ARE HOSPITAL QUALITY INDICATORS CAUSAL?

Amitabh Chandra, Harvard University and NBER
Maurice Dalton, NBER
Douglas O. Staiger, Dartmouth and NBER

This Draft: Sep 29, 2023

Hospitals play a key role in patient outcomes and spending, but efforts to improve their quality are hindered because we do not know whether hospital quality indicators are causal or biased. We evaluate the validity of commonly used quality indicators, such as mortality, readmissions, inpatient costs, and length-of-stay, using a quasi-experimental design where hospital closures reallocate large numbers of patients to hospitals of different quality. This setting allows us to measure whether patient outcomes improve as much as quality indicators predict when a relatively low-quality hospital closes, or decline as predicted when a relatively high-quality hospital closes. Using more than 20 years of Medicare claims for over 30 million patients admitted with five common diagnoses, we find that hospital quality indicators overstate differences in the causal impact of hospitals on mortality and readmission rates by about 7 percent, but overstate differences in the causal impact of hospitals on inpatient cost and length-of-stay measures by closer to 40 percent. On average, hospital closures *reduce* patient mortality by shifting patients to higher quality hospitals, but the effect varies widely depending on the relative quality of the closing hospital.

Keywords: Hospital quality, health care productivity,

We are grateful to Jon Skinner and participants at various seminars for many helpful comments. This research was supported by the NIH through a Pilot grant and a P01 grant to Dartmouth and the NBER (P30AG012810 and P01AG019783). The programs used in the paper can be requested from the authors. We are also able to share hospital-level quality measures, provided it meets CMS' requirements for data-release. Staiger is co-founder and has an equity interest in ArborMetrix, a company that provides performance measurement systems in healthcare. Chandra serves on the Congressional Budget Office's Panel of Health Advisors.

Hospitals are the single largest sector within health care in the US and around the world (OECD, 2021), but there is considerable variation in hospital quality indicators.[1] For example, the 30-day risk-adjusted mortality rate for heart-attack patients ranges from 13 to 26 percent at the hospital level, while average 30-day inpatient costs for the same condition range from $5,000 to $20,000, even in Medicare that uses administratively set prices.[2] Variations of this magnitude have meaningful implications for health and spending and motivate efforts to encourage low-performing hospitals to improve, through pay-for-performance, continuous quality improvement, guidelines, and quality report-cards. There are also reallocation efforts to move patients to higher quality providers, by closing low-quality hospitals, or introducing narrow hospital networks and selective referrals. Improvement efforts that act on unvalidated measures will not improve welfare if these measures are contaminated by unmeasured factors such as health behaviors, disease burden, patient-selection, and socio-economic circumstance. A large, international, enterprise is devoted to constructing more quality indicators but the fundamental challenge of knowing whether a measure is causal is not solved by more indicators (Ivers et al. 2012; Metcalfe et al. 2018; Smith et al. 2010; MacLean et al. 2018). Despite a lack of evidence on the validity of these measures, one recent study notes that US hospitals report data on 162 quality indicators at substantial cost to the institution (Saraswathula et al. 2023).

Researchers have encountered this problem in other settings where observational data are used to assess quality. In education, teacher quality measures based on their students' test-score gains ("value added") have been validated using random assignment of teachers to classrooms (Kane and Staiger, 2008; Kane et al., 2013) and quasi-experimental designs using changes in teaching staff at a school (Chetty et al., 2014; Bacher-Hicks et al., 2015). These studies have found that value added measures have *predictive validity* (students have better outcomes when assigned to higher value-added teachers), and also are *forecast unbiased* (the gain in student test scores is equal to that predicted by teacher value added). If quality indicators are forecast unbiased, implying that actual quality differences are as large as quality indicators suggest, then policies that use these indicators for improvement or reallocation are more likely to succeed.

To determine whether hospital quality indicators are causal, we exploit the reassignment of patients as a result of a hospital closure to create quasi-experimental variation in the quality of care that

[1] There are 5,100 community hospitals in the US which account for over 33.3 million admissions annually (AHA, 2022). Spending on hospitals exceeds $1.3 trillion annually.

[2] See Appendix Figure 1, which illustrates the variation in hospital-level quality indicators such as 30-day mortality, readmissions, cost, and length-of-stay. The quality indicators are empirical-Bayes estimates and adjusted for differences in patient risk factors and for estimation error.

patients receive— if quality indicators are causal, patient outcomes ought to improve when a low-quality hospital closes, and worsen when a high-quality hospital closes (predictive validity). Moreover, the magnitude of the change in patient outcomes should be consistent with the predicted change in hospital quality induced by a closure (forecast unbiasedness). Our analysis parallels Chetty et al. (2014), who use a teacher-switching design to estimate whether test scores improve in a school-grade when a high performance teacher replaces a low performance teacher, or vice versa. In our analysis, we similarly estimate whether hospital quality indicators correctly predict the change in patient outcomes from the year before to the year after closure.

Our approach proceeds in three steps. First, we construct an estimate of hospital quality indicators from observational data (we refer to these observational, unvalidated measures as quality indicators) using empirical-Bayes methods similar to those used in Chetty et al. (2014) and Staiger et al. (2009). Our method is a generalization of the methods Medicare uses to construct hospital quality indicators that allows for hospital quality to depend on patient volume and to drift over time. Second, we use hospital closures as an exogenous change in the hospital quality that patients are exposed to, forming an ex ante prediction of how outcomes would change for patients living in each zip code as a result of a closure if the quality indicators were causal. Closures change predicted outcomes for affected zip codes through three channels-- the quality of the closing hospital, the quality of the non-closing hospitals that absorb patients, and the market share of the closing hospital. Finally, we regress the actual change in outcomes on the predicted change in outcomes, testing whether the regression yields a coefficient of 1, which would happen if the indicators were perfectly forecast unbiased.

Closure induced changes in quality serve as a good instrument for hospital quality for several reasons: they generate large changes in hospital market shares in zip codes previously served by the closing hospital (as seen in Figure 1); they are unlikely to be associated with changes in patient characteristics in the years immediately surrounding the closure (and we can test this); and they identify a policy-relevant LATE, because they capture the impact of restricting hospital choice (similar to narrow networks) and relocating substantial number of patients to alternative hospitals, as opposed to moving a single patient from one hospital to another. We rely on the sharp change in market share induced by hospital closure for identification rather than the change in market share among existing hospitals because changes in market share among existing hospitals are likely to be endogenous; reallocation due to closure is generated by a change in the choice set, whereas reallocation among existing hospitals is generated by the same selection process which may be biasing quality indicators in the first place. Thus, the reallocation induced by hospital closures at the moment of closing generates ideal variation to test the

validity of hospital quality measures in a general-equilibrium policy-relevant setting.

Our analysis uses over 20 years of Medicare claims from over 30 million patients admitted to the hospital in five large diagnostic cohorts: heart attacks, hip fractures, pneumonia, congestive heart failure and strokes. We evaluate hospital quality measures for four patient outcomes: 30-day mortality, 30-day readmissions, 30-day total inpatient (Part A) costs, and length of stay. High mortality and readmission are indicative of low quality care, while high costs and length of stay are indicative of high resource use.

We find that hospital quality indicators using standard claims-based risk adjustment are valid predictors of hospitals' causal effects for all four of these patient outcomes. For mortality and readmissions these measures are approximately forecast unbiased, overstating quality differences by around 7 percent or less depending on the specification. In contrast, while quality indicators for costs and length of stay are significantly related to changes in patient outcomes following closures, they overstate quality differences by around 40 percent.. Thus, while not perfect, risk-adjusted quality measures based on patient claims nevertheless identify substantial causal differences across hospitals. We offer a variety of robustness tests for these results. They are robust to including a variety of controls for pre-closure zip code demographics, and controls for the closing hospital's market share and changes in distance to the nearest hospital, all of which might be expected to influence changes in outcomes following a closure. We also show that quality indicators that are not risk-adjusted have reasonable predictive-validity but are not forecast-unbiased: unadjusted measures overstate mortality by 40 percent, readmissions by 30 percent, and costs and length of stay by 50 percent.

Finally, we find that closures reduce mortality and readmission, even in our sample of high-mortality conditions for which rapid care is necessary for survival: On average, closing hospitals were lower performing relative to the hospitals that their patients went to after closure. Among all patients that were affected by a closure between 1994 and 2013, we estimate that closures reduced mortality by 0.2 percentage points (off a base of 13%) and reduced readmission by 0.1 percentage points (off a base of 18%), while raising costs and length of stay by about 1%. However, the effect of closure varied widely by zip code depending on the relative quality of the closing hospital. For example, mortality among affected patients in zip codes at the 10th percentile fell by about 1.5 percentage points, while at the 90th percentile they rose by about 1 percentage point. Black patients were more likely to be affected by closures, but among black patients affected by a closure the distribution of impacts on patient outcomes was similar to the impacts among all patients. In a counterfactual simulation, we estimate that closing hospitals that were below average quality in their Hospital Referral Region (HRR) on all four of our

quality indicators in 2012 would affect about a quarter of all admissions, and yield a reduction in mortality and readmission rates among these affected admissions of 0.9 percentage points, while reducing costs and length of stay by about 1%. Thus, reallocating patients to higher quality hospitals within a region could improve quality of care while reducing resource use.

Our work builds on recent efforts to measure causal effects in the healthcare industry. It is most closely related to Doyle et al. (2019) who use ambulance referral patterns to validate the predictive validity for a range of hospital performance measures, but do not test if their measures are forecast unbiased.[3] One important difference between our work and prior work validating quality indicators for hospitals (Doyle et al., 2019) and geographic areas (Finkelstein et al., 2016, 2019) is that the earlier work estimated the impact of changing assignment for a single patient through ambulance referrals or migration, which is a partial-equilibrium response, that may be different than the effect of changing allocation for a large group of patients, which is a general-equilibrium response. The partial-equilibrium response may differ from the general-equilibrium response in important ways that depend on the share of the market being reallocated (the market share of the closing hospital), the relative difference in quality between the closing and non-closing hospitals, and the ability of these hospitals to absorb patients without reducing quality. In our setting, we see the typical hospital closure reallocating just over 15 percent of a zip code's patients, which makes our estimates more relevant for policies such as closure and narrow networks that reallocate a substantial portion of patients to different hospitals, where there may be general equilibrium impacts. In this way, our estimates are analogous to Abaluck et al. (2021) who use the exit of health-insurers to validate mortality differences across Medicare Advantage plans.

A second difference for the hospital setting, acknowledged by Doyle et. al (2019), is that it is possible that hospital's chosen by ambulance drivers differ systematically from hospitals selected by non-ambulance drivers (such as patients, family members or referrals from primary-care physicians). About 17 percent of emergency room visits arrive via ambulance with 17 percent of these being admitted to the hospital-- in other words, the ambulance design applies to three percent of all admissions. Our design exploits reassignment of all patients as a result of closures, and is based on patient (rather than ambulance driver) choice of hospital following closure which is more relevant for most policy analyses. As in Abaluck et al. (2021) the causal effect we measure is the policy-relevant effect of a closure, which may be different than the causal effect from randomly assigning patients to hospitals: The two effects will

[3] In healthcare, Finkelstein et al. (2016, 2019) use patient migration to validate geographic variation in costs and mortality, while Fadlon and Van Parys (2020) use exit of physicians to validate primary care quality indicators and Abaluck et al. (2021) use the exit of health-insurers to validate mortality differences across Medicare Advantage plans.

be equal only if closures reallocate patients randomly across hospitals conditional on observed quality (this is the "fallback condition" from Abaluck et. al, 2021 that we discuss later).

Our analysis proceeds as follows. In Section I we introduce the data and setting for our analysis, including the characteristics of closing hospitals. In Section II we describe the estimation of hospital measures. In Section III we discuss how closures can be used to validate these quality measures. Section IV contains results including event studies of closures, and a number of robustness exercises. We discuss the implications of our work for the larger literature on closures and hospital quality in Section .

I. Validating Hospital Quality: Setting and Data

Setting

Hospitals serve the sickest patients who need institutionalized care. They are expensive—hospitals represent the single largest component of US health care spending ($1.3 trillion out of $4.1 trillion) and this is true in all major economies (OECD, 2022).[4] Wide variations in hospital outcomes and spending have been noted in the economics, medical, and public-health literatures (Yasaitis et. al, 2009; Doyle et al., 2015; Chandra and Staiger 2020). It is difficult to reduce these variations, either by improvement or through reallocation, without hospital quality indicators that are causal.

The lack of causal measures also affects competition policy, where there are questions about whether hospitals with higher commercial prices secure these prices by producing better outcomes, selecting healthier patients, or obtaining market-power from consolidation (Cooper et al. 2019; Garthwaite et al. 2022). Ambiguity over whether the quality indicators are causal has caused research on hospital mergers and acquisitions to rely on proxy measures of outcomes, like patient-satisfaction scores (Beaulieu et al. 2020). Our results directly inform these evaluations and antitrust retrospectives, by demonstrating the appropriateness of using direct quality indicators such as mortality, readmissions and length-of-stay.

