# EVALUATING MEASURES OF HOSPITAL QUALITY: EVIDENCE FROM AMBULANCE REFERRAL PATTERNS

Joseph Doyle, John Graves, and Jonathan Gruber*

*Abstract*—Hospital quality measures are crucial to a key idea behind health care payment reforms: "paying for quality" instead of quantity. Nevertheless, such measures face major criticisms largely over the potential failure of risk adjustment to overcome endogeneity concerns when ranking hospitals. In this paper, we test whether patients treated at hospitals that score higher on commonly used quality measures have better health outcomes in terms of rehospitalization and mortality. To compare similar patients across hospitals in the same market, we exploit ambulance company preferences as an instrument for hospital choice. We find that a variety of measures that insurers use to measure provider quality are successful: choosing a high-quality hospital compared to a low-quality hospital results in 10% to 15% better outcomes.

## I. Introduction

THERE is considerable interest in improving the quality and efficiency of health care in the United States. This interest is motivated in part by influential research demonstrating widespread geographic variation in treatment intensity that yields little apparent benefit in terms of patient health outcomes (Fisher, Bynum, & Skinner, 2009; Fisher et al., 2003a, 2003b; Chandra & Skinner, 2011). At the same time, a parallel body of research has documented consistent gaps between the quality of care patients receive and what the medical system could provide if it were productively efficient and operating at its full potential (Chandra & Staiger, 2007; McGlynn et al., 2003).

The contention that the U.S. health care system simultaneously provides too much low-value care and too little high-quality care lies at the heart of many delivery system reform initiatives. A central focus of such initiatives is the creation of more direct linkages between provider reimbursement and measures of quality. The Medicare Hospital Readmission Reduction Program (HRRP), for example, penalizes hospitals with above-average thirty-day readmission rates for certain conditions (Desai et al., 2016; Berenson, Paulus, & Kalman, 2012). Another example is the Hospital Value-Based Purchasing Program (HVBP), which explicitly ties financial incentives to hospital quality performance (Norton et al., 2016; Das et al., 2016). Indeed, the Centers for Medicare and Medicaid Services (CMS) plan to have value-based purchasing become the dominant form of payment (Muhlestein, Saunders, & McClellan, 2017).

Measures of hospital quality remain highly controversial, however, despite their increasingly widespread use (Lilford & Pronovost, 2010; Austin et al., 2015; Gestel et al., 2012; Shahian et al., 2012). A primary concern is the potential inadequacy of risk adjustment to control for patient selection (also known as referral bias). Patients who are in the poorest health may be referred to the hospitals that are of the highest quality, potentially biasing performance assessments and comparisons across hospitals.

Efforts to address this concern through risk adjustment face three challenges. The first is the weak explanatory power of observable variables for health outcomes. The second is that missing or unrecorded data may also be correlated with underlying quality, which could compromise efforts to profile hospital performance and make quality comparisons across hospitals (Ash et al., 2012; Shahian & Normand, 2008). Third, risk adjustment is often made using diagnoses recorded in billing claims. Critically, these recorded diagnoses may capture both underlying patient health as well as the endogenous influence of reimbursement system incentives on coding practices (Finkelstein et al., 2017; Song et al., 2010). All of these concerns place a premium on adjudicating common measures of quality based on comparisons of outcomes among patients exogenously assigned to hospitals. This problem of evaluating and rewarding quality is common, especially in sectors with significant public sector involvement. For example, this problem is akin to the education literature that seeks to substantiate whether controversial quality measures predict better outcomes (Chetty, Friedman, & Rockoff, 2014).

In this paper, we develop an instrumental variables (IV) framework based on earlier work that aims to purge patient selection to different hospitals (Doyle et al., 2015). We do so by leveraging ambulance company referral patterns as an instrument for hospital assignment. Ambulance companies are plausibly exogenously assigned to emergency patients based on availability at the time of the emergency. In most cases, ambulances take patients to a nearby hospital, but there is often a choice made among the set of nearby hospitals. As we have shown in previous work, this means that in some areas, otherwise identical patients can end up in hospitals with very different characteristics depending on which ambulance company is called to transport them (Doyle et al., 2015).

We use this ambulance referral framework to test whether patients treated at hospitals that score well on widely used quality measures achieve better outcomes for patients whose hospital assignment is plausibly exogenous. Our approach provides a compelling lens through which we can evaluate hospital performance measures, at least for emergency care subject to the type of variation we can use to control for patient selection (the emergency conditions we study amount to approximately one-quarter of inpatient care for Medicare

Received for publication January 20, 2017. Revision accepted for publication April 23, 2018. Editor: Amitabh Chandra.

*Doyle: MIT and NBER; Graves: Vanderbilt University; Gruber: MIT and NBER.

We thank Mauricio Caceres for excellent research support and gratefully acknowledge support from the National Institutes of Health R01 AG041794-01 and R01 AG041794-04.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00804.

*The Review of Economics and Statistics*, December 2019, 101(5): 841–852

https://doi.org/10.1162/rest_a_00804

patients nationwide). In contemporaneous work, Hull (2018) develops a model with a more ambitious goal of using this IV framework to not just demonstrate that the widely used CMS mortality measure has a causal relationship with patient outcomes, but to build a better-quality measure. We discuss this work at more length below.

Our primary analyses consider four composite measures constructed to capture different and dimensions of quality: (a) process measures quantifying the frequency with which hospitals provide services that are considered effective in improving patient outcomes; (b) risk-adjusted patient satisfaction scores from patient surveys, which are increasingly used by insurers to "pay for quality"; (c) risk-standardized thirty-day readmission rates among all discharged patients; and (d) risk-standardized thirty-day mortality rates among all admitted patients. For each domain, we estimate how assignment to hospitals that score well on these various measures affects both patient readmission and mortality outcomes.

Using our IV strategy, we find that each of these measures used by the CMS is related to patient outcomes in important ways. First, hospitals with higher process measures of quality have lower long-term mortality for marginal patients. Second, hospitals with lower patient satisfaction scores have a higher likelihood of readmission and death. Third, we find a strong and significant positive effect of hospital readmission rates on the likelihood of subsequent readmission, and an even stronger positive effect of hospital mortality rates on the likelihood of subsequent mortality. Our findings suggest that even correcting for patient selection, the outcome measures used in value-based payment reform efforts by CMS and other payers are useful proxies for hospital quality.

The remainder of this paper proceeds as follows. The next section provides relevant institutional background on hospital quality reporting in the United States, as well as a review of the literature on the relationship between quality report cards and patient outcomes. We then motivate our identification strategy and lay out the key structural equations we seek to estimate. Following that, we discuss the data sources and quality measures we constructed and used. We then present the results and conclude with a discussion of the implications of our findings for hospital reimbursement policy.

## II. Background: Approaches to Measuring Hospital Quality

The measures of hospital quality we consider are defined across three primary dimensions: process measures of timely and effective care, self-reported patient experience of care measures, and risk-standardized rates of patient outcomes. These measures are correlated with hospital characteristics. For example, teaching hospitals and larger hospitals (by patient volume) tend to achieve higher scores (Doyle, Graves, & Gruber, 2017). We discuss the relevant details of quality measurement below.

### A. Process Measures

Process quality measures quantify the rate at which hospitals provide timely and effective care. In this context, effective care constitutes activities with sufficient clinical evidence linking that care to improved patient outcomes. The percentage of acute myocardial infarction (AMI) patients administered aspirin upon arrival, for example, has long been used to assess whether hospitals regularly incorporate high-value, evidence-based care.

The number of process measures used in hospital report cards has grown considerably in recent years. These measures are a key component of the CMS and National Quality Forum's Hospital Compare program, the Leapfrog Group, and U.S. News and World Report's annual hospital rankings. The number of process measures reported on Hospital Compare, for example, rose from twenty in 2005 (the first year of public reporting) to over forty by 2014.

