NBER WORKING PAPER SERIES

EVALUATING MEASURES OF HOSPITAL QUALITY

Joseph J. Doyle, Jr.
John A. Graves
Jonathan Gruber

Working Paper 23166
http://www.nber.org/papers/w23166

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2017

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Evaluating Measures of Hospital Quality
Joseph J. Doyle, Jr., John A. Graves, and Jonathan Gruber
NBER Working Paper No. 23166
February 2017
JEL No. I10

ABSTRACT

In response to unsustainable growth in health care spending, there is enormous interest in reforming the payment system to "pay for quality instead of quantity." While quality measures are crucial to such reforms, they face major criticisms largely over the potential failure of risk adjustment to overcome endogeneity concerns. In this paper we implement a methodology for estimating the causal relationship between hospital quality measures and patient outcomes. To compare similar patients across hospitals in the same market, we xploit ambulance company preferences as an instrument for patient assignment. We find that a variety of measures used by insurers to measure provider quality are successful: assignment to a higher-scoring hospital results in better patient outcomes. We estimate that a two-standard deviation improvement in a composite quality measure based on existing data collected by CMS is causally associated with reductions in readmissions and mortality of roughly 15%.

Joseph J. Doyle, Jr.
MIT Sloan School of Management
100 Main Street, E62-516
Cambridge, MA 02142
and NBER
jjdoyle@mit.edu

John A. Graves
Vanderbilt University
2525 West End Ave.
Suite 1200
Nashville, TN 37203
john.graves@vanderbilt.edu

Jonathan Gruber
Department of Economics, E52-434
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
gruberj@mit.edu

1 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, and Skinner 2009; Fisher et al. 2003a; Fisher et al. 2003b; Chandra and 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 and 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 30-day readmission rates for certain conditions (Desai et al. 2016; Berenson, Paulus, and Kalman 2012). Another example is the Hospital Value-Based Purchasing Program (HVBP), which explicitly ties financial incentives to hospital quality performance (Chen et al. 2016; Das et al. 2016).

Measures of hospital quality remain highly controversial despite their increasingly widespread use (Lilford and 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 performace and make quality comparisons across hospitals (Ash et al. 2012; Shahian and 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 (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.

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).[1] We do so by leveraging ambulance company referral patterns as an instrument for hospital assignment. Ambulance companies are exogenously assigned to emergency patients based on availability at the time of the emergency. In many 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 upon 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 (our sample of emergency conditions amounts to approximately one-quarter of inpatient care for Medicare patients nationwide).[2]

Our primary analyses consider four composite measures constructed to capture different and commonly-measured dimensions of quality: (1) process measures quantifying the frequency with which hospitals provide services that are considered effective in improving patient outcomes; (2) risk-adjusted patient satisfaction scores from patient surveys, which are increasingly used by insurers to "pay for quality"; (3) risk-standardized 30-day readmission rates among all discharged patients; and (4) risk-standardized 30-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 Centers for Medicare and Medicaid Services (CMS) is related to patient outcomes in important ways. First, hospitals with higher process measures of quality have lower long-term mortality for the marginal patient. Second, hospitals with lower patient satisfaction scores have higher odds of readmission and death.

[1] This is akin to the education literature that seeks to substantiate whether or not controversial quality measures predict better outcomes (e.g. Chetty, Friedman, and Rockoff (2014)).

[2] In contemporaneous work, Hull (2016) develops a model with a more ambitious goal to not just demonstrate that the widely used CMS mortality measure has a causal relationship with quality, but to build a better quality measure. We discuss his work at more length below.

Third, we find a strong and significant positive effect of hospital readmission rates on the odds of readmission, and an even stronger positive effect of hospital mortality rates on the odds of mortality. Our findings suggest that, even correcting for patient selection, the outcome measures utilized in value-based payment reform efforts by CMS and other payors are useful proxies for hospital quality.

Our framework also allows us to asses the issue of "competing risks" that may occur when an outcome like mortality precludes other outcomes such as readmission. For example, if hospitals have high readmission rates in part due to those hospitals achieving lower mortality (or more troubling, achieve lower readmission rates in part by having high levels of mortality), then systems that reimburse on one or the other could be at odds with each other. Among our ambulance cases we find modest, though statistically insignificant, evidence of competing risks.

To summarize the magnitude of our findings, we use our regression estimates to create an overall hospital quality measure that incorporates all of the individual indicators included in the analysis, weighted by the relative association between that quality indicator and a given patient outcome. We find that a two standard deviation improvement in this composite readmission quality is associated with a 3.4% reduction in the odds of readmission, which is almost 20% of the mean. We also find that a two standard deviation improvement in composite mortality quality is associated with a 3.2 to 5.4 percentage point (20% to 14% compared to the mean) decrease in the odds of death over 30 days and 1 year, respectively.

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, not only for patients treated for the conditions they measure but for other types of emergency care. This is encouraging as reformers move forward to tie reimbursement to these measures. And our estimates can be useful in assessing the magnitude of the relationship between these indicators and outcomes that can be used by policy-makers to set reimbursement levels.

The remainder of this paper proceeds as follows. The next section provides relevant institutional background on hospital quality reporting in the U.S., 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 then conclude with a discussion of the implications of our findings for hospital reimbursement policy.

2 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. We discuss the relevant details of quality measurement below.

2.1 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 the U.S. News and World Report's annual hospital rankings. The number of process measures reported on Hospital Compare, for example, has risen from 20 in 2005 (the first year of public reporting) to over 40 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 13 measures evolved over an eight-year period (the specific process measures are listed in Table A1). 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 towards 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 be particularly low quality.

2.2 Patient Experience Measures

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

Hospital-specific scores on 11 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 language, and hospital service line used (e.g., surgical vs. medical). These risk-adjusted scores are reported individually on the Hospital Compare website, and more recently a "Five Star" ranking is constructed based on a composite average of hospital HCAHPS performance.

