Estimating Hospital Quality with Quasi-experimental Data*

Peter Hull†

Job Market Paper

Most recent version: http://www.mit.edu/~hull/JMP.pdf

December 16, 2016

Abstract

Non-random sorting can bias outcome-based measures of institutional quality. I develop tractable instrumental variable quality estimators that accommodate non-linear causal effects, institutional comparative advantage, and selection-on-gains. I use this framework to compute empirical Bayes posteriors for U.S. hospital quality that optimally combine estimates from quasi-experimental ambulance company assignment and predictions from observational risk-adjustment models (RAMs). Higher-spending, higher-volume, and privately-owned hospitals have better posteriors, and most markets exhibit positive selection-on-gains. I quantify the effects of selection bias by simulating Medicare reimbursement and consumer guidance policies that use quality posteriors instead of RAMs. The types of hospitals subsidized by performance-linked payment schemes are largely unchanged when quasi-experimental data is incorporated, but existing transfers are magnified. Admission policy simulations highlight the limitations of consumer guidance programs in settings with significant selection on match-specific quality.

*I thank Joshua Angrist, Amy Finkelstein, and Parag Pathak for their invaluable guidance and support, as well as Jonathan Gruber, Joseph Doyle, Douglas Staiger, Christopher Walters, Nikhil Agarwal, Isaiah Andrews, Alberto Abadie, Bruce McGough, Yusuke Narita, Bryan Perry, Nick Hagerty, Evan Riehl, Greg Howard, Brendan Price, C. Jack Liebersohn, William Goulding, Rachael Meager, Serena Canaan, Lindsey Novak, Alexandre Staples, Rebecca Martin, and seminar participants at MIT and NBER for their many helpful comments and suggestions. I am especially thankful to Joseph Doyle, Jonathan Gruber, John Graves, and Samuel Kleiner for sharing code to construct ambulance instruments; to Maurice Dalton, Yunan Ji, Bryan Perry, and Jean Roth for their data expertise; and to emergency service professionals Ben Artin, Mark Millet, Laura Segal, Julia Taylor, and Kevin Wickersham for answering my many institutional questions. I gratefully acknowledge funding from the National Institute on Aging (#T32-AG000186) and the Spencer Foundation (#201600065). All views and errors are my own.

†MIT Department of Economics. Email: hull@mit.edu; website: http://economics.mit.edu/grad/hull

1 Introduction

Outcome-based rankings of institutional quality draw interest in many settings, from school and teacher value-added to the lasting socioeconomic effects of residential, educational, and occupational choice.[1] In the U.S. these measures have begun to play an important policy role, particularly in education and healthcare. Hospitals with low risk-adjusted mortality rates, for example, are now rewarded with higher Medicare reimbursement rates, while providers with poor survival outcomes may be flagged as low-performers. Recent research has found that such quality-based policies shape both hospital incentives and patient admission patterns (Norton et al., 2016; Gupta, 2016; Dranove and Sfekas, 2008; Chandra et al., 2015).

To date, performance-based regulation has relied on observational quality estimators, such as value-added models (VAMs) in education and risk-adjustment models (RAMs) in health. These methods leverage strong selection-on-observables assumptions: that, say, a patient's choice of hospital is as good as random conditional on a set of observed controls. When provider selection is correlated with potential health outcomes, hospital RAMs are prone to systematic bias, and supervisory policies can be distorted. RAM-based admission guidance programs may themselves be a source of this selection bias by encouraging the selection of high-ranked hospitals, as may other intrinsic factors like the medical expertise of a patient's ambulance driver or the non-random location of high-quality providers.

In principle, instrumental variable (IV) techniques offer a solution to selection bias, as in other settings. In practice, researchers hoping to exploit quasi-experimental variation in institutional choice face several methodological challenges. Linear IV methods, including those used by Angrist et al. (2015) to reduce bias in school VAMs, typically depend on an assumption of constant causal effects – for example that switching from the highest- to the lowest-ranked hospital has the same expected health effect for all potential patients.[2] This rules out both institutional comparative advantage and selection-on-gains, two powerful economic forces that are likely important in many settings, including healthcare (Chandra and Staiger, 2007). Moreover, constant effect restrictions are inappropriate for modeling binary outcomes, including the 30-day survival indicators used in hospital RAMs.

This paper develops a new approach for measuring institutional quality with nonlinear causal response functions, selection-on-gains, and quasi-experimental data. Usual nonlinear IV estimators use maximum likelihood methods that can be computationally intractable or require parametric assumptions that are difficult to assess or interpret. Even estimating a nonlinear first stage for institutional sorting requires solving a high-dimensional multinomial choice problem; for decades scholars have grappled with the practical

[1] See, for example, Chetty et al. (2014b), Angrist et al. (2015), Chetty and Hendren (2015), Hoxby (2015), and Card et al. (2013) for recent estimates of the institutional effects of teachers, schools, neighborhoods, colleges, and firms.