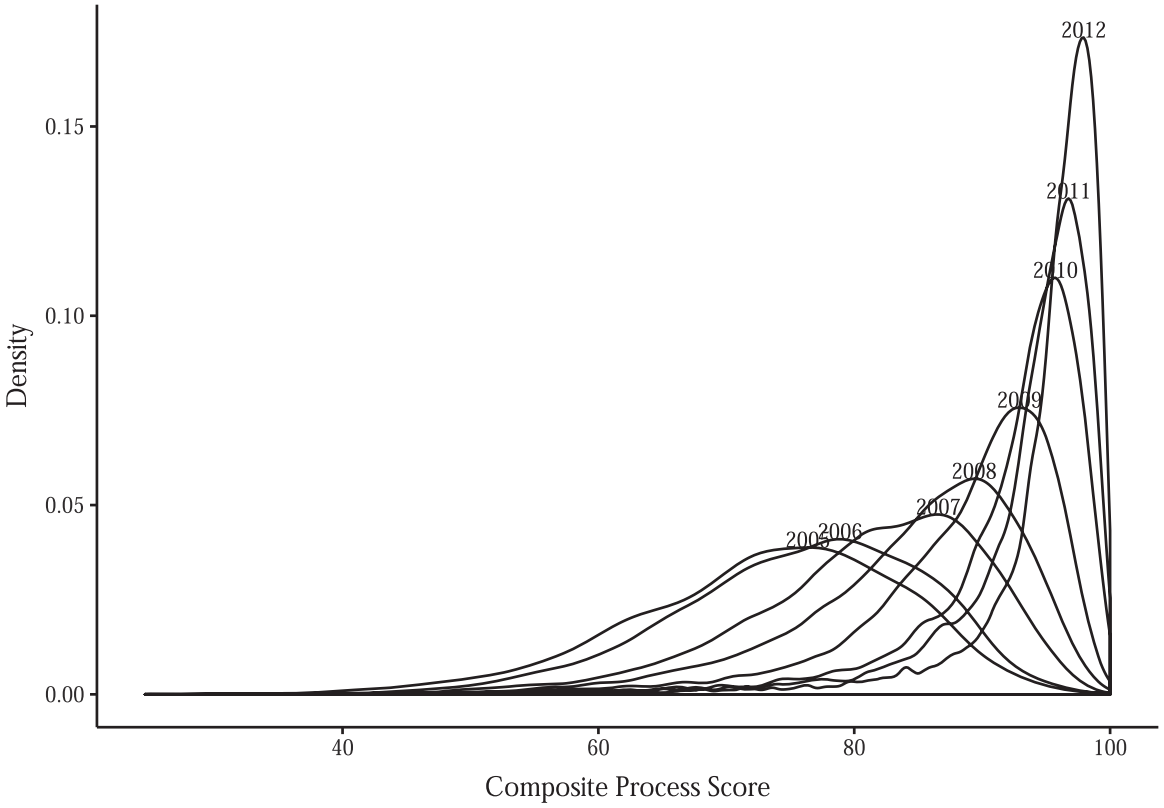
While the number of process quality measures has increased over time, so too has observed hospital performance. For example, figure 1 plots the distribution in each year of our composite process quality measure (described below) between 2005 and 2012. This figure summarizes, for a consistent set of 3,027 hospitals, how their performance on a fixed set of seven measures evolved over an eight-year period (the specific process measures are listed in table A1 in the online supplement). As can be seen in the figure, in the early years of public reporting, there was wide variation in performance. These hospitals collectively performed better over time, as evidenced by the fact that the distribution of scores compresses and shifts toward 1 (the highest possible score). By 2015, the average score was 97.8 (out of a maximum of 100), compared to a mean of 73.2 among the same hospitals in 2005. The amount of variance among hospitals similarly declined, with the standard deviation declining from 10.4 in 2005 to 2.4 in 2015. Figure 1 makes clear that some measures of quality will naturally become less relevant when little variation remains after nearly all hospitals achieve high scores, although laggards may reveal themselves to have particularly low quality.

### B. Patient Experience Measures

Measures of patient experience are captured by the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, which is administered to a sample of patients between 48 hours and 6 weeks after discharge. The survey covers multiple aspects of patient experience, ranging from the cleanliness of facilities and the effectiveness of pain management to how well physicians, nurses, and other hospital staff communicated with the patient.

Hospital-specific scores on eleven patient experience domains, including an overall summary score, are adjusted for the mode of the survey (e.g., phone only or in person) and are risk-adjusted for patient characteristics, including age, education, self-reported health, source of admission, primary

FIGURE 1.—DENSITY DISTRIBUTION OF COMPOSITE PROCESS SCORE BY YEAR




language, and hospital service line used (e.g., surgical versus medical). These risk-adjusted scores are reported individually on the Hospital Compare website; more recently, a "Five Star" ranking is constructed based on a composite average of hospital HCAHPS performance.

### C. *Outcome Measures*

Outcome quality measures compare the observed number of patients who experience a given outcome (e.g., mortality or readmission within thirty days) to the number expected to experience the outcome based on a national risk model (Ash et al., 2012). That is, these measures ask, Is the case-mix-adjusted number of patients who experience the outcome in a given hospital consistent with what would be expected in a hypothetical hospital with the same patient case mix and with average quality?

This "indirect standardization" approach is used to construct hospital report cards by CMS and other organizations, such as U.S. News and World Report. Typically, the basis for these measures is a logistic regression model that includes patient-level measures of clinical acuity (e.g., past diagnoses and comorbidities as recorded in billing claims), demographics (e.g., age and gender), and a hospital-specific random effect that is assumed to be drawn from a known (normal) probability distribution. Some important patient-level attributes (e.g., race, ethnicity, and socioeconomic status) are deliberately excluded from risk adjustment so that the risk-standardized measures do not condition out important racial or socioeconomic disparities in care across facilities.[1]

### D. *Quality Measurement Concerns*

As noted in section I, a common issue with outcome measurement is a concern that a selection-on-observables assumption inherent in risk adjustment is highly controversial. That is, the random-effects model includes patient-level risk adjusters ($\mathbf{x_{ih}}$) but does not fully control for patient health due to the relatively limited number of patient characteristics available in billing data; the deliberate exclusion from the model (on substantive grounds) of certain important confounders like race and socioeconomic status; and growing evidence on the endogeneity of patient-level diagnoses as recorded in hospital billing codes (Finkelstein et al., 2017; Song et al., 2010). In addition, the model does not include hospital-level attributes (e.g., patient volume, teaching status) that may be independent predictors of patient outcomes (Birkmeyer et al., 2002; Daley, 2002; Dudley et al., 2000; Halm, Lee, & Chassin, 2002; Hughes, Hunt, & Luft, 1987; Luft, Hunt, & Maerki, 1987; Shahian & Normand, 2003).

[1] Sensitivity analyses based on comparing outcome rates to hospitals that treat large numbers of Medicaid patients and African American patients yield a similar range of performance relative to other hospitals. See, for example, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/MedicareHospitalQualityChartbook2012.pdf (pp. 23–36).

A more subtle concern with indirect standardization is that an individual hospital's observed performance is compared to the predicted performance for a hypothetical hospital with average quality and with the same patient case mix. Thus, performance comparisons between two hospitals are complicated by the fact that hospitals may not treat the same profile of patients (i.e., they may have a different case mix; Shahian & Normand, 2008). This issue is a manifestation of Simpson's paradox or the observation that a relationship or association observed within individual subgroups can be reversed when data from those subgroups are aggregated.[2]

While interhospital comparisons are frequently made and facilitated by the CMS Hospital Compare website (Ash et al., 2012), an open question is whether risk-adjusted outcomes measures can also be used to guide patient choice of the "best" hospital among local options. Our framework can assess whether high-performing hospitals achieve better outcomes for the marginal patient who is effectively randomized to a local hospital.

## III. Empirical Strategy

### A. *Ambulance Referral Patterns*

Our empirical approach builds on our earlier work that relies on plausibly exogenous sources of hospital assignment determined by ambulance company preferences for certain hospitals (Doyle et al., 2015). The key ingredient is the recognition that the locus of treatment for emergency hospitalizations is, to a large extent, determined by prehospital factors, including ambulance transport decisions and patient location. Critically, areas are often served by multiple ambulance companies, and the ambulance company assignment is effectively random.

Rotational assignment of competing ambulances services—as well as direct competition between simultaneously dispatched competitors—is increasingly common in the United States. In some communities, the opportunity for ambulance transport is broadcast to multiple companies, and whichever arrives there first gets the business. Similarly, in most cities, private ambulance companies work in conjunction with fire departments to provide emergency medical services, EMS; Chiang, David, & Housman, 2006; Johnson, 2001; Ragone, 2012).