These facts motivate the second reason for studying hospitals-- a number of proposals seek to improve the efficiency of hospital care by encouraging patients not to seek care at some hospitals--through narrow networks that impose higher cost-sharing on using them, or by letting hospitals close, or by introducing competition as was done in by the National Health Service in the UK. If the quality measures used to encourage these allocations are flawed, then such proposals are not likely to work as

[4] There are 5,100 community hospitals in the US which accounted for over 33.3 million admissions (AHA, 2022).

promised. On the other hand, if payers have access to causal measures of hospital quality, then it makes it easier to improve outcomes and reduce cost by relying on these measures to construct hospital-networks.

Third, a large industry exists to measure hospital quality but relies on unvalidated methods that might place unfair pressure on hospitals with more complex patients to improve or be penalized, while privileging other hospitals. The most famous of such quality ratings may come from *US News'* annual rating of hospitals whose quality measures combine outcomes like survival, process measures, and patient and expert opinion, with one-third weight to each factor (U.S. News, 2019). Other measures include: hospital-level quality indicators generated by the US Agency for Healthcare Research and Quality; the Hospital Compare project run by the US Department of Health and Human Services; ORYX -- a set of performance measures required by The Joint Commission for accreditation; the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Hospital Survey, a survey measuring adult patients' perspectives on care they experience during a hospital stay; and quality measures from the *Leapfrog Group* an employer-based coalition, which develops measures of hospital quality using safety practices (Leapfrog, 2020, Agency for Healthcare Research and Quality, 2020b, Agency for Healthcare Research and Quality, 2020c, Centers for Medicare & Medicaid Services, 2020, The Joint Commission, 2020). These measures rely on surrogate measures of quality, such as the use of beta-blockers after a heart-attack, or nurse to bed ratios, that may be weakly correlated with direct indicators of quality like mortality, and readmissions. This would imply that the pursuit of improving performance on surrogate measures of quality distracts from the improving more consequential measures.

Fourth, our setting directly informs the effect of hospital closures. These closures are common, in the US and around the world, and often resisted because of the view that some hospital is better than no hospital, especially for acute conditions like heart-attacks and strokes (Holmes et al., 2006, Buchmueller et al, 2006, Joynt et al., 2015, Baicker and Chandra, 2012). There is a large literature on this question but it has not arrived at a consensus estimate. Our findings provide an explanation for the discord: the closure of a relatively high quality hospital (one that is better than the hospitals that patients get reallocated to) harms patients, while the closure of a relatively low quality hospital (one that is worse than the remaining hospitals that patients are now seen at) improves patients health. This heterogeneity has been ignored in prior work on closure.

Data

We use a 100 percent sample of beneficiaries enrolled in Medicare fee-for-service who were

hospitalized between 1992 and 2015 for one of 5 cohorts: heart-attacks, hip-fractures, pneumonia, congestive heart failure (CHF) and stroke. These conditions account for a range of hospitalizations, with heart-attacks and stroke requiring immediate attention, and informing debates about hospital closures being harmful for acute patients who require rapid hospitalization.

As in prior work, index admissions for each cohort were based on patient diagnosis codes and included only those patients that had not been admitted for the same diagnosis within the prior year. We also required enrollees to be in Medicare fee-for-service for 1 year before and after the index event (Chandra et al., 2016; Bekelis et al., 2016; Brauer et al, 2009). For these patients we validate four quality indicators that are commonly used: (1) mortality within 30 days of the index admission; (2) readmission to a hospital within 30 days of a patients discharge (defined only for patients who were discharged alive); (3) Medicare reimbursement for inpatient care within 30 days of hospitalization which is a measure of resource use because Medicare reimburses a fixed amount per service, for example, reimbursement for acute inpatient stays depend on the Diagnosis Related Group (DRG) that the patient was admitted under[5]; and (4) length of stay in the hospital during the index admission.

These restrictions generate a sample of just over 30.24 million patients whose characteristics are described in Table 1. Consistent with our earlier work on these cohorts, we find that on average, patients are over the age of 80, with about 58 percent being females and 11 percent non-White (Chandra et al. 2016a and 2016b). The average patient has just over 1 comorbid condition, with 7 percent having a recorded diagnosis of dementia and 2 percent in nursing facilities. Sixteen percent are dual-eligible, meaning that they are poor enough to be covered by Medicaid which pays for their Medicare copayments and deductibles.

Relative to patients admitted to non-closing hospitals, patients who were admitted to hospitals that closed were 8 percent more likely to be non-white, 25 percent more likely to be dual-eligible, and 13 percent more likely to have a diagnosis for dementia (which is typically Alzheimers). As such, closures may have meaningful implications for racial disparities in care for vulnerable patients. The remaining two columns of Table 1 show patient outcomes separately for hospitals that remained open and hospitals that closed (or converted to a non-acute care hospital) during our sample period.

The disease cohorts are all high-mortality conditions with 30-day mortality averaging 12 percent and ranging from 8 percent for hip-fracture patients to 15 percent for stroke patients. Readmissions rates

[5] Part A costs are taken from the MedPAR file which includes skilled nursing facilities and inpatient stays, including acute, psych and rehab inpatient stays. The payment systems for these facilities differ but share a similar type of payment which reimburses a fixed amount based on the intensity of care received.

are also high, averaging 17 percent, with 13 percent of hip-fracture and stroke patients readmitted within 30 days of the initial hospitalization, while almost 20 percent of heart-attack and congestive heart failure patients readmitted. Readmissions may reflect patients who are unable to comply with discharge orders or problems with the quality of care in the initial hospitalization. Utilization of inpatient medical care is also high among these cohorts, with an average length of stay of 6.5 days and average 30-day inpatient costs of around $13,000.

Defining closures is difficult because traditional sources that list hospital closures have conflicting information about whether a hospital is closed or not (for example, closures identified by the American Hospital Association often differ from closures identified in the CMS Provider of Services files). Many of these discrepancies originate from three different problems. First, hospitals rarely go from being completely open to closed. Rather, they close for some conditions first (e.g. they might stop performing elective surgery). Thus, a hospital may stop treating some groups of patients long before any official closure. The second source of discrepancies is that some hospitals don't close *per se* but may merge with another hospital (and continue operating). These hospitals may continue to operate as before, but stop billing under their identifier and start billing under another hospital's identifier, termed consolidated billing. The third source of discrepancies is that some facilities get converted into long-term care (LTC) facilities, which have an average length of stay over 25 days. These hospitals may continue billing under the same hospital identifier, but no longer provide traditional short-term acute care.

We define a hospital as closing by searching for the last year an index admission or readmission happens for each hospital (this determination is made separately for each patient cohort). This means that for a closure to happen, a hospital should not be admitting patients for all our cohorts.[6] We treat a hospital that converts to a long-term care facility as a closing hospital for it is no longer treating patients in our 5 cohorts. We verify that this closing date is within one year of the hospital closing or conversion to a LTC hospital, with dates reported by either the American Hospital Association (AHA), CMS Provider Of Service File (POS), or CMS Cost report data (HCRIS). The closing dates were assembled with collaborators at the Dartmouth Institute and the file is available as a public resource. This curated file reconciles changes in hospital identifiers due to mergers. If the last year of admission or readmission happens within +1/-1 years of our curated file we use the empirical closure date found in Medicare admissions data. This file is available from the authors as soon as this paper is completed.

With this definition, over 1 million of the 30 million patients in our cohorts were treated at

[6] The last closure that we can use for our analysis is in 2013, because we will examine differences in outcomes from the year before the closure to one year after.

hospitals that closed during our sample period. Relative to all patients in our cohorts, patients at closed hospitals were more likely to be non-White (18%) and dual eligible (21%) suggesting that they served a more disadvantaged population. However, patients at closed hospitals had similar age, gender and comorbidities, and similar mortality and readmission rates, suggesting that they were not markedly sicker than other patients in our sample. Finally, patients at closed hospitals had over one day longer length of stays yet had over $1000 lower 30-day Part A costs. The combination of lower Medicare reimbursement and longer stays could have contributed to financial distress at these hospitals.

Table 2 provides information on the closures in our sample, including the timing of closures. Of the 754 closures in our sample-frame, over 77 percent are from conventional closures and 23 percent are from hospitals that convert to long-term care facilities. The average closing hospital was smaller than average with 98 beds and treated more indigent patients than average with a disproportionate share of 34 percent. 81.7 percent of closing hospitals were located in urban areas. The majority of closures occurred between 1994 and 2004, but there continue to average 10-30 closures per year through 2013.

Key for our identification strategy is that a hospital closure should induce a large change in market-share in the zip codes that the hospital served (the other condition is that there is variation in quality between closing hospitals and absorbing hospitals). Figure 1 plots the change in average zip-code market share for closing hospitals around the time of closure, where each zip code is weighted by the number of patients admitted to the closing hospital, which means that the figure can be interpreted as the experience of the typical patient experiencing a closure. Thus, the average patient admitted to a closed hospital one year before closure lived in a zip code in which the closed hospital had 17% market share during the 3 years before closure, while 75% of patients admitted to a closed hospital one year before closure lived in a zip code in which the closed hospital had less than 23% market share. In other words, there is large variation in the market share of the closing hospital which is a necessary condition for closures to change quality. In our empirical work we will use the two-year change in market share (from the year before to the year after closing) induced by hospital closure to validate our hospital measures—over this window, centered on the closing date, market share drops sharply to zero (market share is not zero in the closing year because the hospital may have been open for part of the year).

II. Estimating Hospital Quality Indicators

We describe our approach for estimating (unvalidated) quality indicators for hospital h in year t, that may or may not capture causal effects. We use a flexible hierarchical Empirical Bayes model that has

been applied previously to estimate teacher performance (Chetty et al., 2014; Bacher-Hicks et al, 2014) and hospital quality (Staiger et al., 2009; Chen et al., 2013). Hierarchical models are commonly used by Medicare and others to derive "shrinkage" estimates of hospital quality, which adjust for reliability of the estimates in order to better predict true quality. Our model differs from those that are most commonly used in two ways. First, we allow for drift in hospital quality over time (as in Chetty et al., 2014), which is empirically important in our data which spans over two decades. Second, we use patient volume to help predict hospital quality, which has been shown to improve the ability of such measures to forecast quality out of sample (Chen et al., 2013). Intuitively, in an empirical-bayes model, noisy estimates for smaller hospitals are shrunk towards the sample average (which effectively pulls up the quality of smaller hospitals); in our implementation these estimates are shrunk towards a volume-specific average. Prior work has found strong associations between patient outcomes and the log of patient volume (Birkmeyer et. al, 2002) but does not require volume to be a causal determinant of outcomes-- in fact, there is evidence that the causality might flow from outcomes to volume as better hospitals attract more patients (Chandra et al. 2016). Instead, the inclusion of volume, or for that matter, any other correlate of hospital quality, improves the predictive power of the quality indicator. As such, adding more data to Medicare claims data-- such as novel data from genomic sequencing, patient satisfaction, or physician reports-- can improve the predictive power of the observational measures.

Statistical Model

We estimate hospital quality indicators, μ_{ht}, for each of our patient cohorts and for each patient outcome in four steps. Our approach closely parallels Chetty et al. (2014), except that we account for patient volume to improve the forecast of our measures. Our estimation is motivated by a hierarchical model in which patient-level outcomes Y^{*}_{it} depend on a hospital-level intercept in each year μ_{ht} that measures hospital quality, and on patient-level risk-adjusters X_{it} that include patient characteristics such as age, sex, race, and comorbidities.

(1) $$Y^{*}_{it} = \mu_{ht} + X_{it}\gamma_{t} + \epsilon_{it}$$

In this equation, μ_{ht} is often used as the causal effect of hospital *h* in time period *t* on patient outcomes.[7] The key concern with quality indicators derived from Equation (1) is that unobserved patient

[7] Following the literature, we have assumed that comorbidities don't affect outcomes in hospital dependent ways. In Chandra and Staiger (2020) we found support for this assumption using chart-data which is much richer than claims data; and there aren't compelling medical reasons to relax this assumption.

characteristics in ϵ_{it} may be correlated with hospital performance (μ_{ht}) due to self-selection of patients into hospitals.

In the first step of our approach, we estimate (1) which produces estimates of the coefficients on the risk-adjusters $\hat{\gamma}_t$, the variance of the patient-level error $\hat{\sigma}^2_{\epsilon}$ in each year, and the first stage residual $O_{it} = Y^*_{it} - X_{it}\hat{\gamma}_t = \mu_{ht} + \epsilon_{it}$ which includes both the patient residual and the hospital effect.

In the second step, we regress the first-stage residual from (1) on the log of hospital volume:

(2) $$O_{it} = \alpha . ln\left(n_{ht}\right) + \delta_{ht}$$

This regression yields an estimate of the coefficient on volume $\hat{\alpha}$. Let $Y_{it} = O_{it} - \hat{\alpha}\ ln\left(n_{ht}\right) = \delta_{ht} + \epsilon_{it}$ be the patient residual from this regression, and the sum of these patient residuals at the hospital-level $\overline{Y_{ht}} = \frac{1}{n_{ht}} \sum_{i=1}^{n_{ht}} \left(Y_{it}\right) = \delta_{ht} + \bar{\epsilon}_{ht}$ be the average residual for hospital h in year t.

