Journal of Health Economics 29 (2010) 866–882

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Returns to physician human capital: Evidence from patients randomized to physician teams☆

Joseph J. Doyle Jr.[a,*], Steven M. Ewer[b], Todd H. Wagner[c]

[a] MIT & NBER, United States
[b] University of Wisconsin-Madison, United States
[c] VA Palo Alto and Stanford, United States

ARTICLE INFO

Article history:
Received 3 February 2010
Received in revised form 12 August 2010
Accepted 16 August 2010
Available online 24 August 2010

JEL classification:
I12
J24

Keywords:
Returns to healthcare
Human capital
Medical school quality
Natural experiment

ABSTRACT

Physicians play a major role in determining the cost and quality of healthcare, yet estimates of these effects can be confounded by patient sorting. This paper considers a natural experiment where nearly 30,000 patients were randomly assigned to clinical teams from one of two academic institutions. One institution is among the top medical schools in the U.S., while the other institution is ranked lower in the distribution. Patients treated by the two programs have similar observable characteristics and have access to a single set of facilities and ancillary staff. Those treated by physicians from the higher ranked institution have 10–25% less expensive stays than patients assigned to the lower ranked institution. Health outcomes are not related to the physician team assignment. Cost differences are most pronounced for serious conditions, and they largely stem from diagnostic-testing rates: the lower ranked program tends to order more tests and takes longer to order them.

1. Introduction

A major question in healthcare is the underlying source of geographic variation in spending: high-spending areas in the U.S. incur costs that are 50% higher than low-spending ones (Fisher et al., 2003). These differences are often ascribed to divergent preferences and training among physicians (Phelps and Mooney, 1993; Eisenberg, 2002; Sirovich et al., 2008). Related evidence suggests that high-spending areas are associated with a greater number of specialists and lower quality care (Baicker and Chandra, 2004; Wennberg et al., 2009). There are also equity concerns that health disparities may result from differences in access to high-quality care (Institute of Medicine, 2002; Chandra and Skinner, 2003; Almond et al., in press).

Estimates of the effects of physicians on cost and quality of care can be confounded by omitted-variable concerns and selection issues. For example, high-risk patients may be referred to or self-select the "best" physicians (referral bias), and as a result the highest-quality physicians can have the highest mortality rates (Glance et al., 2008).[1] Indeed, public report cards that rank providers based on risk-adjusted mortality rates have been controversial due to concerns that patients differ in unobservable ways, and that the reports create incentives for providers to avoid high-risk cases (Marshall et al., 2000; Dranove et al., 2003). In addition, the environments where physicians operate differ, including differences in patient characteristics and complementary physical and human capital.

This paper considers a unique natural experiment in a large, urban Department of Veterans Affairs (VA) hospital, where nearly 30,000 patients (and over 70,000 admissions) were randomly assigned to teams comprised of clinicians from one of two academic

☆ The authors would like to thank Ann Bartel, Amitabh Chandra, Christine Durrance, Sherry Glied, Michael Greenstone, Jonathan Gruber, Lawerence Katz, Steven Levitt, Lars Lefgren, Thomas McGuire, David Meltzer, Abigail Ochberg, Ciaran Phibbs, Joseph Price, Roberto Rigobon, Jonathan Skinner, Douglas Staiger, Thomas Stoker, Tavneet Suri, and Jack Wennberg, as well as two anonymous referees and seminar participants at American Society of Health Economists Annual Meetings, Brigham Young, Carnegie Mellon, Harvard, MIT, NBER Health Care Program, San Francisco VA, Stanford, and Syracuse University, for helpful comments and insights. We also greatly benefited from the efficient programming of Andrew Siroka. All remaining errors are our own.

* Corresponding author. Tel.: +1 6174523761.
E-mail address: jjdoyle@mit.edu (J.J. Doyle Jr.).

[1] This non-random assignment of patients also plagues comparisons across hospitals. Geweke et al. (2003) find that patients with the worst unobservable severity go to high quality hospitals.

0167-6296/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2010.08.004

institutions. As described in more detail later, most VA hospitals are affiliated with one or more academic medical school. In this paper, we analyze data from a VA that has two academic affiliates. One set of physicians is affiliated with an academic institution that is among the top medical schools in the U.S.; the other set is linked with an institution that is ranked lower in the quality distribution.[2] Patient characteristics are similar across the two academic institutions due to the randomization. Meanwhile, the teams have access to the same facilities, the same nursing staff, and the same specialists for consultations. By comparing patient outcomes across these two groups, this paper aims to estimate effects of physicians on costs and health outcomes, i.e. returns to physician human capital.[3]

We find that patients assigned to physicians affiliated with the higher ranked program have 10% lower costs compared to the lower ranked program, and up to 25% lower costs for more complicated conditions. The differences largely stem from diagnostic-testing rates: the lower ranked program tends to order more tests and takes longer to order them. Meanwhile, hospital readmissions and mortality are unrelated to the physician-team assignment.

A main caveat is that the results apply directly to one hospital and two residency training programs, albeit with thousands of physicians that comprise them. The "parent hospital" of the higher ranked institution is similar in treatment intensity to other top teaching hospitals, however. This suggests that practice patterns at the top-ranked institution are similar to other highly ranked institutions as well.

The paper is organized as follows: Section 2 describes the empirical framework and defines the main parameters of interest; Section 3 provides background information on the physician teams and patient assignment, as well as a review of the previous literature; Section 4 describes the data; Section 5 reports the results; and Section 6 concludes.

2. Empirical framework

Consider a health production function that relates mortality, M, to health care inputs and a patient-level severity measure, θ:

$$M = F(H, K; \theta) \tag{1}$$

where H represents human capital of the hospital staff, and K represents physical capital.

We focus on the effects of physician human capital, H, on patient outcomes, as well as differences in treatment intensity. In our empirical application, there are two teams that differ markedly in the screening of physicians that compose each team, including different residents and attending physicians. Let P be an indicator that the patient was assigned to physicians in the lower ranked program, T be a measure of treatment, and X represent observable characteristics of the patients. The main parameters of interest can then be written as:

$$E(T|P=1, X) - E(T|P=0, X) \tag{2a}$$

$$E(M|P=1, X) - E(M|P=0, X) \tag{3a}$$

This gives rise to empirical models of the form:

$$T_i = \alpha_0 + \alpha_1 P_i + \alpha_2 X_i + \varepsilon_i \tag{2b}$$

$$M_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \upsilon_i \tag{3b}$$

where ε and υ are error terms.

A common problem when estimating α_1 or β_1 is that patients are not randomly assigned to physicians. Rather, patients choose or are referred to physicians. A patient's primary physician, who knows more about the illness severity than can be captured in typical data sets, may refer the "toughest" cases to the "best" physicians. This tends to bias against finding survival improvements for physicians with higher levels of human capital.[4] Comparisons across hospitals have the additional confounding factors of differences in technology and support staff, which may have a large impact on patient survival independent of the physician characteristics (Unruh, 2003; Evans and Kim, 2006; Bartel et al., 2009).

The main innovation in this paper is the study of a large number of patients who were randomly assigned to physician teams within the same facility. This should satisfy the identification assumptions that the physician team is mean independent of the error terms: $E(P\varepsilon) = E(P\upsilon) = 0$.

In terms of the standard errors, as in other randomized trials the individual error terms are assumed to be independently and identically distributed. The estimates reported are robust to heteroskedasticity and clustered at the patient level to account for dependence across observations for the same patients treated over time (similar results are found when we restrict the analysis to each patient's initial episode, as described below). These errors are conservative compared to alternatives considered.[5]

3. Background

3.1. Previous literature

Much of the previous work on physician human capital finds that previous test scores, such as undergraduate grade point average or Medical College Admissions Test (MCAT) scores, are positively correlated with later test scores (Case and Swanson, 1993; Glaser et al., 1992; Hojat et al., 1997; Silver and Hodgson, 1997). It is less clear whether physicians with higher scores provide higher quality care. Ferguson et al. (2002) review the literature on predictors of medical school success, and note that little has been done on post-medical school performance. There is some evidence on outcome differences by board-certification status, but it is mixed.[6]

[2] In some ways the top-ranked program's physicians are "stars". Rosen (1981) discusses star physicians, where the potential to be a superstar is limited by the extent of the market—in this case the physician's time to see patients. This time constraint inhibits the scalability of the treatment provided by top physicians.

[3] Gross returns are considered here. The residents studied earn similar wages regardless of their academic institution affiliation, and detailed data linking wages to the quality of medical education do not exist.

[4] In the case of heterogeneous treatment effects, the patients are likely referred based on the expected gain of the assignment: a correlated random coefficient model that can inflate returns to physician human capital (Bjorklund and Moffitt, 1987).

[5] One caveat is that the observations may be correlated within teams that vary over time, although we do not observe team composition. We found that clustering at the month-year level—times when team composition is likely to change—resulted in similar, and often smaller, standard errors. Similarly, when the estimates were jointly estimated using a seemingly unrelated regression, estimated standard errors were again similar and often smaller. Last, we considered correlation within each of the two groups. The outcomes considered here, however, have an intra-class correlation of close to zero (e.g. our cost measures have an intra-class correlation of less than 0.005). As in other randomized trials, these intra-class correlation coefficients imply that correcting the standard errors by clustering at the group level is unnecessary in this context (Moulton, 1986; Angrist and Pischke, 2008).

[6] Certification has been found to be associated with reductions in mortality following heart attacks (Kelly and Hellinger, 1987; Norcini et al., 2000), while other work has found differences in the use of appropriate medications but little difference in mortality (Chen et al., 2006). Licensure examination scores have been found to be related to preventive care and more appropriate prescription medicines (Tamblyn et al., 1998, 2002).

A measure of physician quality directly related to the current study comes from surveys of other physicians in the same market. Hartz et al. (1999) show that surgeons are more likely to be regarded as a "best doctor" in these community surveys if they trained at a prestigious residency or fellowship program. They note that treatment by physicians trained at prestigious programs is not related to mortality, however.

Small-area variation in treatment has received considerable attention, with some evidence that physician quality measures vary across patient groups and may contribute to health disparities (see extensive reviews by Van Ryn (2002) and Bach et al. (2004)). In particular, access to high-quality specialists varies across racial groups, and desegregation has been found to significantly improve health outcomes for African American patients (Mukamel et al., 2000; Chandra and Skinner, 2003; Almond et al., in press). Another reason for the large literature on small-area variation in treatment is that physicians are important cost drivers across areas. Physician characteristics have been found to explain up to 50% of the variation in expenditures, on par with case-mix variables (Pauly, 1978; Burns and Wholey, 1991; Burns et al., 1994; Meltzer et al., 2002; Grytten and Sorensen, 2003).[7]

There is a related literature that estimates the impact of report cards—publicly provided information about physician mortality rates, adjusted for case mix (for reviews, see Marshall et al. (2000); Hofer et al. (1999); and discussions between Hannan and Chassin (2005) and Werner and Asch (2005a,b)). Newhouse (1996) and Cutler et al. (2004) note that such report cards suffer from patient selection problems in ways that can confound estimates of the returns to physician human capital in general. For example, Dranove et al. (2003) found limited access to surgery for high-risk patients following the introduction of report cards: fewer surgeries, more conducted at teaching hospitals, and large increases in adverse health outcomes in the short run.[8]

The empirical strategy in the literature to deal with these selection issues is a selection on observables approach—controlling for illness severity with indicators of comorbidities and patient characteristics such as age. Nevertheless, unobserved (to the researcher) differences in severity may contaminate comparisons. One randomized trial of 1151 patients assigned to resident and staff physicians showed that the staff service had shorter length of stay and costs (Simmer et al., 1991). Previous research that is most similar to ours is Gillespie et al. (1989) that studied 119 patients randomized to two medical school programs in 1984 and 1985. They found little difference in diagnostic testing between the two programs. The analysis excluded patients who received no diagnostic testing, however, which may lead to sample selection bias. The current study will consider nearly 30,000 patients over 13 years. This includes over 72,000 patient encounters to provide a more comprehensive comparison, greater statistical power to detect differences, and a time frame that allows a comparison of long-term outcomes such as 5-year mortality.