[2] Unlike with binary treatments, multi-dimensional linear IV has no local average treatment effect (LATE) interpretation except under strong assumptions (Behaghel et al., 2013; Kirkebøen et al., 2014; Hull, 2015; Blackwell, 2016). Even in these cases, LATE-based quality measures are undesirable, as differences in complier populations could affect the rankings of institutions with the same average effectiveness. As formalized in Section 2, quality differences in my framework reflect average treatment effects, though estimating other parameters, such as average treated effects on the treated, is also possible.

difficulties of fitting these models without unrealistic restrictions on choice substitution patterns (Hausman and Wise, 1978; McFadden, 1989; McColloch and Rossi, 1994; Berry et al., 1995). In an application of state-of-the-art Markov-chain Monte Carlo techniques, Geweke et al. (2003) estimate the quality of 114 Los Angeles County hospitals with relative distance instruments, a multinomial probit model of hospital admissions, and a probit specification for the short-term mortality outcomes of elderly pneumonia patients. To evaluate their likelihood further requires ex ante specification of independent priors for each of the model's 268 free parameters, along with several auxiliary functional form restrictions and calibrations. Characterizing the role of these parameterizations, versus the potentially-exogenous variation in hospital choice generated by the instruments, is far from straightforward.

Rather than fitting a fully-specified likelihood to data, my approach matches a sparse set of moments from a multi-dimensional Roy (1951) selection model to quantities identified by quasi-experimental instrument assignment. This yields a flexible framework, fully non-parametric given sufficiently-rich instrument variation, for estimating institutional effectiveness with comparative advantage and selection-on-gains. Distributional assumptions on the model's latent variables can be used to extrapolate from observed quasi-experimental quantities to structural parameters of interest with more limited variation. A minimum distance procedure easily implements this semi-parametric approach, even when the numbers of individuals, institutions, instruments, and covariates grow large.

I use these methods to estimate hospital quality from a nationally representative sample of U.S. Medicare patients admitted for an emergency condition. Specifically, I fit a multivariate probit model for potential hospital admissions and 30-day survival outcomes using quasi-experimental variation in ambulance company assignment. In a recent paper, Doyle et al. (2015) propose ambulance company instruments as a more credible alternative to distance-based identification strategies, which may be biased by non-random hospital location (e.g. Hadley and Cunningham (2004)). They use ambulance referral variation to instrument for the average Medicare spending of a patient's hospital in linear mortality models, finding large returns to choosing more intensive providers. My nonlinear approach instruments a patient's hospital directly, allowing for violations of the Doyle et al. (2015) exclusion restriction, as well as heterogeneous treatment effects and Roy selection.

The initial analysis yields a set of noisy quality estimates for 1,041 U.S. hospitals with sufficient quasi-experimental data. As in other recent explorations of institutional quality (e.g. Chetty and Hendren (2015)), I use these estimates to fit a hierarchical linear model and compute empirical Bayes quality posteriors that optimally combine quasi-experimental estimates and observational RAM predictions. This procedure reduces overall mean squared prediction error and generates posteriors for the full set of U.S. hospitals in my analysis sample.

Quality posteriors reveal several important dimensions of hospital performance and patient sorting. Higher-volume hospitals and those that spend more per Medicare patient appear to produce better average survival outcomes, while government-run hospitals are systematically lower-performing. Moving a

patient to a provider that would increase her 30-day survival probability by one percentage point places her in a hospital with 1.9% higher spending and 4.3% higher Medicare patient volume, on average. These results are qualitatively similar to what earlier work has found when measuring quality by observational RAMs (Foster et al., 2013; Chandra et al., 2015; Doyle et al., 2015). However, consistent with a broad pattern of better hospitals attracting sicker patients, I show that the strength of these relationships is magnified when they are measured with quasi-experimental data. Comparing quality posteriors and observed survival rates, I moreover find robust evidence for hospital comparative advantage and positive Roy selection-on-gains, with patients admitting to more appropriate hospitals on average. This non-random sorting is only partly explained by differential hospital distance and generates systematic bias in observational RAMs.

To quantify the economic importance of selection bias, I simulate quality-based Medicare reimbursement and patient guidance policies. Ranking hospitals by quality posteriors instead of RAM predictions tends to magnify existing transfers across different types of hospitals rather than changing the distribution of policy winners and losers. Net subsidies paid to privately-owned and teaching hospitals, for example, increase by 9% and 15%. In simulations of quality-based admission policies, I find a typical patient has a 2.8 percentage point higher 30-day survival rate when choosing hospitals on the basis of RAM predictions, rather than admitting at random. Admission to hospitals with the highest quality posteriors yields larger survival rate improvements of between 3.3 and 4.5 percentage points. Nevertheless, the scope for health gains from quality-based admission policies is limited by the extent of positive Roy selection; moving a random patient from her selected hospital to the provider delivering the highest average quality-of-care would *decrease* expected survival by 11 percentage points. This highlights a general issue for performance-based guidance policies that is obscured by a constant-effect quality framework.

The remainder of this paper is organized as follows: the next section develops a general method of moments approach for estimating institutional quality with instrumental variables and discusses non- and semi-parametric identification. I then outline the institutional setting for hospital quality and describe the Medicare analysis sample and estimation procedure in section 3. Next, section 4 discusses my findings on hospital quality, patient sorting, and the consequences of non-random sorting in performance-based healthcare policies. Section 5 concludes.