We are aware of no systematic evidence on the basis for rotational assignment of ambulances. To understand the dispatch process, in Doyle et al. (2015), we conducted a survey of thirty cities with more than one ambulance company serving the area in our Medicare data. The survey revealed that patients can be transported by different companies for two main reasons. First, in communities served by multiple ambulance services, 911 systems often use software that assigns units based on a rotational dispatch mechanism; alternatively, they may position ambulances throughout an area and dispatch whichever ambulance is closest, then reshuffle the other available units to respond to the next call. Second, in areas with a single ambulance company, neighboring companies provide service when the principal ambulance units are busy under so-called mutual aid agreements. Within a small area, then, the variation in the ambulance dispatched is either due to rotational assignment or one of the ambulance companies being engaged on another 911 call. Both sources appear plausibly exogenous with respect to the underlying health of a given patient.

Previous case studies suggest that these ambulances have preferences about which hospital to choose. For example, Skura (2001) studied ambulance assignment in the wake of a new system of competition between public and private ambulances in New York City. He found that patients living in the same postal code as public Health and Hospital Corporation (HHC) hospitals were less than half as likely to be taken there when assigned a private, nonprofit ambulance (29%) compared to when the dispatch system assigned them to an FDNY ambulance (64%). In most cases, the private ambulances were operated by nonprofit hospitals and stationed near or even within those facilities, so they tended to take their patients to their affiliated hospitals.

To operationalize ambulance preferences, we calculate a set of instrumental variables based on the characteristics of hospitals where each ambulance company takes other patients—a leave-out mean approach that helps avoid weak instrument concerns similar to jackknife instrumental variable estimators (Stock, Wright, & Yogo, 2012). For patient $i$ assigned to ambulance $a(i)$, we calculate the average hospital measure (e.g., the readmission rate) among the patients in our analysis sample for each ambulance company:

$$Z_{a(i)} = \frac{1}{N_{a(i)-1} - 1} \sum_{j \neq i}^{N_{a(i)-1}} H_j. \tag{1}$$

This measure is essentially the ambulance company fixed effect in a model for $H_j$ in a model that leaves out patient $i$. Below, we consider values for $H_j$ that include a variety of quality measures, such as the hospital's publicly reported thirty-day readmission rate, its thirty-day mortality rate, or a composite process measure.

[2]For example, suppose there are two types of patients, healthy and sick, and that two hospitals (A and B) treat both types of patients at the same level of quality, with 5% of healthy patients readmitted and 30% of sick patients readmitted. Suppose further that the national risk adjustment model is unbiased and that, on average, healthy patients are expected to be readmitted 5% of the time, while sick patients are expected to be readmitted 20% of the time. In other words, both A and B treat healthy patients as expected but are of equally poor quality when treating sick patients. In this scenario, if hospitals A and B had the same patient case mix (e.g., equal proportions of healthy and sick patients), they would be profiled with identical risk-standardized readmission rates. But if their case mix were different (e.g., if hospital A treated predominantly healthy patients and hospital B treated predominantly sick patients), then hospital B would receive a higher risk-standardized readmission rate simply because it treats more sick patients. Moreover, it would be inappropriate to compare A versus B and conclude that hospital B was of poorer quality. Had patients been randomly assigned to A versus B, we would find no evidence of a difference in patient-level readmission outcomes between the two.

### B. Empirical Model

We use this instrument to estimate the first-stage relationship between hospital quality $H$ and the instrument, $Z$: the hospital measure associated with the ambulance assigned to patient $i$ with principal diagnosis $d(i)$ transported from an origin in postal code $z(i)$ in year $t(i)$:

$$H_i = \alpha_0 + \alpha_1 Z_{a(i)} + \alpha_2 X_i + \alpha_3 A_i + \gamma_{d(i)} + \theta_{z(i)} + \lambda_{t(i)} + \nu_i, \tag{2}$$

where $X_i$ is a vector of patient controls including age, race, and sex, and indicators for seventeen common comorbidities controlled for in the CMS quality scores; $A_i$ represents a vector of ambulance characteristics that summarize the level and scope of treatment provided in the ambulance; indicators for distance traveled in miles; whether the transport utilized advanced life support (e.g., paramedic) capabilities; whether the transport was coded as emergency transport; and whether the ambulance was paid through the outpatient system rather than the carrier system. These ambulance controls are both novel (typically prehospital care is not used as controls) and, in this setting, particularly important given the estimation strategy. In particular, a key concern is that ambulance company assignment may affect patient outcomes directly, not just through its effect on hospital choice. Our analysis directly explores this concern by investigating these observable ambulance characteristics. We cluster standard errors at the hospital service area (HSA) level, as each local market may have its own assignment rules.[3]

We also include a full set of controls for principal diagnosis, year, and postal code $\times$ patient origin fixed effects.[4] This regression, in other words, compares individuals who are transported from similar origins (e.g., at home, in a nursing home, or at the scene of an accident or illness) and who reside in the same postal code but are picked up by ambulance companies with different "preferences" across hospitals with different quality scores. A positive coefficient $\alpha_1$ would indicate that ambulance company preferences are correlated with where the patient actually is admitted.

Our main regression of interest is the relationship between hospital quality on outcomes such as mortality, $M$, for patient $i$:

$$M_i = \beta_0 + \beta_1 H_i + \beta_2 X_i + \beta_3 A_i + \gamma_{d(i)} + \theta_{z(i)} + \lambda_{t(i)} + \epsilon_i. \tag{3}$$

For this regression, we consider various patient outcomes, such as whether they are readmitted to an acute care hospital facility within thirty days of discharge or whether they died within thirty days or one year of admission. Finally, since patient selection is likely to confound this structural model, we estimate equation (3) using two-stage least squares, with the instrument defined as above.

Doyle et al. (2015) discuss at length potential limitations with this strategy and various specification checks that begin to address them. In particular, their study finds that the results are highly robust to controls for both patient characteristics and the characteristics of prehospital care in the ambulance; that selection into the inpatient data from the full ambulance transport sample is not correlated with observable ambulance company characteristics; that the impact of ambulance assignment on health outcomes occurs not on the first day but over longer horizons, which suggests that different (unobserved) levels of care in the ambulance are not driving outcome differences; and that the results do not stem solely from postal codes that are heterogeneous in resident income levels, suggesting that the results are not driven by ambulance companies that specialize in particular subneighborhoods within a postal code.

In contemporaneous work, Hull (2018) addresses a similar question and also builds on the ambulance-instrument approach. That work aims to estimate risk-adjusted mortality measures for each hospital using the quasi-experimental variation in hospital choice.[5] Given the difficulty of precisely estimating these quality measures for thousands of hospitals using thousands of ambulance instruments, Hull develops a valuable, new, semiparametric methodology. In particular, the approach estimates the relationship between the (noisy) quality measures estimated using the quasi-experimental variation and the more precisely estimated but likely biased estimates that come from more traditional random effects models. The resulting shrinkage estimator results in posterior quality estimates for each hospital. This approach relies on distributional assumptions about the latent quality of each hospital and the adequacy of the model that maps the relationship between the two quality estimates, including its application to hospitals where the quasi-experimental variation is not available.

Hull's work is complementary to our paper. If the goal is to test whether existing CMS quality measures causally measure hospital quality, then our approach is a minimally structural approach to doing so. As already emphasized, this is akin to important work in education evaluating teacher quality measures (Chetty et al., 2014). To overcome the precision problems, our approach estimates a standard linear approximation of the relationship between quality measures and patient outcomes, resulting in relatively precise IV estimates. If the goal is to improve on the CMS measure for each hospital using the quasi-experimental variation, then more structure must be imposed to deal with the imprecision inherent in that exercise. Hull (2018) demonstrates how this can be implemented.

[3] As a robustness check, we also cluster standard errors at the ambulance company level, which yields standard errors that are roughly 20% smaller than our preferred HSA-based clustering strategy.

[4] The principal diagnosis is the three-digit ICD-9-CM diagnosis code, as shown in appendix table A2.

[5] The empirical strategy differs slightly in its implementation from the current one. For example, it incorporates the patient's distance from the headquarters of ambulance companies to smooth the first-stage relationship.

## IV. Data

### A. Medicare Claims Data

Our primary sources of patient-level data are Medicare claims between 2008 and 2012, the time period where we observe the CMS quality measures under investigation. We use these data to identify an uncensored sample of patients admitted to an acute care hospital after being transported by ambulance to the emergency department.

