Order Signed is a binary variable equal to one if at least one diagnostic order was signed and zero otherwise. log(Diagnostic Orders Signed) is the natural logarithm of the number of diagnostic orders signed and is undefined when zero orders are signed (for instance, if the patient required stitches but did not receive an X-Ray prior to the procedure). Standard errors are clustered by physician. See text and notes to Tables 1.3 and 1.4 for additional details.

Table A3: When does Supervision Occur and Change for Complex Patients?

	Fraction Orders Signed by Resident			
Years in Program	0.047***	0.008	0.010	-0.000
	(0.008)	(0.011)	(0.009)	(0.006)
Period	1st Hour	2nd Hour	Middle	Last Hour
DepVar Mean	0.430	0.505	0.451	0.176
Num Orders	11.8	4.1	11.1	4.3
Obs	9,196	8,632	7,648	9,196

Notes: Regressions of the fraction of orders signed by the resident during various portions of the patient's stay in the ED on the number of fractional years in the program. If the patient stay is less than or equal to two hours, the second hour is counted only as part of the Last Hour. "Middle" includes all hours after hour three and prior to the last hour before inpatient upgrade (for admitted patients) or discharge (for discharged patients). The dependent variable mean is listed, as is the mean number of orders signed during the period. Most of the change occurs in the first hour, which is also when the bulk of the orders are signed. I select a similar set of patient covariates as in the binned scatterplots of Figure 1.3, but additionally include the number of simultaneous patients managed by the resident and its square. Standard errors are clustered by physician.

Table A4: Allocation of Complex Patients: Congestion

	Patient Assigned to 1st Year Resident			
# Complex Pt in ED	0.976	0.972	0.973	0.973
	(0.017)	(0.018)	(0.018)	(0.018)
# Pt in Waiting Room	0.986	0.971***	0.970***	0.969***
	(0.011)	(0.010)	(0.010)	(0.010)
Likely Handoff				0.421**
				(0.149)
# other EM1 FE	Y	Y	Y	Y
Month FE		Y	Y	Y
Patient Condition FE			Y	Y
Other Controls				Y
Obs	6,903	6,903	6,896	6,896

Notes: Odds ratios reported. Going from the 25th to 75th percentile of patients in the waiting room lowers the probability of assignment to a first year resident by 15 percentage points. # of other EM1 FE are fixed effects for the number of other first year residents on shift at the time of patient allocation: clearly it is more likely to assign a patient to a first year resident when there are more of them working. Not shown are the coefficients for patient condition fixed effects, which include interactions between ex-ante triage nurse estimated severity and chief complaint. Standard errors are clustered by physician.

Table A5: Dynamic Results: Full Estimates

Specification	β	L^*	RMSE	L^* Implied Hrs	Graduating Skill (Hrs)
Linear	0.90	-1.650	.0090	6.598	6.349
	0.95	-1.649	.0081	6.594	6.354
	0.99	-1.646	.0080	6.585	6.363
Quadratic	0.90	-10.936	.0104	6.614	6.351
	0.95	-10.925	.0077	6.611	6.349
	0.99	-10.907	.0076	6.605	6.356
Log	0.90	0.195	.0099	6.645	6.355
	0.95	0.197	.0085	6.636	6.359
	0.99	0.197	.0085	6.636	6.359

Notes: This table shows the full estimation results for the three functional forms of hospital utility for patient length of stay and three values of the discount rate β. The first two columns show the estimated lower bound of patient quality L^* in utils, as well as the model's root mean squared error compared to the observed patient assignment shares. L^* implied hours coverts the utils to hours, and I also show the graduating skill of the resident, also in hours per patient. There is not much of a difference between specifications.

Table A6: Distribution of Average Number of Complex Patients Per Shift, by Quarter

Academic Quarter	Mean Pt/Shift	SD Pt/Shift
1	1.573	.717
2	1.587	.775
3	1.659	.658
4	1.736	.790

Notes: This table shows the mean and standard deviation for the number of complex patients seen per shift by residents during each academic quarter. The mean increases, as expected (see Figure 1.2), but the standard deviation is relatively stable. The stability in the standard deviation suggests that patient allocation patterns are not mean-reverting in the sense that it is likely that allocation patterns are similar in the hospital where I do not have data from. If they were, then I would expect a lower standard deviation in later academic quarters, where due to the law of large numbers, variation in patients seen due to exogenous factors such as ED congestion should be more similar across residents.