In the third step, we use the results from the first two steps to estimate the variance σ^2_{δ} and correlation ρ_k of the hospital quality residuals δ_{ht}. As in Chetty et. al (2014), we assume that each hospital's residual quality follows a common stationary stochastic process in which $corr(\delta_{ht}, \delta_{ht-k}) = \rho_k$ and $var(\delta_{ht}) = \sigma^2_{\delta}$ (stationary drift). In other words, we assume that the variance in residual quality across hospitals is constant over time and the correlation in residual hospital quality between any two years is only determined by the length of time between the two years. We estimate the variance of the quality residual as $\hat{\sigma}^2_{\delta} = Var\left(Y_{it}\right) - \hat{\sigma}^2_{\epsilon}$, since $Var\left(Y_{it}\right) = \sigma^2_{\delta} + \sigma^2_{\epsilon}$. To estimate the correlation of the quality residuals we use $\hat{\rho}_k = \frac{Cov\left(\overline{Y}_t, \overline{Y}_{t-k}\right)}{\hat{\sigma}^2_{\delta}}$, since $Cov\left(\overline{Y}_t, \overline{Y}_{t-k}\right) = Cov\left(\delta_t, \delta_{t-k}\right)$.

The final step assembles these pieces to form best linear predictions of hospital quality measures ($\hat{\mu}_{ht}$) for each year and hospital using only data from other years (a leave-out approach), that is needed for the validation exercise. Let the vector $\{\overline{Y}\}_h^{-t} = (\overline{Y_{h1}}, ..., \overline{Y_{h,t-1}}, \overline{Y_{h,t+1}}, ..., \overline{Y_{h,T}})'$ be a vector of average residual outcomes for hospital *h* in all years available but excluding the year *t* (more generally, in our

validation exercise we will exclude a window of years around year t). The best linear predictor of hospital h's quality in year t is:

$$(3) \qquad \widehat{\mu_{ht}} = \hat{\alpha}\ ln\left(n_{ht}\right) + \sum_{s \neq t} \psi_{hs} \overline{Y_{hs}} = \hat{\alpha}\ ln\left(n_{ht}\right) + \psi_h' \{\overline{Y}\}_h^{-t} = E(\mu_{ht} |\ ln\left(n_{ht}\right),\ \{\overline{Y}\}_h^{-t})$$

where ψ_{hs} is the *hospital-specific* weight put on the average residual outcome for year s at hospital h, and ψ_h is the vector of these weights.[8] The weight placed on the average hospital residual from other years depends on both the correlation ($\hat{\rho}_k$) between years (intuitively, placing less weight on more distant years), and the reliability of the average hospital residual (intuitively, placing less weight on smaller sample estimates). Thus, equation (3) is a generalization of the usual shrinkage estimator that incorporates drift ($\hat{\rho}_k$)and shrinks back to a conditional mean ($\hat{\alpha}\ ln\left(n_{ht}\right)$). Note that $\hat{\mu}_{ht}$ can be thought of as a jack-knifed estimator of hospital quality because of *omitting the current year* for estimation. Ignoring this approach, and regressing patient outcomes in year *t* on hospital measures that were estimated without leaving out the data from year *t,* would introduce mechanical correlation and bias the forecast validation.

To provide intuition for the weights, consider the case of predicting hospital quality in year t using its value from a single year k years earlier (e.g. $k = t - s$). In this simple case, the *hospital-specific weight* on quality from year s simplifies to:

$$(4) \qquad \psi_{hs} = \hat{\rho}_{t-s} \frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_\epsilon^2 / n_{hs}}$$

The first term in equation (4) accounts for drift in outcomes between periods t and s, while the second term is the ratio of signal variance to total variance in the estimate $\overline{Y}_{hs}$. Thus, equation (4) collapses to standard shrinkage weights from Empirical Bayes when $\rho = 1$.

Figure 2 illustrates the variation in hospital quality indicators ($\hat{\mu}_{ht}$) for the heart-attack cohort in the year 2000 and for our 4 quality indicators we examine - 30-day mortality, readmissions over 30 days, Part A spending over 30 days, and length of stay. We estimated (3) for each outcome in each year, and the figure plots the hospital quality measure against hospital volume (in Appendix Figure 1, we present similar plots for all outcomes and all cohorts in 2000; these relationships are very similar in any year and

[8] In principle, we could use information from patient cohort j to influence with the estimation of hospital indicators for disease cohort k, but there is evidence from the health-services literature that this approach works best when the two cohorts are similar in terms of the care teams that treat them -- e.g. knee and hip replacement patients, or stents for heart-attacks and stents for unstable coronary disease (see Scally et al., 2015, Dimick et al., 2012, Staiger et al., 2009).

there is nothing unique about the year 2000). Three features of these plots are worth emphasizing. First, there is considerable variation in quality indicators across hospitals for each of the outcomes. For example, we predict a nearly two-fold difference in mortality rates across hospitals. Second, there is a strong average relationship between volume and patient outcomes (the solid line), particularly at low volume (the reason for using log volume): higher volume is associated with lower mortality and readmissions, and associated with higher Part A costs and LOS. Third, our approach predicts substantial variation in hospital quality beyond that predicted by volume alone, as has been found in prior work (Staiger et al, 2009; Chen et al., 2013).

Our estimates of quality indicators rely on a drift process where outcomes in year *t* can be differentially predicted by outcomes in *t-1, t-2, t-3*, etc. In Appendix Figure 2 we illustrate the drift parameters for all outcomes and all cohorts. As expected, the correlation declines at longer lags: more distant quality data is less correlated with current quality data. In the Appendix figure we superimpose an AR(1) process on the non-parametric drift process to illustrate that the data are richer than a simple AR(1) but also to highlight that assuming an AR(1) process, which might be necessary in other settings because of less data, would still be improvement over the two extremes of assuming that the quality indicators are uncorrelated or fixed over time.

III. Validating Quality Indicators

We now describe how we use hospital closures to validate the indicators. The key idea is that hospital closure reallocates patients to other hospitals. Because we know the market-share of the closing hospital prior to closure (which is observed), and have estimates of the quality of the closing hospital and non-closing hospitals (which we have as $\widehat{\mu_{ht}}$) , we can form a prediction of how much quality should change as a result of a closure.

To formalize our approach, first, let $\hat{\mu}_{c,h,t+1}^{-\{t-1,t,t+1\}}$ and $\hat{\mu}_{c,h,t-1}^{-\{t-1,t,t+1\}}$ represent the hospital quality estimate for disease-cohort c in hospital h and years t+1 and t-1, constructed as described in Section II and based on data from all years *excluding* the years *t-1, t* and *t+1*. Let $\overline{Q}_{z,c,t+1}$ and $\overline{Q}_{z,c,t-1}$ be the patient-weighted average of these quality measures in each zip code *z* in years *t+1* and *t-1*, i.e. the average predicted quality of hospitals treating patients from cohort *c* and zip code *z*, and $\Delta\overline{Q}_{z,c,t}$ be the difference between the two measures, which is the change in predicted hospital quality from year *t-1* to *t+1* based on

where patients received their care in each of these years. Similarly, let $\overline{O}_{z,c,t+1}$ and $\overline{O}_{z,c,t-1}$ be the patient-weighted average of risk-adjusted patient outcomes (O_{it} as defined in Section II) in each zip code z in years *t-1* and *t+1*, and let $\Delta\overline{O}_{z,c,t}$ represent the change in risk-adjusted patient outcomes from year *t-1* to *t+1*. We can test for forecast bias by regressing changes in risk-adjusted patient outcomes on changes in predicted hospital quality:

$$(5) \qquad \Delta\overline{O}_{z,c,t} = \beta_0 + \beta_1 \Delta\overline{Q}_{z,c,t} + \pi_{c,t} + \varepsilon_{z,c,t}$$

where $\pi_{c,t}$ is a set of cohort x time fixed effects that capture differential trends in outcomes by cohort. Note that since $O_{it} = \mu_{ht} + \epsilon_{it}$, the left hand side of equation (5) is simply the change in the market-share weighted average of true hospital effects ($\Delta\overline{\mu}_{z,c,t}$) plus estimation error. The key explanatory variable is the zip-code change in hospital quality which is a function of hospital quality indicators weighted by their market share.

If hospitals were randomly assigned market shares in *t-1* and *t+1*, then OLS estimates of the coefficient β_1 would determine the validity of hospital quality measures. There are two reasons to be concerned that OLS estimation on (5) is biased. First, changes in market shares are likely to be endogenous because hospitals with better quality have larger market shares and gain more market share over time (Chandra et. al. 2016). Moreover, if patients self-select into hospitals then some of the change in market share over time may reflect changes in patient characteristics. Second, some of the estimated changes in quality for individual hospitals may capture underlying changes in unobserved patient characteristics, e.g. zip codes becoming healthier will appear to have improved quality hospitals because of inadequate risk-adjustment. These effects bias the OLS estimate of β_1 towards 1, but we emphasize that there is nothing mechanical about estimating (5) with OLS, because the key independent variable uses data from outside the time period used to construct the dependent variable.[9]

Identification

Ideally, we would randomly assign patients to hospitals to break the correlation between patient characteristics and market-shares. We simulate the experimental assignment of patients to hospitals by relying on the sudden change in which hospitals treat patients following a hospital closure, i.e. using the

[9] To illustrate, the dependent variable for t=1999 measures the change in outcomes between 1998 and 2000, and includes data from both years, whereas the explanatory variable excludes data from 1998, 1999, 2000.

expected change in hospital quality in each zip code due to hospital closure as an instrument. More specifically, let $\hat{\mu}^{-\{t-1,t,t+1]}_{c,closed,t-1}$ be the hospital quality estimate from cohort c in year t-1 for a hospital that closed in year t, and let $\overline{Q}^{\ -closed}_{z,c,t-1}$ be the patient-weighted average of performance measures in year t-1 across the remaining hospitals in each zip code (which did not close). Thus, $\overline{Q}^{\ -closed}_{z,c,t-1} - \hat{\mu}^{-\{t-1,t,t+1]}_{c,closed,t-1}$ represents the predicted difference in quality between the average non-closed hospital serving the zip code and the closed hospital. This difference captures the predicted impact on the outcomes of patients admitted to the closed hospital in year t-1 had they instead been admitted to an average of the other hospitals serving their zip code. For example, if the closed hospital had higher mortality than average this difference is negative, and those patients would be expected to have lower mortality if they were treated at the other hospitals. Multiplying this difference in outcomes by the closed hospital's market share in the year before closure ($\omega^{\ closed}_{c,z,\,t-1}$) yields the expected change in hospital quality (across all patients) in each zip code due to hospital closure in year t ($E_{z,c,t}$, our instrument for $\Delta\overline{Q}_{\ z,c,t}$ in equation 5):

$$(6) \qquad E_{z,c,t} = \omega^{\ closed}_{c,z,\,t-1} * \left\{ \overline{Q}^{\ -closed}_{z,c,t-1} - \hat{\mu}^{-\{t-1,t,t+1]}_{c,closed,t-1} \right\}$$

The expected change in hospital quality following a closure depends on two components that are worth understanding separately. First, it depends on the market share of the closing hospital in each zip code in the year before closure, which was estimated to be about 16 percent in Figure 1 for the typical patient affected by a closure. Thus, because closed hospitals account for a substantial market share in the zip codes they serve, closure has the potential to impact patient outcomes. The second component in equation (6) is the difference in quality between the remaining hospitals and the closing hospital; the larger this difference, in either direction, the larger the effect of a closure.

Figure 3 graphs the distribution of the quality difference between the closed hospital and the remaining hospitals in zip codes with closing hospitals (this is the part of (6) in parenthesis, but graphed as the negative of this quantity), weighting each zipcode by the number of patients who were seen at the closing hospital in year t-1. There are many zip codes where the difference in mortality between closing hospitals and surrounding hospitals was 5 percentage points or larger. In these zipcodes the instrument will predict that a closure will reduce mortality because of reallocating patients to lower mortality hospitals. Similarly, the gap in Part A spending between closing and non-closing hospitals exceeds 50 percent in many zip codes. These histograms would yield the right instrument if the market share of the closing hospital was 100 percent, but we know from Figure 1 that the typical closure affects about 17

percent of patients. The correct instrument ($E_{z,c,t}$) therefore, is the product of the closing hospitals' market share with the variation in Figure 3.

The expected change in hospital quality due to closure ($E_{z,c,t}$) is a valid instrument if it is uncorrelated with the error in equation (5). There are two ways in which this may break down. Our strategy assumes that changes in unobserved patient risk factors in a zip code are uncorrelated with whether a relatively high or low quality hospital closed. This will be true if closure is exogenous to changes in patient risk (unlike normal changes in hospital market share, which may respond to changes in patient risk due to patient sorting). We evaluate this assumption by (1) looking at whether changes in observed patient risk-factors are correlated with our instrument; and (2) estimating event studies to see whether the instrument is correlated with changes in patient outcomes prior to closure. Neither of these checks suggest that the quality of the closing hospital is associated with changes in patient risk.