CMS reimburses ambulance companies using two systems captured by the carrier file and the outpatient claims file. We can access carrier and outpatient claims for a 20% random sample of beneficiaries. Most ambulance claims are paid via the carrier claims, and we increase our sample by 6% by including the outpatient claims—claims that are affiliated with a hospital or other facility file. We link each ambulance patient's claims to her inpatient claims in the Medicare Provider Analysis and Review (MEDPAR) files, which record pertinent information on date of admission, primary and secondary diagnoses, and procedures performed. Diagnoses and procedures recorded in each patient's claims for the year prior to (but not including) the ambulance admission are then mapped to hierarchical condition codes (HCC) to construct a set of comorbidity measures. We also link each ambulance patient to a Medicare denominator file that contains other information on age, race, and gender. Finally, the claims data also include the postal code of the beneficiary, where official correspondence is sent; in principle, this could differ from the patient's home postal code. In addition, vital statistics data that record when a patient dies are linked to these claims, which also allow us to measure mortality at different time frames, such as thirty days or one year.

### B. Sample Construction

We rely on two primary analytic samples. The primary sample consists of patients admitted to the hospital after an ambulance transport to the emergency room with 29 "nondiscretionary" conditions. According to Doyle et al. (2015), these are conditions where selection into the health care system is largely unavoidable (i.e., femur fracture, poisoning, and stroke).[6]

An advantage of this sample is that it provides the broadest possible sample to which our key instrument is well matched. The disadvantage is that it extends far beyond the three conditions that are embedded in the CMS quality measures we examine, which are measured for patients with diagnoses of acute myocardial infarction (AMI), pneumonia (PN), or heart failure (HF). We therefore also extend our results to a second sample: ambulance-transported patients admitted via the emergency department for these three conditions.

For this three-condition sample, we include all patients who had not been admitted for any of these within the previous 365 days. We also exclude patients who are 100 miles or more from their residential postal code to focus on emergency patients who are close to home at the time of their episode. Finally, our sample exclusion criteria removed patients treated at hospitals with fewer than thirty episodes (in the 20% Medicare files), as well as patients whose ambulance company transported thirty or fewer patients over the study period. These criteria resulted in 546,700 patient episodes for the primary sample and 171,246 patients for the three-condition sample. In addition, for regressions that consider one-year mortality outcomes, we use the subsample of 451,503 nondiscretionary patients and 142,424 CMS condition patients with uncensored one-year outcomes (those treated between 2008 and 2011).

Appendix table A2 shows the distribution of admissions across these diagnostic categories for the primary sample. These conditions represent 39% of the hospital admissions via the emergency room, 61% of whom arrived by ambulance. Given that roughly 60% of all Medicare admissions originate in the emergency room, these conditions constitute approximately 23% of all hospital admissions for Medicare patients in the United States, and when we restrict our work to those transported by ambulance, the sample represents 14% of inpatient admissions. Moreover, these conditions are particularly expensive—for example, the most costly inpatient condition in the United States.

The reliance on ambulance transports allows us to focus on patients who are less likely to decide whether to go to the hospital. This sample is slightly older and has a higher 365-day mortality rate (37%) compared to all Medicare patients who enter the hospital via the emergency room (20%). These are relatively severe health shocks, and the estimates of the effects of hospital types on mortality apply to these types of episodes, so the applicability of our results to less emergent hospitalizations may be limited. We discuss this limitation further in section VI.

### C. Hospital Quality Measures

Our primary source for hospital quality measures is archived Hospital Compare data from CMS. Hospital Compare began publicly reporting process measures in 2005; thirty-day mortality measures and patient satisfaction scores were added in 2008 and the thirty-day readmission measures were added in 2009.

Reported process measures generally have a one-year time lag, while HCAHPS scores are typically reported after a one- to two-year lag. Risk-standardized readmission and mortality outcome rates are based on claims from a pooled three-year sample of fee-for-service Medicare and Veterans Health Administration patients, with a one-year lag between the most

[6] Discretionary admissions see a marked decline on the weekend, but particularly serious emergencies do not. Following Dobkin (2003) and Card, Dobkin, and Maestas (2007), diagnoses whose weekend admission rates are closest to 2/7ths reflect a lack of discretion as to the timing of the hospital admission. Using our Medicare sample, we chose a cutoff of all conditions with a weekend admission rate that was as close or closer to 2/7ths as hip fracture, a condition commonly thought to require immediate care.

recent claims year used and the public reporting date. Thus, public reporting of hospital quality does not capture concurrent quality, but rather the quality of care received by patients treated within (approximately) four years prior to the reporting date.

For our main analyses, we maintain these temporal lags and assign each hospital-year its three-year average for each of the composite quality measures. This choice is guided by the observation that a patient choosing a hospital on a given date would do so based on publicly reported measures, which would reflect the quality of care rendered to patients in prior years. The use of a lagged measure also ensures that our quality measures are "leave out" with respect to the patient associate with that regression observation. That is, when we consider the impact of standardized hospital mortality rates on the likelihood that patient X dies, we are not including patient X in the calculation. In robustness analyses, we also consider measures of quality that run concurrent to the patient treatment date (e.g., the hospital quality measure for a patient treated in 2008 will reflect the quality of care provided to patients in that hospital in 2008).

Our composite process score is based on the pooled average of seven individual measures of timely and effective care for acute myocardial infarction (AMI), pneumonia (PN), and heart failure (HF) patients (see appendix table A1; Yasaitis et al., 2009).[7] Similarly, the thirty-day mortality and readmission composite measures are based on averaging the mortality and readmission rates for AMI, PN, and HF for each hospital. A full listing of all CMS quality measures that go into our composite scores is provided in appendix table A1.

For all of our regressions, the quality measures enter as a continuous measure that has been demeaned and standardized by 2 standard deviations to facilitate interpretation and comparison across measures. Thus, each has an overall mean of 0 and a standard deviation of 0.5. This standardization procedure is designed so that the coefficients can be interpreted as if they were estimated on a binary low versus high "quality" measure. For context, the main results provide both the unstandardized mean and standard deviation for each composite measure.

### *D. Outcome Measures for Assessing Validity of Quality Measures*

In order to assess the validity of quality measures, we want to measure their impact on welfare-relevant outcomes. In this paper, we consider two outcomes that are commonly used, including on the Hospital Compare website. The first is the rate of hospital readmissions over the thirty days after the initial admission. This is a key outcome since hospital readmissions are often viewed as a signal of inefficiencies in the delivery of hospital care. The thirty-day patient readmission outcomes we consider are defined as unplanned readmission to any hospital within thirty days of live discharge from the indexing visit.

The second outcome is mortality. While CMS and other payers do not yet directly reimburse providers based on patient mortality rates alone, public reporting on mortality is a common feature of hospital quality report cards. These measures typically use mortality within thirty days of hospital admissions. A disadvantage of this approach, however, is that hospital actions can manipulate this measure of mortality over shorter time horizons. In particular, Maxwell et al. (2014) show a discontinuity in patient mortality at thirty days for cardiac surgery, suggesting the possibility of hospital manipulation of mortality at shorter horizons. For this reason, we show mortality results at thirty days but also focus on a longer horizon, assessing the impact of quality on mortality over the first year postadmission.

## V. Do CMS Quality Measures Actually Measure Quality?

### *A. Balance*

To evaluate the relationship between measured hospital performance and patient outcomes, we rely on an IV approach that assumes patients are quasi-randomly assigned to ambulances in an emergency. To test whether this is plausible along observable dimensions, we regressed each control variable on our instrument for the hospital's thirty-day mortality rate, a measure that we emphasize in our results. The models use the same specification as the first stage shown above.

Table 1 reports the coefficients, which should be regarded as an increase of 2 standard deviations in the instrument.[8] The table shows that our sample is remarkably balanced on observable demographic and health characteristics. The age, race, sex, and comorbidities of patients are nearly identical across patients assigned to ambulances that transport other patients to low- versus high-quality hospitals. Among the ambulance controls, they show similar emergency transport rates (using their lights and sirens), as well as advanced life support capability. Miles traveled is one element that shows some meaningful differences, a difference of 0.4 to 1 mile compared to a mean of 7 miles, although the sign is unstable, which supports the idea that the measure is not systematically related to hospital quality. The primary diagnoses are also balanced across the instrument values as shown in the appendix. Similar results are found across quartiles of an inpatient reimbursement-based instrument, as documented in Doyle et al. (2015).