3.2. *Training at the VA*

Physician training programs offer a way to accumulate human capital largely through learning by doing, and such training can have an effect on patient outcomes (Huckman and Barro, 2005).[9] One of the most common training grounds for physicians is the VA medical system.

The VA operates the largest integrated health care system in the US, with 155 medical centers and over 850 community-based outpatient clinics. Veterans can receive a range of services from general medical care to specialized services. In 2007, VA treated over 5 million unique patients, and some health care reform experts use the VA as a model (Ibrahim, 2007). The VA is organized around 21 regions, known as VISNs (Veterans Integrated Service Networks). Operating funds are distributed from Washington DC to each VISN, which then distributes the money to its hospitals, community clinics and outreach facilities. The financing system is based on a capitated risk-adjustment model.

Graduate medical education is part of the VA's statutory mission, and VA medical centers are located near academic medical centers to enhance training. One hundred and seven of the 126 medical schools in the U.S. are affiliated with a VA medical center. The primary physicians for patients at VA hospitals are thus residents, particularly from internal medicine and general surgery training programs. Residents rotate through the VA system and treat many low income and disabled veterans—patients who provide valuable variation across a wide range of diseases. Each year, 31,000 residents (30% of all residents in the U.S.) and 17,000 medical students train in VA facilities (Chang, 2005; VHA, 2005).

This study considers a VA hospital in a large urban area that has affiliations with two medical schools.[10] This VA hospital is a full-service teaching hospital that provides over 3500 surgical procedures each year. It has an intensive care unit and what are considered excellent laboratory facilities, including the ability to conduct magnetic resonance imaging and angiography. In addition to the main hospital, there are some smaller satellite hospitals elsewhere in the city that handle mental health, substance use treatment and long-term care.

3.3. *The residency programs*

The medical and surgical residency training programs compared vary substantially in terms of their ranking: one is regarded as a top program in the U.S., whereas the other is ranked lower in the quality distribution. In the remainder of the paper, the higher ranked institution will be referred to as Program A, and the lower ranked institution will be referred to as Program B.

To establish the difference in credentials, Table 1 reports some summary characteristics of the two programs. First, the residency programs are affiliated with two different medical schools where the attending physicians that supervise and train the residents are faculty members. These medical schools differ in their rankings. Some years, the school affiliated with Program A is the top school in the nation when ranked by the incoming students' MCAT scores, and it is always near the top. In comparison, the lower ranked program that serves this VA hospital is near the median of medical schools. Similar differences are found in the rankings of medical schools with respect to their National Institutes of Health funding levels.

Second, each training program is affiliated with another teaching hospital in the same city, in addition to the VA hospital. Program A's "parent hospital" is ranked among the top 10 hospitals in the country according the U.S. News and World Report Honor Roll rankings of hospitals. Out of 15 specialties ranked by U.S. News, Program A's hospital is among the top 10 hospitals in the country for nearly half of them, and among the top 20 in nearly all of them (U.S. News

[7] Not all studies find significant effects of physicians on costs, however. Hayward et al. (1994) find that residents and attending physicians in one hospital do not explain much of the variation in length of stay (on the order of 1–2%).

[8] See also Schneider and Epstein (1996) and Omoigui et al. (1996).

[9] See Marder and Hough (1983) for an early discussion on supply and demand for such opportunities.

[10] We have chosen to keep the name of the VA hospital confidential out of respect for the patients and medical schools.

Table 1
Residency program comparisons.

		Program A	Program B
Affiliated medical school rankings (out of 126 schools):	Medical College Admissions Test (MCAT) Ranking	Top 5	Top 50
	NIH Funding Ranking	Top 5	Top 80
Affiliated hospital	US News Honor Roll (overall)	Top 10	Not listed
Resident characteristics	% with MD from Top 10 Medical School (US News rankings)	30%	3%
	% with MD from Top 25 Medical School (US News rankings)	50%	9%
	% with MD from Top 10 Medical School (NIH Funding rankings)	25%	2%
	% with MD from Top 25 Medical School (NIH Funding rankings)	40%	8%
	% Foreign Medical School	10%	20%
Board certification:	American Board of Internal Medicine	99% (95th percentile)	85% (20th percentile)
Residency program pass rate	American Board of Surgery	85% (75th percentile)	60% (20th percentile)

Figures are approximate out representative of rankings over the past 20 years. *Sources*: US News and World Report rankings, various years; American Board of Internal Medicine; American Board of Surgery; AMA Masterfile, 1993–2005.

and World Report, 2007). Meanwhile, Program B's parent hospital is not a member of this Honor Roll overall or ranked among the top hospitals in terms of subspecialties. The treatment intensity across the two parent hospitals is similar to one another, however, as described below.

Third, the residents themselves can be compared using data from the AMA Masterfile. Approximately 30% of residents who were trained in Program A received their MD from a medical school in the top 10 of the U.S. News and World Report rankings in 2004, compared to 3% of those trained in Program B. Approximately half of Program A's residents graduated from a top-25 medical school compared to less than 10% for Program B. Similar differences are seen when the residents' medical schools are ranked by NIH funding levels. In addition, twice as many of Program B's physicians earned their medical degree from a medical school outside of the U.S.[11]

Perhaps the most striking evidence comes from Board scores. At the end of the residency program students will often take board-certification exams, and the major Boards publish the pass rate for each residency program among those who were taking the exam for the first time. The two most relevant exams are given by the American Board of Internal Medicine and the American Board of Surgery. Table 1 shows that the pass rate for Internal Medicine is close to 100% for the residents in Program A compared to a pass rate of approximately 85% for Program B (a rate that is in the bottom quintile of the 391 programs listed).[12] The pass rate for General Surgery is lower, 85% for Program A and 60% for Program B. These scores place Program A in the top quartile, and Program B in the bottom quintile, of residency programs in the U.S.[13]

In sum, the physicians in Program A perform substantially better on exams than physicians in Program B. These differences are stable over time, as a survey in the early 1970s asking medical school faculty to rank programs included Program A in its top 10, whereas Program B was ranked near the median of the rankings (Cole and Lipton, 1977).

3.4. The clinical teams

Discussions with physicians familiar with the programs revealed similarities and differences across the teams at this VA Medical Center. The clinical and teaching teams conduct independent rounds each day during which they discuss their patients. The timing of these rounds does not differ systematically between the two institutions. This parallel structure allows a comparison of the two groups' treatment decisions and patient outcomes.[14] The patients assigned to each team are interspersed throughout the floors and share a common pool of nursing and ancillary staff. In particular, the two teams have access to the same specialists for consultations. There is a single set of clinical laboratories and imaging facilities for use by both teams, and our investigation of the programs suggests that neither institution receives favorable treatment from these ancillary service providers. We have also found that the overall philosophies of care do not differ substantially across the two programs, and the amount of resident oversight at the VA is thought to be similar across the two programs.[15] This is described in more detail below.

Members of the clinical team include attending physicians, interns, senior residents and medical students, all of whom are affiliated with the parent teaching hospital. The intern, also known as a first-year resident, is the primary physician assigned to the patient, and this role includes evaluating patients, prescribing medicines, ordering diagnostic studies, performing bedside procedures, interacting with nursing staff and consultants, and writing the notes that make up the bulk of the medical record. The senior resident directly supervises the work of the intern, leads the team on daily rounds during which clinical care and teaching are afforded, and serves as a backup for the intern. The attending physician serves as the official provider of medical care and oversees the work of all other members of the team. This person typically does not attend the daily rounds of the team, but rather sees patients separately and discusses cases with the senior resident, confirming the clinical decision making of the team. The medical students, not yet physicians, are not allowed to write orders or officially contribute to the medical record. They work alongside residents to evaluate patients, and any contribution to decision making must go through the residents. This distribution of work is representative for teams in both Program A and Program B.

The size of the two physician teams is similar, consistent with the equal assignment of patients to the two teams. At a given time, Program A has four medicine teams, each consisting of one attending physician, one senior resident and one intern. Program B likewise has four medicine teams composed of one attending and one senior resident, but one difference across the two teams is that

[11] These data also include primary specialty, and two of the most common are internal medicine and pediatrics. Physicians who trained in Program B listed these 20% of the time, compared to 13% for Program A.

[12] American Board of Internal Medicine. Figures for 2005–2007. http://www.abim.org/pdf/pass-rates/residency-program-pass-rates.pdf.

[13] American Board of Surgery, 5-year pass rate from 2002–2007. http://home.absurgery.org/default.jsp?prog_passreport.

[14] Other VA Medical Centers that are served by multiple residency training programs generally allow the teams to mix, with rounds attended by all of the residents.

[15] Historically, VA hospitals were thought to provide less attending supervision than other teaching hospitals. In the 1990s, this was addressed and has continued to increase. For example, in 2004 the VA required an attending to be present for all major elective surgeries (Chang, 2005).

Program B teams include two interns. In practice, the implication of this difference in team size is that Program B has an advantage in total residents (12 vs. 8). The senior-resident to patient ratios are the same across the groups, however, and we consider the effects of different intern-to-patient ratios below. Last, in terms of senior resident experience, Program B again has an advantage with second and third-year residents at this facility compared to exclusively second-years from Program A.

3.5. *Patient assignment*

To ensure an equitable distribution of cases and overall workload, the patients are randomly assigned to each institution: patients with social security numbers ending in an odd number are assigned to Program A whereas those with even social security numbers are assigned to Program B. This randomization method ensures that there is no crossover-if a patient is readmitted, the patient is assigned to the same physician group. Discussions with physicians at the VA hospital suggest that this randomization process was established when the facility was constructed in the 1950s.

As part of our investigation of this natural experiment, we found three exceptions to the randomization. First, the randomization only occurs at the main teaching facility, not at satellite facilities. Second, not all subspecialties use the randomization. For example, neurology patients are not randomized; rather all of the patients are assigned to one team. Third, the medical intensive care unit is headed by a single attending physician that oversees patients assigned to both teams. We will employ these groups of patients in specification checks below.

4. Data description

We used the VA Patient Treatment Files (PTF) to identify inpatient encounters from 1993–2006. We restrict the main analysis to patients admitted to the main hospital facility, and patients who did not have a major diagnostic category of "nervous system"—these cases are less likely to enter the randomization. This results in an analysis data set of over 72,000 inpatient stays and nearly 30,000 patients. The main results include the information in all of the episodes and the standard errors are clustered by patient to take into account dependence within these observations as described above. Results will be shown for a sample restricted to patients' first episodes in the database as well.

The PTF includes the patient's age at admission, race, sex, marital status, and ZIP code of residence.[16] Data from the 2000 Census of Population were matched to the data to characterize the patient ZIP code, including the median household income, population density, and education, race, and age composition. Time and date of admission are also available, and the models include day-of-week, month, and year indicators, as well as indicators for 6-h time-of-day blocks.