Table A7: Characters Added to Notes vs. Time Spent Editing for Men and Women

	log(note length, character count)					
log(minutes editing)	0.474***	0.319***	0.317***	0.268***	0.264***	0.263***
	(0.028)	(0.017)	(0.017)	(0.018)	(0.017)	(0.017)
Female	0.019	0.058	0.062	-0.043	-0.040	-0.045
	(0.115)	(0.065)	(0.066)	(0.065)	(0.065)	(0.064)
Female X log(mins editing)	0.007	-0.007	-0.007	0.019	0.020	0.018
	(0.038)	(0.020)	(0.020)	(0.018)	(0.019)	(0.018)
Patient Ex-Ante Complexity			0.071***	0.062***	0.062***	0.061***
			(0.007)	(0.007)	(0.007)	(0.007)
Note Type FE		X	X	X	X	X
Specialty FE				X	X	X
Grad Decade FE					X	X
Med School Rank FE						X
Obs	100,765	100,765	100,765	100,765	100,765	100,765
R-squared	0.389	0.607	0.611	0.646	0.650	0.653
Adj R-squared	0.389	0.607	0.611	0.646	0.650	0.653

Notes: This table shows results from regressions on the natural logarithm of total character count of notes on the natural logarithm of minutes the notewriter spent editing it, an indicator for female, the interaction of those two terms, a continuous measure of ex-ante patient complexity, and fixed effects for note and physician characteristics. The measure of complexity is constructed with patient attributes that are either immutable or observed prior to any physician involvement by the emergency department triage nurse. Therefore, it is not endogenous to the physicians who subsequently care for the patient. The sample is all notes that are edited by a single attending physician. Results suggest that men and women have similar rates of note production: the interaction of female and time spent editing is close to zero and statistically insignificant. Therefore, on average, because women are spending longer writing notes, they are writing longer notes as a result. See text for additional details.

Table A8: Encounter-Level Clinical Impact of Longer Notes: without log(Inpatient Days)

	All Orders, Total		All Orders, Day		All Orders, Night	
Frac. Edits by Women	-0.049**	-0.076***	-0.022	-0.050***	-0.064**	-0.081***
	(0.023)	(0.015)	(0.025)	(0.016)	(0.031)	(0.028)
	Diagnostic, Total		Diagnostic, Day		Diagnostic, Night	
Frac. Edits by Women	-0.059*	-0.085***	-0.043	-0.067**	-0.072**	-0.093***
	(0.031)	(0.025)	(0.033)	(0.026)	(0.029)	(0.028)
	Therapeutic, Total		Therapeutic, Day		Therapeutic, Night	
Frac. Edits by Women	-0.044*	-0.068***	-0.022	-0.048***	-0.056*	-0.068**
	(0.023)	(0.016)	(0.025)	(0.017)	(0.030)	(0.028)
log(Inpatient Days)		X		X		X
Diagnosis Category FE	X	X	X	X	X	X
LASSO controls	X	X	X	X	X	X
Obs	7,373	7,373	7,373	7,375	7,373	7,373

Notes: This table is just like Table 3.9, except that it compares specifications with and without the log(Inpatient Days) control in Equation (3.7). See text and notes to Table 3.9 for additional details.

Table A9: Daily Clinical Impact of Longer Notes: Heterogeneity

(a) Less Complex Patients

	All Orders	Diagnostic	Therapeutic	Meds	Monitoring
OLS (IHS)	0.036	0.035	0.033	0.031	0.027
	(.010)	(0.011)	(0.011)	(0.012)	(0.010)
Wald Coef (IHS)	-0.474	-0.301	-0.693	-0.725	0.075
	(0.328)	(0.324)	(0.393)	(0.410)	(0.281)
Wald Effect Size (IHS)	-0.044	-0.028	-0.064	-0.067	0.007
Daily Mean (levels)	8.7	1.6	4.1	3.1	1.3
Mean Daily Effect (levels)	-0.38	-0.04	-0.26	-0.21	0.01