2.3 Outcome Measures

Outcome quality measures compare the observed number of patients who experience a given outcome (e.g., mortality or readmission within a 30 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 by 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, 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 socio-economic status) are deliberately excluded from risk adjustment so that the risk-standardized measures do not condition out important racial or socio-economic disparities in care across facilities.[3]

[3]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-

More formally, define as a binary indicator of whether patient treated at hospital experienced the outcome, and define as a vector of patient-level characteristics. The current CMS model assumes the following:

(1)

$$Y_{ih}|\alpha_h, \beta, \mathbf{x_{ih}} \overset{ind}{\sim} \text{Bernoulli}(\alpha_h + \beta \mathbf{x_{ih}})$$

(2)

$$\alpha_h|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$$

Based on this model, the risk-standardized rate for hospital h (RSR_h) is estimated as

(3)

$$\text{RSR}_h(\mathbf{x_{ih}}) = \frac{\sum_{i=1}^{n_h} E(Y_{ih}|\alpha_h, \mu, \beta, \mathbf{x_{ih}}, \sigma^2)}{\sum_{i=1}^{n_h} E(Y_{ih}|\mu, \beta, \mathbf{x_{ih}}, \sigma^2)} \bar{Y}$$

where $\bar{Y}$ is the national average outcome and n_h is the number of patients treated in hospital h (Ash et al. 2012).

In practice, equation (1) is estimated as a hierarchical Bayes logistic regression model that includes a hospital random effect (α_h). The underlying patient-level data and the estimated model parameters ($\hat{\alpha}_h, \hat{\mu}, \hat{\beta}, \hat{\sigma}^2$) are used to construct predicted values, which are then fed into equation (3) to obtain the risk-standardized rate for each hospital.[4]

2.4 Concern: Risk-Adjustment is Insufficient

As noted in the introduction, a common issue with outcome measurement is a concern that a selection-on-observables assumption inherent in risk adjustment is highly controversial. Further, outcomes are assumed independent of hospital-level attributes conditional on observable patient characteristics, $\mathbf{x_{ih}}$ in Equation 1 above. That is, the random effects model assumes that the patient-level risk adjustors capture the relevant clinical characteristics such that other hospital level

Instruments/HospitalQualityInits/Downloads/MedicareHospitalQualityChartbook2012.pdf (pp. 23-36).

[4]The numerator in equation (3) is simply the sum of predicted values for each patient in the hospital, where the predicted values are based on observed patient values, $\hat{\mu}, \hat{\beta}$, and the estimated hospital random effect ($\hat{\alpha}_h$). The denominator is similarly estimated as the sum of patient-level predictions, but only $\hat{\mu}$ – the estimated national mean of $\hat{\alpha}_h$ – is used. This measure is then multiplied by $\bar{Y}$ to place the risk standardized measure on the same scale as the outcome. In effect, outcome-based quality assessments are determined primarily by how much each hospital's estimated random effect ($\hat{\alpha}_h$) deviates from $\hat{\mu}$. The hypothetical reference hospital in the denominator of equation (3) has the same patient case-mix as the hospital in question, however it has average quality, as reflected by the use of only $\hat{\mu}$ to summarize hospital-level factors. In this way each hospital is evaluated against a hypothetical reference hospital with an identical case-mix and with average quality.

attributes (e.g., patient volume, teaching status, etc.) are not independent predictors of patient outcomes. This is a strong assumption, particularly in light of the large medical and health services literature linking hospital attributes such as volume to improved patient outcomes (Birkmeyer et al. 2002; Daley 2002; Dudley et al. 2000; Halm, Lee, and Chassin 2002; Hughes, Hunt, and Luft 1987; Luft, Hunt, and Maerki 1987; Shahian and Normand 2003); 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 socio-economic status; and growing evidence on the endogeneity of patient-level diagnoses as recorded hospital billing codes (Song et al. 2010).

2.5 Concern: Inter-Hospital Comparisons Are Inappropriate

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 those hospitals may not treat the same profile of patients (Shahian and Normand 2008).

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 the above scenario, if hospital 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 vs. B and conclude that hospital B was of poorer quality; had patients been randomly assigned to A vs. B we would find no evidence of a difference in patient-level readmission outcomes between the two.

In short, the indirect standardization method does not, by construction, facilitate inter-hopsital

comparisons, despite the fact that such comparisons are consistently made by even seasoned experts, and are even facilitated by the CMS website entitled "Hospital Compare" (Ash et al. 2012). Thus, an open question is whether outcome rates can also be used to guide patient choice of the "best" hospital among local options. Fortunately, this is a question our instrumental variables approach can answer because we observe patients who are effectively randomly assigned to different hospitals with varying quality and case mix. That is, our framework can assess whether high performing hospitals achieve better outcomes for the marginal patient who is effectively randomized to a local hospital.

3 Empirical Strategy

3.1 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 pre-hospital 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 U.S. 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, and 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 30 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 ZIP code as public Health and Hospital Corporation (HHC) hospitals were less than half as likely to be taken there when assigned a private, non-profit ambulance (29%) compared to when the dispatch system assigned them to an FDNY ambulance (64%). In most cases, the private ambulances were operated by non-profit 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, and 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:

(4)
$$Z_{a(i)} = \frac{1}{N_{a(i)-1} - 1} \sum_{j \neq i}^{N_{a(i)-1}} H_j$$

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 30-day readmission rate, its 30-day mortality rate or a composite process measure.

3.2 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 from an origin in ZIP code $z(i)$ in year $t(i)$:

(5)

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

where X_i is a vector of patient controls including age, race, and sex, and indicators for 17 common comorbidities controlled for in the CMS quality scores; A_i represents a vector of ambulance characteristics that summarise 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. We cluster standard errors at the Hospital Service Area (HSA) level, as each local market may have its own assignment rules.[5]

We also include a full set of controls for principal diagnosis, year and ZIP code × patient origin fixed effects.[6] 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 ZIP code, but who are picked up by ambulance companies with different "preferences" across hospitals with different quality scores. A positive coefficient α_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:

(6)

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

For this regression we consider various patient outcomes, such as whether they are readmitted to another acute care hospital facility within 30 days of discharge, or whether they died within 30 days or one year of admission. Finally, since patient selection is likely to confound this structural model, we estimate equation (6) using two-stage least squares, with the instrument defined as above.

[5] 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.

[6] The principal diagnosis is the 3-digit ICD-9-CM diagnosis code, as shown in Appendix Table A2.

Doyle et al. (2015) discusses at length potential limitations with this strategy and various specification checks that begin to address them. In particular, that study finds that the results are highly robust to controls for both patient characteristics and the characteristics of pre-hospital 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 are robust to the level of heterogeneity in patient characteristics across ZIP codes, which is inconsistent with potential locational bias in ambulance assignment within a ZIP.