The PTF data also include ICD-9 diagnosis and procedure codes. This allows us to compare treatment across primary diagnoses, and nine secondary diagnoses will be used to characterize the comorbidities of the patient. It is possible that Programs A and B code diagnoses differently. This is testable in our data, as the sample sizes within diagnoses can be compared across the two programs. These diagnosis codes are recorded for the benefit of patient histories and ongoing care rather than for billing purposes and, therefore, should not be affected by financial incentives to code patients into more profitable diagnoses (Dafny, 2005). Records can be coded by physicians or support staff, which would handle coding for both Programs A and B.

The VA PTF uses a scrambled social security number as the patient identifier. We linked this identifier to the last digit of the patient's true social security number to compare patients assigned to the different teams. The PTF does not have physician or resident identifiers to verify that all even numbered patients were indeed assigned to Program B, for example. We do not expect patients with even-numbered social security numbers to be assigned to Program A apart from the exceptions listed in the background section.

There are four main measures of treatment provided. The patient's length of stay in the hospital is observed for all years in our dataset. Longer stays represent greater time for supervision and additional care. Length of stay can also measure the ability of physicians to process administrative tasks such as scheduling a series of treatments. The VA strove to decrease length of stays in the mid-1990s by decentralizing power to geographic regions, changing ambulatory care benefits and creating incentives that reward medical center directors for shorter lengths of stay (Ashton et al., 2003). These policy changes would have been uniformly applicable to both Programs A and B, although we can test for differences in the response to these initiatives.

In addition to length of stay, two cost measures are available as well. Accounting cost data (using step-down accounting methods) comes from the Decision Support System (DSS) and the Health Economics Resource Center databases. These data are reliable from 2000 to 2006. A related summary measure is the Health Economics Resource Center Average Cost Data. These data are considered available from 1998 onwards, and uses non-VA (largely Medicare) relative value weights to estimate expenditures for VA care (Phibbs et al., 2003; Wagner et al., 2003). One limitation of these estimated expenditures is that they are geared toward assigning average costs for patients with similar diagnoses and procedures, and are, therefore, less precise than DSS and can miss outlier costs (Wagner et al., 2003). Costs were standardized to 2006 dollars using the general urban consumer price index from the Bureau of Labor Statistics.

The fourth treatment measure is the number and timing of procedures, based on ICD-9 procedure codes and dates. Physicians' use of diagnostic tests in particular can shed light on practice differences between Programs A and B.

There are two health outcomes that we consider. First, readmissions to the VA hospital within 30 days or 1 year of the date of admission are identified. A limitation of these readmissions is that they do not include admissions to non-VA hospitals. If lower quality care drove patients from the VA system into a non-VA facility, then lower readmission rates could signal lower quality care. Still, many veterans depend on the free care provided by the VA, and we will generally regard readmissions as a negative outcome for patients. Another limitation is that any differences in initial length of stay will change the time at risk for a 30-day readmission, for example. When the measure was 30 days from discharge (as opposed to days from admission), nearly identical results were found, however. Two related readmission measures are the costs of these readmissions, and readmissions with the same major diagnosis as the initial episode.

The second outcome is more straightforward: mortality. The main results will focus on 30-day, 1-year, and 5-year mortality, and these measures were calculated for patients whose measures are not right censored. For example, 5-year morality was calculated for patients admitted to the VA hospital at least 5 years earlier than the end of the sample period. These measures are taken from the VA vital status files and cover deaths occurring outside of the hospital

[16] Of these variables, the definition of race changed over time, as did its collection method (from admission-clerk assignment to self-report). This suggests that some caution is warranted with regard to this control. The non-white patients are strongly correlated with the fraction African American in the patient's ZIP code.

Table 2
Summary statistics.

		Assigned to Program A (odd SSN)	Assigned to Program B (even SSN)	*p*-Value
Demographics	Age	63.0	62.8	0.35
	18–34	0.019	0.022	0.15
	35–44	0.074	0.075	0.80
	45–54	0.186	0.186	0.94
	55–64	0.229	0.229	0.92
	65–69	0.134	0.131	0.50
	70–74	0.149	0.146	0.57
	75–84	0.179	0.184	0.39
	84+	0.030	0.027	0.24
	Male	0.976	0.978	0.19
	White	0.466	0.472	0.42
	Married	0.443	0.446	0.65
	Divorced	0.271	0.269	0.80
Comorbidities	Charlson index = 0	0.294	0.290	0.52
	Charlson index = 1	0.274	0.278	0.37
	Charlson index = 2	0.433	0.432	0.91
Admission time	Midnight–6am	0.096	0.098	0.56
	6 am–12 noon	0.237	0.233	0.29
	12 noon–6 pm	0.420	0.425	0.28
	6 pm–midnight	0.247	0.245	0.59
Day of the week	Weekend	0.163	0.162	0.72
ZIP code characteristics	Median HH income	33,714	33,945	0.24
	Fraction HS dropout	0.249	0.247	0.18
	Fraction HS only	0.317	0.318	0.34
	Fraction some college	0.271	0.272	0.024*
	Fraction White	0.628	0.633	0.48
	Fraction Black	0.331	0.327	0.52
	Fraction aged 19–34	0.214	0.213	0.21
	Fraction aged 35–64	0.368	0.369	0.38
	Fraction aged 65+	0.141	0.141	0.22
	Population per 1000 m^2	1.102	1.072	0.09
	Observations (discharges)	35,932	36,434	

p-Values calculated using standard errors clustered by patient.
* Significant at 5%.

as well as in-hospital mortality. These data have been shown to be highly accurate in terms of sensitivity and specificity (Arnold et al., 2006). Other measures of mortality, such as 10-hour mortality, will be compared as well.

To describe the data available and compare patients assigned to the two groups, Table 2 reports summary statistics. The two columns of means are for patients with odd or even social security numbers: patients assigned to Program A and Program B, respectively. We do not believe that patients are aware of the dichotomy of physician teams and the difference in the quality of the residency programs, but to the extent that patients know they will be assigned to one of the two programs, sample selection could be an issue. If selection were a factor, then the observable characteristics may differ across the two groups as well as the frequency of observations.

Table 2 shows that the means are nearly identical across the two teams. For example, the average ages are 63.0 and 62.8. The most common age is between 55 and 64, with smaller fractions of patients over the age of 65 when Medicare provides access to non-VA hospitals.[17] Still, there are many older patients in the sample, and the fraction of patients that no longer visit the VA hospital after the age of 65 does not vary systematically across the two physician teams.

Nearly all of the patients are male, an artifact of the older, veteran population. Forty-seven percent are White, 44% are married, and 43% have a Charlson severity score of 2—an aggregation of the secondary diagnoses that is strongly associated with mortality (Quan et al., 2005). Most patients are admitted to the hospital between 12 noon and 6 pm (42%), the average patient's ZIP code has a median household income of $34,000 and 63% of its population is White. The number of observations is similar across the two groups, with Program B treating 50.3% of the patients (35,932 vs. 36,434).[18] It appears that the patients who enter the VA hospital are randomly assigned to the two programs and that differential selection into the VA is unlikely to drive differences in treatment or health outcomes.

5. Results

5.1. Treatment differences

A first look at how the two programs' treatment levels differ can be seen in Fig. 1A–C. In each figure, the vertical axis reports one of

[17] Demand for VA care appears inelastic with regard costs of visiting a VA hospital. Mooney et al. (2000) find that patients over the age of 65 are more inelastic with respect to distance to the VA hospital compared to those under the age of 65, despite access to Medicare for the older group.

[18] With the large sample size, this difference is marginally significantly different from 0.5 (*p*-value = 0.06). When first episodes are considered, the fraction assigned to Program B is 0.5002 (*p*-value = 0.92).

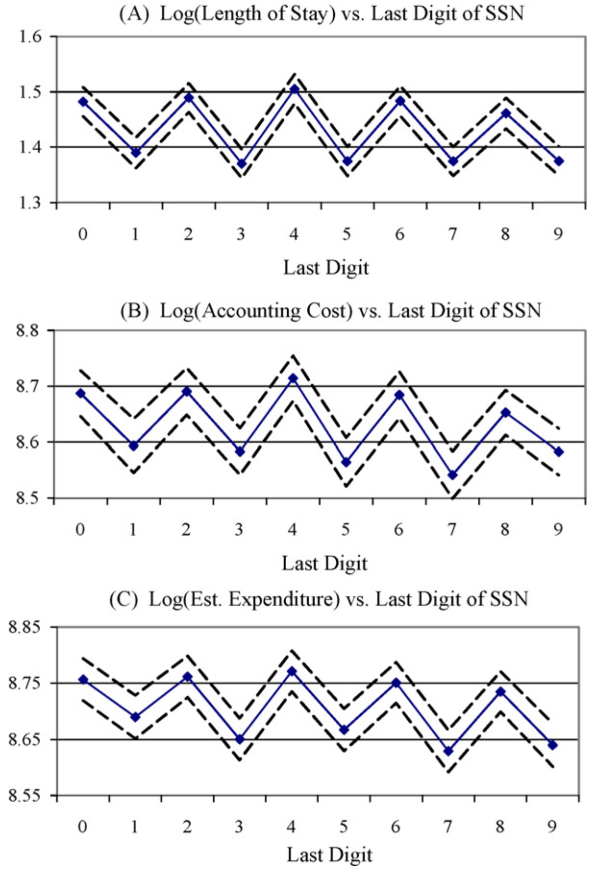

Fig. 1. (A) Log(length of stay) vs. last digit of SSN. (B) Log(accounting cost) vs. last digit of SSN. (C) Log(est. expenditure) vs. last digit of SSN.

the three summary measures of treatment: length of stay, accounting cost, and estimated expenditures. These data are right skewed and each measure was transformed using the natural logarithm. The means of the three measures are 1.43 log days (or 4.17 days), 8.63 log costs (or $5600 in 2006 dollars), and 8.71 log estimated expenditures (or $6100). The horizontal axis in each figure is the last digit of the patient's social security number. We would expect similar measures for each odd (or even) digit if differences in the physician team assignment were responsible for any differences as opposed to sampling variation.

Fig. 1A–C show a pronounced sawtooth pattern, with length of stay and the two cost measures 10 log points higher for patients with an even-numbered social security number compared to patients with an odd-numbered social security number; patients treated by Program B have higher costs. This difference is seen for each digit, as the means are similar for all even (or odd) last digits.

To aggregate the data up to the program level and introduce controls in the spirit of estimating equation (2b), Table 3 reports results from ordinary least squares regressions for the three cost measures. Similar results were found when the length of stay was estimated as a count variable using a negative binomial model. Each column represents a separate regression. The first model reported includes no controls and the 10–11 log point differences shown in Fig. 1 have a standard error of close to 1 log point.[19] Results were similar, although slightly smaller, when the estimates were re-transformed and heteroskedasticity was taken into account (Manning, 1998).[20]

The second model includes 3-digit primary-diagnosis fixed effects to estimate differences in treatment within disease classes and offer a first look at potential differences in diagnoses across the two groups. The models reported in Table 3 show that the results are largely unchanged when these effects are incorporated, although the estimates are slightly larger for accounting costs (12 log points).

The last column for each dependent variable includes the controls in Table 2, as well as year, month, and day-of-week indicators. The results are nearly identical to the model without the additional controls. This is consistent with the randomization effectively balancing the observable characteristics across the two groups, as shown in Table 2.

Part of the cost difference is due to the longer length of stay, but we find substantial differences in costs even when controlling for length of stay. In models with full controls, the main coefficient of interest is 0.068 (S.E. = 0.008) for accounting costs and 0.058 (S.E. = 0.007) for estimated expenditures.