[7] These measures are simply the arithmetic mean of the individual outcome scores. In earlier versions of this work, we also constructed a weighted composite based on a generalized least squares estimator to maximize the amount of information contributed by each measure (Anderson, 2012). We found that the generalized least squares method yielded nearly identical results as the simple arithmetic mean, so have opted for the more straightforward (mean) measure here.

[8] A means comparison is available in the working paper version of this paper: Doyle et al. (2017).

TABLE 1.—BALANCE OF PATIENT CHARACTERISTICS BY QUALITY INSTRUMENT

| Measure | Timely and Effective Care | Patient Satisfaction | 30-Day Mortality Rate | 30-Day Readmission Rate |
|---|---|---|---|---|
| Age 70–74 | −0.0011 | 0.0051 | −0.0025 | 0.0050 |
| Age 75–79 | −0.0047 | −0.0075 | −0.0091* | 0.0010 |
| Age 80–84 | 0.0094 | 0.0029 | 0.0015 | −0.0080 |
| Age 85–89 | −0.0078 | −0.0016 | 0.012** | −0.0039 |
| Age 90–94 | −0.0028 | 0.00072 | −0.00040 | 0.0019 |
| Age 95+ | 0.0025 | −0.0013 | −0.00031 | 0.0027 |
| Gender: Male | 0.0056 | −0.0043 | 0.0076 | 0.0095 |
| Race: Black | −0.0013 | −0.012*** | −0.0027 | 0.025*** |
| Race: Other | −0.0062** | −0.015*** | −0.0065** | 0.016*** |
| Ambulance: Miles traveled with patient | 0.31*** | 0.94*** | 0.72*** | −0.90*** |
| Ambulance: Advanced life support | 0.070*** | 0.082*** | 0.042*** | −0.080*** |
| Ambulance: Emergency traffic | 0.0024 | 0.015*** | 0.0066** | 0.0076*** |
| Ambulance: Payment | 14.00*** | −2.0 | 6.0*** | −5.7*** |
| Comorbidity: Hypertension | 0.0062 | 0.0056 | −0.0017 | −0.0046 |
| Comorbidity: Stroke | 0.0013 | −0.000065 | 0.0018 | −0.00096 |
| Comorbidity: Cerebrovascular disease | 0.00047 | −0.00034 | −0.0032 | 0.0025 |
| Comorbidity: Renal failure disease | −0.0061 | −0.0035 | −0.0059 | −0.00098 |
| Comorbidity: Dialysis | −0.00039 | 0.00067 | 0.00093 | −0.0010 |
| Comorbidity: COPD | −0.0046 | −0.011** | 0.0038 | 0.0047 |
| Comorbidity: Pneumonia | 0.00027 | 0.00024 | −0.00029 | −0.0011 |
| Comorbidity: Diabetes | 0.0048 | −0.0035 | 0.00057 | 0.0048 |
| Comorbidity: Protein calorie malnutrition | 0.0028 | 0.0000069 | −0.0053** | −0.00075 |
| Comorbidity: Dementia | −0.013*** | −0.013*** | 0.000034 | 0.0072* |
| Comorbidity: Paralysis | −0.00054 | −0.0026 | −0.0034 | −0.0017 |
| Comorbidity: Peripheral vascular disease | 0.0051 | 0.0046 | 0.00060 | −0.0011 |
| Comorbidity: Metastatic cancer | 0.00070 | 0.0018 | −0.0033* | 0.0029 |
| Comorbidity: Trauma | 0.0061* | 0.0042 | −0.0068** | 0.0031 |
| Comorbidity: Substance abuse | −0.00096 | 0.0024 | 0.0031 | −0.0040 |
| Comorbidity: Major psychological disorder | −0.00096 | −0.00095 | −0.00026 | 0.0064** |
| Comorbidity: Chronic liver disease | 0.0023* | 0.0040*** | −0.0012 | −0.00062 |

The table shows balance across controls used in regressions. Balance is assessed using pairwise regressions of each characteristic on each instrument based on the specification in equation (5). The reported estimates (rows) show the coefficient on the instrument from each of these regressions fit separately for each instrument (columns). Sample size = 546,700. Significant at ***0.001, **0.01, and *0.05.

TABLE 2.—OLS RESULTS

| Outcome | | 30-Day Readmission | 30-Day Mortality | 365-Day Mortality |
|---|---|---|---|---|
| Timely and effective care composite | 93.40 [5.01] | −0.0068 (0.001) | −0.0035 (0.001) | −0.0042 (0.002) |
| Patient experience composite | 66.89 [5.02] | −0.011 (0.002) | −0.0046 (0.002) | −0.0064 (0.002) |
| 30-day mortality rate composite | 12.69 [1.29] | −0.0040 (0.002) | 0.011 (0.002) | 0.014 (0.002) |
| 30-day readmission rate composite | 20.98 [1.48] | 0.014 (0.002) | −0.0023 (0.002) | 0.006 (0.003) |

Each cell reports ordinary least squares (OLS) coefficient estimates for a separate regression. Quality measures have been demeaned and standardized by 2 standard deviations so they can be interpreted like binary (low-to-high) indicators. The underlying mean and standard deviation of each quality measure are provided in the first column to facilitate interpretation on the original scale. Outcome means: 30-day readmission = 15.0%, 30-day mortality = 17.0%, 365-day mortality = 37.2%. Sample sizes: 546,700 (30-day outcomes), 451,503 (1-year mortality). All models include patient demographic and ambulance controls as listed in table 1, as well as the diagnosis controls as listed in table A2. Models also include postal code–patient origin fixed effects. Standard errors, clustered at Health Service Area (HSA) level, are reported in parentheses.

### *B. Quality Measures and Patient Outcomes: OLS Conditional Correlations*

We begin, in table 2, by showing the results for OLS estimates of the CMS quality measures. The OLS results show correlations between the composite quality measures and the outcomes of interest, focusing separately on thirty-day readmissions, thirty-day mortality, and one-year mortality. The means of the key dependent variables, as reported in the table footnote, are 15.0% for thirty-day readmission, 17.0% for thirty-day mortality, and 37.2% for one-year mortality. The first column shows the (raw) means and standard deviation of each composite quality measure. Each cell represents a separate regression; for example, the first row of the second column shows the OLS relationship between the quality process score and thirty-day readmission.

We find that the CMS process measure of quality is only modestly correlated with readmission rates and mortality. For example, regarding one-year mortality rates, a 2 standard deviation increase in the process score (i.e., a change of 10.2 points) was associated with a 0.4 percentage point decline in one-year mortality, which is about 1% of the mean and is not statistically significant.

The next row of table 2 shows the findings for composite patient experience score. Patients treated in high-performing hospitals have a marginally lower (and statistically significant) likelihood of readmission and mortality. In particular, a 10 point increase in the patient experience score (which averages 67), is associated with a 1.1 percentage point reduction in such readmissions. The same change in patient experience index is found to result in a 0.5 percentage point reduction in thirty-day mortality and a 0.6 percentage point reduction in one-year mortality, with standard errors close to 0.2 percentage points.

The next two rows focus on composite performance scores based on 30-day mortality and readmission rates. There is no mechanical relationship between these measures and the associated outcomes, as they are leave-out means of the relevant

TABLE 3.—FIRST-STAGE RESULTS

| Quality Measure Instrument | No Comorbidity Controls (1) | With Comorbidity Controls (2) |
|---|---|---|
| Ambulance average: Timely and effective care composite | 0.623 (0.005) | 0.622 (0.005) |
| Ambulance average: Patient experience composite | 0.604 (0.004) | 0.603 (0.004) |
| Ambulance average: 30-day mortality rate composite | 0.549 (0.003) | 0.549 (0.003) |
| Ambulance average: 30-day readmission rate composite | 0.576 (0.003) | 0.576 (0.003) |

Each cell reflects a separate first-stage regression of the ambulance instrument on the quality measure. Quality measures have been demeaned and standardized by 2 standard deviations so they can be interpreted like binary (low-to-high) indicators. The underlying mean and standard deviation of each quality measure are provided in the first column to facilitate interpretation on the original scale. Outcome means: 30-day readmission = 15.0%, 30-day mortality = 17.0%, 365-day mortality = 37.2%. Sample sizes: 546,700 (30-day outcomes), 451,503 (1-year Mortality). All models include patient demographic and ambulance controls as listed in table 1, as well as the diagnosis controls as listed in table A2. Models also include postal code–patient origin fixed effects. Comorbidity controls are listed in table 1. Standard errors, clustered at health service area (HSA) level, are reported in parentheses.