2 Quality identification

2.1 The Quasi-experimental Setting

Suppose we observe outcomes Y_i for each individual i attending one of many possible institutions $j = 1, \ldots, J$. We indicate institutional choice by a set of dummy variables D_{ij}, collected in the vector D_i. For example $D_{ij} = 1$ may denote patient i's admission to hospital j, while $Y_i = 1$ if she survives the first 30 days following admission. Corresponding to each institutional alternative is a potential outcome Y_{ij}; these are linked to

observed outcomes by

$$Y_i = \sum\nolimits_j Y_{ij} D_{ij}. \tag{1}$$

Policymakers aim to rank institutions by quality, defined as $q_j = E[Y_{ij}]$. This represents the expected outcome from sending a random individual to institution j, so that institutional quality comparisons avoid any bias from non-random sorting that would cause Y_{ij} and D_{ij} to be correlated. Let $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$, the difference in average selected and potential outcomes, quantify this selection bias for institution j.

Along with choices and outcomes, suppose we observe an individual's assignment to a discretely-valued instrument Z_i. Without loss of generality we let Z_i be a vector of indicators $Z_{i\ell}$ for the set of L possible instrument values and denote vectors in the support of Z_i by z_ℓ. For example, in the hospital application, $Z_{i\ell} = 1$ (and $Z_i = z_\ell$) if ambulance company ℓ is dispatched to individual i. Attending institution j after being assigned to the ℓth instrument value generates latent utility $U_{ij}(z_\ell)$, and individuals choose the institution that maximizes these payoffs. Institutional selection is thus given by

$$D_{ij} = \mathbf{1}[U_{ij}(Z_i) \geq U_{ik}(Z_i), \forall k]. \tag{2}$$

Equations (1) and (2) structure the vector of observed outcomes, institutional choices, and instrument assignments, $(Y_i, D_i', Z_i')'$, by a generalized multi-dimensional Roy (1951) selection model (Heckman et al., 2008). This model asserts the existence of counterfactual outcomes Y_{ij} and latent utilities $U_{ij}(z_\ell)$ with a conventional stable unit treatment value assumption (Imbens and Rubin, 2015) and adopts an implicit exclusion restriction, that the instrument only affects outcomes through the choice of institution. Importantly, the model does not limit the possibility of either institutional comparative advantage or endogenous selection on potential outcomes. The causal effects $Y_{ij} - Y_{ik}$ need not be constant across individuals, and potential outcomes may be correlated with the latent utilities governing institutional choice, generating "essential heterogeneity" in the language of Heckman et al. (2006).

A conditional independence assumption completes the quasi-experimental framework: that, given a set of auxiliary controls X_i, the instrument Z_i is as good as randomly assigned with respect to the vector of latent outcomes and utilities:

Assumption 1 (*Independence*): $\Big((Y_{ij}, (U_{ij}(z_\ell))_{\ell=1,\ldots,L})_{j=1,\ldots,J}\Big) \perp\!\!\!\perp Z_i \mid X_i$.

Quasi-random instrument assignment ensures that while institutional choice itself may be correlated with potential outcomes, there is variation in conditionally-exogenous factors $Z_{i\ell}$ that can affect sorting by changing the frontier of latent payoffs, $U_{ij}(Z_i)$. My framework leverages this variation with knowledge or first-step non-parametric estimation of the conditional expectation functions $p_\ell(X_i) = E[Z_{i\ell}|X_i]$. I refer to these as instrument "propensity scores" and maintain throughout an assumption of common support: that $p_\ell(X_i) > 0$ for each ℓ with probability one. All individuals thus face some non-zero risk of assignment to each of the L instrument values.

2.2 Non-parametric Identification

Quasi-experimental instrument assignment is a powerful restriction that is sufficient for non-parametric estimation of certain moments of the model's latent variables, Y_{ij} and $U_{ij}(z_\ell)$. Namely, the following auxiliary result shows that Assumption 1 identifies both the first-stage shares of individuals who would choose each institution j if assigned to each instrument value ℓ (what I refer to as "choice probabilities") and the means of any function of potential outcomes for individuals who would select j under this assignment (termed "mean selected outcomes"):

Lemma 1 (*Identification of choice probabilities and mean selected outcomes*): Let $f(\cdot)$ be any measurable function of Y_i. Under Assumption 1,

$$Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) = E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right] \tag{3}$$

$$E[f(Y_{ij})|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = E\left[\frac{f(Y_i)D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right] / E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \tag{4}$$

Proof: See the econometric appendix.

Note that without controls (so that the instrument is unconditionally randomly assigned, as in a randomized control trial) choice probabilities and mean selected outcomes are given by the moments $E[D_{ij}|Z_{i\ell} = 1]$ and $E[f(Y_i)|D_{ij} = 1, Z_{i\ell} = 1]$. The formulas in Lemma 1 use the non-parametrically identified propensity scores to appropriately re-weight the data so that it mimics this idealized experimental setting.