(b) More Complex Patients

	All Orders	Diagnostic	Therapeutic	Meds	Monitoring
OLS (IHS)	0.064	0.065	0.042	0.040	0.059
	(.007)	(0.009)	(0.008)	(0.009)	(0.008)
Wald Coef (IHS)	-0.407	-0.230	-0.548	-0.564	-0.234
	(0.159)	(0.177)	(0.183)	(0.189)	(0.162)
Wald Effect Size (IHS)	-0.056	-0.032	-0.076	-0.078	-0.032
Daily Mean (levels)	11.6	2.4	5.3	4.1	1.7
Mean Daily Effect (levels)	-0.65	-0.08	-0.40	-0.32	-0.05

Notes: This table is similar to Table 3.8, except that the sample of encounters is split into those with below- and above-median predicted ex-ante complexity. Panel (a) shows the results for the less complex 50% of encounters, which comprise 8,017 encounter-days. The first stage coefficient of the Wald Estimator is 0.0925 (0.033). Panel (b) shows the results for the more complex 50% of encounters, which comprise 14,599 encounter-days. The first stage coefficient of the Wald Estimator is 0.138 (0.024). See text and notes to Table 3.8 for additional details.

Table A10: Sum of Notes and Orders: Less Complex Patients

(a) Orders Signed

	All Orders, Total		All Orders, Day		All Orders, Night	
Frac. Edits by Women	-0.022	-0.108***	0.002	-0.083***	-0.018	-0.073*
	(0.039)	(0.024)	(0.040)	(0.025)	(0.053)	(0.041)
	Diagnostic, Total		Diagnostic, Day		Diagnostic, Night	
Frac. Edits by Women	0.019	-0.090**	0.046	-0.062	-0.017	-0.072*
	(0.050)	(0.037)	(0.051)	(0.038)	(0.043)	(0.038)
	Therapeutic, Total		Therapeutic, Day		Therapeutic, Night	
Frac. Edits by Women	-0.027	-0.099***	-0.012	-0.077***	-0.023	-0.057
	(0.038)	(0.025)	(0.041)	(0.026)	(0.050)	(0.041)
log(Inpatient Days)		X		X		X
Diagnosis Category FE		X		X		X
LASSO controls		X		X		X
Obs	3,688	3,686	3,688	3,688	3,688	3,686

(b) Patient Outcomes

	log(Inpatient Days)		30-day Readmissions		
Frac. Edits by Women	0.105***	0.043	0.005	-0.003	-0.003
	(0.031)	(0.030)	(0.013)	(0.013)	(0.013)
log(Inpatient Days)					X
Diagnosis Category FE		X		X	X
LASSO controls		X		X	X
Obs	3,688	3,688	3,688	3,688	3,688

Notes: This table is similar to Table 3.9, except that it shows results for the below-median ex-ante complexity sample of encounters. Results for the above-median complexity encounters are shown in Appendix Table A11. See text and notes to Table 3.9 for additional details.

Table A11: Sum of Notes and Orders: More Complex Patients

(a) Orders Signed

	All Orders, Total		All Orders, Day		All Orders, Night	
Frac. Edits by Women	-0.102***	-0.033*	-0.090**	-0.012	-0.151***	-0.071*
	(0.039)	(0.020)	(0.042)	(0.021)	(0.053)	(0.038)
	Diagnostic, Total		Diagnostic, Day		Diagnostic, Night	
Frac. Edits by Women	-0.161***	-0.067*	-0.147***	-0.054	-0.185***	-0.096**
	(0.054)	(0.034)	(0.056)	(0.036)	(0.054)	(0.041)
	Therapeutic, Total		Therapeutic, Day		Therapeutic, Night	
Frac. Edits by Women	-0.105***	-0.031	-0.098**	-0.017	-0.137***	-0.054
	(0.039)	(0.021)	(0.042)	(0.022)	(0.052)	(0.038)
log(Inpatient Days)		X		X		X
Diagnosis Category FE		X		X		X
LASSO controls		X		X		X
Obs	3,687	3,687	3,687	3,687	3,687	3,687

(b) Patient Outcomes

	log(Inpatient Days)		30-day Readmissions		
Frac. Edits by Women	-0.022	0.006	0.004	-0.016	-0.016
	(0.033)	(0.030)	(0.015)	(0.016)	(0.016)
log(Inpatient Days)					X
Diagnosis Category FE		X		X	X
LASSO controls		X		X	X
Obs	3,687	3,687	3,687	3,687	3,687

Notes: This table is similar to Table 3.9, except that it shows results for the above-median ex-ante complexity sample of encounters. Results for the below-median complexity encounters are shown in Appendix Table A10. See text and notes to Table 3.9 for additional details.