To place a 10 log-point difference in these treatment measures in context, Appendix Table A1 provides estimates for selected covariates. Such a difference is akin to an increase in age category from 45–54 to 65–69. Treatment levels for patients with a Charlson severity score of 2 are 11–13 log points higher compared to patients with a score of 1—a difference in severity that leads to substantial health outcome differences as described below. Admissions during business hours also accrue higher costs. Meanwhile, there is little relationship with day of admission, and married patients have 7–9% lower treatment levels compared to single patients.

Much of the remainder of the paper considers how the different programs differ in terms of procedures and across different types of patients to explore the mechanisms that drive the difference in the summary treatment measures. Before the sources of the treatment differences are explored, the next section reports tests of differences in health outcomes.

5.2. Health outcomes

Given the results in Fig. 1, it is possible that Program A discharges patients prematurely, and they may have worse long-term health outcomes. It is also possible that Program A provides higher quality care in less time and at lower expense. Fig. 2 reports estimates of mean outcomes by the last digit of the social security number, and no differences are found across the patients in terms of 30-day readmissions, as well as 1-year and 5-year mortality.

Again to introduce controls and place the results in context, Tables 4A and 4B reports the results of OLS regressions of the readmission and mortality indicators on the program assignment and controls (equation 3b). Results are similar when probit and logit models were used instead, partly because the dependent variables are sufficiently far from zero: 13% and 43% readmission rates at the 30-day and 1-year intervals, respectively, as well as 30-day, 1-year, and 5-year mortality rates of 6.4%, 24% and 51%.

Tables 4A and 4B shows that the program assignment is unrelated to readmissions and mortality, with coefficients that are not statistically or economically significant. For example, Program B is associated with a 0.6% increase in 1-year readmissions, or 1.4%

[19] The different samples for the cost measures are due to the different time periods when they are available.

[20] For models with full controls, when interpreting the estimates in terms of percentages rather than log points, a smearing factor (the ratio of the average exponentiated residuals in the regressions for each group) is applied and the estimated difference in length of stay is 10%; the difference in accounting cost is 9% and the difference in estimated expenditure is 8%.

Table 3
Treatment differences.

Dependent variable:	Log(length of stay)			Log(accounting cost)			Log(estimated expenditure)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assigned to Program B	0.108 [0.0086]**	0.114 [0.0075]**	0.113 [0.0072]**	0.113 [0.0136]**	0.123 [0.0116]**	0.125 [0.0114]**	0.100 [0.0120]**	0.102 [0.0104]**	0.104 [0.0099]**
Diagnosis fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Full controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	72,366			34,098			42,518		
Mean of dep. var.	1.43			8.63			8.71		
Exp(mean of dep. var.)	4.17			5600			6100		

Models estimated using OLS. Robust standard errors in brackets, clustered by patient. Full controls include variables listed in Table 1, as well as month, year, and day-of-the-week indicators. Cost measures are in 2006 dollars.

** Significant at 1%.

of the mean. When 1-year readmissions with the same major diagnostic code as the previous major diagnosis are compared, Program B is associated with a 0.3% increase or 1.5% of the mean.

In terms of mortality, Program B is associated with a 0.1 percentage-point reduction in 30-day mortality (or 1.1% of the mean), a 0.7 percentage-point reduction in 1-year mortality (or 2.9% of the mean), and a 0.3 percentage-point reduction in 5-year mortality (or 0.6% of the mean). The results are fairly precise as well. For 1-year mortality the 95% confidence interval is [−0.0155, 0.0016], and 5-year mortality the confidence interval is [−0.0162, 0.0106]. This difference is small compared to a 5-year mortality rate of over 50%, and the precision of the estimate largely rules out survival benefits from assignment to the highly ranked program.[21] If anything, across all of the results the lower ranked program appears to achieve modestly better outcomes.

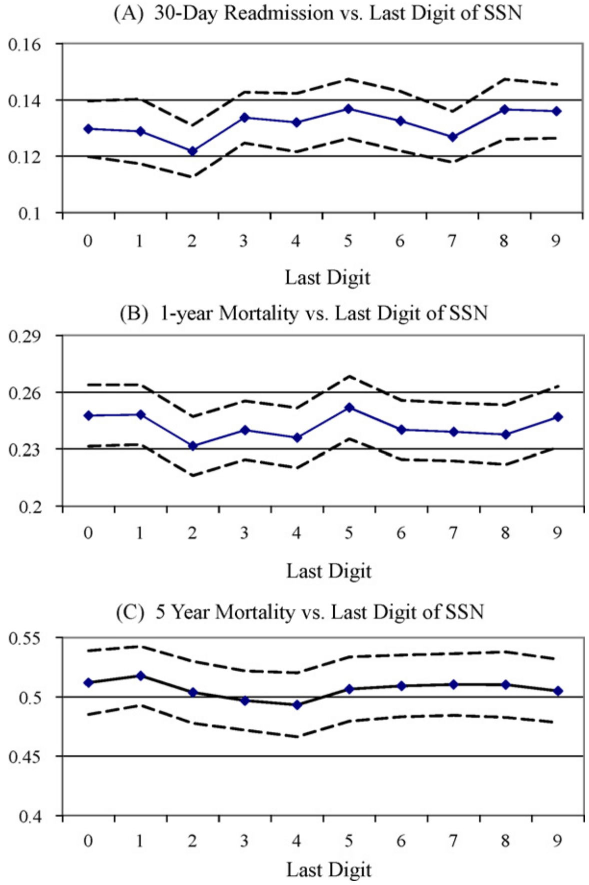

Fig. 2. (A) 30-day readmission vs. last digit of SSN. (B) 1-year mortality vs. last digit of SSN. (C) 5-year mortality vs. last digit of SSN.

To place these small differences in mortality in context, other covariates are associated with higher mortality, as shown in Appendix Table A1. Men have 18% higher mortality rates, a Charlson severity score of 2 is associated with a 50% higher mortality compared to a score of 1, and mortality is strongly associated with the age of the patient. Another way to consider the difference in mortality is the cost of saving a statistical life year. While we prefer not to associate the only difference between the two groups as the difference in average costs, Program B is associated with costs that are on the order of $1000 higher and a 1-year mortality rate that is 0.7 ppt lower. This would imply a $140,000 cost per life year saved.[22] This cost rises with more severe conditions, which are explored in the next section.

5.3. *Mechanisms*

5.3.1. *Across diagnoses*

To compare the robustness of the results across diagnoses and investigate whether the differences arise in more complex cases, Table 5 reports results from models estimated separately across common diagnoses. First, the top 10 most frequent diagnoses are compared.[23] Two rows are presented for each diagnosis: estimates from a model for log length of stay—the resource measure that is available for the full time period, and 1-year mortality. Similar results were found for the other measures as well. The means of the dependent variables are listed, and they vary widely across the diagnoses.

The results show that for some serious conditions with high 1-year mortality rates, such as heart failure, chronic obstructive pulmonary disease (COPD), and pneumonia, treatment differences are between 20 and 25 log points. Smaller differences in treatment are found for less serious conditions such as chronic ischemic heart

[21] Across the six measures, the lower limit on the 95% confidence intervals are less than 7% of their respective means, and the upper limits are less than 5% of their means.

[22] Average costs are approximately $10,000 and Program B is associated with approximately 10% higher costs. $140,000 = $10,000 × 0.1/0.007.

[23] The top 10 diagnoses were determined by calculating the frequency of patients in 3-digit ICD-9 diagnosis codes, as well as more general definitions of gastrointestinal bleeding (Volpp et al., 2007) and chronic obstructive pulmonary disease.

Table 4A
Differences in VA hospital readmissions.

Dependent variable:	30-Day readmission			1-Year readmission			1-Year readmission same major diagnosis		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assigned to lower ranking program	−0.0019 [0.0032]	−0.0019 [0.0031]	−0.0021 [0.0030]	0.0057 [0.0058]	0.0057 [0.0053]	0.0055 [0.0051]	0.0032 [0.0045]	0.0032 [0.0039]	0.0033 [0.0039]
Diagnosis fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Full controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	71,954			66,938			66,998		
Mean of dep. var.	0.132			0.429			0.204		

Table 4B
Differences in mortality.

Dependent variable:	30-Day mortality			1-Year mortality			5-Year mortality		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assigned to lower ranking program	−0.0006 [0.0020]	−0.0006 [0.0019]	−0.0007 [0.0019]	−0.0067 [0.0051]	−0.0061 [0.0045]	−0.0072 [0.0044]	−0.0016 [0.0085]	0.0001 [0.0072]	−0.0028 [0.0068]
Diagnosis fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Full controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	71,954			66,938			47,337		
Mean of dep. var.	0.0642			0.242			0.507		

Models estimated using OLS on a sample that includes patients seen 30 days, 1 year, or 4 years from the end of the sample period. Robust standard errors in brackets, clustered by patient.

disease, with a difference closer to 10%. Acute myocardial infarction (AMI) has a 25% 1-year mortality rate, and a difference in log length of stay of 9 points.

To summarize all of the diagnoses, the 3-digit primary diagnosis codes were divided into quartiles based on their mortality rates.[24] No difference in treatment is found for the lowest quartile. This is a group with a 4% mortality rate and the treatment may be more standardized for less serious conditions. Eleven and 12 log-point differences in length of stay are found for the 2nd and 3rd quartiles, and the most seriously ill patients have a 14 log-point difference in length of stay when the two programs are compared. These cases are likely more complicated, as they have higher costs in addition to the higher mortality rates.

In terms of outcomes, the estimates are less precisely estimated within particular diagnoses given the smaller sample sizes, but the point estimates are unstable in sign and generally small in magnitude. The largest differences are found for AMI and cardiac dysrhythmias, with Program B associated with mortality rates that are 12–18% lower than the sample mean. These differences are not statistically significant, however, and no difference in 30-day readmissions is found for these diagnoses. In addition, no difference in 5-year mortality is found for AMI patients.[25] Program A is associated with lower mortality for pneumonia patients (5% lower compared to the sample mean); again the difference is not statistically significant. Overall, even at the extremes of our confidence intervals, a hypothesis that Program A is associated with lower mortality is not supported by these data.

Table 5 also reports the fraction of patients treated by Program B for each diagnosis, along with a *p*-value from a test that the fraction of patients seen within a diagnosis equals 0.5. This tests whether the programs differ when recording the primary diagnosis. Some of the principal diagnoses show differences that are statistically significantly different from 0.5, with Program A more likely to categorize patients as having chronic obstructive pulmonary disease, and Program B more likely to categorize patients as having respiratory and chest symptoms, as well as diabetes. The rates are close to 0.5 across diagnoses once we aggregate the conditions into the four mortality quartiles.

5.3.2. *Differences in types of care*

The summary measures of treatment can be disaggregated to better understand the types of care that differ across the two sets of physicians. Table 6 reports the results of nine such models. The first is a simple count of the number of procedures, which averages 1.7. Patients assigned to Program B are found to receive 0.25 additional procedures on average. In terms of the types of procedures, column (2) shows that there is little difference in the number of surgeries. Much of the overall difference stems from differences in diagnostic procedures, and these differences will be explored further below.

The next six columns use the accounting cost segments, which sum to the total accounting cost measure described above. Levels (instead of logs) are used to avoid dropping observations with zero costs in a particular segment. Surgery costs are found to be

[24] The mortality-rate quartiles could be affected by differences in the programs' diagnoses and their effectiveness, but when the conditions are scanned, they are similar to severity rankings when an independent dataset, the Nationwide Inpatient Sample, is used to characterize diagnoses by their mortality rates.