Without further parameterizations of the model, equations (3) and (4) are enough to estimate institutional quality from rich quasi-experimental data. Intuitively, by varying the instrument $Z_{i\ell}$ and setting $f(Y_i) = Y_i$ we non-parametrically observe average outcomes at institution j across different groups of individuals for whom utility is maximized at j when $Z_i = z_\ell$. We can moreover rank these averages by the fraction that each group represents of the population, $Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)$. If the number of observed instrument values grows with the sample, we may expect to find assignments that bring this choice probability arbitrarily close to one; in the limit we could thus estimate the population $E[Y_{ij}] = q_j$ by constructing averages of estimated mean selected outcomes $E[Y_{ij}|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]$ that place more weight on z_ℓ with the highest choice probabilities.

Formally, given any consistent set of propensity score estimators $\hat{p}_\ell(\cdot)$, we have the following result:

Proposition 1 (*Local linear quality identification*): For each j, collect the set of choice probabilities $G_{j\ell} = Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)$ in the vector G_j. If the support of $G_j'Z_i$ has a supremum of 1 then under Assumption 1 $\hat{q}_j \xrightarrow{p} q_j$ where given N independent, identically-distributed draws of $(Y_i, D_i', Z_i', X_i')'$,

$$\hat{q}_j = \arg_1 \min_{q,b} \sum\nolimits_{\ell:\hat{G}_{j\ell} \geq \hat{c}_j} \hat{w}_{j\ell} \left(\hat{H}_{j\ell} - q - (1 - \hat{G}_{j\ell})\right)^2, \tag{5}$$

for $\hat{G}_{j\ell} = \frac{1}{N}\sum_{i=1}^N \frac{D_{ij}Z_{i\ell}}{\hat{p}_\ell(X_i)}$, $\hat{H}_{j\ell} = \sum_{i=1}^N \frac{Y_iD_{ij}Z_{i\ell}}{\hat{p}_\ell(X_i)} / \sum_{i=1}^N \frac{D_{ij}Z_{i\ell}}{\hat{p}_\ell(X_i)}$, and where $\hat{c}_j$ and the $\hat{w}_{j\ell}$ are scalars with $\hat{c}_j \leq \max_\ell(\hat{G}_{j\ell})$, $\hat{c}_j \xrightarrow{p} 1$, $\hat{w}_{j\ell} > 0$, and $\sum_\ell \hat{w}_{j\ell} = 1$.

Proof: Let $\widehat{\ell^*}(j)$ be an arbitrary element from the set of instrument values ℓ maximizing the sample choice probabilities $\hat{G}_{j\ell}$. Under the assumptions, $\hat{G}_{j\widehat{\ell^*}(j)} \xrightarrow{p} 1$ and $\hat{H}_{j\widehat{\ell^*}(j)} \xrightarrow{p} E[Y_{ij}]$ by the Weak Law of Large Numbers. Thus $\hat{q}_j \xrightarrow{p} q_j$, provided the bandwidth $\hat{c}_j$ approaches 1 and the weights $\hat{w}_{j\ell}$ are convex. □

The local linear regression estimator $\hat{q}_j$ is consistent for institution j's quality when $G_j'Z_i$, the choice probability of the instrument assigned to individual i, has sufficiently large support. This result follows a broad literature on non-parametric identification of Roy models, including Heckman and Honore (1990), Lewbel (2007), and D'Haultfoeuille and Maurel (2013). In fact, the estimator in Lewbel (2007) is also consistent for q_j under a somewhat stronger support condition than the one used in Proposition 1.[3] Other estimators can be obtained by adding higher-order polynomials or other transformations of the regressor $1 - \hat{G}_{j\ell}$. Characterizing the optimal choice of weights, bandwidths, and local regressors for non-parametric quality estimation is left for future research.

2.3 Semi-parametric Quality Estimation

Limited variation in choice probabilities makes the estimator in Proposition 1 inconsistent. When institutional quality is not non-parametrically identified, further restrictions on the selection model can substitute for rich quasi-experimental data. Intuitively, there are parametrizations of the joint distribution of latent variables Y_{ij} and $U_{ij}(z_\ell)$ that render the moments identified by Lemma 1 functions of some finite-dimensional parameter vector, θ_0. A quasi-experimental design generating sufficient variation in the moments may pin down these structural parameters and thus the marginal means of latent Y_{ij} – that is, quality. A minimum distance procedure (Ferguson, 1958), which is computationally simple relative to earlier likelihood-based IV methods, implements this semi-parametric approach.