Table A12: Minutes Reading Notes vs. Minutes Writing Notes for Men and Women

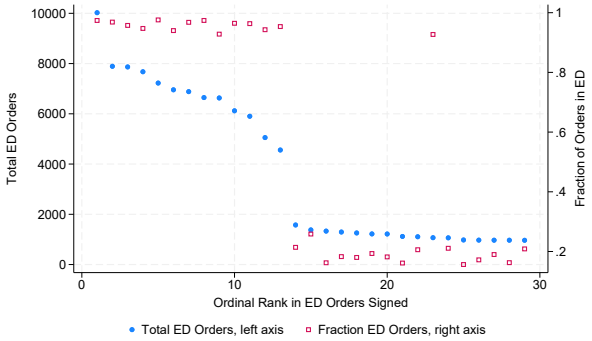
	log(minutes viewing by others within one week)					
log(minutes editing)	0.330***	0.312***	0.311***	0.276***	0.274***	0.276***
	(0.019)	(0.015)	(0.016)	(0.020)	(0.020)	(0.020)
Female	-0.074	-0.014	0.001	-0.058	-0.063	-0.067
	(0.080)	(0.066)	(0.067)	(0.062)	(0.060)	(0.059)
Female X log(mins editing)	0.025	0.014	0.011	0.023	0.023	0.024
	(0.028)	(0.023)	(0.023)	(0.021)	(0.021)	(0.021)
Patient Ex-Ante Complexity			0.175***	0.183***	0.183***	0.182***
			(0.010)	(0.009)	(0.009)	(0.009)
Note Type FE		X	X	X	X	X
Specialty FE				X	X	X
Grad Decade FE					X	X
Med School Rank FE						X
Obs	80,318	80,318	80,318	80,318	80,318	80,318
R-squared	0.134	0.241	0.258	0.275	0.277	0.277
Adj R-squared	0.134	0.241	0.258	0.274	0.276	0.276

Notes: This table shows regression results corresponding to Figure 3.3, which depicts the specification in Column 2. The dependent variable is the natural logarithm of the sum of minutes viewing notes by other clinical staff (physicians, residents, nurse practitioners, physician assistants, and nurses) within one week after the note was finalized as a proxy for clinical utilization of the note. The sample is single-authored notes written by attending physicians belonging to any specialty. See text for additional details.