[25] For 30-day readmissions, the coefficient for the cardiac dysrhythmia sample is −0.006 compared to a mean of 13% and the coefficient for the AMI sample is −0.01 compared to a mean of 16%. The coefficient for 5-year mortality is −0.06 compared to a mean of 52% for cardiac dysrhythmias and −0.006 compared to a mean 49% for AMI.

Table 5
Results across diagnoses.

Top 10 most common diagnoses	Dependent variable	Coeff. on assignment to Program B	S.E.	Mean of dep. var.	Program B fraction	p-Value: fraction = 0.5	Obs.
Heart failure	Log(length of stay)	0.252	[0.0272]**	1.53	0.520	0.018	3598
	1-Year mortality	0.005	[0.0210]	0.349			3249
Chronic ischemic heart disease	Log(length of stay)	0.083	[0.0299]**	0.85	0.514	0.15	2662
	1-Year mortality	−0.013	[0.0125]	0.0794			2368
Acute myocardial infarction	Log(length of stay)	0.089	[0.0372]*	1.61	0.505	0.62	2187
	1-Year mortality	−0.030	[0.0201]	0.248			2071
Respiratory and chest symptoms	Log(length of stay)	0.175	[0.0302]**	0.77	0.518	0.092	2142
	1-Year mortality	−0.004	[0.0133]	0.0914			1828
Chronic obstructive pulmonary disease	Log(length of stay)	0.191	[0.0343]**	1.36	0.457	<0.001	2137
	1-Year mortality	0.001	[0.0256]	0.294			1965
Diabetes	Log(length of stay)	0.131	[0.0456]**	1.61	0.544	0.001	2097
	1-Year mortality	−0.025	[0.0198]	0.184			1920
Cardiac dysrhythmias	Log(length of stay)	0.145	[0.0392]**	1.41	0.494	0.56	2034
	1-Year mortality	−0.039	[0.0205]	0.213			1899
GI bleed	Log(length of stay)	0.163	[0.0370]**	1.40	0.493	0.53	1974
	1-Year mortality	−0.015	[0.0221]	0.218			1856
Pneumonia	Log(length of stay)	0.210	[0.0364]**	1.50	0.516	0.15	1944
	1-Year mortality	0.015	[0.0232]	0.307			1749
Other acute and subacute forms of ischemic heart disease	Log(length of stay)	0.129	[0.0372]**	1.33	0.512	0.32	1843
	1-Year mortality	−0.027	[0.0151]	0.0895			1821
Pr(mortality diagnosis), bottom quartile	Log(length of stay)	0.023	[0.0167]	1.13	0.508	0.16	8767
	1-Year mortality	−0.004	[0.0047]	0.0412			8250
Pr(mortality diagnosis), 2nd quartile	Log(length of stay)	0.112	[0.0131]**	1.18	0.510	0.012	17,153
	1-Year mortality	−0.008	[0.0056]	0.101			15,765
Pr(mortality diagnosis), 3rd quartile	Log(length of stay)	0.119	[0.0116]**	1.48	0.493	0.030	26,420
	1-Year mortality	−0.009	[0.0068]	0.230			24,424
Pr(mortality diagnosis), top quartile	Log(length of stay)	0.142	[0.0141]**	1.72	0.510	0.0035	20,026
	1-Year mortality	−0.005	[0.0090]	0.466			18,499

Top 10 most frequent diagnoses based on 3-digit ICD-9 diagnosis codes, with the exception GI bleed and COPD defined by a group of diagnosis codes. Models estimated using OLS. All models include full controls and diagnostic fixed effects. Robust standard errors in brackets, clustered by patient.

* Significant at 5%.

** Significant at 1%.

$123 lower for Program B on average, or 9% of the sample mean. In all of the other categories, Nursing, Radiology, Lab, Pharmacy, and "all other" costs, Program B is associated with similarly higher costs in comparison to the mean for each segment, ranging from 7% of the mean for nursing care to 13% of the mean for laboratory costs.

One explanation for the lower costs associated with Program A is that these physicians may rely more heavily on outpatient care

Table 6
Differences by types of care.

			Accounting cost segments:						
Dependent Variable:	Number of Procedures	Number of Surgeries	Nursing	Surgery	Radiology	Lab	Pharmacy	All other	Outpatient referral
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Assigned to Program B	0.250 [0.0143]**	−0.002 [0.0036]	292 [88.2776]**	−123 [30.5502]**	40 [12.1013]**	53 [8.8733]**	112 [48.6039]*	253 [46.0791]**	−0.009 [0.0039]*
Observations	72,366	72,366	34,098	34,098	34,098	34,098	34,098	34,098	72,366
Mean of dep. var	1.68	0.290	4145	1354	483	415	982	2431	0.793

Models estimated using OLS. All models include full controls and diagnostic fixed effects. Robust standard errors in brackets, clustered by patient. Cost measures are in 2006 dollars.

* Significant at 5%.

** Significant at 1%.

Table 7
Use of diagnostic tests and non-surgical procedures.

Comparison:	Procedure rate		# | any		Days to procedure | ordering		Days to procedure	
	Program A	Program B	Program A	Program B	Program A	Program B	Hazard ratio (Program B:Program A)	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(A) All cases								
Any diagnostic	68.4%	73.1%**	2.99	3.25**	1.41	1.55**	0.993	0.0069
X-ray	22.4%	25.1%**	1.77	1.77	3.04	3.17	0.948	0.0075**
Chest X-ray	6.3%	7.5%**	1.11	1.13*	4.39	4.69*	0.930	0.0077**
Endoscopy	5.2%	5.7%**	1.26	1.30**	4.90	4.89	0.921	0.0078**
Angiography	8.1%	8.3%	2.70	2.67	3.16	3.53**	0.915	0.0077**
Cardiac stress test	6.4%	7.8%**	1.02	1.02	3.96	4.39**	0.925	0.0078**
Other cardiac test (incl. echo.)	12.7%	15.0%**	1.12	1.11	1.39	2.21**	0.933	0.0079**
Observations	35,932	36,434					72,366	
(B) Heart failure								
Any diagnostic	78.6%	82.7%	2.92	3.33*	1.10	1.34**	0.937	0.025*
Angiography	5.6%	6.3%	2.80	2.75	4.81	7.26**	0.747	0.026**
Cardiac stress test	11.4%	13.6%*	1.03	1.03	3.42	4.52**	0.771	0.026**
Other cardiac test (incl. echo.)	29.7%	33.2%*	1.09	1.15	0.93	1.62**	0.821	0.027**
Observations	1728	1870					3598	
(C) Acute myocardial infarction								
Any diagnostic	90.7%	93.2%*	3.88	4.18**	1.26	1.36	0.951	0.031
Angiography	46.6%	46.3%	3.01	3.00	3.04	3.36	0.911	0.037*
Cardiac stress test	20.6%	29.6%**	1.03	1.03	5.43	5.33	1.010	0.042
Other cardiac test (incl. echo.)	33.2%	38.0%**	1.15	1.13	2.01	3.02**	0.904	0.037*
Observations	1082	1105					2187	
(D) Chronic obstructive pulmonary disease								
Any diagnostic	84.3%	87.1%	3.26	3.30	0.59	0.94**	0.909	0.028**
X-ray	16.0%	18.1%	1.52	1.54	2.93	3.58	0.825	0.033**
Chest X-ray	9.9%	11.6%	1.09	1.07	2.91	3.66	0.838	0.034**
Observations	1160	977					2137	
(E) GI bleed								
Any diagnostic	75.0%	79.4%*	2.68	2.98*	0.74	0.94**	0.951	0.033
Endoscopy	59.0%	62.8%	1.29	1.35*	2.19	2.28	0.848	0.034**
Observations	1001	973					1974	

Columns (1) and (2) report the fraction of patients who received the procedure at least once; Columns (3) and (4) report the number of procedures conditional on having at least one; Columns (5) and (6) report the mean number of days to the first time the procedure is conducted conditional on having the procedure; Column (7) reports hazard ratios of the duration to the first time a procedure is conducted: results are from Cox proportional hazard models with full controls. Standard errors are clustered at the patient level.

* Significant at 5%.

** Significant at 1%.

as a substitute for inpatient care. Our data describes whether an outpatient referral is made, which happens in most cases when a patient was admitted to the hospital (79% of the time). Program B is associated with a 1 percentage-point lower outpatient referral rate, which suggests that such substitution does not drive the inpatient cost differences.

5.3.3. *Differences in diagnostic testing*

Table 6 suggests that the difference in costs stems from differences in diagnostic testing. Table 7 explores this question overall and for particular diagnoses. Columns (1) and (2) report the frequency with which each program orders particular tests. For example, patients assigned to physicians from Program B are more likely to undergo diagnostic tests compared to patients treated by Program A (73% vs. 68%). This difference is found among common diagnostic tests including X-rays and stress tests. Columns (3) and (4) report the number of tests conditional on ordering any tests. Even conditional on ordering some tests, Program B is found to order 8% more than Program A (3.25 vs. 2.99). Within procedures, the frequency of tests is more likely to be similar—a cardiac stress test, for example, is only conducted once (on average) in both groups if it is conducted at all.

Another source of variation in treatment is the timing of diagnostic tests. Table 7 shows that Program B is 10% slower, on average, to order the first test conditional on ordering one (1.55 days vs. 1.41 days). To account for the time at risk for procedures and include all observations, Cox proportional hazard models estimates show that for individual procedures, Program B is approximately 8% slower to order a test Program A. These differences are seen for X-rays, angiography, and cardiac tests.

The differences in Panel A may mask differences within particular diagnoses. Four common diagnoses were chosen that have fairly standard diagnostic tests. The differences are less likely to be statistically significant due to the smaller sample sizes, but large point estimates point to patterns, especially the longer duration to the first test.

Panel B reports results for congestive heart failure, a chronic condition that is a common source of hospital admission. Higher test rates are found for Program B (5% higher overall; 19% higher for stress tests). Program B orders 14% more tests conditional on any (3.33 vs. 2.92). In terms of timing, they take 21% longer to order the first test (1.34 days vs. 1.10 days), 51% longer to order an angiography if one is ordered (7.26 days vs. 4.81 days), 32% longer to order a cardiac stress test, and 74% longer to order other cardiac tests (including echocardiograms). Hazard ratios that take into account patients that did not receive the test as well show somewhat smaller but still economically and statistically significant differences: hazard ratios of 0.75 and 0.77 for angiography and cardiac stress tests, for example.

Table 8
Specification and robustness checks.

	Dependent variable	Coeff. on assignment to Program B	S.E.	Mean of dep. var.	Obs.
Sample: nervous system patients	Log(length of stay)	0.047	0.048	1.34	1353
	30-Day readmission	−0.011	0.022	0.191	1345
	1-Year mortality	−0.040	0.021	0.153	1284
Sample: outside main facility	Log(length of stay)	−0.012	0.014	1.89	70,775
	1-Year mortality	0.0050	0.004	0.141	63,299
Intensive care unit	Admission to ICU	−0.0020	0.0033	0.181	72,366
	Log(length of stay in ICU)	−0.0169	0.015	0.806	13,110
	Died in the ICU	−0.0023	0.0037	0.047	13,110
White veterans	Log(length of stay)	0.0759	0.012**	1.48	33,923
	1-Year mortality	−0.0060	0.0066	0.239	33,923
Non-White veteran (or missing race)	Log(length of stay)	0.1380	0.011**	1.39	38,443
	1-Year mortality	−0.0048	0.0070	0.245	33,015
Readmission outcomes	30-Day readmission: same major diagnosis	−0.0020	0.0021	0.071	71,954
	30-Day readmission costs	20.3	89.4	1653	42,106
	1-Year readmission costs	243	155	4868	37,090
Mortality outcomes	10-Hour mortality	−0.00042	0.0004	0.0025	72,366
	Died in the hospital	0.0020	0.0014	0.040	72,366
Transfers	Transfer to another hospital	−0.0028	0.0016	0.040	72,366
Sample: first episode	Log(length of stay)	0.096	0.0097**	1.40	29,391
	30-Day readmission	−0.010	0.0033**	0.091	29,278
	1-Year mortality	−0.0037	0.004	0.173	27,581
	5-Year mortality	−0.0040	0.006	0.391	20,882

All models include full controls, including 3-digit diagnosis indicators. Robust standard errors in brackets, clustered by patient.
** Significant at 1%.