A second problematic correlation would be if the instrument is correlated with changes in the unforecastable component of hospital quality effects. We must assume that patients displaced from a relatively high quality hospital do not choose alternative (fallback) hospitals that differ systematically from patients displaced from relatively low quality hospitals in terms of this unforecastable component – what Abaluck et al. (2021) label the "fallback condition." Failure of the fallback condition would occur, for example, if patients displaced from high quality hospitals were better able to discern true hospital quality than those displaced from relatively low quality hospitals (so had less of a decline in true hospital quality than indicated by the change in their observed hospital quality). While failure of the fallback condition would introduce bias in the second stage, the reduced form equation would still be of policy interest – providing estimates of the net effect of an expected change in hospital quality due to closure ($E_{z,c,t}$) on patient outcomes.

Estimation

We report OLS and 2SLS estimates of equation (5) for each outcome (mortality, readmission, log of Part A costs, and log of LOS). For each outcome, we stack the cohorts and estimate a pooled model across cohorts that includes HRR-year-by-cohort fixed-effects (and in some specifications year-cohort fixed effects) to account for differential time and geographic trends in patient outcomes across the cohorts. The equation is estimated at the zip-code level which allows a given hospital to affect patients in multiple zip codes. By focusing on variation at the zip code level we avoid bias due to self-selection of patients into hospitals. To minimize potential bias due to patients self-selecting into zip codes the regression is specified in changes to remove location fixed-effects, and zip code in both *t-1* and *t+1* is

based on each patient's zip code of residence in *t-1*. As noted before, by excluding patient outcomes from the years *t-1, t* and *t+1* in constructing our hospital quality indicators, we avoid any mechanical correlation between the quantities on the left- hand side and right-hand side of equation (5).

If $\beta_1 > 0$ in the 2SLS results, quality indicators have *predictive validity* for they are informative about how exogenous changes in hospital market share will affect patient outcomes. If the 2SLS results do not allow us to reject $\beta_1 = 1$ with informative confidence-intervals, then the performance indicators are *forecast unbiased* because they accurately forecast the magnitude of the change in patient outcomes (i.e. are well-calibrated). Note that even if a quality indicator is forecast unbiased, it may not predict all of the systematic variation due to the true hospital effects $\Delta\bar{\mu}_{z,c,t}$ because there is remaining forecast error that could be reduced using alternative forecasts of hospital quality incorporating more information.

IV. Results

Estimates of Forecast Bias

Table 3 presents our main results estimating Equation (5) to test whether quality indicators are forecast unbiased. Each panel of the table contains estimates for a different outcome: mortality, readmission, log Part A costs, and log length of the initial hospital stay. For every outcome, we present the OLS and 2SLS estimates of the slope coefficient (β_1) from Equation (5), which regresses changes in risk-adjusted patient outcomes $\left(\Delta\overline{O}_{z,c,t}\right)$ on the change in hospital quality $\left(\Delta\overline{Q}_{z,c,t}\right)$, using the expected change in hospital quality following closure $\left(E_{z,c,t}\right)$ as the instrument for 2SLS. We also report estimates of the first-stage (regressing $\Delta\overline{Q}_{z,c,t}$ on the instrument $E_{z,c,t}$) and reduced form (regressing $\Delta\overline{O}_{z,c,t}$ on the instrument $E_{z,c,t}$). In these regressions, the number of observations reflects the total number of unique cohort-hrr-zipcode-year cells in the regression.[10] Finally, all regressions are weighted by the number of patients in the cell, and cluster standard errors at the level of 306 Hospital Referral Regions, that are commonly used health-care markets that are built up from zip-codes, based on patient zip code in t-1 (Chandra and Staiger 2007). This format is also used to illustrate the robustness of our results (in Tables 4 (Panel A and B) and Table 5), where we focus on the 2SLS results but report the others for completeness.

Four facts are apparent in Table 3. First, the OLS estimates yield precisely estimated coefficients

[10] Note that a cell must have at least one admission in year t-1 and year t+1 to be included in the regression. As a result, the number of observations are slightly smaller for readmissions (since a patient must be discharged alive to be counted for readmission rates).

near one for three of the outcomes and .8 for readmission, which would suggest that most of the measures are approximately forecast unbiased.[11] However, OLS estimates are unlikely to be causal because of the endogeneity of market-share. In particular, year-to-year changes in market share are likely to reflect the same patient-selection into hospitals that results in bias in the observational measures, leading OLS estimates to be biased towards one.

Second, and most importantly, the 2SLS estimates of Equation (5) yield coefficient estimates that are less than one but significantly greater than zero for all outcomes, suggesting the indicators have predictive validity but have some forecast bias: actual changes in patient outcomes are somewhat smaller than expected based on changes in average hospital quality following a closure. The 2SLS estimates for mortality and readmissions are smaller than one (around 0.93) and we are unable to reject the null that these hospital quality measures are forecast-unbiased (although confidence intervals are large). For the remaining two indicators of resource utilization we find stronger evidence of bias and smaller coefficients (around 0.60) that are more precisely estimated and significantly below one. All these indicators are still strongly positively associated with actual outcomes ($p<.001$) and therefore have predictive validity.

Third, for each outcome the reduced-form coefficient on the expected change in hospital quality following closure is statistically significantly greater than zero. In other words, the expected change in hospital quality predicts how patient outcomes will change in zip codes affected by the closure in the year following closure. From a pure policy perspective this is an important result: Hospital quality measures and market shares can be used to forecast the impacts of potential hospital closures on patient outcomes. We exploit this result in the last section of the paper.

Finally, for each outcome there is a strong first stage, with t-statistics on the instrument of at least 10 (so first-stage F-statistic exceeds 100). The expected change in hospital quality following closure ($E_{z,c,t}$, which is based on performance and market share from the year before closure) is a reasonably accurate proxy for the actual change in hospital quality from the year before to the year after closure $\left(\Delta\overline{Q}_{z,c,t}\right)$, with a coefficient ranging from 0.56 to 0.64.

Overall, the 2SLS estimates suggest that outcome indicators that reflect direct patient outcomes such as mortality and readmissions are less biased than utilization measures such as length of stay and spending, even though all measures have predictive validity. This finding challenges the conventional wisdom in health-services research that it's easier to risk-adjust measures of utilization than measures of

[11] Note that the OLS coefficient on readmissions is relatively smaller than the other OLS estimates. This is because there is a difference between the market-share of admissions and discharges-- which the readmissions indicator uses as a denominator-- and that creates measurement-error. This measurement-error goes away in the 2SLS.

health outcomes.

To illustrate our approach graphically, Figure 4 (Panel A) presents binned scatter plots that represent non-parametric versions of the 2SLS results for each of the outcomes.[12] Panel B of Figure 4 uses the same setup to demonstrate that the instrument is not correlated with changes in predicted outcomes based on how observable patient characteristics changed in the zipcode. In both panels, we ranked zip codes using the prediction from the first stage ($\widehat{\Delta\overline{Q}}_{z,c,t}$) and formed 20 equal sized bins (weighting each zip-code by the number of patients so that each bin has the same number of patients). This is the instrument and is on the x-axis of both panels. We then graphed the bin-level association between this variable and the change in patient outcomes (Panel A) and change in predicted outcomes using only patient characteristics (Panel B) after controlling for year-cohort-hrr fixed effects. Three facts are apparent from these figures. First, the changes in actual patient outcomes match the predicted changes in hospital quality reasonably well throughout the distribution, and the magnitudes from the bin-scatter relationships are similar to the 2SLS estimation in Table 3. In other words, the results in Table 3 are not driven by idiosyncratic data-points. Second, the predicted and actual changes are quite large for the bins at the top and bottom of the distribution. For example, the first stage predicted that mortality would rise by about one percentage point (relative to average mortality of 12% across all the cohorts) for zip codes in the top 5%, and actual mortality rose by a bit more than one percentage point. Similarly, the first stage predicted that Part A costs and length of stay would increase by about .1 log points in the top bin, while actual costs and length of stay increased by about .08 log points. Third, the lack of association in Panel B between the instrument and changes in patient characteristics that predict outcomes in Panel B is reassuring: The lack of association of the instrument with changes in observable patient characteristics supports the assumption that the instrument is not associated with changes in unobservable patient characteristics.

Robustness

One concern with using hospital closures as an instrument is that closures may reflect pre-existing trends in the hospital's market area, such as declining utilization due to growing pressure from local payers to reduce reimbursement. This could lead to bias in the 2SLS estimates of Equation (5) if our instrument (based on closures) was correlated with these local trends. In Table 4A we report estimates similar to the 2SLS estimates in Table 3, but that further control for baseline characteristics of patients from each zipcode (measured in year t-1) that could proxy for such market-area trends: average age,

[12] These plots are equivalent to the second stage from a 2SLS model that uses dummies for 20 bins of $E_{z,c,t}$ as instruments.

%male, %black, and the average number of comorbidities. These estimates are very similar to the 2SLS estimates in Table 3, suggesting that our estimates are not biased by differential trends affecting the types of zip codes served by higher and lower quality closing hospitals.

A second concern with using hospital closures as an instrument for changes in hospital quality is that closures may affect patient outcomes through channels other than the quality of the hospitals that treat patients in a given zip code. First, even if the closed hospital was average quality, closure may lead to worse patient outcomes in zip codes that relied heavily on the closed hospital for their care because patients are forced to switch providers (a disruption effect). Second, hospital closures may lead to worse patient outcomes for zip codes in which the closed hospital was the nearest hospital because patients are forced to travel further to receive urgent care (a delay effect). In Table 4B we control for the market share of the closed hospital in each zip code in the year prior to closure (a proxy for the disruption effect) and for the change in the distance to the nearest hospital for each zip code due to the closure (a proxy for the delay effect). These estimates are again broadly similar to the 2SLS estimates in Table 3: The coefficient on mortality moves closer to 1 and readmission a bit further from 1, with the other coefficients on the utilization measures remaining the same as in Table 3. Surprisingly, distance has no significant effect for any outcome. However, the estimates are imprecise because most of the closures have little impact on distance to the nearest hospital (distance increases by less than 4 miles for 95% of zip codes affected by a closure, and does not increase at all for over three quarters of these zip codes). The market share of the closed hospital has a significant but small negative effect only on part A costs (costs decline by 1.7 percent for zip codes in which the market share of the closed hospital was 1), providing little evidence of a disruption effect.

Event Studies

Our identification relies on hospital closures suddenly reallocating patients across hospitals. Figure 1 provided evidence that market share of the closed hospitals fell suddenly at the time of closure. However, we also want to verify that patient outcomes changed suddenly in the expected direction at the time of closure. That is, in the year immediately after the closure, outcomes should improve for patients living in zip codes most affected by the closure of low performing hospitals, while outcomes should worsen for patients living in zip codes most affected by the closure of high performing hospitals.

Figure 5 plots an event study showing the effect of closure on patient outcomes in the zipcodes for which outcomes were expected to improve or worsen most. We compare the outcomes of patients who live in zip codes in the top and bottom 10% in terms of the expected impact of hospital closure (the

extreme two bins at either end in Figure 4A[13]), from three years before to two years after closure. We normed each trendline to have a value of zero in the year of closing and controlled for HRR-cohort-year fixed effects. To avoid dropping large numbers of patients from this analysis, we calculated each zip-code's outcome measure using patients in that zip code in that year (rather than fix patient zipcodes to t-1 as was done in our 2SLS estimates to minimize bias from patient mobility across zip codes).

As expected, mortality rose suddenly around the time of the closure of a low mortality hospital and fell suddenly around the time of the closure of a high mortality hospital. The pattern is similar for the other three outcomes. Moreover, the gap that opens up between outcomes in the two types of zip codes after closure are large -- roughly 1-2 percentage points in mortality and readmissions, and roughly 5-6 log points in Part A costs and length of stay. Importantly, changes are largest in the window immediately surrounding closure, although there is some evidence of a pre-trend in the length-of-stay- measure.

Does Risk-Adjustment Matter?

Prior studies have generally found that better risk adjustment tends to reduce the variation in measured performance across hospitals, but that even crude measures with no risk adjustment are correlated with more fully risk adjusted measures (Dimick et al., 2012). In other words, crude performance measures may be correlated with true (causal) hospital performance and therefore have predictive validity, but they tend to overstate the differences across hospitals and may not be forecast unbiased. In Table 5 we test whether performance measures without risk-adjustment are forecast unbiased: they are not. The 2SLS estimates in Table 5 are identical to those in Table 3, except that both the change in hospital quality $\left(\Delta\overline{Q}_{z,c,t}\right)$ and the expected change in hospital quality following closure $\left(E_{z,c,t}\right)$ were constructed with no risk-adjustment.[14] As anticipated by the prior literature, performance measures that are not risk adjusted continue to have predictive validity (the coefficients are positive and statistically significant for all outcomes) but have more forecast bias -- all of the estimates in Table 5 are closer to zero than the corresponding estimates in Table 3, suggesting that unadjusted measures overstate causal differences across hospitals relative to fully risk-adjusted measures.

Of course, the fact that estimates of hospital quality are valid predictors of the causal impact of a

[13] The bins were defined slightly differently than in Figure 4. In Figure 4 we defined bins using the original instrument ($E_{z,c,t}$), which was based on performance measures that left out data from t-1, t, and t+1. For Figure 6, to avoid using patient outcomes from the event window, we reconstructed $E_{z,c,t}$ using performance measures that left out data from the years t-3 through t+2.