I first outline the proposed minimum distance quality estimator for a generic identified parameterization of the latent variables; I then establish and characterize identification for a particular multivariate probit specification which is later used to estimate hospital quality. Suppose for some known distribution function $F(\cdot)$ we have

$$\Big((Y_{ij}, (U_{ij}(z_\ell))_{\ell=1,\dots,L})_{j=1,\dots,J}\Big) \sim F(\theta_0), \tag{6}$$

so that the various choice probabilities and mean selected outcomes identified under Assumption 1 are also known functions of θ_0. Let $m(\cdot)$ be a vector collecting some subset of these functions and $\hat{m}$ be the sample analogues of the corresponding formulas of Y_i, D_i, Z_i, and $p_\ell(X_i)$ from Lemma 1, constructed with some consistent non-parametric propensity score estimators $\hat{p}_\ell(\cdot)$. Under mild regularity conditions (see, e.g., Hirano et al. (2003)), we then have $\sqrt{N}(\hat{m} - m(\theta_0)) \Rightarrow N(0, Q)$, where Q is a non-parametrically

[3]Namely, note that we can write $D_{ij} = \mathbf{1}[0 \leq M_{ij}^* + V_{ij} \leq A_i^*]$ where for independent $M_i \sim U[0,1]$ and $g_j = \min_\ell G_{j\ell}$, we let $M_{ij}^* = -M_i + g_j$, $V_{ij} = G_j'Z_i - g_j$, and $A_i^* = 1 - M_i$. This corresponds to equation (1) in Lewbel (2007) and his support condition is satisfied if $G_j'Z_i$ continuously varies over $[p, 1]$.

identified asymptotic variance matrix. If the structural parameters in θ_0 are uniquely determined by the quasi-experimental variation in $m(\cdot)$, a consistent minimum distance estimator is then given by

$$\hat{\theta} = \arg\min_{\theta}(\hat{m} - m(\theta))'\hat{A}(\hat{m} - m(\theta)), \tag{7}$$

for some weight matrix $\hat{A}$. Furthermore, under the same conditions for the asymptotic normality of $\hat{m}$, we have

$$\sqrt{N}(\hat{\theta} - \theta_0) \Rightarrow N\left(0, (M'AM)^{-1}M'AQAM(M'AM)^{-1}\right) \tag{8}$$

where $M = \frac{\partial m(\theta)}{\partial \theta}|_{\theta_0}$ and $\hat{A} \xrightarrow{p} A$. As usual with such extremum estimators, the asymptotic variance of $\hat{\theta}$ is minimized by setting $\hat{A} = \hat{Q}^{-1}$ for some consistent variance estimator $\hat{Q} \xrightarrow{p} Q$, in which case $\sqrt{N}(\hat{\theta} - \theta_0) \Rightarrow N\left(0, (M'Q^{-1}M)^{-1}\right)$. Note that with Q non-parametrically identified, this estimator can be formed in a single step and its asymptotic variance is consistently estimated by $\left(\hat{M}'\hat{Q}^{-1}\hat{M}\right)^{-1}$ for $\hat{M} = \frac{\partial m(\theta)}{\partial \theta}|_{\hat{\theta}}$. The choice of quasi-experimental moment estimator $\hat{m}$ thus entirely determines the relative efficiency of both $\hat{\theta}$ and, applying the Delta method to the formulas implied by equation (6), the corresponding estimates of quality $E[Y_{ij}]$. When the model is overidentified, an omnibus specification test statistic can be formed from the estimator's minimized criterion function:

$$\hat{T} = N(\hat{m} - m(\hat{\theta}))'\hat{Q}^{-1}(\hat{m} - m(\hat{\theta})). \tag{9}$$

Under the joint null hypothesis of Assumption 1 and the correct specification of $F(\cdot)$, this statistic will have an asymptotic chi-squared distribution.

Computing minimum distance estimates is relatively straightforward, even as the number of institutions J, instrument values L, and/or controls in X_i grows large. Each element of $\hat{m}$ is determined by one of $L-1$ propensity scores which do not depend on the model's structural parameters and may be separately approximated by standard techniques (e.g. Geman and Hwang (1982)). Given $\hat{m}$, evaluating the estimator's objective function requires computing at most $((D+1)J-1)L$ nonlinear functions for each candidate parameter vector θ, where D is the dimension of the outcome function $f(\cdot)$.[4] Importantly these functions do not depend on the data, so unlike with likelihood-based estimators the difficulty of the nonlinear computation does not increase with the sample size. In some cases, including the multivariate probit model considered below, $m(\theta)$ will take a form that is straightforward to evaluate by standard statistical software packages (see the econometric appendix). Simulation methods can solve more exotic parameterizations; again the fact that the simulated objects are non-stochastic makes this procedure fast relative to typical applications of the simulated minimum distance approach of McFadden (1989) and Pakes and Pollard (1989).

The separation of quasi-experimental data in $\hat{m}$ from the structural assumptions underlying $m(\theta)$ also helps establish and characterize identification of semi-parametric quality models. I illustrate this with a

[4]Namely, there are at most $(J-1)L$ linearly-independent choice probabilities and DJL mean selected outcomes.

multivariate probit specification for the latent variables, which produces my benchmark hospital quality estimates. Let h_{ij} denote the latent health of emergency patient i upon admission to hospital j, and assume patients survive the first 30 days following admission when their health is above some arbitrary threshold, here normalized to zero:

$$Y_{ij} = \mathbf{1}[h_{ij} \geq 0]. \tag{10}$$

With the vector h_i collecting the J health indices, the observed outcome equation (1) becomes

$$Y_i = \mathbf{1}[h_i' D_i \geq 0]. \tag{11}$$

The random coefficients in h_i retain the feature of institutional comparative advantage from the general selection model: some individuals may be more likely to survive when moved from hospital j to hospital k, while for others such a move may result in worse health outcomes.