As noted, the goal of this empirical strategy is to develop a causal framework through which we can assess the validity of the widely used CMS quality measures. In contemporaneous work, Hull (2016) builds on the ambulance-instrument approach to address a more ambitious goal: revising the quality measure towards one which more accurately reflects patient outcomes. In particular, the paper uses the closest ambulance company office location for each hospital as a proxy for hospital affiliation to arrive at one instrument for each of the 2,357 hospitals in the estimation sample. Given the difficulty of precisely estimating quality at each hospital, Hull develops a valuable new methodology that uses a Roy model to investigate nonlinear selection on gains (e.g. patients are often sorted to hospitals based on their expected improvement in outcomes such as sending patients to trauma centers).[7] To overcome precision concerns that arise from estimating thousands of effects using instruments based on thousands of ambulance companies, the semi-parametric approach developed in the paper shrinks the raw estimates based on the quasi-experimental variation due to ambulance-company assignment using distributional assumptions on the latent quality variable, as well as incorporating estimates generated with the more precisely estimated but likely biased estimates that come from more traditional random-effects models to result in a lower mean squared prediction error. This shrinkage estimator results in posterior quality estimates for each hospital.

Thus, we view Hull's work as complementary to our approach. If the goal is to demonstrate convincingly that existing CMS quality measures causally measure hospital quality, then our approach is a minimally structural approach to doing so; as emphasized earlier, this is akin to important work in education evaluating existing teacher quality measures. If the goal is to improve

[7] We can also estimate nonlinear relationship between the quality measure and subsequent outcomes as in Doyle (2012), but such estimates are not particularly precisely estimated. Hull's important insight is that a parametric structure on the problem can help resolve this imprecision problem.

on the CMS measure, given the imprecision in the data, then more structure must be imposed, and Hull provides an innovative way of doing so. In practice, Hull's new quality measure is highly correlated with the CMS measure (0.68 correlation within Health Service Areas), but there are major differences in magnitudes. This combined set of results confirms that CMS quality measures are predictive, and that they can be improved. As one step towards that goal, we consider adding other quality measures to the mix in our empirical work below.

4 Data

4.1 Medicare Claims Data

Our primary source of patient-level data are Medicare claims between 2008-2012: 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 records 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 ZIP code of the beneficiary, where official correspondence is sent; in principle, this could differ from the patient's home ZIP 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 timeframes, such as 30 days or one year.

4.2 Sample Selection

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 "non-disretionary" conditions. As described in Doyle et al. (2015), these are conditions where selection into the health care system is largely unavoidable (i.e., femur fracture, poisoning and stroke). 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.

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 ZIP 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 30 episodes (in the 20% Medicare files), as well as patients whose ambulance company transported 30 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 utilize the sub-sample of 451,503 non-discretionary patients and 142,424 CMS condition patients with uncensored one-year outcomes (i.e., those treated between 2008-2011).

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 which arrived by ambulance. Given that roughly 60% of all Medicare admissions originate in the emergency room, these condiditions constitute approximately 23% of all hospital

care for Medicare patients in the U.S. Moreover these conditions are particularly expensive, such as sepsis – 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 or not 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 point further in the conclusion.

4.3 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, while 30-day mortality measures and patient satisfaction scores were added in 2008 and the 30-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 3-year sample of fee-for-service Medicare and Veterans Health Administration patients, with a one year lag between the most 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) 4 years prior to the reporting date.

The use of a lagged measure 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 odds that patient X dies, we are not including patient X him/herself in the calculation.

Our composite process score is based on the pooled average of 13 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). Similarly, the 30-day mortality and readmission composite measures are based on the average mortality/readmission rates for AMI, PN and HF patients.

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 vs. high "quality" measure. For context, the main results provide both the unstandardized mean and standard deviation for each composite measure.

4.4 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 such outcomes. The first is the rate of hospital readmissions over the 30 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. As a result, CMS introduced the readmissions penalty as part of the Affordable Care Act, as noted earlier. The 30-day patient readmission outcomes we consider are defined as unplanned readmission to any hospital within 30 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 30 days of hospital admissions. A disadvantage of this approach, however, is that hospital actions can impact mortality over shorter time horizons. In particular, Maxwell et al. (2014) show that there is a discontinuity in patient mortality at 30 days for cardiac surgery, suggesting the possibility of hospital manipulation of morality at shorter horizons. For this reason, we show mortality results at 30 days but also focus on a longer horizon, assessing the impact of quality on mortality over the first year post-admission.

5 Do CMS Quality Measures Actually Measure Quality?

5.1 Balance

To evaluate the relationship between measured hospital performance and patient outcomes we rely on an instrumental variables approach that assumes patients are quasi-randomly assigned to ambulances in an emergency. If this assumption holds then our empirical approach should purge

endogeneity stemming from patient-level selection into different hospitals. To test whether this is plausible along observable dimensions, Table 1 shows means of patient-level demographic and health measures across those whose ambulances tend to transport patients to hospitals with high versus low 30-day mortality rates: a measure that we emphasize in our results below. In particular, the data are divided into quartiles based on the distribution of the ambulance instrument for 30-day mortality after it has been de-meaned at the ZIP code level.

The table shows that our sample is remarkably balanced on observable demographic and health characteristics. The age distribution of patients whose ambulances are more likely to take patients to hospitals across these quartiles is nearly identical, with 12.5 to 12.6% of patients aged 70-74 for example. Likewise, patients in the lowest quartile are transported by ambulance an average of 6.9 miles, versus 7.1 miles in the highest quartile. Balance is similar even across measured dimensions of health as captured by our comorbidity measures: 28.2% of patients in the lowest quartile have an indicator of hypertension in their Medicare claims for the previous year – the same percentage with hypertension in the highest quartile. Similar results are found across quartiles of an inpatient reimbursement-based instrument, as documented in Doyle et al. (2015).

5.2 Quality Measures and Patient Outcomes – OLS 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 30-day readmissions, 30-day mortality, and one-year mortality. The means of the key dependent variables, as reported in the Table note, are 15.0% for 30-day readmission, 17.0% for 30-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; e.g. the first row of the second column shows the OLS relationship between the quality process score and 30-day readmission.

We find that the CMS process measure of quality is uncorrelated with readmission rates, and with 30 day mortality. There is a modest correlation with one-year mortality rates, with a two standard-deviation increase in the process score (i.e., a change of 10.2 points) associated with a 0.4 percentage point decline in one-year mortality, which is only about 1% of the mean.