[14] We continue to use fully risk-adjusted patient outcomes for the dependent variable for comparability to Table 3, although using patient outcomes that are not risk adjusted yield similar results.

hospital on patient outcomes does not imply that they are the optimal forecast of hospital quality given all the information which may be available to private and public payers with access to even more data (this point is underscored by our finding in Table 5 that it is only risk-adjusted measures that are forecast-unbiased). Even when quality measures are forecast unbiased there remains forecast error: some hospitals are better and some worse than the forecast suggests. One could improve on forecasts of hospital quality by using more information, as we did by using patient volume or more generally by creating composite measures using patient volume and outcomes in related conditions (Staiger et al., 2009). In principle, there is also information from electronic-health-records on laboratory values, prescription drug use, adherence, genetic tests, that could all be used to improve estimates of hospital measures. In surgical settings direct observation of surgical skill has been found to be a strong predictor of patient outcomes (Birkmeyer et al., 2013). All of this additional information can be used to reduce forecast error and the resulting quality measures can be validated with quasi-experimental methods such as ours.

V. Implications for Hospital Closure

The use of hospital closures to reallocate patients and shift their quality of care generates a treatment effect that directly informs the large literature on the effect of hospital closures. In Table 6 we use our estimates of the impacts of closures to explore the distributional impacts of hospital closures – both across zip codes and across patient populations. For this exercise, we use the reduced form estimates from Table 3 to predict changes in patient outcomes in each zip code as a function of our instrument $\left(E_{z,c,t}\right)$. This provides an estimate of how closures in each year affected outcomes in each zip code. To convert this estimate into a predicted impact on patients affected by the closure, we divide this estimate by the market share of the closed hospital in each zip code – resulting in a treatment-on-treated estimate. We then calculated the mean and percentiles of this impact across zip codes weighting by the number of patients going to a closed hospital in each zip code in year t-1. Because Black patients are more likely to receive care at hospitals that closed, we repeated this exercise weighting by the number of Black patients going to the closed hospital in each zip code (assuming that the treatment-on-treated effect was the same for Black patients as for others living in the same zip code).

The first panel of Table 6 reports the results for the actual closures that occurred between 1994 and 2013. Each year, on average, 0.22% of patients went to a hospital that closed in the subsequent year. Thus, cumulatively over the 20 years in our sample approximately 4.4% of all patients would eventually be impacted by a closure. Among the patients affected by a closure, the average impact on mortality was

-.184 percentage points; for every 1000 patients who would have gone to a closed hospital, we estimate 1.84 fewer deaths as a result of the closure. This average impact masks wide variation across zip codes: at the 10th percentile of patients impacted by a closure (zip codes served by low quality closing hospitals) closure reduced mortality by 1.46 percentage points, while at the 90th percentile (zip codes served by high quality closing hospitals) closure increased mortality by 1.03 percentage points. Black patients were more likely to be impacted by hospital closure (.32% per year vs .22% per year for whites), but conditional on being affected by a closure, the distribution of impacts on mortality were similar to that for all patients, with some patients benefiting and others harmed by the closure. The impacts on readmission were similar to mortality, with a small average decline in readmission and wide variation across zip codes. Finally, closures increased resource use (both 30 day costs and length of stay) by about 1% (.01 in logs), with a decline of 5-6% at the 10th percentile and an increase of 7-8% at the 90th percentile. Again, the impact distribution for Black patients was similar to that for all patients.

This evidence suggests that the relative quality of a closing hospital is key to understanding the impact of a hospital closure on the patients it serves. One reason that the earlier literature did not find evidence of harm to patients is because it did not distinguish between the relative performance of the hospitals being closed (see Joynt et. al, 2015 for an example of this and a rich summary of the literature). In fact, we find that most hospital closures reduce the mortality of affected patients because neighboring hospitals are of higher quality (lower mortality) than the closed hospital. These findings stand in sharp contrast to the belief that hospital closures harm patients because of treatment delays for acute conditions like heart-attacks and stroke (Baicker and Chandra, 2012).

Our findings suggest that closure *per se* does not harm patients, but rather closing the wrong (high-quality) hospitals is what harms patients. The second panel of Table 6 reports a counterfactual simulation of the impact on patient outcomes of closing only hospitals that were worse than the HRR average on all four quality indicators in 2012. This scenario would achieve gains similar to those that a narrow network could achieve by excluding low quality hospitals in their region. We estimate that 24% of all patients are at hospitals that are below average quality in their HRR. For those patients that would be affected by closures, we estimate that mortality and readmissions would drop by about .9 percentage points, while costs would drop by about 1%. For the 1.16m patients in our sample in 2012, this would have resulted in about 2500 fewer deaths and savings of roughly $43m in 2012.

VI. Conclusions

Our findings have several implications for literatures in economics and medicine. Most directly, our results support the view that variation across hospitals in risk-adjusted mortality and costs reflect large differences in productivity rather than unmeasured differences in patient characteristics. Understanding the source of such large productivity differences and reasons for their persistence is central to developing policies that will improve patient welfare (Chandra and Staiger 2007, 2020; Chandra et al., 2016a; Skinner and Staiger, 2009).

In the medical literature it is common to measure quality using process measures (such as beta-blockers or volume) instead of direct measures such as mortality because of the view that the direct measures reflect selection (Baicker and Chandra 2004; Skinner and Staiger 2009; Chiche et al. 2015; Epstein 2002). We find that measures calculated with standard claims data are causal and can be used for quality measurement, which lowers the need for process measures of health care quality. Our measures were based on claims-based risk-adjustment but had considerable predictive power, suggesting that our results may be generalizable to other populations. Future work should verify whether hospital quality measures constructed in other datasets and using other methods are also forecast unbiased.

Our work also speaks to the role of market-forces in health care: Chandra et.al (2016b) found patients choose lower mortality hospitals, but what was not known from that work is whether patients select hospitals that are lower mortality in the causal sense, or whether they mistakenly chose hospitals with healthier patients because of conflating outcomes with risk-adjusted outcomes. Our work fills this gap by noting that market-learning, even in healthcare, reflects accurate learning about risk-adjusted quality. Understanding the exact channels by which markets disseminate information about quality remains a fruitful area for future research. It has large implications for racial-disparrites in care as well, for there is evidence that the largest component of racial-disparies in care today is that minority patients are treated at lower quality hospitals (Chandra, Kakani, Sacarny, 2020).

The availability of causal hospital quality indicators and their relative ease of computational use means that it becomes possible, at least in principle, to design narrow networks and use cost-sharing to discourage the use of providers that are less effective. Our results also suggest that standard risk-adjusted measures of mortality and costs can be used more broadly to measure the effect of hospital consolidation on outcomes which is a central question in antitrust determinations, and the effect of payment reforms such as Accountable Care Organizations (ACOs) on health (as opposed to only on spending), replacing or augmenting the current emphasis on indirect measures of quality such as process-measures and patient-satisfaction (Beaulieu et al. 2020). It is important to recognize that our measures are valid up to

the size of reallocations induced by closures (but it is also the case that unvalidated measures, such as process measures or patient satisfaction scores, may not be robust to larger reallocations).

Finally, while our focus is on hospital quality in this paper, our approach is general and can be used to understand the causal effect of many aspects of health care delivery and technology where clinical trials are difficult. Chandra and Staiger (2007, 2020) and Hull (2020) allow for comparative advantage in hospital performance and patient selection (the idea that a hospital might specialize in a particular type of treatment which generates a 'match specific' component to the hospital assignment). Future work might explore whether estimates of hospital quality that account for comparative advantage and selection yield hospital performance estimates that reduce forecast bias. Our approach can also be used to validate the causal effect of surgeon quality (where individual effects are large relative to facility effects), outpatient facilities, and combinations of drugs and medical technologies where the causal effect of a bundle of technologies is not known. As long as it is possible to estimate quality measures for providers, facilities, or technologies and exogenously change the use of these inputs (perhaps through closures or movers), the causal effect of these combinations can be determined without conducting an RCT.

REFERENCES

Agency for Healthcare Research and Quality. “About CAHPS.” 2020a. Accessed June 15, 2020. http://www.ahrq.gov/cahps/about-cahps/index.html.

Agency for Healthcare Research and Quality. “Quality Indicators.” 2020c. Accessed June 15, 2020. https://www.qualityindicators.ahrq.gov/.

Agency for Healthcare Research and Quality. “Trends in Hospital Costs/Utilization - HCUP Fast Stats.” 2020a. Accessed June 12, 2020.

American Hospital Association. Fast Facts on US Hospitals, 2022. Available at https://www.aha.org/statistics/fast-facts-us-hospitals

Baicker, Katherine, and Amitabh Chandra. “The Health Care Jobs Fallacy.” New England Journal of Medicine 366, no. 26 (June 28, 2012): 2433–35. https://doi.org/10.1056/NEJMp1204891.

Baicker, Katherine, and Amitabh Chandra. 2004. “Medicare Spending, the Physician Workforce, and Beneficiaries’ Quality of Care.” *Health Affairs* Suppl Web Exclusives (January): W4–184 – 97.

Birkmeyer, J.D., Finks, J.F., O'reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J., Banerjee, M. and Birkmeyer, N.J., 2013. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, *369*(15), pp.1434-1442.

Chandra, Amitabh, David Cutler, and Zirui Song. 2011. “Who Ordered That? The Economics of Treatment Choices in Medical Care.” *Handbook of Health Economics*. https://doi.org/10.1016/b978-0-444-53592-4.00006-2.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson. 2016a. “Productivity Dispersion in Medicine and Manufacturing.” *The American Economic Review* 106 (5): 99–103.

———. 2016b. “Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector.” *The American Economic Review* 106 (8): 2110–44.

Chandra, Amitabh, and Douglas O. Staiger. 2020. “Identifying Sources of Inefficiency in Healthcare*.” *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjz040.

Chiche, Jean-Daniel, Rui P. Moreno, Christian Putensen, and Andrew Rhodes. 2015. *Patient Safety and Quality of Care in Intensive Care Medicine*. MWV.

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen. 2019. “THE PRICE AIN’T RIGHT? HOSPITAL PRICES AND HEALTH SPENDING ON THE PRIVATELY INSURED.” *The Quarterly Journal of Economics* 134 (1): 51–107.

Epstein, Arnold M. 2002. “Volume and Outcome--It Is Time to Move Ahead.” *The New England Journal of Medicine* 346 (15): 1161–64.

Skinner, Jonathan, and Douglas Staiger. 2009. “Technology Diffusion and Productivity Growth in Health Care.” https://doi.org/10.3386/w14865.

Collard-Wexler, Allan, and Jan De Loecker. 2015. “Reallocation and Technology : Evidence from the US Steel Industry.” *American Economic Review* 105 (1): 131–71.

Buchmueller, Thomas C., Mireille Jacobson, and Cheryl Wold. “How Far to the Hospital?: The Effect of Hospital Closures on Access to Care.” Journal of Health Economics 25, no. 4 (July 1, 2006): 740–61. https://doi.org/10.1016/j.jhealeco.2005.10.006.

Centers for Medicare & Medicaid Services. “Quality Measures.” 2020b. Accessed June 15, 2020.https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson. “Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector.” American Economic Review 106, no. 8 (August 2016): 2110–44. https://doi.org/10.1257/aer.20151080.

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.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." American Economic Review, 104 (9): 2633-79.

Dimick, Justin B., Amir A. Ghaferi, Nicholas H. Osborne, Clifford Y. Ko, and Bruce L. Hall. “Reliability Adjustment for Reporting Hospital Outcomes with Surgery.” Annals of Surgery 255, no. 4 (April 2012): 703–7. https://doi.org/10.1097/SLA.0b013e31824b46ff.

Doyle, Joseph, John Graves, Jonathan Gruber, and Samuel Kleiner. “Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns.” The Journal of Political Economy 123, no. 1 (01 2015): 170–214. https://doi.org/10.1086/677756.

Holmes, George M, Rebecca T Slifkin, Randy K Randolph, and Stephanie Poley. “The Effect of Rural Hospital Closures on Community Economic Health.” Health Services Research 41, no. 2 (April 2006): 467–85. https://doi.org/10.1111/j.1475-6773.2005.00497.x.

Joynt, Karen E., Paula Chatterjee, E. John Orav, and Ashish K. Jha. “Hospital Closures Had No Measurable Impact On Local Hospitalization Rates Or Mortality Rates, 2003–11.” Health Affairs 34, no. 5 (May 1, 2015): 765–72. https://doi.org/10.1377/hlthaff.2014.1352.

Kane, Thomas J., and Douglas O. Staiger. “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” NBER Working Papers. NBER Working Papers. National Bureau of Economic Research, Inc, December 2008. https://ideas.repec.org/p/nbr/nberwo/14607.html.

Kane, Thomas J., and Douglas O. Staiger. “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” NBER Working Papers. NBER Working Papers. National Bureau of Economic Research, Inc, December 2008. https://ideas.repec.org/p/nbr/nberwo/14607.html.