Panel C reports the results for myocardial infarction. Program B is associated with 40% higher rates of cardiac stress tests (30% vs. 21%) and higher rates of "other cardiac tests including echocardiograms. Stress tests are often used to provide evidence that the patient is safe to be discharged, and the difference is consistent with Program B relying on such additional information at a much higher rate. Conditional on ordering the tests, they order 8% more and have a 7% longer duration to the first test, including 50% more time before tests such as an echocardiogram is taken (3 days vs. 2 days). The hazard ratios are closer to 0.90 for angiography and other cardiac tests.

Panel D reports the results for another common admission: chronic obstructive pulmonary disease. Overall, diagnostic-testing rates are similar across the programs, although Program B is 17% more likely to order a chest X-ray and 13% more likely to order any X-ray compared to Program A. The main difference within this diagnosis is the time to the first test: 59% longer for Program B on average (0.94 days vs. 0.59 days), and approximately 25% longer for an X-ray (hazard ratios of 0.91 and 0.82). Panel E reports similar results for gastrointestinal bleeding, with 6% higher test rates, 11% more tests conditional on ordering any, and 27% longer duration before the first test (0.94 days vs. 0.74 days), with a hazard ratio for endoscopy of 0.85.

In summary, Program B tends to order more diagnostic tests, and they take longer to order tests.

5.4. Robustness and specification checks

This section offers tests to verify the experimental nature of the setting and offer more clues to the sources of the differences in costs. It also considers heterogeneity in effects across patient groups.

5.4.1. Placebo tests

Three placebo checks were conducted. Table 8 shows results for patients with the major diagnostic category of "nervous system"—a group that is less likely to enter the randomization—a much smaller treatment difference is found (coefficient of 0.047), and the difference is not statistically significant. Second, when patients admitted to a satellite facility (where randomization does not take place) were compared, again there is no difference in length of stay or 1-year mortality. This is consistent with similar comorbidity levels across the two groups, similar outcomes for the two groups of patients, and no difference in the reliance on outpatient care across the two physician teams.

Third, the other area where the randomization has less of an effect is when a patient is admitted to the intensive care unit, which is overseen by a single attending from one of the programs at any given point in time. We also did not find a difference in the rate of transfer to the ICU across the two groups. Once in the ICU, the length of stay and mortality rates were similar. For patients who were transferred out of the ICU to another hospital bed, their post-ICU length of stay was significantly different (not shown). Further, when patients who did not use an intensive care unit were analyzed, the treatment differences were somewhat larger in magnitude, and no outcome differences were found.

5.4.2. Heterogeneity across patients

Part of the interest in estimating the returns to physician human capital is the concern that minority patients may lack access to top physicians. The natural experiment here allows us to compare the treatment and outcome differences for White vs. non-White patients.[26] Table 8 shows that the difference in treatment is larger

[26] The non-White category includes missing race (Sohn et al., 2006). Racial composition in the patent's ZIP code is associated with the race listed in the patient treatment file, however, which suggests that the race variable is informative: we

for non-White patients (14 log point difference in length of stay compared to 8 log points for White patients). 1-Year mortality is similar across Whites and non-Whites at 24%, and the Program assignment is unrelated to this outcome. Results were also similar for patients over and under the age of 65—the latter group has alternative insurance coverage through Medicare.

5.4.3. *Alternative outcome measures*

Table 8 reports results for additional outcomes, and the results are robust. This includes outcomes such as 30-day readmissions for the same major diagnostic category and readmission costs. In terms of mortality, both in-hospital mortality and 10-hour mortality—measures that perhaps have the most direct influence of the resident team, especially those that require faster decisions—are similar across the teams.

An explanation for the shorter stays associated with Program A could be that these physicians are more likely to transfer patients to another hospital, potentially to perform a surgery that is not conducted at the VA such as a coronary artery bypass. Table 8 shows that Program B is associated with a slightly lower transfer rate: 0.3% compared to a mean transfer rate of 4%. This difference cannot by itself explain the difference in length of stay.[27] Further, when (the small number of) transferred patients were dropped from the analysis, the results are essentially the same as the main results (see Appendix Table A2).

5.4.4. *June vs. July: heterogeneity in resident experience*

One limitation of the analysis of residents is that the practice styles and outcomes may converge or diverge as the physicians gain experience later in their careers. Future analysis will use Medicare data to track these physicians into their careers. In the data available here, we can compare patients in June vs. July—the month when new residents begin training and the pool of residents has nearly one less year of experience. This 2-month comparison also controls for seasonal differences in the types of conditions encountered. Given the smaller sample sizes, results should be taken with some caution, as the differences between June and July are not statistically significant. That said, we find that the magnitude of the treatment differences is smaller in June when the residents are more seasoned (7% difference). Patients assigned to Program B when the residents are relatively inexperienced in July have lengths-of-stay that are an additional 5% longer (see Appendix Table A3). The outcome results are more mixed for readmissions and mortality, but the differences continue to be small.

5.4.5. *Differences when workloads differ*

One difference between the two teams is that Program A's teams have one intern, whereas Program B's teams have two. One way to test whether these different intern-to-patient ratios are driving the results is to estimate the effect of caseload on treatment intensity across the two groups. Each program sees approximately 50 patients per week on average. Busier times were generally associated shorter lengths of stay and lower mortality rates, which likely reflects lower levels of illness severity at these times. When loads are higher for Program B, they are somewhat more treatment intensive compared to busier times for Program A. The effects of patient load on treatment appear too small for the difference in intern-to-patient ratios to explain the main results, however. In particular, when we control for the admission load in the week prior to any given admission, the difference between the two groups is largely unchanged.[28]

5.4.6. *Additional robustness tests*

The results were also similar when the sample is restricted to the initial episode (in our data), especially for 5-year mortality. Other tests were conducted that are not shown in Table 8 include similar results when date fixed effects were included[29]; when probit models were used for outcomes and count models were used for length of stay; and when hours in care were compared rather than days. In addition, results were robust to the time period, with large cost differences each year.

5.5. *Interpretation*

5.5.1. *Competing explanations*

Program B is found to have higher costs, yet the health outcomes we measure are similar compared to Program A. One interpretation is that the additional tests and wait times by Program B are unnecessary. If this were the case, it may be possible for the lower ranked physicians to achieve similar outcomes at substantial savings. For example, Program A may be better at administrative tasks that reduce costs, such as test scheduling, but are unrelated to patient health.

An alternative explanation is that the physicians from the lower ranked program may require the additional tests and input from consultants to achieve the same results as the higher ranked program. If this were the case, then the decision-making ability of the physicians in the higher ranked program would not be so easily replicable. This explanation finds some support from the results that the larger differences in treatment appear for more complicated diagnoses. In addition, the longer duration before the first test for Program B suggests that these physicians may need more time and advice before ordering tests.

Another set of explanations of the treatment differences may be more bureaucratic hurdles faced by Program B and not Program A. For example, attending physicians in Program B could provide more oversight, which takes more time to administer. If a mechanical rule that all tests had to be approved by the attending led to the cost differences, we would expect differences in treatment even for less serious cases, but that was not found (Table 5). Of course, the greater supervision may be requested only in more serious cases. In some ways, additional supervision may capture important differences in the two programs if the physicians in the lower ranked program require additional advice. Importantly, physicians familiar with the training at this VA do not believe that the level of attending supervision or other bureaucratic differences, such as access to lab results, are substantially different across the two groups.

divided the sample into quartiles based upon the fraction white in the patient's ZIP code. Patients in the bottom quartile are recorded as white 9.5% of the time compared to 72% in the top quartile. When treatment and outcomes are compared, the bottom quartile shows the largest difference in log length of stay (16 log points), and a model without controls suggests that Program B is associated with mortality that is 2 percentage points lower compared to a mean of 25% in this quartile.

[27] For this difference in transfer rate to explain the 10% difference in length of stay, those patients more likely to remain due to Program B assignment would have to stay for 139 days compared to a mean of 4.4.

[28] Models for log costs and log length of stay were re-estimated with the addition of a measure of admission load: the number of admissions for each group in the 7 days prior to the observation, as well this load measure interacted with the Program B indicator, full controls, and date fixed effects. The difference in length of stay across the two groups, evaluated at the sample mean of the admission count, was 0.10, similar to the main results. When the estimates are evaluated when Program A has half the number of admissions (so that the intern-to-patient ratios are the same), the magnitude of the length-of-stay difference increases to −0.11. For log costs, these estimates are even more similar: 0.117 and 0.119, respectively.

[29] See Table A2.

The structure of the two groups is somewhat different, with Program B having 2 interns per team compared to 1 for Program A. Another explanation for faster treatment among the smaller teams in the higher ranked program could be lower coordination costs, but teams do not coordinate care across the interns for any given patient.[30] In addition, senior residents in Program A may be more likely to take admissions during busy times compared to Program B, which has two interns to handle a higher load. While we have not found evidence that this is the case, it would change the interpretation to combine the effects of a higher ranked program and a more-experienced team leading to substantially lower costs. That said, the effects of caseload itself appear too small to drive the differences in costs, as discussed above.

5.5.2. Limitations

There are a number of additional limitations in the current study. First, the randomization applies to two residency programs at one teaching hospital, which raises questions of external validity. These programs are comprised of thousands of physicians over the 13 years considered, however, and the results are robust to the particular set of physicians at any given time. One reason to believe that there may be wider applicability is that Program A's parent hospital is fairly similar to other U.S. News and World Report's Honor Roll Hospitals according to the Dartmouth Atlas. In terms of average number of hospital days and the number of physician visits in the last two years of life between 2001 and 2005, the parent hospital is in the middle of the distribution of these hospitals. It appears that other top hospitals provide similar levels of treatment intensity as the higher ranked program. In comparison, the parent hospital affiliated with Program B has similar treatment intensity measures as the parent hospital for Program A—both are higher than the national average, but not at the extremes like some Honor Roll hospitals.[31] Such results may not apply to community hospitals where the goals and incentives of physicians may differ.

Second, variations in delivery of health care can be explained both by differences in the selectivity of the programs and clinical training during residency (Weiss, 1995; Semeijn et al., 2005). It is difficult to separate the two effects here, but it appears that the training is qualitatively similar. We found that the program curriculum, teaching philosophy, approach to clinical care, as well as treatment intensity in the parent hospitals of the two programs, are generally similar across the two institutions.