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 signficant) likelihood of readmission within 30-days. They exhibit no differences in 30-day mortality outcomes, but show a small and significant impact on one-year mortality.

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

Here, we find that in OLS there is a stronger positive correlation between the CMS outcomes measures and patient outcomes in our sample. We find that a two standard deviation increase in the composite readmission rate (i.e., a difference of 3.0 points) is associated with a 1.5 percentage point increase in the probability of readmission. Likewise, we find that a two standard deviation increase (2.6) in the composite 30-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 modest links between the CMS measures and patient outcomes. But these results are only correlations. It is possible that the associated relationships merely reflect patient selection and not true underlying differences in quality. 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.

5.3 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 30-day mortality that is two standardized deviations higher is strongly linked with patients being treated at higher 30-day mortality hospitals, and 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.[8]

The fact that our first stage estimates are less than one is informative of the nature of variation the estimation strategy is using. For example, when an ambulance company is dispatched to

[8]The standard error estimates do not correct for the fact that the measure is a leave out mean, although this should have little impact given the large number of observations per ambulance companies: an average of TK.

help a nearby community through a "mutual aid" agreement, the first-stage estimate implies that the ambulance company is more likely to transport the patient back to the hospitals where the ambulance company usually operates, but not at the same rate that it transports patients living where the company usually operates. This results in a strong positive correlation, but one that is not one-to-one.

5.4 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. 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 effect on readmissions. But we find large impacts on mortality at one year: a two standard deviation improvement in quality measured along this dimension 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 sizeable, for mortality; at one year, the effect is 2.8 percentage points (8% of the mean). This measure also has a sizeable and significant impact on readmission probability; a two-standard deviation increase in patient experience quality leads to a 1.9 percentage point (13%) reduction in the rate of readmission.

In addition, our 2SLS results show strong effects for the patient outcome rate measures. We find that hospitals with a high readmission rate are also much more likely to readmit the marginal patient controlling for patient selection. A two 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 2.9 percentage point rise in the odds of 30 day readmission among patients with non-deferrable conditions (19% compared to the mean).

We also find that hospitals with a high 30-day mortality rate are much more likely to have patients die within 30 days of admission (after controlling for patient selection). The effect size is large with a reduction of 2.1 percentage points compared to a baseline 30-day mortality rate of 17%. 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. Hospitals that perform well on timely and effective care processes, better patient experience, and lower hospital mortality rates (for other patients) are associated with a significant and meaningful decline in the odds that randomly assigned patient dies in the subsequent year. And better patient experience and lower readmission rates (for other patients) are strongly associated with a lower odds that the randomly assigned patient is readmitted to a hospital.

5.5 Sensitivity Checks

One possibility is that the 2SLS results in Table 4 suffer from residual confounding, since the specifications used for the regressions rely on our ambulance framework for identification and do not include a large array of controls for (potentially endogenous) comorbidity measures. Moreover, there could be important cross-correlations among the quality measures with, for example, hospitals with low mortality rates also scoring well on patient experience of care surveys. We investigate the interdependence of the measures in this section.

Table 5 considers several sensitivity checks on our results. In the first panel, we add comorbidity (diagnosis) controls to our regressions to assess sensitivity. By and large, the results are not very sensitive to these controls, which is consistent with the exogeneity of our ambulance instrument.

In the second panel, 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 the final panel, 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 non-discretionary conditions sample, while the third, Congestive Heart Failure, is not considered non-discretionary where we expect our instrument to be most appropriate. That said, the instrument values are well balanced on observable characteristics this secondary sample, similar to the balance achieved in the main sample shown in Table 1.

For this sample, we find results that are very similar, albeit generally larger, than for the 29

non-discretionary 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.

5.6 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 30 day mortality rate lowers the odds of readmission, which suggests competing risks. But this result is not robust to the various changes we make in Table 5. In Panel C of Table 5 we also find some evidence of competing risks for 30-day readmission outcomes in the CMS condition sample, with point estimates that offset each other for the 30-day readmission rate (-2.3 percentage points) and 30-day mortality rate (+2.4 percentage points) composite measures. However, both of these coefficients are not statistically signficiant. Similarly, we also find partially offseting, though somewhat noisy, coefficients for the 30-day mortality outcome.

5.7 Composite Quality Measure

Thus far in the paper we have considered a variety of widely-used quality measures. Each of these measures help explain our readmission and mortality outcomes. Using our 2SLS results we can build a composite quality index for each outcome based on the four quality measures. More specifically, we use the coefficients from the horse-race regressions in Panel B of Table 5 to calculate relative weights and create a weighted average quality measure for each of the three main outcomes of interest:

readmission, 30-day mortality and 1-year mortality. These measures incorporate any competing risk considerations of the type raised above.

Table 6 shows the results from using these composite quality measures. We show only "on-diagonal" estimates because the composites were constructed to create measures most likely to be related to each particular outcome. Interpretation of the off-diagonal estimates may be misleading, as the mortality outcome composite, for example, has information on the readmission quality measure contained within it.

In the first row we show the results of using our weighted average readmission quality measure. We find that a two standard deviation increase in this composite [+2.0] leads to 3.4 percentage point reduction in readmissions, or about 22% of the baseline mean. In the second and third rows, we show that a mortality composite quality indicator has very large impacts on mortality outcomes. A two standard deviation increase in the 30 day mortality composite lowers patient mortality by 3.2 percentage points, or about 19% of the mean; a two standard deviation increase in the one-year mortality composite lowers one-year mortality by 5.4 percentabe points, or 15% of the mean.

5.8 Quality Scores for Different Types of Hospitals

Given that hospital 30-day mortality rates appear informative of quality even after controlling for patient selection, it becomes important to understand which hospital characteristics are most associated with high patient quality. In Table 7, we provide an initial exploration of this question for the sample of 2,116 hospitals in our sample in 2012. In the first panel of the Table, each cell shows the regression coefficient from a regression of the quality measure listed at the top of each column on a set of hospital characteristics. In these regressions, the dependent variable is expressed in terms of standard deviations.

We first consider whether the hospital is a teaching hospital, defined as having an accredited residency program. We then include a measure of standardized hospital size based on the number of beds. Finally, we include ownership type, including dummies for for-profit (i.e., investor-owned) and public hospitals, relative to the omitted category of non-profit hospitals.