Leapfrog. “Leapfrog.” 2020. Accessed June 12, 2020. https://www.leapfroggroup.org/.

McClellan, Mark, and Doug Staiger. “The Quality of Health Care Providers.” NBER Working Paper. National Bureau of Economic Research, Inc, August 1999.

https://econpapers.repec.org/paper/nbrnberwo/7327.htm.

OECD. Health Expenditure and Financing. 2022. Available at https://stats.oecd.org/Index.aspx?DataSetCode=SHA.

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." The Quarterly Journal of Economics 125, no. 1 (February 1, 2010): 175–214. https://doi.org/10.1162/qjec.2010.125.1.175.

Scally, Christopher P., Jyothi R. Thumma, John D. Birkmeyer, and Justin B. Dimick. "Impact of Surgical Quality Improvement on Payments in Medicare Patients." Annals of Surgery 262, no. 2 (August 2015): 249–52. https://doi.org/10.1097/SLA.0000000000001069.

Staiger, Douglas O., Justin B. Dimick, Onur Baser, Zhaohui Fan, and John D. Birkmeyer. "Empirically Derived Composite Measures of Surgical Performance." Medical Care 47, no. 2 (2009): 226–33.

The Joint Commission. "Accreditation - ORYX." 2020. Accessed June 15, 2020. https://www.jointcommission.org/measurement/Reporting/Accreditation-ORYX.

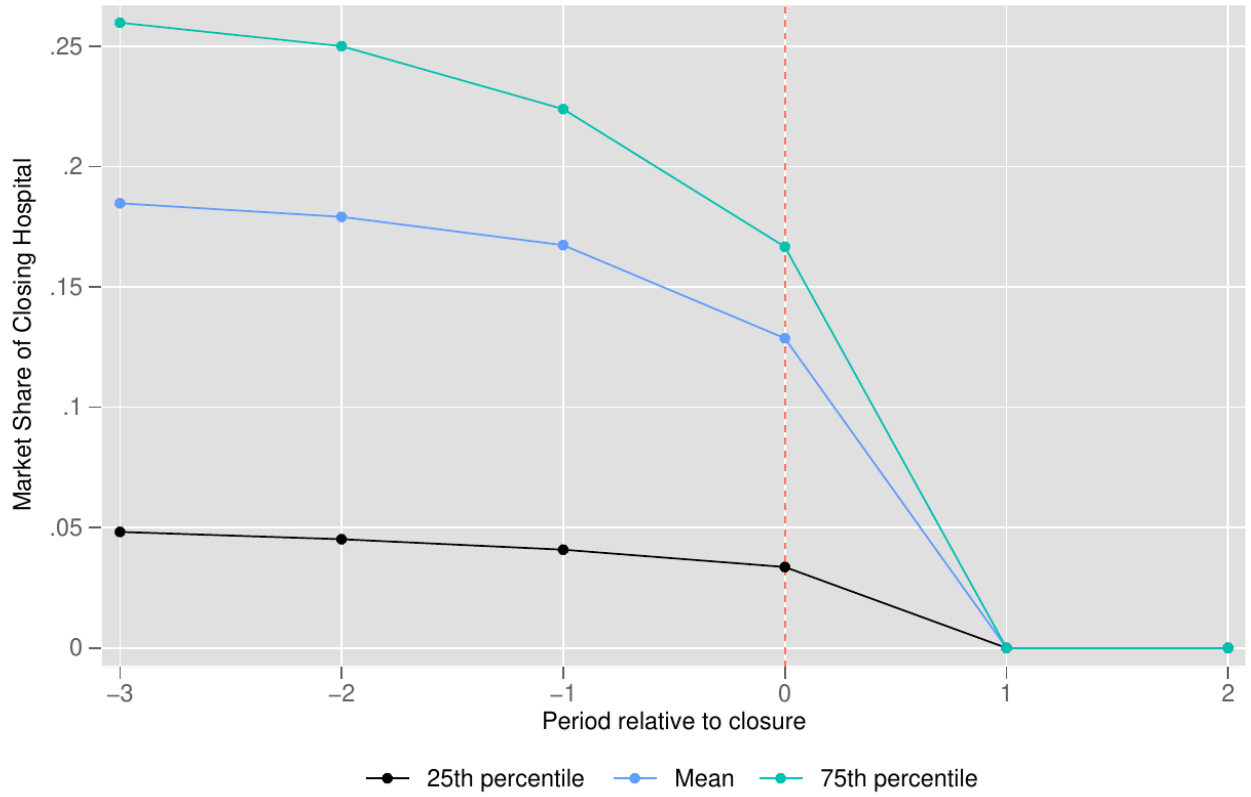
U.S. News. "FAQ: How and Why We Rank and Rate Hospitals." July 29, 2019. https://health.usnews.com/health-care/best-hospitals/articles/faq-how-and-why-we-rank-and-rate-hospitals.

Yasaitis, Laura, Elliott S. Fisher, Jonathan S. Skinner, and Amitabh Chandra. "Hospital Quality And Intensity Of Spending: Is There An Association?" Health Affairs 28, no. Supplement 1 (January 1, 2009): w566–72. https://doi.org/10.1377/hlthaff.28.4.w566.

Bekelis K, Marth NJ, Wong K, Zhou W, Birkmeyer JD, Skinner J. Primary Stroke Center Hospitalization for Elderly Patients With Stroke: Implications for Case Fatality and Travel Times. JAMA Intern Med. 2016;176(9):1361-1368. doi:10.1001/jamainternmed.2016.3919

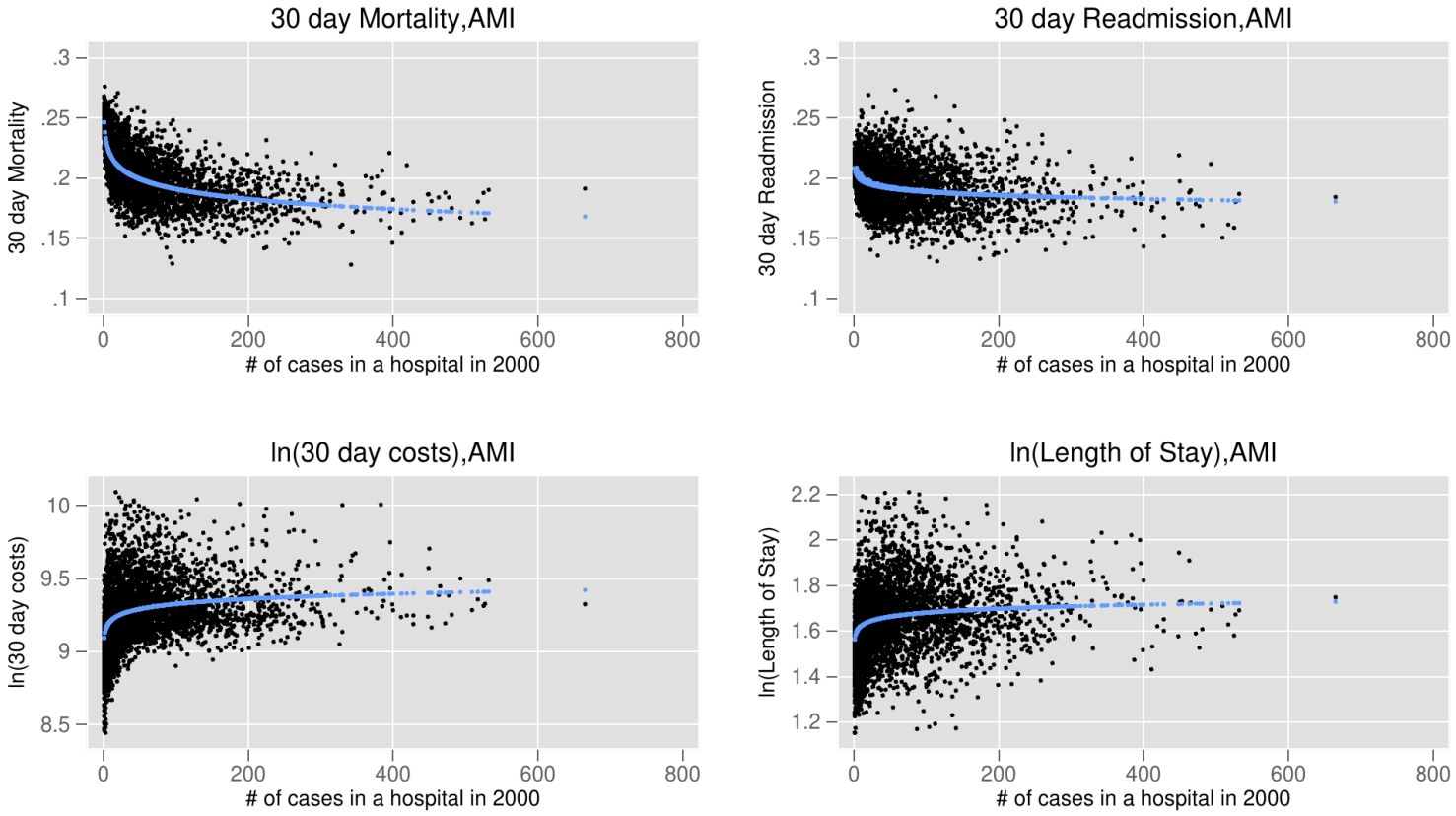
Brauer CA, Coca-Perraillon M, Cutler DM, Rosen AB. Incidence and Mortality of Hip Fractures in the United States. JAMA. 2009;302(14):1573–1579. doi:10.1001/jama.2009.1462

Figure 1: Market shares for closing hospitals



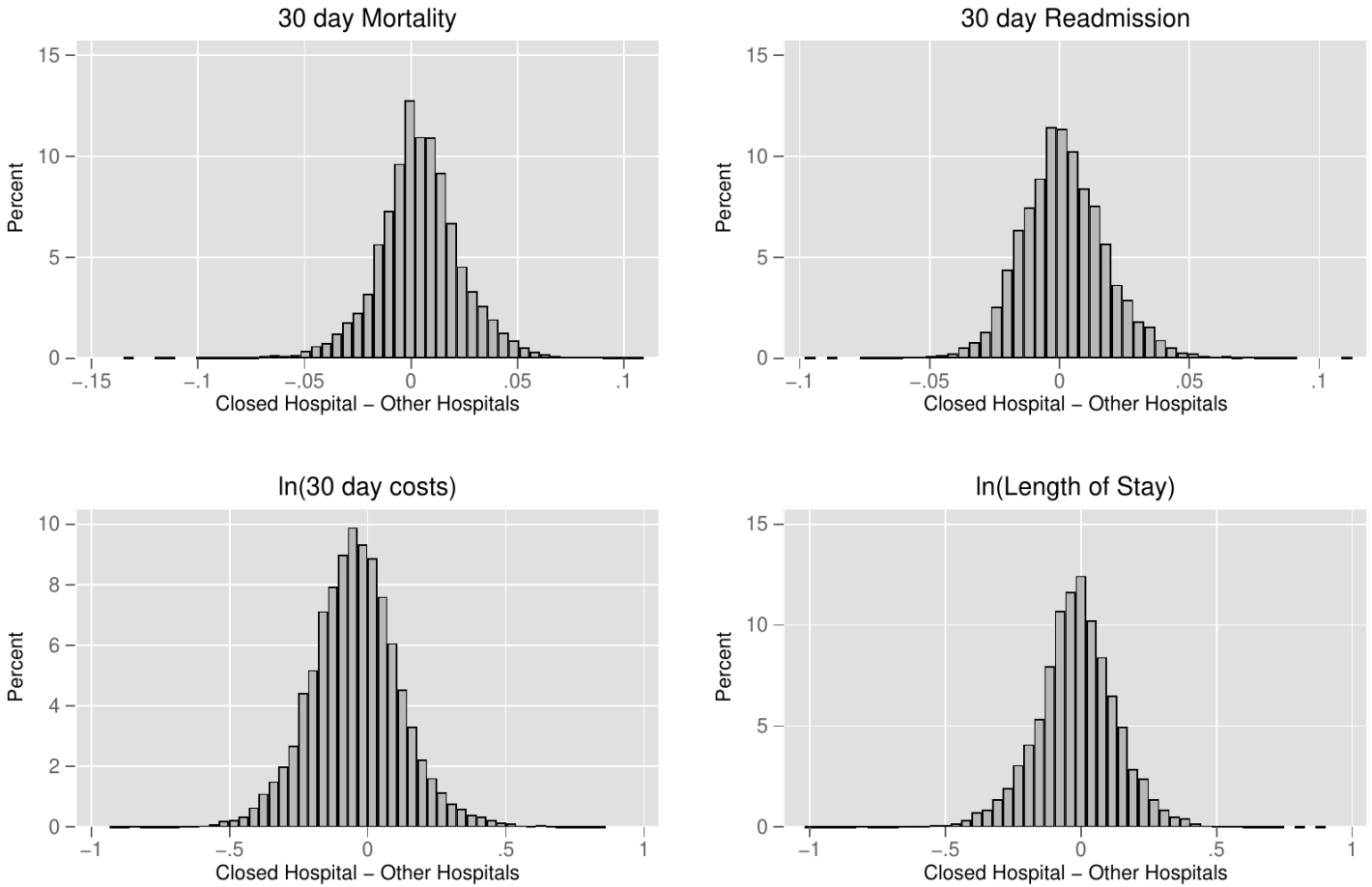
Note: The graph reports the distribution of closed hospitals' market share across zip codes and dependent variables (mean and 25th/75th percentile) in the years before and after closure (0 is the year of hospital closure). The distribution is weighted by the number of patients treated by the closed hospital in each zip code in the year before closure. Thus, the average patient admitted to a closed hospital one year before closure lived in a zip code in which the closed hospital had 17% market share, while 75% of patients admitted to a closed hospital one year before closure lived in a zip code in which the closed hospital had 23% market share.