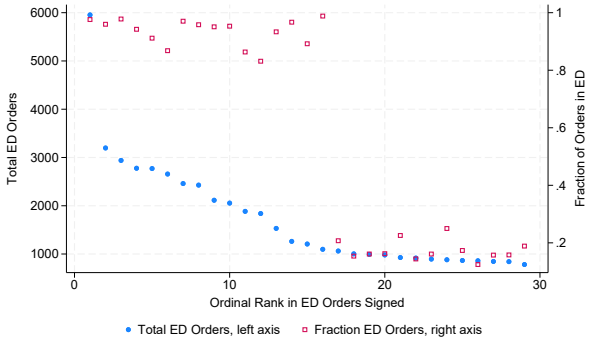
A.4 Additional Figures

Figure A1: Identifying EM Residents Based on Orders Signed

(a) 2017 Cohort



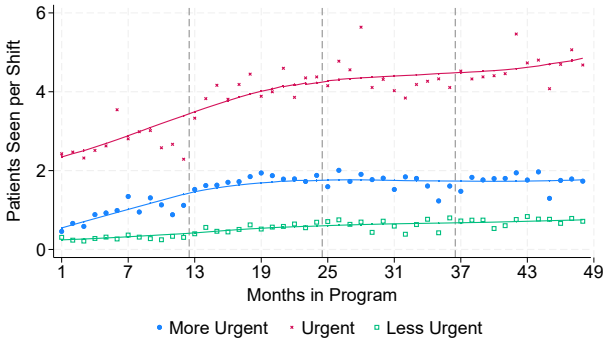
(b) 2018 Cohort



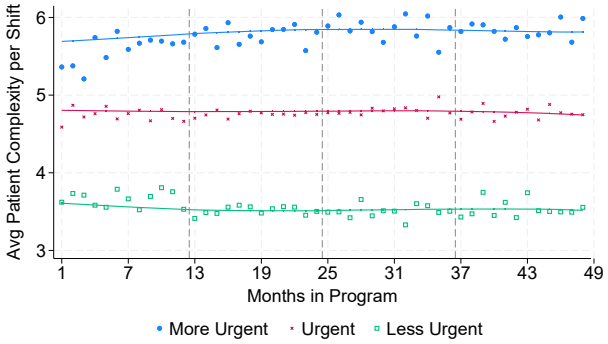
Notes: These figures illustrate the discontinuity in total orders signed in the ED and fraction of orders signed in the ED for residents belonging to two cohorts. Residents are ordered based on the number of ED orders signed, and for each resident both the total number of orders (blue circles) and fraction of orders (hollow red squares) signed in the ED are plotted. Each vertical pair of markers represents one individual resident. Panel (a) shows the relationship for the 2017 cohort, whereas Panel (b) shows the relationship for the 2018 cohort. See text for additional details.

Figure A2: Patients Seen Per Shift

(a) Patients Seen per Shift



(b) Average Ex-Ante Complexity of Patients



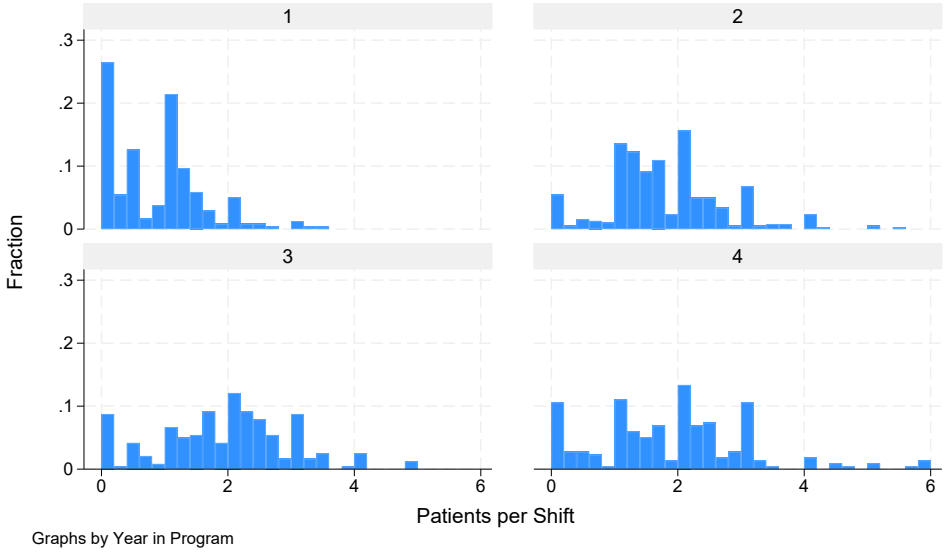
Notes: Panel (a) shows the number of patients seen per shift split by three groups of ex-ante severity, as assigned by the triage nurse upon patient arrival. The blue circles represent the most urgent patients, and comprise about 20% of all arriving patients. We see growth in these complex patients during the first year that levels off after residents enter their second year. The red X's represent the middle category of urgency, which comprise about 60% of all patients, and growth continues throughout the program. The hollow green squares represent the least urgent patients, who comprise the remaining 20% of patients. Growth in these patients is minimal. Panel (b) shows the average predicted ex-ante complexity of patients assigned to residents in the same categories. The complexity measure corresponds to variation in patient severity within ESI category and can be thought of as the "intensive margin" of complexity assignment to residents. "Complexity" is a prediction of patient severity based on ex-ante and immutable patient covariates developed in Chu, et al. (2023). This figure shows that with the exception of residents in the first six months of the program getting slightly simpler patients in the highest complexity category, averages are stable across experience. This means that residents are not assigned less complex patients as they increase the number of patients they see simultaneously with experience.