The results in this top panel show that there are strong associations between these quality measures and hospital characteristics. Teaching hospitals are more likely to score better on process measures and mortality outcomes measures; for example, teaching hospitals score 0.43 of one

standard deviation better on process quality, and 0.14 of one standard deviation better on mortality quality. Larger hospitals have better outcomes on process quality and mortality, but score much worse on patient experience and 30 day readmission. For-profit hospitals deliver higher process quality (relative to non-profit hospitals), but score worse on every other quality measure; the effect is particularly large for patient experience, where for-profit hospitals score two standard deviations worse than others. Finally, government hospitals score much worse on process and mortality, with no significant effect on the other quality indicators.

6 Conclusions

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 both of these criticisms using an instrumental variables strategy that controls for patient selection. We find within this framework that quality scores are very informative. In general, higher scores on each of the quality metrics that we study are associated with better outcomes along a variety of dimensions. And we do not find consistent and compelling evidence that competing risks undo the validity of these measures.

Overall, our findings suggest that these quality scores may be useful metrics for new payment models. There are at least two important caveats, however. First, the results here apply to emergency situations where ambulance transport is involved, though it is worth noting that nearly 60% of inpatient stays in the Medicare program initiate through the ED. The results may be most appropriately applied to commonly discussed payment models that pay for quality within episodes of care associated with particular health conditions and events such as those studied here. Second, once payments are made for quality scores, there is concern that hospitals may game the scores through patient selection, changes in coding behavior that affect the risk adjustment, or creative accounting. When such models are put into place, it will be important to emphasize the use of risk

adjusters that are less likely to be gamed and carefully monitor changes in the patient mix as part of the compensation model.

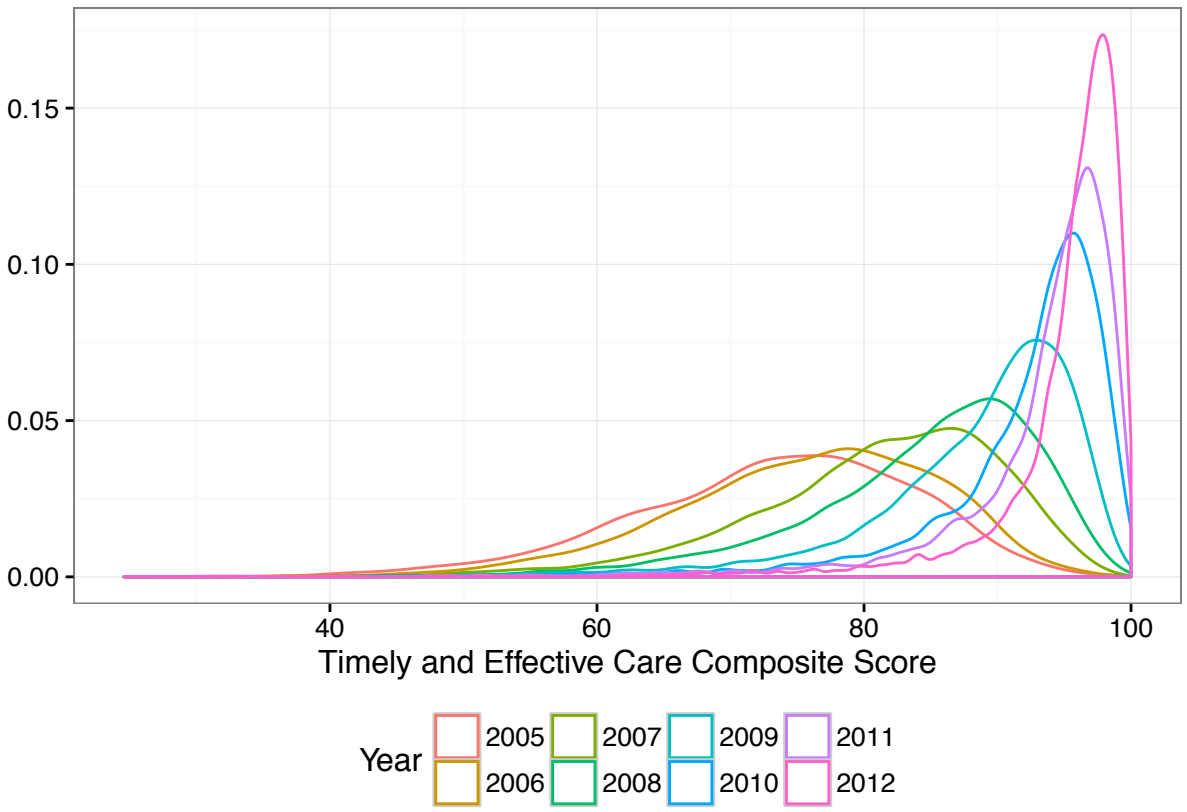

Figure 1: Density Distribution of Composite Process Score, By Year

Table 1: Balance of Patient Characteristics by 30-Day Mortality Rate Instrument Quintile

Measure	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Age 70-74	0.125	0.126	0.126	0.126
Age 75-79	0.166	0.166	0.167	0.165
Age 80-84	0.221	0.221	0.221	0.220
Age 85-89	0.224	0.225	0.225	0.225
Age 90-94	0.134	0.133	0.132	0.134
Age 95+	0.049	0.048	0.048	0.049
Gender: Male	0.372	0.374	0.375	0.375
Race: Black	0.085	0.086	0.086	0.085
Race: Other	0.042	0.042	0.041	0.041
Comorbidity: Hypertension	0.282	0.280	0.280	0.282
Comorbidity: Stroke	0.015	0.014	0.015	0.015
Comorbidity: Cerebrovascular Disease	0.037	0.036	0.036	0.036
Comorbidity: Renal Failure Disease	0.122	0.121	0.120	0.121
Comorbidity: Dialysis	0.011	0.010	0.010	0.011
Comorbidity: COPD	0.119	0.119	0.120	0.120
Comorbidity: Pneumonia	0.063	0.062	0.062	0.063
Comorbidity: Diabetes	0.134	0.132	0.132	0.133
Comorbidity: Protein Calorie Malnutrition	0.037	0.036	0.036	0.037
Comorbidity: Dementia	0.101	0.097	0.097	0.101
Comorbidity: Paralysis	0.038	0.036	0.036	0.037
Comorbidity: Peripheral Vascular Disease	0.080	0.079	0.079	0.080
Comorbidity: Metastatic Cancer	0.022	0.022	0.022	0.021
Comorbidity: Trauma	0.067	0.065	0.065	0.065
Comorbidity: Substance Abuse	0.038	0.038	0.039	0.039
Comorbidity: Major Psychological Disorder	0.032	0.031	0.031	0.033
Comorbidity: Chronic Liver Disease	0.007	0.007	0.007	0.007
Ambulance: Miles Traveled with Patient	6.938	6.909	6.998	7.120
Ambulance: Emergency Traffic	0.957	0.958	0.958	0.955
Ambulance: Advanced Life Support	0.711	0.725	0.727	0.731
Ambulance: Payment	369	371	372	375