Figure 2: Hospital Quality and Hospital Volume, Heart-Attacks in 2000



Notes: Each figure plots a hospital performance indicator ($\widehat{\mu}_{ht}$) from 2000 for heart attack patients against hospital volume in 2000. The graph reports hospital level effects for 30-day mortality, 30-day readmissions, log of 30-day costs, and log of length of Stay. The blue line indicates the component of hospital performance that is predicted by ln(volume). See Section II for details on estimation of the performance measures. Appendix A2 provides plots for all other conditions.

Figure 3: Quality differences between closing hospital and other hospitals in a patient's zipcode



Notes: Each histogram plots the distribution across zipcodes of the quality difference between the closed hospital and the remaining hospitals (pooling estimates from all five patient cohorts). Each zip code is weighted by the number of patients who were seen at the closing hospital in the year before closure. This difference when multiplied by the closing hospitals market-share is the instrument.

Figure 4A: Relationship between Change in Outcomes and Predicted Change from Closure

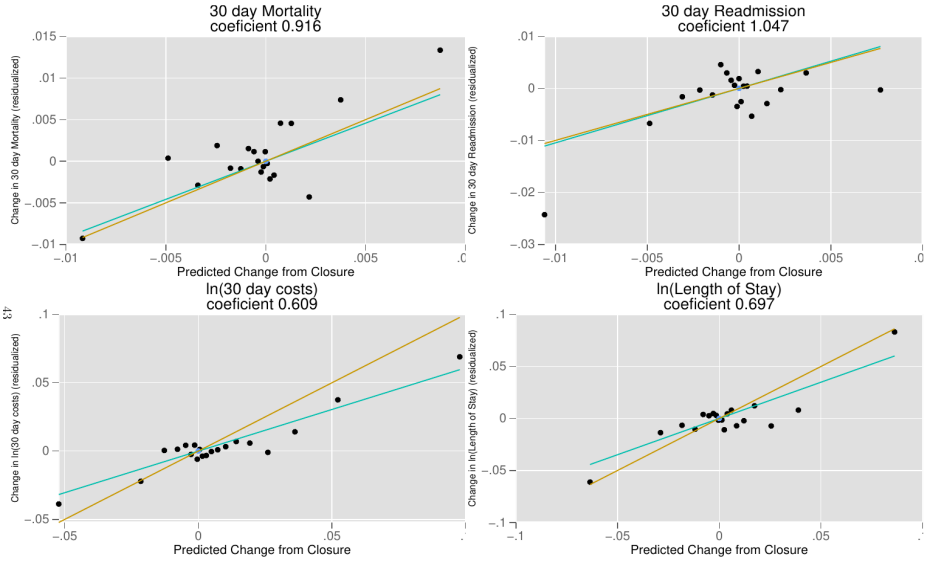
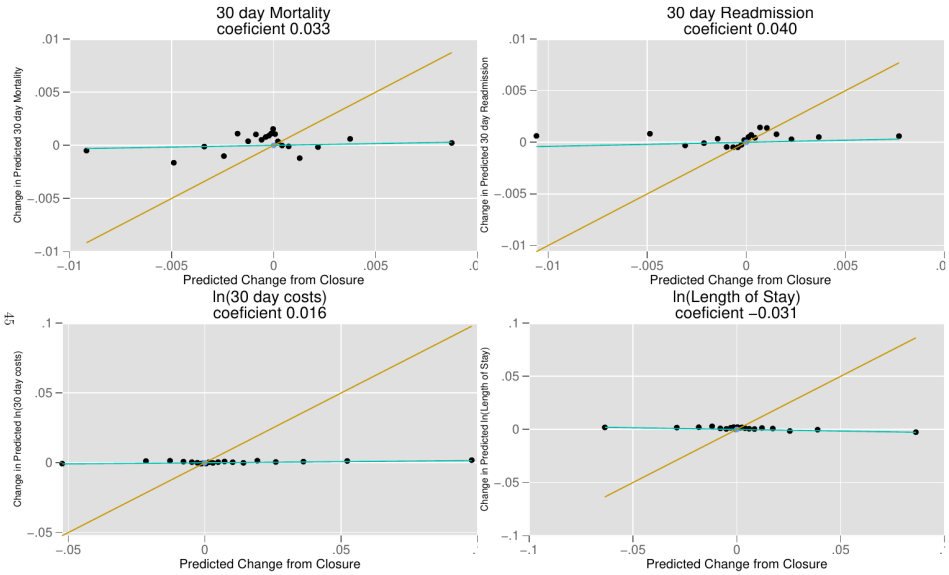
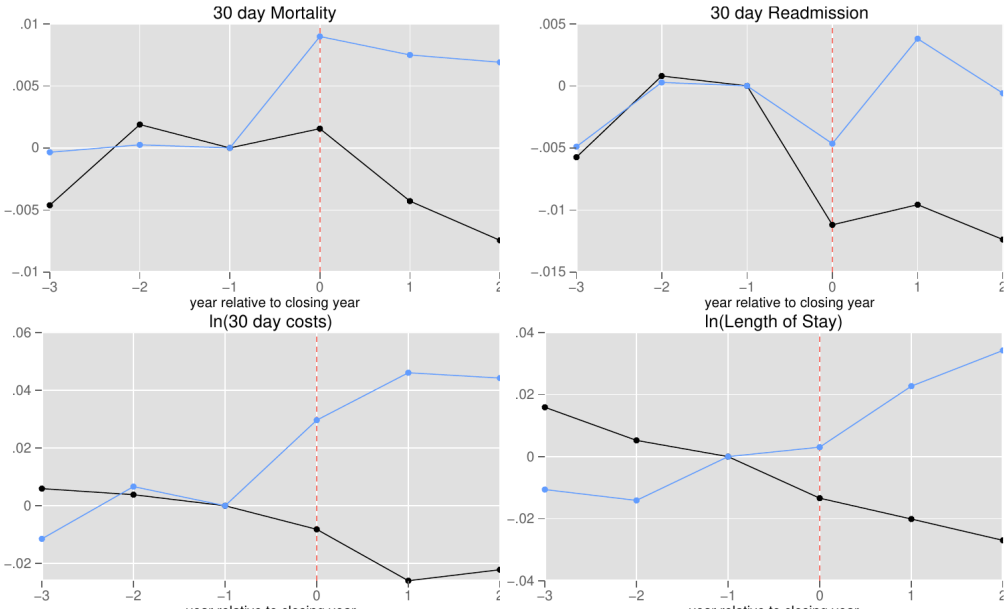


Figure 4B: Relationship between Change in Predicted Outcomes and Predicted change from Closure



Notes: Panel A illustrates the 2SLS relationship between changes in patient outcomes and the predicted change in outcomes from the first stage, binning expected outcomes into ventiles (20 bins), and controlling for HRR-cohort--year effects. One additional bin is included for zip codes unaffected by closure. The regression coefficient for the binned data is reported at the top of each figure, along with the fitted line (in blue). A 45-degree line denoting a coefficient of 1 is graphed for comparison (in brown). Panel B is organized in an identical manner, but illustrates a falsification exercise, plotting the relationship between the change in predicted outcomes using patient characteristics and ventiles of the predicted change from closure. Changes in patient characteristics should not be correlated with the instrument.

Figure 5: Event study analysis of the effect of hospital closures on patient outcomes, zip codes in the top and bottom deciles of expected change in outcomes



Note: Each figure plots average patient outcomes from 3 years before to two years after hospital closure for patients who live in zip codes in the top (blue) and bottom (black) deciles of the expected impact of hospital closure (the instrument). We normed each trendline to have a value of zero in the year of closing and controlled for hrr-cohort-year fixed effects. To avoid dropping large numbers of patients from this analysis, we calculated each zip-code's outcome measure using patients in that zip code in that year (rather than fix patient zipcodes to t-1 as was done in our 2SLS estimates).

Table 1: Patient Level Summary Statistics by Hospital Closure

	All	Admitted to hospital that did not close	Admitted to closed hospitals
Patient level characteristics			
Patient age	80.45	80.45	80.69
Proportion female	0.58	0.58	0.61
Proportion white	0.89	0.89	0.82
# of comorbidities	1.03	1.04	1.02
Dementia HCC-based Krumholz	0.07	0.07	0.08
Institutionalized	0.02	0.02	0.03
Dual Eligible	0.16	0.16	0.21
Outcomes			
30 day Mortality	0.12	0.12	0.13
30 day Readmission	0.17	0.17	0.18
Length of Stay	6.50	6.46	7.81
30 day costs	13013	13058	11733
Number of Patients in cohort			
AMI	4585314	4452287	133027
Hip Fracture	4067749	3949815	117934
CHF	7389163	7119611	269552
Pnuemonia	8665595	8317480	348115
Stroke	5540734	5378001	162733
Observations	30248555	29217194	1031361

Note: We use a 100 percent sample of beneficiaries enrolled in Medicare fee-for-service who were hospitalized between 1992 and 2015 for one of 5 cohorts: heart-attacks, hip-fractures, pneumonia, congestive heart failure (CHF) and stroke. Patients admitted to a closed hospital include patients who ever went to a hospital before it closed.

Table 2: Characteristics of Hospital Closures

Hospital Closure Events, 1992-2013	
Closure or Conversion	754
Closure Only	580
Conversion Only	174
Hospital Closed Years	
1994-1996	143
1997-2000	204
2001-2004	173
2005-2008	131
2009-2011	53
2012-2013	50
Closed Hospital Characteristics	
Mean Number of Beds	98.2
Mean Disproportionate Share Percent	33.7
% of Hospital in Urban	*81.7*
Closed Hospital Ownership Type	
Private	220
Religious	75
Government	142
Proprietary	411

Note: For a hospital to be defined as closed it should not be admitting patients for all our cohorts. We treat a hospital that converts to a long-term care facility (labeled conversion) as a closing hospital for it is no longer treating patients in our 5 cohorts. We verify that this closing date is within one year of the hospital closing or conversion to a LTC hospital, with dates reported by either the American Hospital Association (AHA), CMS Provider Of Service File (POS), or CMS Cost report data (HCRIS).

Table 3: OLS and 2SLS estimates of the effect of hospital quality on patient outcomes

	A:Δ 30 day Mortality				B:Δ 30 day Readmission			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	1.005***	0.927**			0.756***	0.934**		
	(0.016)	(0.346)			(0.021)	(0.401)		
$\Delta E_{z,c,t}$			0.532**	0.574***			0.598**	0.640***
			(0.202)	(0.031)			(0.261)	(0.026)
HRR-YR-CHRT	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2369146	2369146	2369146	2369146	2316922	2316922	2316922	2316922

	C:Δ 30 day Cost				D:Δ Length of Stay			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	0.969***	0.584***			0.935***	0.626***		
	(0.006)	(0.093)			(0.006)	(0.099)		
$\Delta E_{z,c,t}$			0.326***	0.557***			0.384***	0.613***
			(0.067)	(0.051)			(0.064)	(0.027)
HRR-YR-CHRT	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2369146	2369146	2369146	2369146	2369146	2369146	2369146	2369146

Note: For each quality measure, we present the OLS and 2SLS estimates of the slope coefficient from Equation (5), which regresses changes in patient outcomes on the contemporaneous change in hospital quality. In the 2SLS we instrument this change with the expected change in hospital quality following closure. We also report estimates of the reduced form (regressing outcomes on the instrument $E_{z,c,t}$) and the first-stage (regressing hospital quality on the instrument $E_{z,c,t}$). All regressions pool data from the five patient cohorts and include cohort-by-hrr-by-year fixed effects. The number of observations reflects the total number of unique cohort-hrr-zipcode-year cells in the regression. Regressions are weighted by the number of patients in the cell, and standard errors are clustered at the level of 306 Hospital Referral Regions (HRRs).

Table 4A: OLS and 2SLS estimates of the effect of hospital quality on patient outcomes, Robustness to additional controls

	A:Δ 30 day Mortality				B:Δ 30 day Readmission			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	1.003***	0.915**			0.759***	0.922**		
	(0.016)	(0.347)			(0.021)	(0.403)		
$\Delta E_{z,c,t}$			0.524**	0.572***			0.591**	0.640***
			(0.202)	(0.031)			(0.262)	(0.026)
Observations	2369146	2369146	2369146	2369146	2316922	2316922	2316922	2316922

	C:Δ 30 day Cost				D:Δ Length of Stay			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	0.969***	0.572***			0.935***	0.634***		
	(0.006)	(0.095)			(0.006)	(0.100)		
$\Delta E_{z,c,t}$			0.319***	0.557***			0.388***	0.612***
			(0.068)	(0.052)			(0.065)	(0.027)
HRR-YR-CHRT	Y	Y	Y	Y	Y	Y	Y	Y
$Age_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Male_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Black_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$NumComorbidities_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2369146	2369146	2369146	2369146	2369146	2369146	2369146	2369146

Notes: See notes to Table 3. In addition, these estimates control for the following average characteristics of patients from each zip code (measured in year t-1): average age, %male, %Black, and the average number of comorbidities.