Figure A3: Average Fraction of Complex Patients Seen, by Role in Each Calendar Quarter

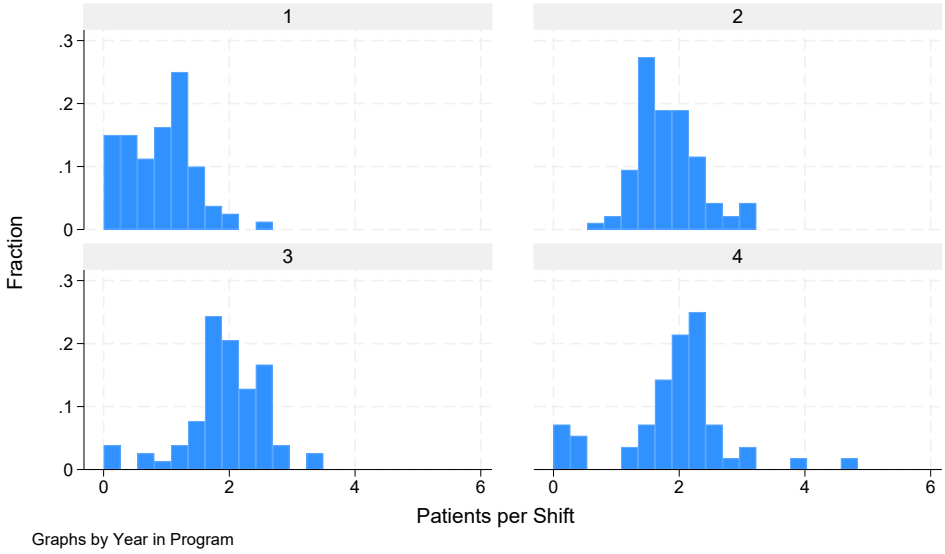
Notes: This figure shows the average fraction of complex patients seen by role for each calendar quarter. The figure shows that the shares are relatively stable across academic years, delineated by the vertical dashed lines, providing evidence supporting the assumption that in the steady-state, the hospital trains the same amount each academic year. See text and notes to Figure 1.4 for more details.

Figure A4: Cross-Sectional Variation in Average Complex Patients Seen per Shift

(a) Two-Week Averages

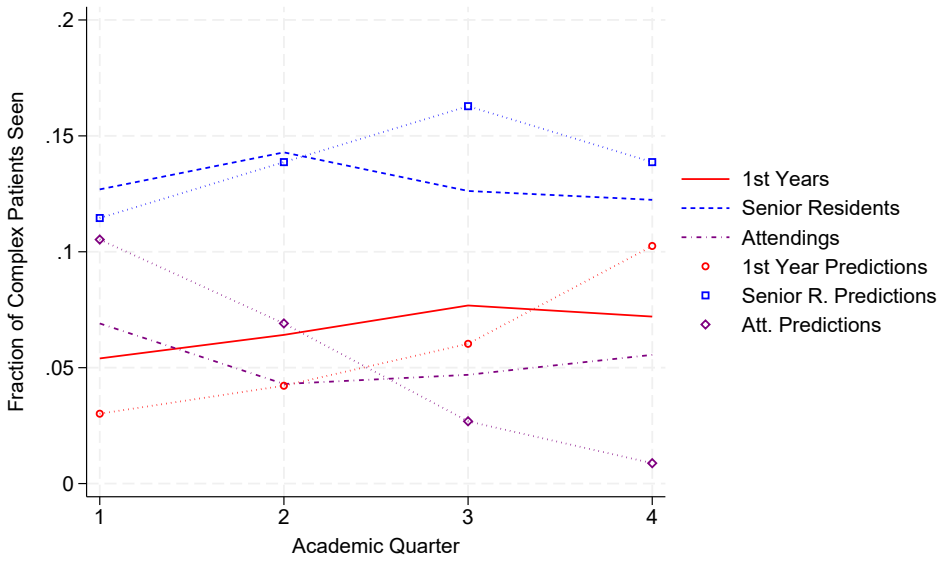


(b) Quarterly Averages



Notes: This figure shows the average number of patients per shift over two different periods of aggregation. Panel (a) shows this over two-week periods, while Panel (b) shows this over calendar quarters. Each observation is a resident-period and variation is shown separately by the resident's year in the program.

Figure A5: Model Fit by Quarter



Notes: This figure plots the actual allocation of complex patients with the model predicted allocation. The scale is different than in Figure 1.4 and Appendix Figure A3 because I have normalized each quarter to have a mass of 0.25 patients. The data are represented by the heavier lines without markers, and the model predictions are the lighter lines with markers on each quarter. The model fits the qualitative patterns well but the gradient on patients allocated to attendings is steeper than in the data. See text for additional details.