Table shows balance across covariates used in regressions. Columns correspond to quartiles of the 30-day CMS mortality rate instrument after netting out a ZIPxpatient origin fixed effect. All estimates are expressed in terms of the fraction of the sample with each characteristic. Sample size = 546,700

Table 2: OLS Results

Outcome:		30D Readmission	30D Mortality	365D Mortality
Quality Measure	Mean [SD]	(1)	(2)	(3)
Timely and Effective Care Composite	92.63	0.001	-0.001	-0.004
	[5.08]	(0.002)	(0.002)	(0.002)
Patient Experience Composite	66.41	-0.006	-0.003	-0.006
	[5.04]	(0.002)	(0.002)	(0.002)
30-Day Mortality Rate Composite	12.68	-0.003	0.011	0.014
	[1.29]	(0.002)	(0.002)	(0.002)
30-Day Readmission Rate Composite	20.96	0.015	-0.002	0.006
	[1.48]	(0.002)	(0.002)	(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 is provided in the first column to facilitate interpretation on the original scale. Outcome means: 30D Readmission = 15.0%, 30D Mortality = 17.0%, 365D 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. Standard errors, clustered at Health Service Area (HSA) level, are reported in parentheses.

Table 3: First Stage Results

	No Comorbidity Controls	With Comorbidity Controls
Quality Measure Instrument	(1)	(2)
Ambulance Avg: Timely and Effective Care Composite	0.426	0.426
	(0.004)	(0.004)
Ambulance Avg: Patient Experience Composite	0.558	0.558
	(0.004)	(0.004)
Ambulance Avg: 30-Day Mortality Rate Composite	0.550	0.549
	(0.003)	(0.003)
Ambulance Avg: 30-Day Readmission Rate Composite	0.572	0.572
	(0.003)	(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 is provided in the first column to facilitate interpretation on the original scale. Outcome means: 30D Readmission = 15.0%, 30D Mortality = 17.0%, 365D 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. 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:		30D Readmission	30D Mortality	365D Mortality
Quality Measure	Mean [SD]	(1)	(2)	(3)
Timely and Effective Care Composite	92.63	0.003	-0.017	-0.038
	[5.08]	(0.011)	(0.011)	(0.014)
Patient Experience Composite	66.41	-0.019	-0.005	-0.028
	[5.04]	(0.008)	(0.008)	(0.011)
30-Day Mortality Rate Composite	12.68	-0.013	0.021	0.025
	[1.29]	(0.007)	(0.008)	(0.010)
30-Day Readmission Rate Composite	20.96	0.029	-0.007	0.005
	[1.48]	(0.007)	(0.008)	(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 is provided the first column to facilitate interpretation on the original scale. Outcome means: 30D Readmission = 15.0%, 30D Mortality = 17.0%, 365D 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. Standard errors, clustered at Health Service Area (HSA) level, are reported in parentheses.

Table 5: 2SLS Results

Outcome:	Mean [SD]	30D Readmission	30D Mortality	365D Mortality
Panel A. Add Comorbidity Controls				
Timely and Effective Care Composite	92.63	0.002	-0.018	-0.040
	[5.08]	(0.011)	(0.011)	(0.015)
Patient Experience Composite	66.41	-0.018	-0.005	-0.027
	[5.04]	(0.008)	(0.008)	(0.011)
30-Day Mortality Rate Composite	12.68	-0.011	0.025	0.032
	[1.29]	(0.007)	(0.007)	(0.011)
30-Day Readmission Rate Composite	20.96	0.028	-0.008	0.003
	[1.48]	(0.008)	(0.007)	(0.010)
Panel B. Horse-Race				
Timely and Effective Care Composite	92.63	0.008	-0.017	-0.034
	[5.08]	(0.011)	(0.011)	(0.015)
Patient Experience Composite	66.41	-0.014	-0.004	-0.021
	[5.04]	(0.008)	(0.009)	(0.011)
30-Day Mortality Rate Composite	12.68	-0.010	0.024	0.030
	[1.29]	(0.007)	(0.007)	(0.011)
30-Day Readmission Rate Composite	20.96	0.024	-0.009	-0.004
	[1.48]	(0.008)	(0.008)	(0.011)
Panel C. CMS Condtion Sample (AMI,PN,HF)				
Timely and Effective Care Composite	92.57	-0.008	-0.031	-0.047
	[5.12]	(0.031)	(0.026)	(0.034)
Patient Experience Composite	66.33	-0.023	-0.041	-0.063
	[5.05]	(0.018)	(0.020)	(0.026)
30-Day Mortality Rate Composite	12.66	-0.023	0.041	0.046
	[1.31]	(0.017)	(0.017)	(0.025)
30-Day Readmission Rate Composite	21.01	0.024	-0.018	-0.000
	[1.49]	(0.019)	(0.018)	(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' specificiation 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. 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 is provided the first column to facilitate interpretation on the original scale. Outcome means for non-discretionary condition sample: 30D Readmission = 15.0%, 30D Mortality = 17.0%, 365D Mortality = 37.2%. Sample sizes for non-discretionary condition sample: 546,700 (30-Day outcomes), 451,503 (1-Year Mortality). Outcome means for CMS condition sample: 30D Readmission = 19.0%, 30D Mortality = 15.8%, 365D 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.