To the extent that the results are driven by different residents, as opposed to different attending physicians, a related limitation is that differences could fade (or increase) over time as physicians gain experience. The June vs. July comparisons described above suggest that treatment differences may converge somewhat, although the outcome differences were similar when the residents were relatively inexperienced.[32]

Fourth, the results apply to a veteran population, and the results may not apply to a wider set of patients. Still, this population is particularly policy relevant given the concerns that differing access to high-quality physicians may lead to health disparities among low-income groups. Here, we have just such a group that has an equal chance of being treated by a top physician team or one ranked much lower. Further, medical schools join with VA medical centers partly because the patients present with a wide range of illnesses—an advantage here in that we can compare the results across these diagnoses as well.

Fifth, a usual limitation of randomized trials is that they do not incorporate the value of matching physicians to patients. Here, the lack of a health outcome difference suggests that such triage is less likely to be necessary. In addition, if the cost savings would be greater with matching, then the magnitude of the cost differences that we find can be viewed as a lower bound.

Last, the difference in costs reported here is in terms of hospital resources devoted to a given patient. It is possible that lower cost physicians extract part of this through higher wages. While the wages for residents do not vary across the residents considered here, they may diverge later in the physicians' careers, and therefore may have different implications in non-teaching environments. Unfortunately, physician-income surveys do not include physicians' medical school or residency training program to test such a relationship.

6. Conclusions

Physicians play a major role in determining the cost of health care, and there are concerns that limitations on the supply of physicians and disparities in access to high-quality physicians and facilities can affect health outcomes. Comparisons of physicians are often confounded by differences in the patients they treat and the environments where they work. We study a unique natural experiment where nearly 30,000 patients were randomized to two sets of residency training programs in the same hospital. One is consistently ranked among the top programs in the country, whereas the other is ranked lower in the distribution according to measures such as the pass rate for Board exams.

We find patients randomly assigned to the highly ranked program incur substantially lower costs: 10% overall and up to 25% depending on the condition. This difference is driven largely by variation in diagnostic testing, where Program B orders more tests and takes longer to order them. No difference is found for health outcomes, however.

The results suggest a number of potential implications. First, physician effects on costs can be substantial, as expected but usually difficult to quantify. Second, if the results apply more broadly, inequality in access to top-ranked physicians may lead to differences in the use of specialists and testing but may not lead to health disparities. This suggests that a relaxation of accreditation standards for medical schools, for example, may not adversely affect quality of care, but may raise costs despite a greater supply of physicians.[33] Third, the results are consistent with previous evidence that high-cost areas are associated with a greater use of diagnostic tests and reliance

[30] The interns "scatter" during the start of their shifts to provide care to each of their patients. The interns do round together, but the difference in composition is not expected to result in substantially different amounts of time spent on rounds. Recently, Program B switched to a 3-team system described in an earlier version of this paper, but the change is outside of our sample period.

[31] We thank Jack Wennberg for this suggestion. According to the Dartmouth Atlas performance reports for 2001–2005, the average hospital days per Medicare beneficiary during the last 2 years of life—a preferred measure of utilization that controls for the health of the patient and is not directly affected by price differences—is nearly identical for the two parent hospitals. They also have similar facility capacity in terms of total beds and ICU beds—measures that have been found to be associated with treatment intensity (Fisher et al., 1994).

[32] One study that compares residents and attending that we are aware of found that their practice patterns to be similar: Detsky et al. (1986) examined a strike by residents in 1980 and found that the volume of tests performed did not change when the attendings provided the care instead.

[33] A classic study by Friedman and Kuznets (1945) attributed relatively high salaries among physicians, relative to dentists, to more stringent licensing requirements. This study suggests a countervailing effect of higher resource use among marginal entrants.

on specialists with little difference in health outcomes. This additional care may be unnecessary—providing a basis for innocuous cost containment. The results here suggest an alternative interpretation is possible as well: that higher cost areas may require greater treatment intensity to achieve similar outcomes. It remains to be tested whether high-cost areas are able to replicate the higher quality care associated with the low-cost areas.

Appendix A.

See Tables A1–A3.

Table A1
Selected covariates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	Log(length of stay)	Log(accounting cost)	Log(estimated cost)	30-Day readmission	1-Year readmission	30-Day mortality	1-Year mortality
Assigned to Program B	0.1125	0.1251	0.1039	−0.0021	0.0055	−0.00073	−0.0072
	[0.0072]**	[0.0114]**	[0.0099]**	[0.0030]	[0.0051]	[0.0019]	[0.0044]
Midnight–6 am	0.0474	0.2142	0.1847	−0.0175	−0.029	−0.0228	−0.0401
	[0.0133]**	[0.0205]**	[0.0177]**	[0.0052]**	[0.0077]**	[0.0037]**	[0.0062]**
6 am–12 noon	0.1658	0.0808	0.1065	−0.0091	−0.0112	−0.0098	0.0038
	[0.0121]**	[0.0177]**	[0.0153]**	[0.0048]	[0.0071]	[0.0034]**	[0.0058]
12 noon–6 pm	0.241	0.1297	0.1738	−0.0096	−0.0038	−0.0046	0.0127
	[0.0123]**	[0.0180]**	[0.0156]**	[0.0049]	[0.0074]	[0.0036]	[0.0060]*
Wednesday (vs. Saturday)	0.0327	−0.0454	−0.0082	−0.0018	−0.0078	−0.0065	−0.0017
	[0.0134]*	[0.0226]*	[0.0194]	[0.0054]	[0.0080]	[0.0038]	[0.0062]
Married	−0.0893	−0.0763	−0.07	0.0034	0.0058	−0.0067	−0.0264
	[0.0091]**	[0.0143]**	[0.0125]**	[0.0038]	[0.0063]	[0.0024]**	[0.0056]**
Male	0.061	−0.0275	0.0864	0.006	0.0205	0.0111	0.0451
	[0.0225]**	[0.0315]	[0.0296]**	[0.0087]	[0.0163]	[0.0042]**	[0.0129]**
White	0.0158	0.0308	0.0115	−0.0062	−0.0004	0.0033	0.0065
	[0.0112]	[0.0199]	[0.0157]	[0.0046]	[0.0076]	[0.0031]	[0.0069]
Charlson index = 1	0.0884	0.0695	0.0974	0.0201	0.066	0.0032	0.0351
	[0.0091]**	[0.0145]**	[0.0129]**	[0.0034]**	[0.0058]**	[0.0019]	[0.0040]**
Charlson index = 2	0.202	0.2054	0.2248	0.0555	0.1422	0.0352	0.1584
	[0.0099]**	[0.0158]**	[0.0140]**	[0.0039]**	[0.0063]**	[0.0025]**	[0.0053]**
Age: 35–44	0.181	0.1336	0.092	0.0115	0.0391	0.004	0.0044
	[0.0295]**	[0.0659]*	[0.0500]	[0.0117]	[0.0212]	[0.0038]	[0.0137]
45–54	0.2452	0.1913	0.1134	0.0101	0.0653	0.0104	0.0276
	[0.0284]**	[0.0616]**	[0.0466]*	[0.0110]	[0.0205]**	[0.0037]**	[0.0135]*
55–64	0.3328	0.2839	0.1319	0.0106	0.0666	0.0216	0.0621
	[0.0284]**	[0.0617]**	[0.0468]**	[0.0110]	[0.0205]**	[0.0038]**	[0.0138]**
65–69	0.3598	0.2533	0.0969	0.0061	0.0773	0.0303	0.0998
	[0.0292]**	[0.0634]**	[0.0483]*	[0.0113]	[0.0208]**	[0.0043]**	[0.0144]**
70–74	0.372	0.3103	0.1074	0.0111	0.0819	0.0409	0.1283
	[0.0292]**	[0.0629]**	[0.0480]*	[0.0114]	[0.0209]**	[0.0043]**	[0.0145]**
75–84	0.3894	0.2958	0.0775	0.0281	0.0823	0.0573	0.18
	[0.0290]**	[0.0622]**	[0.0474]	[0.0114]*	[0.0209]**	[0.0043]**	[0.0145]**
84+	0.3873	0.2803	0.0338	0.0164	0.0562	0.0973	0.3124
	[0.0344]**	[0.0673]**	[0.0533]	[0.0136]	[0.0243]*	[0.0085]**	[0.0200]**
Constant	1.3466	8.3545	8.6239	0.0388	0.043	0.0943	0.1759
	[0.1792]**	[0.2980]**	[0.2563]**	[0.0730]	[0.1199]	[0.0484]	[0.1107]
Observations	72,366	34,098	42,518	71,954	66,938	71,954	66,938
R-squared	0.22	0.25	0.26	0.03	0.07	0.11	0.22
Mean of dep. var.	1.43	8.63	8.71	0.1315	0.4287	0.0642	0.2418

Models also included year, month, day-of-week, and divorced indicators, as well as ZIP code characteristics. Robust standard errors in brackets.

* Significant at 5%.

** Significant at 1%.

Table A2
Additional checks.

	Dependent variable	Coeff on assignment to Program B	S.E.	Mean of dep. var.	Obs.
Model: probit (marginal effects)	30-Day readmission	−0.002	0.0030	0.133	71,373
	1-Year mortality	−0.008	0.0048	0.244	66,230
Model: OLS w/date fixed effects	Log(length of stay)	0.109	0.007**	1.43	72,366
	30-Day readmission	−0.003	0.003	0.131	71,954
	1-Year mortality	−0.007	0.004	0.242	66,938
Sample: drop transferred patients	Log(length of stay)	0.114	0.007**	1.42	69,451
	30-Day readmission	−0.003	0.003	0.129	69,047
	1-Year mortality	−0.007	0.004	0.241	64,177

All models include lull controls, including 3-digit diagnosis indicators. Robust standard errors in brackets, clustered by patient.

** Significant at 1%.

Table A3
Effects of experience: June vs. July.

Dependent variable:	Log(length of stay) (1)	30-Day readmission (2)	1-Year mortality (3)
Assigned to Program B	0.069 [0.0221][**]	−0.0091 [0.0091]	0.0025 [0.0110]
July	−0.0008 [0.0213]	−0.0081 [0.0086]	−0.0055 [0.0101]
Assigned to Program B × July	0.049 [0.0302]	0.017 [0.0122]	−0.0010 [0.0143]
Observations	12,256	12,256	11,286
Mean of dep. var.	1.39	0.134	0.244

Sample limited to patients admitted in June or July. Models estimated using OLS with full controls. Robust standard errors in brackets, clustered by patient.

[**] Significant at 1%.

References

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, New Jersey.

Almond, D., Chay, K., Greenstone, M., in press. Civil rights, the war on poverty, and Black–White convergence in infant mortality in the rural south and Mississippi. American Economic Review.

Arnold, N., Sohn, M., Maynard, C., Hynes, D.M., 2006. VA-NDI Mortality Data Merge Project. VIReC Technical Report 2. VA Information Resource Center, Edward Hines, Jr. VA Hospital, Hines, IL.

Ashton, C.M., Souchek, J., Petersen, N.J., Menke, T.J., Collins, T.C., Kizer, K.W., Wright, S.M., Wray, N.P., 2003. Hospital use and survival among veterans affairs beneficiaries. New England Journal of Medicine 349 (17), 1637–1646.

Bach, P.B., Phram, H.H., Schrag, D., Tate, R.C., Hargraves, J.L., 2004. Primary care physicians who treat Blacks and Whites. New England Journal of Medicine 351 (6), 575–584.

Baicker, K., Chandra, A., 2004. Medicare spending, the physician workforce, and beneficiaries quality of care. Health Affairs W4, 184–197.

Bartel, A., Phibbs, C., Stone, P., Beaulieu, N., (2009), Human Capital and Productivity: The Case of Nursing Teams, Working paper.

Bjorklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. The Review of Economics and Statistics 69 (February(1)), 42–49.

Burns, L.R., Wholey, D.R., 1991. The effects of patient, hospital, and physician characteristics on length of stay and mortality. Medical Care 29 (3), 251–271.

Burns, L.R., Chilingerian, J.A., Wholey, D.R., 1994. The effect of physician practice organization on efficient utilization of hospital resources. Health Services Research 29 (5), 583–603.

Case, S.M, Swanson, D.B., 1993. Validity of the NBME Part I and Part II scores for selection of residents in orthopaedic surgery, dermatology, and preventive medicine. Academic Medicine 68, S51–S56.

Chandra, A., Skinner, J., 2003. Geography and Racial Health Disparities. NBER Working Paper No. 9513.

Chang, B.K., 2005. Resident supervision in VA teaching hospitals. ACGME Bulletin (September), 12–13.

Chen, J., Rathore, S.S., Wang, Y., Radford, M.J., Krumholz, H.M., 2006. Physician board certification and the care and outcomes of elderly patients with acute myocardial infarction. Journal of General Internal Medicine 21 (3), 238–244.

Cole, J.R., Lipton, J.A., 1977. The reputations of American medical schools. Social Forces 53 (3), 662–684.

Cutler, D.M., Huckman, R.S., Landrum, M.B., 2004. The role of information in medical markets: an analysis of publicly reported outcomes in cardiac surgery. American Economic Review 94 (2), Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association, May, pp. 342–346.

Dafny, L., 2005. How do hospitals respond to price changes. American Economic Review 95 (December(5)), 1525–1547.

Detsky, A.S., McLaughlin, J.R., Abrams, H.B., L'Abbe, K., Markel, F.M., 1986. Do interns and residents order more tests than attending staff? Results of a house staff strike. Medical Care 24 (6), 526–534.

Dranove, D., Daniel, K., McClellan, M., Satterthwaite, M., 2003. Is more information better? The effects of "Report Cards" on health care providers. Journal of Political Economy 111 (3), 555–588.

Eisenberg, J.M., 2002. Physician utilization: the state of research about physicians' practice patterns. Medical Care 40 (11), 1016–1035.

Evans, W.N., Kim, B.S., 2006. Patient outcomes when hospitals experience a surge in admissions. Journal of Health Economics 25 (2), 365–388.

Fisher, E.S., Wennberg, J.E., Stukel, T.A., Sharp, S.M., 1994. Hospital readmission rates for cohorts of medicare beneficiaries in Boston and New Haven. New England Journal of Medicine 331, 989–995.

Fisher, E., Wennberg, D., Stukel, T., Gottlieb, D., Lucas, F., Pinder, E., 2003. Implications of regional variations in Medicare spending. Part 2. Health outcomes and satisfaction with care. Annals of Internal Medicine 138 (4), 288–298.

Ferguson, E., David, J., Madeley, L., 2002. Factors associated with success in medical school: systematic review of the literature. British Medical Journal 324, 952–957.

Friedman, M., Kuznets, S., 1945. Income from Independent Professional Practice. National Bureau of Economic Research, New York.

Geweke, J., Gowrisankaran, G., Town, R.J., 2003. Bayesian inference for hospital quality in a selection model. Econometrica 71 (4), 1215–1238.

Gillespie, K.N., Romeis, J.C., Virgo, K.S., Fletcher, J.W., Elixhauser, A., 1989. Practice pattern variation between two medical schools. Medical Care 27 (5), 537–542.

Glance, L.G., Dick, A., Mukamel, D.B., Li, Y., Osler, T.M., 2008. Are high-quality cardiac surgeons less likely to operate on high-risk patients compared to low-quality surgeons? Evidence from New York State. Health Services Research 43 (1), 300–312.

Glaser, K., Hojat, M., Velkoski, J.J., Blacllow, R.S., Goepp, C.E., 1992. Science, verbal, or quantitative skills: which is the most important predictor of physician competence? Educational and Psychological Measurement 52, 395–406.

Grytten, J., Sorensen, R., 2003. Practice variation and physician-specific effects. Journal of Health Economics 22, 403–418.

Hannan, E.L., Chassin, M.R., 2005. Publicly Reporting Quality Information 293 (24), 2999–3000.

Hartz, A.J., Kuhn, E.M., Pulido, J., 1999. Prestige of training programs and experience of bypass surgeons as factors in adjusted patient mortality rates. Medical Care 37 (1), 93–103.

Hayward, R.A., Manning Jr., W.G., McMahon Jr., L.F., Bernard, A.M., 1994. Do attending and resident physician practice styles account for variations in hospital resource use? Medical Care 32 (8), 788–794.

Hofer, T.P., et al., 1999. The unreliability of individual physician 'Report Cards' for assessing the costs and quality of care of a chronic disease. Journal of the American Medical Association 281, 2098–2105.

Hojat, M., Gonnella, J.S., Erdmann, J.B., Jon, V.J., 1997. The fate of medical students with different levels of knowledge: are the basic medical sciences relevant to physician competence. Advances in Health Sciences Education 1, 179–196.

Huckman, R., Barro, J., 2005. Cohort Turnover and Productivity: The July Phenomenon in Teaching Hospitals. NBER Working Paper No. 11182.

Ibrahim, S.A., 2007. The Veterans Health Administration: a domestic model for a national health care system? American Journal of Public Health 97 (December(12)), 2124–2126.

Institute of Medicine, 2002. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. National Academies Press, Washington D.C.

Kelly, J.V., Hellinger, F.J., 1987. Heart disease and hospital deaths: an empirical study. Health Services Research 22 (August(3)), 369–395.

Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17, 283–295.

Marshall, M.N., Shekelle, P.G., Leatherman, S., Brook, R.H., 2000. The public release of performance data: what do we expect to gain? A review of the evidence. Journal of the American Medical Association 283, 1866–1874.

Marder, W.D., Hough, D.E., 1983. Medical Residency as Investment in Human Capital. Journal of Human Resources 18 (1), 49–64.

Meltzer, D., Manning, W.G., Morrison, J., Shah, M.N., Jin, L., Guth, T., Levinson, W., 2002. Hospitalists and the costs and outcomes of hospital care. Annals of Internal Medicine 137, 1–15.

Mooney, C., Zwanziger, J., Phibbs, C., Schmitt, S., 2000. Is travel distance a barrier to veterans' use of VA hospitals for medical surgical care? Social Science and Medicine 50 (12), 1743–1755.

Moulton, B., 1986. Random group effects and the precision of regression estimates. Journal of Econometrics 32, 385–397.

Mukamel, D.B., Murthy, A.S., Weimer, D.L., 2000. Racial differences in access to high-quality cardiac surgeons. American Journal of Public Health 90 (11), 1774–1777.

Newhouse, J., 1996. Reimbursing health plans and health providers: efficiency in production vs. selection. Journal of Economic Literature 34 (3), 1236–1263.

Norcini, J.J., Kimball, H.R., Lipner, R.S., 2000. Certification and specialization: do they matter in the outcome of acute myocardial infarction? Academic Medicine 75, 1193–1198.

Omoigui, N.A., Miller, D.P., Brown, K.J., Annan, K., Cosgrove, D., Bytle, B., Loop, F., Topol, E.J., 1996. Outmigration for coronary bypass surgery in an era of public dissemination of clinical outcomes. Circulation 93 (1), 27–33.

Pauly, M.V., 1978. Medical staff characteristics and hospital costs. Journal of Human Resources 13 (S), 77–111.

Phelps, C., Mooney, C., 1993. Variations in medical practice use: causes and consequences. In: Richard, A., Robert, R., White, W. (Eds.), Competitive Approaches to Health Care Reform. The Urban Institute Press, Washington, DC.

Phibbs, C.S., Bhandari, A., Yu, W., Barnett, P.G., 2003. Estimating the costs of VA ambulatory care. Medical Care Research and Review 60 (3), 54S–73S.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.C., Saunders, L.D., Beck, C.A., Feasby, T.E., Ghali, W.A., 2005. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical Care 43 (11), 1073–1077.

Rosen, S., 1981. The economics of superstars. American Economic Review 71 (5), 845–858.

Schneider, E.C., Epstein, A.M., 1996. Influence of cardiac surgery performance reports on referral practices and access to care. New England Journal of Medicine 335 (4), 251–256.

Semeijn, J., Van der Velden, R., Heijke, H., Van der Vleuten, C., Boshuizen, H., 2005. The role of education in selection and allocation in the labour market: an empirical study in the medical field. Education Economics 13 (4), 449–477.

Silver, B., Hodgson, C.S., 1997. Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution. Academic Medicine 72 (5), 394–396.

Simmer, T.L., Nerenz, D.R., Rutt, W.M., Newcomb, C.S., Benfer, D.W., 1991. A randomized, controlled trial of an attending staff service in general internal medicine. Medical Care 29, JS31–JS40.

Sirovich, B., Gallagher, P.M., Wennberg, D.E., Fisher, E.S., 2008. Discretionary decision making by primary care physicians and the cost of U.S. health care. Health Affairs 27 (3), 813–823.

Sohn, M.W., Arnold, N., Maynard, C., Hynes, D.M., 2006. Accuracy and completeness of mortality data in the department of veterans affairs. Population Health Metrics 4 (2). Available at: http://www.pophealthmetrics.com/content/4/1/2.

Tamblyn, R., Abrahamowicz, M., Brailovsky, C., Grand'Maison, P., Lescop, J., Norcini, J., Girard, N., Haggerty, J., 1998. Association between licensing examination scores and resource use and quality of care in primary care practice. Journal of the American Medical Association 280, 989–996.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Hanley, J.A., John, N., Nadyne, G., Grand'Maison, P., Brailovsky, C., 2002. Association between licensure examination scores and practice in primary care. Journal of the American Medical Association 288, 3019–3026.

Unruh, L., 2003. Licensed nurse staffing and adverse events in hospitals. Medical Care 41, 142–152.

U.S. News and World Report, 2007. Best Hospitals 2007. Accessed via web at: http://health.usnews.com/usnews/health/best-hospitals/honorroll.htm.

Van Ryn, M., 2002. Research on the provider contribution to race/ethnicity disparities in medical care. Medical Care 40 (1), 140–151.

VHA (Veterans Health Administration), 2005. Report to the Secretary of Veterans Affairs. Accessed at: http://www.va.gov/oaa/archive/FACA_Report_2005.pdf.

Volpp, K.G., Rosen, A.K., Rosenbaum, P.R., Romano, P.S., Even-Shoshan, O., Wang, Y., Bellini, L., Behringer, T., Silber, J.H., 2007. Mortality among hospitalized medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. Journal of the American Medical Association 298, 975–983.

Wagner, T.H., Chen, S., Barnett, P.G., 2003. Using average cost methods to estimate encounter-level costs for medical-surgical stays in the VA. Medical Care Research and Review 60 (3), 15S–36S.

Weiss, A., 1995. Human capital vs. signaling explanations of wages. Journal of Economic Perspectives 9 (4), 133–154.

Wennberg, J.E., Bronner, K., Skinner, J.S., Fisher, E.S., Goodman, D.C., 2009. Inpatient care intensity and patients' ratings of their hospital experiences. Health Affairs 28 (1), 103–112.

Werner, R.M., Asch, D.A., 2005a. Publicly reporting quality information—reply. Journal of the American Medical Association 293 (24), 3000–3001.

Werner, R.M., Asch, D.A., 2005b. The unintended consequences of publicly reporting quality information. Journal of the American Medical Association 293 (10), 1239–1244.