In my application, emergency patients are referred to hospitals by ambulance, with $Z_{i\ell}$ indicating the quasi-experimental assignment of ambulance company ℓ to patient i.[5] As shown in Doyle et al. (2015), differences in ambulance referral preferences may generate variation in hospital admissions. The multivariate probit specification structures this first-stage variation by a monotonicity assumption, as in the identification of local average treatment effects and related causal parameters (Imbens and Angrist, 1994; Heckman et al., 2006):

Assumption 2 (*Monotonicity*): $\forall \ell,m,j$, either $Pr(U_{ij}(z_\ell) \geq U_{ij}(z_m)) = 1$ or $Pr(U_{ij}(z_\ell) < U_{ij}(z_m)) = 1$.

To the extent ambulance companies have different preferences for referring to each hospital j, they are fixed over different subpopulations of patients when Assumption 2 holds. Indeed, monotonicity implies an additively-separable model for latent utility:

$$U_{ij}(z_\ell) = \pi_{j\ell} + \eta_{ij}. \tag{12}$$

In my application, $\pi_{j\ell} - \pi_{k\ell}$ represents ambulance company ℓ's relative preference for referring to hospital j over hospital k, while η_{ij} denotes the latent utility from admitting at hospital j for patient i, which may also reflect common ambulance company preferences. With the vector π_j collecting the $\pi_{j\ell}$ parameters, the admissions process in equation (2) becomes

$$D_{ij} = \mathbf{1}[\pi_j' Z_i + \eta_{ij} \geq \pi_k' Z_i + \eta_{ik}, \forall k]. \tag{13}$$

A final parametric assumption defines the multivariate probit specification, along with equations (10)

[5] One could instead imagine using geographic instruments, such as indicators for a patient's home ZIP code, in place of the ambulance company design. Assumption 1 would then require a patient's location to be conditionally-independent from her latent health and admission utility at each hospital, as with the relative distance instruments used in Geweke et al. (2003). IV estimates would be biased if, for example, hospital quality is endogenously determined by local patient characteristics; Hadley and Cunningham (2004) offer evidence for this kind of non-random location for so-called "safety net" providers. The importance of minimizing travel time for treating emergency conditions also brings into question the exclusion restriction for such models.

and (12): joint-normality of latent health and utility,[6]

Assumption 3 (*Normality*): $(h_i', \eta_i')' \sim N(\mu, \Sigma)$.

Many parameterizations of the model will be observationally equivalent under Assumptions 1-3 for any amount of quasi-experimental data. Namely, without loss of generality we can normalize $E[\eta_i] = 0$, $Var(h_{ij}) = 1$, $\forall j$, and $Var(\eta_i) = I_J$, where I_x is an identity matrix of size x, and restrict attention to the vector of relative utilities $U_{ij}(z_\ell) - U_{i\bar{j}}(z_\ell)$ for a fixed reference hospital $\bar{j}$. The relevant structural parameter vector θ_0 then consists of J quality index coefficients $\beta_j = E[h_{ij}] = \Phi^{-1}(q_j)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $J(J-1)$ health-utility correlations $\rho_{jk} = Corr(h_{ij}\eta_{ik} - \eta_{ij})$, and $(J-1)L$ relative ambulance company preferences $\pi_{j\ell} - \pi_{\bar{j}\ell}$, for a total of $J^2 + (J-1)L$ parameters.[7]

For Bernoulli outcomes, quasi-experimental ambulance company assignment offers at most $(2J-1)L$ linearly-independent moments identified by Lemma 1, for any choice of $f(\cdot)$. The order condition for identifying θ_0 is thus satisfied with $L \geq J$ (in my setting, as many ambulance companies as hospitals) and the rank condition holds when ambulance company preferences are unique:

Proposition 2 (*Multivariate probit identification*): In the multivariate probit model, suppose Π, the $J \times L$ matrix of preference parameters $\pi_{j\ell}$, has no redundant columns and that Assumption 1 holds. Then all quality parameters q_j are identified if $L \geq J$.

Proof: For each instrument value ℓ, the J choice probabilities identified by Lemma 1 are uniquely determined by $J-1$ relative preferences $\pi_{j\ell} - \pi_{\bar{j}\ell}$ by Assumptions 2-3. With these parameters solved, the L mean selected outcomes for each institution j are determined by one quality parameter q_j and $J-1$ correlations ρ_{jk}, uniquely so when the columns of Π are unique. Identification thus follows if $L \geq J$. □

Estimating each institution's quality by Proposition 2 will generally use the full set of choice probabilities. In practice, with many small institutions or rare instrument value assignments, some of the associated $E\left[D_{ij}Z_{i\ell}/p_\ell(X_i)\right]$ may be infeasibly or poorly approximated in finite samples, thereby rendering all quality estimates unreliable. This concern is particularly relevant in my hospital application: the distribution of hospital volume in administrative claims data is right-skewed, with many small providers.[8] A more attractive estimation approach leverages alternative-specific instruments of the kind traditionally found in multinomial choice applications (Keane, 1992). Suppose we can partition the instrument vector Z_i into J subvectors Z_{ij} whereby moving across different values in the support of Z_{ij} only affects the latent utility generated

[6]Note that under joint-normality a patient's utility from care can be written as a linear function of potential health, as in the classic Grossman (1972) healthcare demand model, with an independent normal error term.