TABLE 4.—2SLS RESULTS

| Outcome Quality Measure | Mean [SD] | 30-Day Readmission (1) | 30-Day Mortality (2) | 365-Day Mortality (3) |
|---|---|---|---|---|
| Timely and effective care composite | 93.40 [5.01] | −0.0035 (0.007) | −0.014 (0.008) | −0.038 (0.014) |
| Patient experience composite | 66.89 [5.02] | −0.021 (0.007) | −0.0057 (0.008) | −0.028 (0.011) |
| 30-day mortality rate composite | 12.69 [1.29] | −0.012 (0.007) | 0.021 (0.008) | 0.025 (0.010) |
| 30-day readmission rate composite | 20.98 [1.48] | 0.027 (0.007) | −0.0078 (0.007) | 0.005 (0.010) |

Each cell reports two-stage least squares (2SLS) coefficient estimates for a separate regression. Quality measures have been demeaned and standardized by 2 standard deviations so they can be interpreted like binary (low-to-high) indicators. The underlying mean and standard deviation of each quality measure are provided in the first column to facilitate interpretation on the original scale. Outcome means: 30-day readmission = 15.0%; 30-day mortality = 17.0%; 365-day mortality = 37.2%. Sample sizes: 546,700 (30-day outcomes), 451,503 (1-year Mortality). All models include patient demographic and ambulance controls as listed in table 1, as well as the diagnosis controls as listed in table A2. Models also include postal code–patient origin fixed effects. Standard errors, clustered at health service area level, are reported in parentheses.

measures measured in an earlier time period. Here, we find that OLS estimates suggest a stronger positive correlation between the CMS outcomes measures and patient outcomes in our sample. We find that a 2 standard deviation increase in the composite readmission rate (i.e., a difference of 3.0 points) is associated with a 1.4 percentage point increase in the probability of readmission. We also find that a 2 standard deviation increase (2.6) in the composite thirty-day mortality rate at the hospital is associated with a 1.1 percentage point higher probability of death within 30 days, or more than 6% of baseline value.

The results in table 2 suggest links between the CMS measures and patient outcomes. It is possible that these conditional correlations are biased, however. In particular, if those with worse risk-adjusted quality scores document more comorbidities through "upcoding," then the comparisons controlling for comorbidities may bias the coefficients toward 0. On the other hand, to the extent that hospitals that have higher quality scores treat patients in worse (unobservable) health, we would expect 2SLS estimates to be larger in magnitude.

### C. *First Stage*

We now turn to showing that ambulance assignment is associated with hospital assignment-our first stage. Table 3 shows that assignment to an ambulance company that takes other patients to hospitals with an average risk-adjusted thirty-day mortality that is 2 standardized deviations higher is strongly linked with patients being treated at higher thirty-day mortality hospitals; the estimate is similar with and without patient and ambulance controls. We find similarly strong first-stage effects for our other quality measures. The only noticeable difference is a slightly weaker first stage for the reported process quality measure. All of the estimates are highly statistically significant.

The first-stage coefficients range from 0.55 to 0.62. These are significantly less than 1 due to the use of postal code fixed effects, which may exacerbate the influence of noise in the estimation, but the estimates are also consistent with our understanding of the natural experiment. When an ambulance company is dispatched to help in the other company's community, the company appears more likely to transport the patient back to its usual hospitals, but not at the same rate that it transports the patients living in its primary service area. This results in a strong, positive correlation, but one that is not one-to-one.

### D. *Quality Measures and Patient Outcomes: 2SLS Estimates*

In order to address the potential correlation with patient selection, we turn now to 2SLS estimates based on our ambulance instrument. The results are shown in table 4, which parallels table 2 in format.[9] Once again, each cell is from a separate regression. Overall, this 2SLS strategy largely confirms the OLS results, although with point estimates that are larger in magnitude.

We begin with process measures. For the process quality measure, we continue to find no statistically significant effect on readmissions. But we find large impacts on mortality at one year: a 2 standard deviation improvement in quality measured along this dimension (an increase of 10 points) leads to a 3.8 percentage point (10%) reduction in one-year mortality.

For the patient experience measure, we find slightly smaller effects, although still sizable, for mortality; at one year, the effect of a 2 standard deviation increase in the score (an increase again of 10 points) results in a 2.8 percentage point reduction in mortality (8% of the mean). This measure also has a sizable and significant impact on readmission probability; a 10 point increase in the score leads to a 2.1 percentage point (14%) reduction in the rate of readmission.

In addition, our 2SLS results show strong effects for the patient outcome measures. We find that hospitals with a high readmission rate are also much more likely to readmit the

[9] Coefficients from models where the measures are not standardized are reported in table A4.

marginal patient, controlling for patient selection: a 2 standard deviation increase in the rate of readmissions (i.e., a difference of 3.0 points on the composite readmission rate scale) leads to a 2.7 percentage point rise in thirty-day readmission among patients with nondeferrable conditions (18% compared to the mean).

We also find that hospitals with a high thirty-day mortality rate are much more likely to have patients die within thirty days of admission (after controlling for patient selection). The effect size is large, with a 2-standard deviation increase of 1.6 points in the risk-adjusted mortality scale, resulting in a 2.1 percentage point reduction in thirty-day mortality, or 12% of the mean. The effect on one-year mortality is only slightly larger, so that it is only about 7% of the baseline mean.

To conclude, we continue to find that in general, the types of quality measures used by CMS are strongly associated with patient outcomes; these quality measures do appear informative when a random patient is making a hospital choice.[10]

### *E. Sensitivity Checks*

Table 5 considers several sensitivity checks on our results. In panel A, we add (potentially endogenous) comorbidity indicators to our regression models. By and large, the results are not very sensitive to these controls, which is consistent with the exogeneity of our ambulance instrument. In panel B, we move from using separate regressions for each quality measure to a "horse race" framework, where we include all of the quality measures together in one regression with all patient controls. This allows us to account for cross-correlations across quality measures in interpreting their effects. In fact, we find that the results are remarkably consistent, even conditional on including the other quality measures.

In panel C, we turn to a more limited sample that consists of just the three conditions that are incorporated into the CMS quality measures themselves. Two of these conditions are included in our larger nondiscretionary conditions sample, while the third, congestive heart failure, is not considered nondiscretionary where we expect our instrument to be most appropriate. That said, the instrument values are well balanced on observable characteristics in this secondary sample, similar to the balance achieved in the main sample shown in table 1.[11] For this sample, we find results that are very similar, albeit generally larger, than for the 29 nondiscretionary conditions sample. The difference is particularly striking for the effect of patient experience on mortality, where the effects more than double from the larger sample results. This may reflect a closer correspondence between the quality measures and the sample or the change in conditions studied, although the standard errors are larger as well.