Table 6: Overall Quality Composite Measure: 2SLS Results

Outcome:		30D Readmission	30D Mortality	365D Mortality
Quality Measure	Mean [SD]	(1)	(2)	(3)
Composite Quality Measure				
30-Day Readmission Outcome Composite	-0.00	-0.032		
	[1.00]	(0.008)		
30-Day Mortality Outcome Composite	0.00		-0.030	
	[1.00]		(0.007)	
1-Year Mortality Outcome Composite	-0.00			-0.048
	[1.00]			(0.010)
Year and Diagnosis Controls		Yes	Yes	Yes
Full Controls		Yes	Yes	Yes

Each cell reports ordinary 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 is provided the first column to facilitate interpretation on the original scale. Outcome means: 30D Readmission = 15.0%, 30D Mortality = 17.0%, 365D Mortality = 37.2%. Sample sizes: 546,700 (30-Day outcomes), 451,503 (1-Year Mortality). 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. Standard errors, clustered at Health Service Area (HSA) level, are reported in parentheses.

Table 7: Association Between Composite Quality Scores and Structural Hospital Characteristics, 2012 Values

	Timely and Effective Care	Patient Satisfaction	30D Mortality	30D Readmission
	(1)	(2)	(3)	(4)
Teaching Hospital	0.427	−0.027	−0.140	0.014
	(0.166)	(0.239)	(0.066)	(0.076)
Hospital Beds (Standardized)	0.353	−0.518	−0.214	0.175
	(0.078)	(0.112)	(0.031)	(0.035)
Ownership: Investor-Owned	1.013	−2.105	0.263	0.356
	(0.175)	(0.251)	(0.070)	(0.080)
Ownership: Public	−1.590	0.002	0.587	0.016
	(0.211)	(0.303)	(0.084)	(0.096)
Constant	96.540	69.790	12.880	20.930
	(0.099)	(0.143)	(0.040)	(0.045)

Table shows results of a "horse race" OLS regression of each quality measure on hospital characteristics. The dependent variable in each regression is expressed in terms of standard deviations. Sample size = 2,116 general acute care hospitals.

Table A1: CMS Quality Measures Used to Construct Composite Domain Scores

Domain	Measure
Timely and Effective Care	
	Heart Failure Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction (LVSD)
	AMI Patients Given Aspirin at Discharge
	Heart Failure Patients Given Assessment of Left Ventricular Function (LVF)
	Heart Failure Patients Given Discharge Instructions
	Pneumonia Patients Given the Most Appropriate Initial Antibiotic(s)
	Surgery Patients Who Received Preventative Antibiotic(s) One Hour Before Incision
	Surgery Patients Whose Preventative Antibiotic(s) are Stopped Within 24 hours After Surgery
Patient Experience of Care	
	Doctors always communicated well
	Nurses always communicated well
	Pain was always well controlled
	Patients always received help as soon as they wanted – Patients who gave a rating of 9 or 10 (high)
	Room was always clean
	Room always quiet at night
	Staff always explained
	Yes, patients would definitely recommend the hospital
	Yes, staff did give patients this information
30-Day Mortality Rates	
	AMI 30-Day Mortality Rate
	Pneumonia 30-Day Mortality Rate
	Heart Failure 30-Day Mortality Rate
30-Day Readmission Rates	
	AMI 30-Day Readmission Rate
	Pneumonia 30-Day Readmission Rate
	Heart Failure 30-Day Readmission Rate

Source: Archived Hospital Compare website data. The process and outcome measures are defined for specific patient conditions (e.g., AMI, pneumonia, heart failure) whereas the HCAHPS patient satisfaction surveys are distributed to a wider range of patients.

Table A2: Balance: Non-Deferrable Condition Sample

Measure	Quartile 1	Quartile 2	Quartile 3	Quartile 4
038 Septicemia	0.154	0.154	0.153	0.152
162 Malignant neoplasm of trachea, bronchus, and lung	0.007	0.008	0.008	0.008
197 Secondary malignant neoplasm of respiratory and digestive systems	0.004	0.004	0.004	0.004
410 Acute myocardial infarction	0.083	0.083	0.084	0.081
431 Intracerebral hemorrhage	0.013	0.013	0.013	0.013
433 Occlusion and stenosis of precerebral arteries	0.009	0.009	0.009	0.009
434 Occlusion of cerebral arteries	0.081	0.083	0.083	0.082
435 Transient cerebral ischemia	0.028	0.029	0.029	0.029
482 Other bacterial pneumonia	0.012	0.012	0.012	0.012
486 Pneumonia, organism unspecified	0.103	0.104	0.105	0.106
507 Pneumonitis due to solids and liquids	0.038	0.036	0.037	0.038
518 Other diseases of lung	0.050	0.052	0.051	0.051
530 Diseases of esophagus	0.011	0.010	0.011	0.010
531 Gastric ulcer	0.010	0.010	0.010	0.010
532 Duodenal ulcer	0.007	0.007	0.008	0.007
557 Vascular insuciency of intestine	0.007	0.008	0.007	0.007
558 Other and unspecified noninfectious gastroenteritis and colitis	0.008	0.008	0.008	0.008
560 Intestinal obstruction without mention of hernia	0.024	0.024	0.023	0.022
599 Other disorders of urethra and urinary tract	0.083	0.082	0.082	0.083
728 Disorders of muscle, ligament, and fascia	0.009	0.009	0.009	0.009
780 General symptoms	0.085	0.084	0.084	0.085
807 Fracture of rib(s), sternum, larynx, and trachea	0.007	0.007	0.007	0.007
808 Fracture of pelvis	0.017	0.016	0.016	0.016
820 Fracture of neck of femur	0.128	0.127	0.124	0.128
823 Fracture of tibia and fibula	0.005	0.005	0.005	0.005
824 Fracture of ankle	0.011	0.010	0.011	0.010
959 Injury, other and unspecified	0.002	0.002	0.002	0.002
965 Poisoning by analgesics, antipyretics, and antirheumatics	0.003	0.003	0.003	0.002
969 Poisoning by psychotropic agents	0.002	0.002	0.002	0.002

Table shows balance across diagnoses in non-deferrable condtiion sample. Columns correspond to quartiles of the 30-day CMS mortality rate instrument after netting out a ZIPxpatient origin fixed effect. All estimates are expressed in terms of the fraction of the sample with each characteristic. Sample size = 546,700.

Ash, Arlene S., Stephen F. Fienberg, Thomas A. Louis, Sharon-Lise T. Normand, Therese A. Stukel, and Jessica Utts. 2012. "Statistical Issues in Assessing Hospital Performance." 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. 2015. "National Hospital Ratings Systems Share Few Common Scores And May Generate Confusion Instead Of Clarity." *Health Affairs* 34 (3): 423–30. doi:10.1377/hlthaff.2014.0201.