[7]In general the cross-institution health correlations will not be identified, nor are they necessary for quality identification.

[8]The difficulty of estimating hospital quality models due to the presence of small providers is well-known: both federal policymakers and Geweke et al. (2003) remove patients admitted to low-volume hospitals from their analysis samples, though this practice likely induces selection bias. The separable identification result I provide in Proposition 3 overcomes this issue without endogenous sample selection.

by institution j and not any other alternatives. This would be the case in the stylized hospital quality example if each ambulance company has at most one preferred hospital – for example, the one based closest to company offices – but otherwise has no preferences that would differentially shift patients between other local hospitals. In this case, the following result shows we may separately identify the quality of each hospital using only a subset of the choice probabilities:

Proposition 3 (*Multivariate probit identification with alternative-specific instruments*): For a given j in the multivariate probit model, suppose $\pi_{k\ell} = \bar{\pi}_k$ for all instrument values ℓ in the support of an alternative-specific instrument vector Z_{ij} and all $k \neq j$. Then q_j is identified under Assumption 1 if the subvector of π_j corresponding to Z_{ij} has $L_j \geq J$ distinct values.

Proof: Under the assumptions the $J-1$ relative preference parameters $\bar{\pi}_k$ are identified by J choice probabilities involving D_{ik} for $k \neq j$ and any $Z_{i\ell}$ in Z_{ij}. With these known, the $2L_j$ choice probabilities and mean selected outcomes involving D_{ij} and the $Z_{i\ell}$ in Z_{ij} are uniquely determined by institution j's quality q_j, $J-1$ correlations ρ_{jk}, and L_j relative preferences $\pi_{j\ell}$, when the latter are non-redundant. □

Alternative-specific instruments thus provide a method for estimating the quality of only a subset of institutions for which choice probabilities and mean selected outcomes are likely to be well estimated, leaving the quality of other hospitals with less quasi-experimental data underidentified.

Minimum distance quality estimators based on results like Propositions 2 and 3 use a low-dimensional parameterization of the distribution of potential outcomes and latent utility to extrapolate from a discrete set of non-parametric instrumental variable moments to the structural parameters of interest. This is in the spirit of Brinch et al. (2012), who directly parameterize conditional marginal treatment effect curves in the binary treatment case; here both the extrapolation and number of instruments needed for identification are guided by a multiple-treatment Roy model and do not depend on the distribution of quasi-experimental controls except through the set of non-parametric instrument propensity scores.[9]

The parametric extrapolation of reduced-form moments is most clearly seen in the case of $J = 2$ institutions, for which the model given by Assumptions 2 and 3 is a bivariate probit and the conditions for identification in Propositions 2 and 3 coincide. Without loss of generality, we may then normalize $\pi_{2\ell} = \eta_{i2} = 0$ and drop j subscripts from the latent utility parameters for institution 1 to write

$$Y_i = \mathbf{1}[h_{i1}D_{i1} + h_{i2}D_{i2} \geq 0] \tag{14}$$

$$D_{i1} = \mathbf{1}[\pi' Z_i + \eta_i \geq 0] = 1 - D_{i2}, \tag{15}$$

[9]Brinch et al.'s approach requires estimating the functions $E[Y_i|D_{ij} = 1, Z_{i\ell} = 1, X_i = x]$ for each j, ℓ, and value in x in the support of the quasi-experimental controls X_i. In practice this can be infeasible when the controls are continuous or take on many discrete values, as in my setting. Standard asymptotic theory may also provide only poor approximations for the sampling distribution of estimators based on many stratified conditional means, an issue discussed in Robins and Ritov (1997) and Angrist and Hahn (2004) and that motivates Hirano, Imbens, and Ridder's (2003) inverse propensity score weighting approach for efficiently estimating average treatment effects.

where, under Assumption 3, $(h_{i1}, h_{i2}, \eta_i)' \sim N((\beta_1, \beta_2, 0)', \Sigma)$. Here the covariance matrix Σ has two health-utility correlations, ρ_1 and ρ_2, which along with β_1, β_2, and π yield $L+4$ parameters in θ_0. Under Assumption 1 we observe L sets of linearly-dependent choice probabilities and $2L$ mean selected outcomes by the formulas in Lemma 1, and $L \geq 2$ ambulance companies satisfies the order condition.