We also investigated whether results differed when we relate contemporaneous measures to patient outcomes as opposed to the lagged measures that are commonly used today. We find that the results are remarkably similar to our main results, as shown in the appendix.[12]

TABLE 5.—2SLS RESULTS

| Outcome | Mean [SD] | 30-Day Readmission | 30-Day Mortality | 365-Day Mortality |
|---|---|---|---|---|
| A. Add comorbidity controls | | | | |
| Timely and effective care composite | 93.40 [5.01] | −0.0052 (0.008) | −0.015 (0.008) | −0.040 (0.015) |
| Patient experience composite | 66.89 [5.02] | −0.020 (0.007) | −0.0063 (0.008) | −0.027 (0.011) |
| 30-day mortality rate composite | 12.69 [1.29] | −0.011 (0.007) | 0.025 (0.007) | 0.032 (0.011) |
| 30-day readmission rate composite | 20.98 [1.48] | 0.026 (0.008) | −0.0090 (0.007) | 0.003 (0.010) |
| B. Horse race | | | | |
| Timely and effective care composite | 93.40 [5.01] | 0.005 (0.012) | −0.013 (0.011) | −0.034 (0.015) |
| Patient experience composite | 66.89 [5.02] | −0.016 (0.009) | 0.001 (0.010) | −0.021 (0.011) |
| 30-day mortality rate composite | 12.69 [1.29] | −0.007 (0.008) | 0.029 (0.008) | 0.030 (0.011) |
| 30-day readmission rate composite | 20.98 [1.48] | 0.024 (0.009) | −0.004 (0.009) | −0.004 (0.011) |
| C. CMS condition sample (AMI, PN, HF) | | | | |
| Timely and effective care composite | 92.57 [5.12] | −0.0082 (0.031) | −0.031 (0.026) | −0.047 (0.034) |
| Patient experience composite | 66.33 [5.05] | −0.023 (0.018) | −0.041 (0.020) | −0.063 (0.026) |
| 30-day mortality rate composite | 12.66 [1.31] | −0.023 (0.017) | 0.041 (0.017) | 0.046 (0.025) |
| 30-day readmission rate composite | 21.01 [1.49] | 0.024 (0.019) | −0.018 (0.018) | −0.00012 (0.024) |

In panel A, each cell reports two-stage least squares (2SLS) coefficient estimates for a separate regression. In panel B, each column reports 2SLS estimates for a "horse race" specification that includes all quality measures in a single regression. In panel C, each cell reports two-stage least squares (2SLS) coefficient estimates for a separate regression using a sample of only acute myocardial infarction (AMI), pneumonia (PN), and heart failure (HF) patients. All models include patient demographic, comorbidity and ambulance controls as listed in table 1, as well as the diagnosis controls as listed in table A2. Models also include postal code–patient origin fixed effects. Quality measures have been demeaned and standardized by 2 standard deviations so they can be interpreted like binary (low-to-high) indicators. The underlying mean and standard deviation of each quality measure are provided in the first column to facilitate interpretation on the original scale. Outcome means for nondiscretionary condition sample: 30-day readmission = 15.0%, 30-day mortality = 17.0%, 365-day mortality = 37.2%. Sample sizes for nondiscretionary condition sample: 546,700 (30-day outcomes), 451,503 (1-year Mortality). Outcome means for CMS condition sample: 30-day readmission = 19.0%, 30-day mortality = 15.8%, 365-day mortality = 41.0%. Sample sizes for CMS condition sample: 171,246 (30-day outcomes), 142,424 (1-year mortality). Standard errors, clustered at health service area (HSA) level, are reported in parentheses.

[10] We can also create a combined quality measure that weights the four dimensions of quality included in the analysis by the relative association between that quality indicator and a given patient outcome using our 2SLS estimates. Not surprisingly, this improves the performance of the measure by construction. To give a sense of magnitudes, a 2 standard deviation improvement in composite quality measure aimed at identifying hospitals that perform well on patient mortality measure would result in approximately a 15% reduction in one-year mortality compared to the mean. See Doyle et al. (2017) for more on this calculation.

[11] We also considered our primary results estimated on a nondiscretionary condition sample that excludes the CMS conditions used to define the quality measure and are included in our set of nondiscretionary conditions (AMI and pneumonia). Here, again, we find broadly similar results (though slightly larger standard errors) as compared with those in table 4, as shown in appendix table A6.

[12] Indeed, a movement from a low- to a high-quality hospital based on the hospital's risk-adjusted mortality measure has the same coefficient (0.021). One of the only coefficients that changes is the relationship between a risk-adjusted mortality quality measure and thirty-day readmissions, where the coefficient declines from −0.12 to 0.007 (both with SE of 0.007).

### *F. Competing Risks*

One important question about quality measurement is possible bias arising from competing risks concerns. For example, suppose that hospitals that perform well on readmissions do so by raising mortality risk. This would suggest that a lower readmission rate is not a strong measure of better hospital outcomes because this may come at the expense of another outcome of even greater importance. The issue of competing risks therefore poses yet another challenge to the usefulness of quality measures.

Reviewing tables 4 and 5, we find modest, but not consistent, evidence for competing risks. In no specifications do we find statistically significant evidence that admission to a hospital with a higher readmission rate leads to lower mortality; the effects are particularly small when focusing on one-year mortality. We do find marginally significant evidence in table 4 that admission to a hospital with a higher thirty-day mortality rate lowers the likelihood of readmission, which suggests competing risks. We choose not to emphasize this result, however, as it is not particularly robust to specification checks. In addition, hospitals could perform well on one measure and not on another for reasons other than competing risks, such as employing technologies that are particularly suited to affect one, but not the other, dimension of quality.

## VI. Conclusion

The use of quality scores to guide consumer choice or as a central part of a move toward paying for quality instead of quantity is controversial. Providers take on risk when evaluated in this way, especially for outcomes that they do not fully control, such as readmissions and mortality. A primary criticism of the scores is that patients differ across hospitals in ways that are difficult to control using comorbidities and other patient characteristics. Another criticism of readmissions measures is that higher mortality can improve a hospital's readmission score, and we would not want to set up our quality measurements to reward such an outcome. We address these criticisms using an instrumental variables strategy that controls for patient selection.

There are a number of limitations to the approach. First, the estimates are most appropriate for emergency care and say less about quality for the health care of more chronic conditions. Second, the approach is interested in whether a randomly assigned patient achieves better outcomes by choosing a hospital that scores higher on these quality measures. While informative about whether these controversial measures are informative, patients may want to choose a hospital that is more tailored to their level of illness (Chandra & Staiger, 2007; Hull, 2018). Third, the results attempt to circumvent one problem with existing measures: nonrandom patient selection, but other concerns remain. In particular, if measures are risk adjusted, then we should be concerned with the endogeneity of measures used to risk-adjust through upcoding (Finkelstein et al., 2017; Dafny, 2005; Song et al., 2010). Perhaps more concerning is that hospitals may respond by selecting which patients to accept, which can have adverse effects for patients (Dranove et al., 2003). Efforts should be made to construct quality measures that reduce the scope for such gaming, such as a careful selection of risk adjusters and careful monitoring of access to hospitals as part of the compensation model.

We find that hospitals that perform well on timely and effective care processes, better patient experience, and lower hospital mortality rates (for other patients) achieved a significant and meaningful decline in the likelihood that an effectively-randomly assigned patient dies in the subsequent year. In addition, better patient experience and lower readmission rates (for other patients) are strongly associated with a lower likelihood that such a patient is readmitted to a hospital. We do not find consistent and compelling evidence that competing risks undo the validity of these measures.

We conclude that the measures used today by CMS to reimburse and rate hospitals on their quality are reliable and valid indicators of hospital quality for patients admitted to the hospital for a wide range of emergency care, and our estimates can be useful in assessing the magnitude of the relationship between these indicators and outcomes that policymakers can use to set reimbursement levels. This is encouraging as reformers move forward to tie reimbursement to these measures, especially for commonly discussed payment models that pay for quality within episodes of emergency care such as those studied here.

## REFERENCES

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association* 103 (2012), 1481–1495, http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000841.

Ash, Arlene S., Stephen F. Fienberg, Thomas A. Louis, Sharon-Lise T. Normand, Therese A. Stukel, and Jessica Utts, "Statistical Issues in Assessing Hospital Performance" (2012), http://escholarship.umassmed.edu/qhs_pp/1114/.

Austin, J. M., A. K. Jha, P. S. Romano, S. J. Singer, T. J. Vogus, R. M. Wachter, and P. J. Pronovost, "National Hospital Ratings Systems Share Few Common Scores and May Generate Confusion Instead of Clarity," *Health Affairs* 34 (2015), 423–430, doi:10.1377/hlthaff.2014.0201.

Berenson, Robert A., Ronald A. Paulus, and Noah S. Kalman, "Medicare's Readmissions-Reduction Program—A Positive Alternative," *New England Journal of Medicine* 366 (2012), 1364–1366, doi:10.1056/NEJMp1201268.

Birkmeyer, John D., Andrea E. Siewers, Emily V. A. Finlayson, Therese A. Stukel, F. Lee Lucas, Ida Batista, H. Gilbert Welch, and David E. Wennberg, "Hospital Volume and Surgical Mortality in the United States," *New England Journal of Medicine* 346 (2002), 1128–1137, doi:10.1056/NEJMsa012337.