Berenson, Robert A., Ronald A. Paulus, and Noah S. Kalman. 2012. "Medicare's Readmissions-Reduction Program — A Positive Alternative." *New England Journal of Medicine* 366 (15): 1364–6. 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. 2002. "Hospital Volume and Surgical Mortality in the United States." *New England Journal of Medicine* 346 (15): 1128–37. doi:10.1056/NEJMsa012337.

Card, David, Carlos Dobkin, and Nicole Maestas. 2007. "Does Medicare Save Lives?" Working Paper 13668. National Bureau of Economic Research. http://www.nber.org/papers/w13668.

Chandra, Amitabh, and Jonathan S. Skinner. 2011. "Technology Growth and Expenditure Growth in Health Care." NBER Working Paper 16953. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/16953.html.

Chandra, Amitabh, and Douglas O. Staiger. 2007. "PRODUCTIVITY SPILLOVERS IN HEALTHCARE: EVIDENCE FROM THE TREATMENT OF HEART ATTACKS." *The Journal of Political Economy* 115: 103–40. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2311510/.

Chen, Lena, Anup Das, Jun Li, and Edward Norton. 2016. "Moneyball in Medicare." *Working Paper*.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates †." *American Economic Review* 104 (9): 2593–2632. doi:10.1257/aer.104.9.2593.

Chiang, Arthur J., Guy David, and Michael Gene Housman. 2006. "The Determinants of Urban Emergency Medical Services Privatization." *Critical Planning* 13. http://papers.ssrn.com/

sol3/papers.cfm?abstract_id=918525.

Daley, Jennifer. 2002. "Invited Commentary: Quality of Care and the Volume-Outcome Relationship—What's Next for Surgery?" *Surgery* 131 (1): 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. 2016. "Adding A Spending Metric To Medicares Value-Based Purchasing Program Rewarded Low-Quality Hospitals." *Health Affairs* 35 (5): 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. 2016. "Association Between Hospital Penalty Status Under the Hospital Readmission Reduction Program and Readmission Rates for Target and Nontarget Conditions." *JAMA* 316 (24): 2647. doi:10.1001/jama.2016.18533.

Dobkin, Carlos. 2003. "Hospital Staffing and Inpatient Mortality." Unpublished Working Paper.

Doyle, Joseph J., John A. Graves, Jonathan Gruber, and Samuel A. Kleiner. 2015. "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns." *Journal of Political Economy* 123 (1): 170–214. doi:10.1086/677756.

Dudley, R. Adams, Kirsten L. Johansen, Richard Brand, Deborah J. Rennie, and Arnold Milstein. 2000. "Selective Referral to High-Volume Hospitals: Estimating Potentially Avoidable Deaths." *Jama* 283 (9): 1159–66. http://jama.jamanetwork.com/article.aspx?articleid=192451.

Fisher, Elliott S., Julie P. Bynum, and Jonathan S. Skinner. 2009. "Slowing the Growth of Health Care Costs — Lessons from Regional Variation." *New England Journal of Medicine* 360 (9): 849–52. doi:10.1056/NEJMp0809794.

Fisher, Elliott S., David E. Wennberg, Thrse A. Stukel, Daniel J. Gottlieb, F. L. Lucas, and Étoile L. Pinder. 2003a. "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care." *Annals of Internal Medicine* 138 (4): 273–87. doi:10.7326/0003-4819-138-4-200302180-00006.

———. 2003b. "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care." *Annals of Internal Medicine* 138 (4): 288–98. 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. 2012. "The Hospital Standardized Mortality Ratio Fallacy: A Narrative Review." *Medical Care* 50 (8): 662–67. doi:10.1097/MLR.0b013e31824ebd9f.

Halm, Ethan A., Clara Lee, and Mark R. Chassin. 2002. "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature." *Annals of Internal Medicine* 137 (6): 511–20. http://annals.org/article.aspx?articleid=715648.

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

Hull, Peter. 2016. "Estimating Hospital Quality with Quasi-Experimental Data." http://www.mit.edu/~hull/JMP.pdf.

Johnson, Robin. 2001. *The Future of Local Emergency Medical Service Ambulance Wars 1 Or Public-Private Truce?* Reason Public Policy Institute.

Lilford, R., and P. Pronovost. 2010. "Using Hospital Mortality Rates to Judge Hospital Performance: A Bad Idea That Just Won't Go Away." *BMJ* 340 (apr19 2): c2016–c2016. doi:10.1136/bmj.c2016.

Luft, Harold S., Sandra S. Hunt, and Susan C. Maerki. 1987. "The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns?" *Health Services Research* 22 (2): 157. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065430/.

Maxwell, Bryan G., Jim K. Wong, D. Craig Miller, and Robert L. Lobato. 2014. "Temporal Changes in Survival After Cardiac Surgery Are Associated with the Thirty-Day Mortality Benchmark." *Health Services Research* 49 (5): 1659–69. doi:10.1111/1475-6773.12174.

McGlynn, Elizabeth A., Steven M. Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A. Kerr. 2003. "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine* 348 (26): 2635–45. doi:10.1056/NEJMsa022615.

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

Shahian, David M., and Sharon-Lise T. Normand. 2003. "The Volume-Outcome Relationship:

From Luft to Leapfrog." *The Annals of Thoracic Surgery* 75 (3): 1048–58. http://www.sciencedirect.com/science/article/pii/S0003497502043084.

———. 2008. "Comparison of 'Risk-Adjusted' Hospital Outcomes." *Circulation* 117 (15): 1955–63. http://circ.ahajournals.org/content/117/15/1955.short.

Shahian, David M., Gregg S. Meyer, Elizabeth Mort, Susan Atamian, Xiu Liu, Andrew S. Karson, Lawrence D. Ramunno, and Hui Zheng. 2012. "Association of National Hospital Quality Measure Adherence with Long-Term Mortality and Readmissions." *BMJ Quality & Safety* 21 (4): 325–36. doi:10.1136/bmjqs-2011-000615.

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

Song, Yunjie, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E. Wennberg, and Elliott S. Fisher. 2010. "Regional Variations in Diagnostic Practices." *New England Journal of Medicine* 363 (1): 45–53. doi:10.1056/NEJMsa0910881.

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

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