Table 4B: OLS and 2SLS estimates of the effect of hospital quality on patient outcomes, Robustness to additional controls including market share and changes in distance

	A:Δ 30 day Mortality				B:Δ 30 day Readmission			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	1.003***	0.997**			0.759***	0.865**		
	(0.016)	(0.362)			(0.021)	(0.404)		
mshare^{closed}	0.004	0.004	0.002	−0.002**	−0.003	−0.003	−0.003	0.001
	(0.004)	(0.004)	(0.004)	(0.001)	(0.005)	(0.005)	(0.005)	(0.001)
$\Delta distance$	0.004	0.004	0.004	0.000	−0.002	−0.002	−0.002	0.000
	(0.004)	(0.004)	(0.004)	(0.000)	(0.005)	(0.005)	(0.005)	(0.000)
$\Delta E_{z,c,t}$			0.555**	0.557***			0.559**	0.646***
			(0.205)	(0.029)			(0.266)	(0.027)
Observations	2369146	2369146	2369146	2369146	2316922	2316922	2316922	2316922

	C:Δ 30 day Cost				D:Δ Length of Stay			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	0.969***	0.639***			0.936***	0.632***		
	(0.006)	(0.104)			(0.006)	(0.104)		
mshare^{closed}	−0.032***	−0.017*	−0.006	0.018**	−0.018	−0.013	−0.004	0.013**
	(0.008)	(0.010)	(0.013)	(0.008)	(0.014)	(0.014)	(0.014)	(0.006)
$\Delta distance$	−0.004	−0.001	−0.003	−0.003	0.004	0.009	0.010	0.001
	(0.007)	(0.008)	(0.009)	(0.005)	(0.009)	(0.009)	(0.009)	(0.004)
$\Delta E_{z,c,t}$			0.335***	0.524***			0.378***	0.598***
			(0.069)	(0.048)			(0.067)	(0.028)
HRR-YR-CHRT	Y	Y	Y	Y	Y	Y	Y	Y
$Age_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Male_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Black_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$NumComorbidities_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2369146	2369146	2369146	2369146	2369146	2369146	2369146	2369146

Notes: See notes to Table 3. In addition, these regressions control for the market share of the closed hospital in each zip code (mshare^{closed}) in year t-1 and the change in the distance to the nearest hospital from year t-1 to year t+1.

Table 5: OLS and 2SLS estimates of the effect of hospital quality on patient outcomes, using no-risk adjustment for quality measures

	A:Δ 30 day Mortality				B:Δ 30 day Readmission			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	0.775***	0.601**			0.639***	0.688**		
	(0.012)	(0.274)			(0.017)	(0.328)		
$\Delta E_{z,c,t}$			0.350**	0.583***			0.442**	0.642***
			(0.163)	(0.030)			(0.213)	(0.026)
Observations	2369146	2369146	2369146	2369146	2316922	2316922	2316922	2316922

	C:Δ 30 day Cost				D:Δ Length of Stay			
	(1) OLS	(2) 2SLS	(3) reduced form	(4) 1st stage	(5) OLS	(6) 2SLS	(7) reduced form	(8) 1st stage
$\Delta Q_{z,c,t}$	0.882***	0.446***			0.824***	0.511***		
	(0.005)	(0.087)			(0.006)	(0.103)		
$\Delta E_{z,c,t}$			0.241***	0.540***			0.297***	0.582***
			(0.061)	(0.048)			(0.063)	(0.028)
HRR-YR-CHRT	Y	Y	Y	Y	Y	Y	Y	Y
$Age_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Male_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$Black_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
$NumComorbidities_{z,c,t-1}$	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2369146	2369146	2369146	2369146	2369146	2369146	2369146	2369146

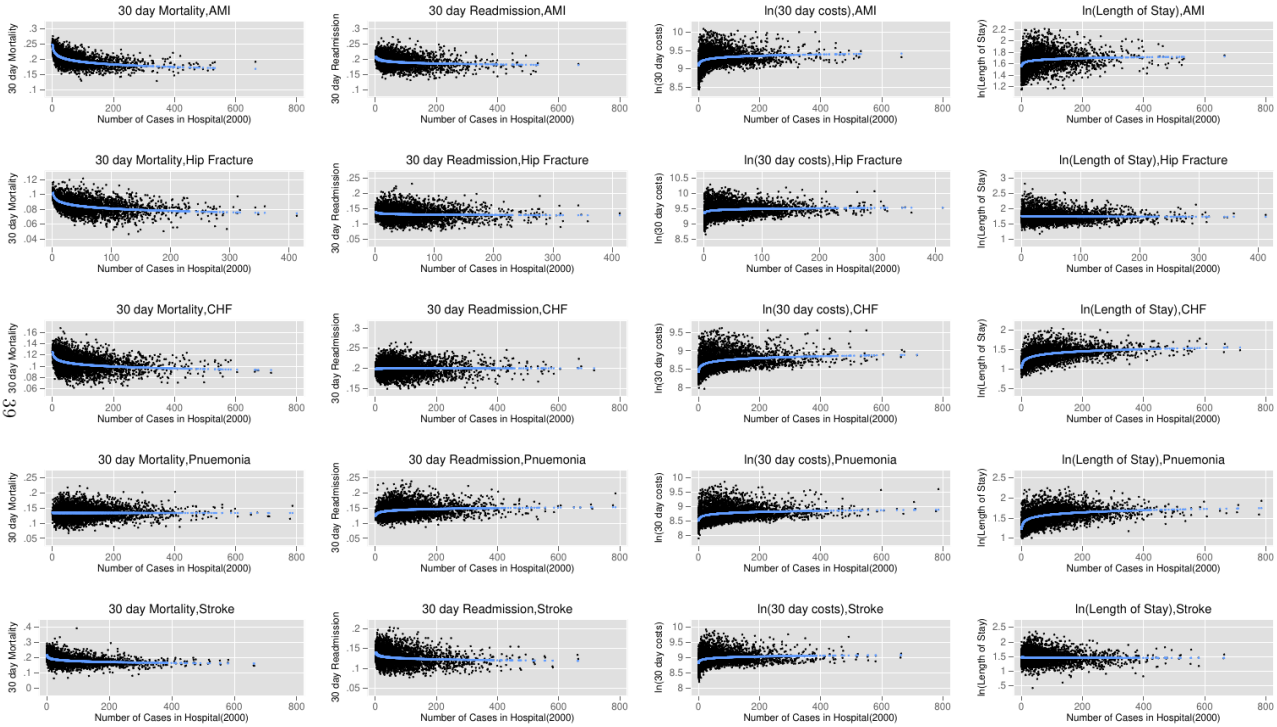
Notes: See notes to Table 3 and 4. In these regressions both the change in hospital quality $\Delta Q_{z,c,t}$ and the expected change in hospital quality following closure $E_{z,c,t}$ were constructed without risk-adjustment.

Table 6. Actual and Simulated Impact of Hospital Closures on Patient Outcomes

	All Patients				Black Patients			
	% patient impacted each year	Average Impact on affected patients	10th Percentile Impact on affected patients	90th Percentile Impact on affected patients	% patient impacted each year	Average Impact per affected patient	10th Percentile Impact on affected patients	10th Percentile Impact on affected patients
Estimated impact of actual hospital closures 1994-2013								
30 day Mortality (%)	0.220	−0.184	−1.461	1.025	0.322	−0.161	−1.379	0.959
30 day Readmission (%)	0.201	−0.096	−1.338	1.098	0.322	−0.013	−1.224	1.153
log of 30 day costs	0.220	0.013	−0.052	0.079	0.322	0.014	−0.046	0.078
log of Length of Stay	0.220	0.006	−0.062	0.076	0.322	−0.001	−0.069	0.067
Simulated impact of closing hospitals that are below HRR average on all four outcomes in 2012								
30 day Mortality (%)	24.228	−0.884	−2.051	0.386	26.469	−0.893	−1.746	−0.104
30 day Readmission (%)	22.691	−0.888	−2.016	0.217	26.469	−0.973	−2.016	0.122
log of 30 day costs	24.228	−0.013	−0.097	0.063	26.469	−0.012	−0.137	0.052
log of Length of Stay	24.228	−0.009	−0.077	0.060	26.469	−0.000	−0.070	0.094

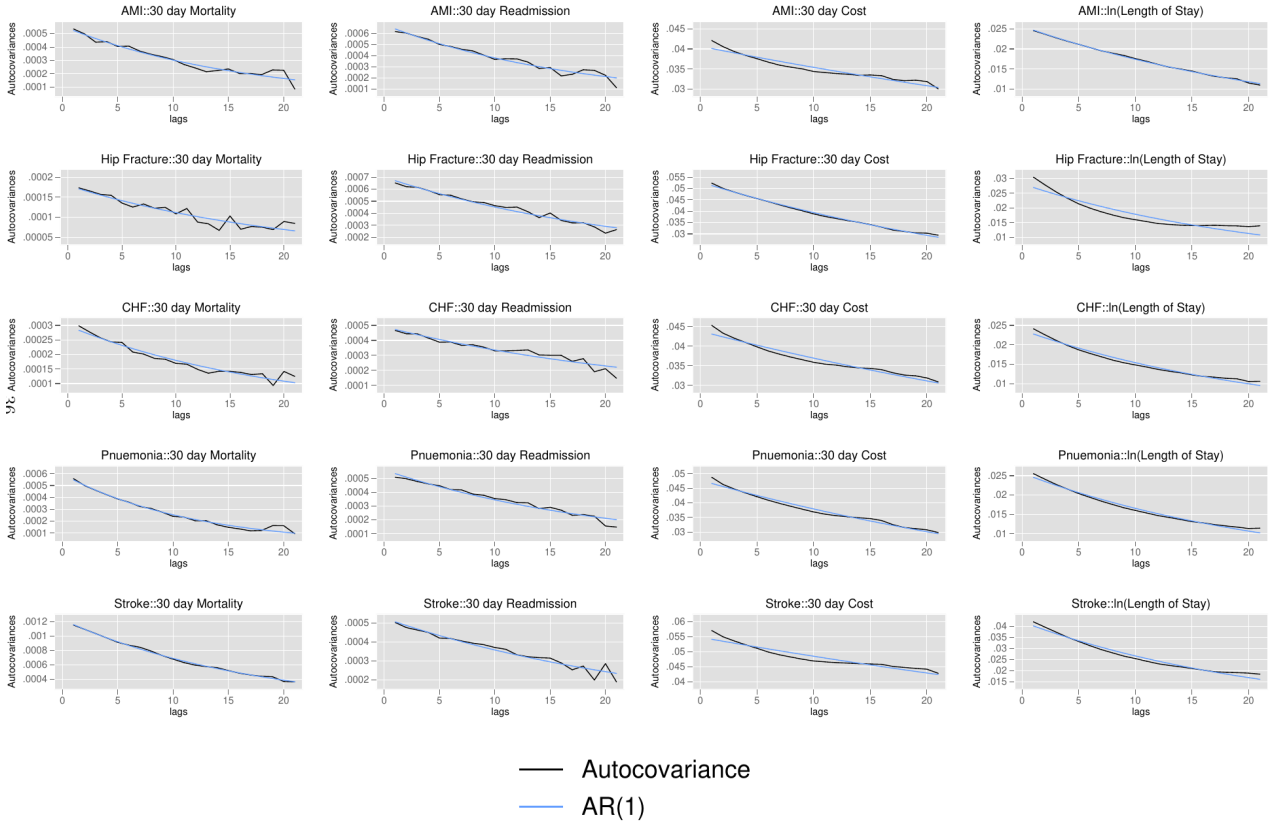
Notes: Reduced form estimates from Table 3 were used to predict changes in patient outcomes in each zip code as a function of our instrument $\left(E_{z,c,t}\right)$, and divided by market share of the closed hospital in each zip code to yield the predicted impact on patients affected by the closure (treatment on treated). The mean and percentiles of this impact across zip codes reported above were weighted by the number of patients (or Black patients) going to a closed hospital in each zip code in year t-1.

Appendix 1: Hospital Quality Indicators in 2000, by disease cohort



Notes: Each figure plots a hospital performance indicator ($\hat{\mu}_{ht}$) from 2000 for a given cohort and outcome against hospital volume in 2000. The blue line indicates the component of hospital performance that is predicted by ln(volume). See Section II for details on estimation of the performance measures.

Appendix 2: Autocovariagram for each hospital quality indicator by disease cohort.



Notes: each figure plots estimates of the covariance of the hospital residuals at each lag (the drift parameter in Section II). For comparison, we include the estimated autocovariances predicted from an AR(1) fit to these data.