Card, David, Carlos Dobkin, and Nicole Maestas, "Does Medicare Save Lives?" NBER working paper 13668 (2007), http://www.nber.org/papers/w13668.

Chandra, Amitabh, and Jonathan S. Skinner, "Technology Growth and Expenditure Growth in Health Care," NBER working paper 16953 (2011), https://ideas.repec.org/p/nbr/nberwo/16953.html.

Chandra, Amitabh, and Douglas O. Staiger, "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy* 115 (2007), 103–140, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2311510/.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review* 104 (2014), 2593–2632, doi:10.1257/aer.104.9.2593.

Chiang, Arthur J., Guy David, and Michael Gene Housman, "The Determinants of Urban Emergency Medical Services Privatization," *Critical Planning* 13 (2006), 5–22, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=918525.

Dafny, Leemore S., "How Do Hospitals Respond to Price Changes?" *American Economic Review* 95 (2005), 1525–1547.

Daley, Jennifer, "Invited Commentary: Quality of Care and the Volume-Outcome Relationship—What's Next for Surgery?" *Surgery* 131:1 (2002), 16–18, doi:10.1067/msy.2002.120237.

Das, A., E. C. Norton, D. C. Miller, A. M. Ryan, J. D. Birkmeyer, and L. M. Chen, "Adding a Spending Metric to Medicare's Value-Based Purchasing Program Rewarded Low-Quality Hospitals," *Health Affairs* 35 (2016), 898–906, doi:10.1377/hlthaff.2015.1190.

Desai, Nihar R., Joseph S. Ross, Ji Young Kwon, Jeph Herrin, Kumar Dharmarajan, Susannah M. Bernheim, Harlan M. Krumholz, and Leora I. Horwitz, "Association between Hospital Penalty Status under the Hospital Readmission Reduction Program and Readmission Rates for Target and Nontarget Conditions," *JAMA* 316 (2016), 2647, doi:10.1001/jama.2016.18533.

Dobkin, Carlos, "Hospital Staffing and Inpatient Mortality," unpublished working paper (2003).

Doyle, Joseph J., John A. Graves, and Jonathan Gruber, "Evaluating Measures of Hospital Quality," NBER working paper 23166 (2017), doi:10.3386/w23166.

Doyle, Joseph J., John A. Graves, Jonathan Gruber, and Samuel A. Kleiner, "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy* 123 (2015), 170–214, doi:10.1086/677756.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers," *Journal of Political Economy* 111 (2003), 555–588.

Dudley, R. Adams, Kirsten L. Johansen, Richard Brand, Deborah J. Rennie, and Arnold Milstein, "Selective Referral to High-Volume Hospitals: Estimating Potentially Avoidable Deaths," *JAMA* 283 (2000), 1159–1166. http://jama.jamanetwork.com/article.aspx?articleid=192451.

Finkelstein, Amy, Matthew Gentzkow, Peter Hull, and Heidi Williams, "Adjusting Risk Adjustment—Accounting for Variation in Diagnostic Intensity," *New England Journal of Medicine* 376 (2017), 608–610, doi:10.1056/NEJMp1613238.

Fisher, Elliott S., Julie P. Bynum, and Jonathan S. Skinner, "Slowing the Growth of Health Care Costs: Lessons from Regional Variation," *New England Journal of Medicine* 360 (2009), 849–852, doi:10.1056/NEJMp0809794.

Fisher, Elliott S., David E. Wennberg, Thrse A. Stukel, Daniel J. Gottlieb, F. L. Lucas, and Étoile L. Pinder, "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care," *Annals of Internal Medicine* 138 (2003a), 273–287, doi:10.7326/0003-4819-138-4-200302180-00006.

——— "The Implications of Regional Variations in Medicare Spending: Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine* 138 (2003b), 288–298, doi:10.7326/0003-4819-138-4-200302180-00007.

Gestel, Yvette R. B. M. van, Valery E. P. P. Lemmens, Hester F. Lingsma, Ignace H. J. T. de Hingh, Harm J. T. Rutten, and Jan Willem W. Coebergh, "The Hospital Standardized Mortality Ratio Fallacy: A Narrative Review," *Medical Care* 50 (2012), 662–667, doi:10.1097/MLR.0b013e31824ebd9f.

Halm, Ethan A., Clara Lee, and Mark R. Chassin, "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature," *Annals of Internal Medicine* 137 (2002), 511–520, http://annals.org/article.aspx?articleid=715648.

Hughes, Robert G., Sandra S. Hunt, and Harold S. Luft, "Effects of Surgeon Volume and Hospital Volume on Quality of Care in Hospitals," *Medical Care*, 25 (1987), 489–503, http://www.jstor.org/stable/3765332.

Hull, Peter, "Estimating Hospital Quality with Quasi-Experimental Data" (2018), http://www.mit.edu/∼hull/RAM_022018.pdf.

Johnson, Robin, *The Future of Local Emergency Medical Service Ambulance Wars 1 or Public-Private Truce?* (Los Angeles: Reason Public Policy Institute, 2001).

Lilford, R., and P. Pronovost, "Using Hospital Mortality Rates to Judge Hospital Performance: A Bad Idea That Just Won't Go Away," *BMJ* 340 (2010), c2016–c2016, doi:10.1136/bmj.c2016.

Luft, Harold S., Sandra S. Hunt, and Susan C. Maerki, "The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns?" *Health Services Research* 22 (1987), 157, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065430/.

Maxwell, Bryan G., Jim K. Wong, D. Craig Miller, and Robert L. Lobato, "Temporal Changes in Survival after Cardiac Surgery Are Associated with the Thirty-Day Mortality Benchmark," *Health Services Research* 49:5 (2014), 1659–1669, doi:10.1111/1475-6773.12174.

McGlynn, Elizabeth A., Steven M. Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A. Kerr, "The Quality of Health Care Delivered to Adults in the United States," *New England Journal of Medicine* 348:26 (2003), 2635–2645, doi:10.1056/NEJMsa022615.

Muhlestein, David, Robert Saunders, and Mark McClellan, "Growth of ACOs and Alternative Payment Models in 2017," *Health Affairs Blog* (2017), https://www.Healthaffairs.Org/Do/10.1377/Hblog20170628607 19.

Norton, Edward C., Jun Li, Anup Das, and Lena M. Chen, "Moneyball in Medicare," NBER working paper 22371 (2016), doi:10.3386/w22371.

Ragone, Michael, "Evolution or Revolution: EMS Industry Faces Difficult Changes," *JEMS: A Journal of Emergency Medical Services* 37:2 (2012), 34–39, http://europepmc.org/abstract/med/22365034.

Shahian, David M., Gregg S. Meyer, Elizabeth Mort, Susan Atamian, Xiu Liu, Andrew S. Karson, Lawrence D. Ramunno, and Hui Zheng, "Association of National Hospital Quality Measure Adherence with Long-Term Mortality and Readmissions," *BMJ Quality and Safety* 21 (2012), 325–336, doi:10.1136/bmjqs-2011-000615.

Shahian, David M., and Sharon-Lise T. Normand, "The Volume-Outcome Relationship: From Luft to Leapfrog," *Annals of Thoracic Surgery* 75 (2003), 1048–1058, http://www.sciencedirect.com/science/article/pii/S0003497502043084.

——— "Comparison of 'Risk-Adjusted' Hospital Outcomes," *Circulation* 117 (2008), 1955–1963, http://circ.ahajournals.org/content/117/15/1955.short.

Skura, Barry, "Where Do 911 System Ambulances Take Their Patients? Differences between Voluntary Hospital Ambulances and Fire Department Ambulances" (New York: City of New York Office of the Comptroller, 2001).

Song, Yunjie, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E. Wennberg, and Elliott S. Fisher, "Regional Variations in Diagnostic Practices," *New England Journal of Medicine* 363 (2010), 45–53, doi:10.1056/NEJMsa0910881.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics* 25 (2012), 518–529, http://amstat.tandfonline.com/doi/abs/10.1198/073500102288618658.

Yasaitis, Laura, Elliott S. Fisher, Jonathan S. Skinner, and Amitabh Chandra, "Hospital Quality and Intensity of Spending: Is There an Association?" *Health Affairs* 28 (2009), w566–w572, http://content.healthaffairs.org/content/28/4/w566.short.