Bivariate probit mean selected outcomes, $E[Y_{ij}|\pi_\ell + \eta_i \geq 0]$, are monotone in the first-stage parameters π_ℓ.[10] Thus any two instrument values ℓ and m for which $\pi_\ell > \pi_m$ inform the sign of selection bias at each institution. If, for example, we learn by Lemma 1 that $Pr(\pi_\ell + \eta_i \geq 0) > Pr(\pi_m + \eta_i \geq 0)$ and $E[Y_{i1}|\pi_\ell + \eta_i \geq 0] < E[Y_{i1}|\pi_m + \eta_i \geq 0]$, we would know that patients with *lower* utility admissions utility η_i, who only select hospital 1 when assigned to ambulance company ℓ (that is, in the language of Imbens and Angrist (1994), the ambulance company "compliers"), have *worse* health outcomes at hospital 1 than those who would be admitted by either ambulance company (the quasi-experiment's "always-takers"). By normality, hospital 1's average potential outcome in the population of patients (i.e., its quality $E[Y_{i1}]$) is therefore lower than that of patients who actually choose hospital 1: $E[Y_{i1}|D_{i1} = 1] - E[Y_{i1}] > 0$, so that hospital 1 is positively selected.

Along with the direction of selection bias, joint-normality prescribes a particular translation of admitted patient health to the population. In the bivariate model, the quality index $\beta_j = \Phi^{-1}(q_j)$ can be written as a linear combination of the health of patients who would be admitted by the two ambulance companies: e.g.,

$$\beta_1 = E[h_{i1}|\pi_\ell + \eta_i \geq 0]\omega + E[h_{i1}|\pi_m + \eta_i \geq 0](1 - \omega) \tag{16}$$

for

$$\omega = 1/\left(1 - \frac{\phi(\pi_\ell)}{\Phi(\pi_\ell)}/\frac{\phi(\pi_m)}{\Phi(\pi_m)}\right), \tag{17}$$

where $\phi(\cdot)$ denotes the standard normal probability density function. The inverse Mills ratio $-\phi(\pi_\ell)/\Phi(\pi_\ell)$ is increasing in the first-stage parameters, so with $\pi_\ell > \pi_m$ we have $\omega > 1$, and the non-convex weighting scheme given by equation (16) extrapolates in the direction of the larger patient subpopulation. This is illustrated in panel A of Figure 1, in the case of positive selection bias ($\rho_1 > 0$). The two vertical dashed lines show the inverse Mills ratio for two ambulances' first-stage parameters, while the two horizontal dashed lines show the associated average health of patients who would be admitted by each company. The downward-sloping blue line that intercepts the maximum inverse Mills ratio of zero at β_1 (with a slope of $-\rho_1$) gives the extrapolation from these two patient subpopulations to population health.

When $L > 2$ in the bivariate probit model, any two ambulance companies with different referral preferences identify hospital quality in this way, and the minimum distance quality estimator given by equation

[10] Namely, $E[Y_{ij}|x + \eta_i \geq 0] = Pr(h_{ij} \geq 0|x + \eta_i \geq 0) = \int_{-\infty}^{x} \Phi\left((\beta_j - \rho_j t)/\sqrt{1-\rho_j^2}\right) \frac{\phi(t)}{\Phi(x)} dt$ when h_{ij} and η_i are normally distributed. The derivative of this function with respect to x is proportional to $\Phi\left((\beta_j - \rho_j x)/\sqrt{1-\rho_j^2}\right) - \int_{-\infty}^{x} \Phi\left((\beta_j - \rho_j t)/\sqrt{1-\rho_j^2}\right) \frac{\phi(t)}{\Phi(x)} dt \gtrless 0 \iff \rho_j \lessgtr 0$.

(7) efficiently aggregates all pairwise comparisons. If the relative preference parameters π were known, this would amount to solving a variance-weighted nonlinear least squares problem of fitting estimated mean selected outcomes to a particular parametric curve. An example of this is plotted in panel B of Figure 1, using data simulated from the same probit specification used in panel A. The nonlinear curve of best fit is parameterized by an intercept, hospital quality q_j, along with a shape parameter ρ_j that determines the sign and extent of selection bias. The R-squared for the curve's fit informs the overidentification test statistic $\hat{T}$ from equation (9).

The same extrapolative logic applies to estimation of multi-institution models, when $J > 2$. Each new institution adds a shape parameter to the multivariate probit curve, thereby necessitating an additional mean selected outcome point. Other parameterizations of the selection model would yield other curves with different quasi-experimental data requirements. Note that, unlike with linear IV (see, e.g., Angrist (1991) and the references therein), the least-squares interpretation of these nonlinear estimators no longer holds when the π_j are estimated, as the IV moment vector is not linear in the first-stage parameters. Nevertheless, a procedure wherein the first stage is initially obtained from the set of choice probabilities and then used to fit appropriate parametric curves to mean selected outcomes will yield consistent semi-parametric estimates.

3 Estimating Hospital Quality

3.1 Data and RAMs

I use the preceding framework to estimate the quality of U.S. hospitals according to their effects on short-term patient mortality. Policymakers currently base observational hospital RAMs on three-year windows of emergency Medicare claims (YNHHSC/CORE, 2013); correspondingly, I draw a sample of 405,173 Medicare fee-for-service beneficiaries brought to an acute-care hospital by an ambulance for one of 29 emergency conditions in 2010-2012.[11] Observations come from a nationally-representative 20% sample of administrative inpatient claims from the Centers of Medicare and Medicaid Services (CMS) and include information on basic patient demographics (such as age, sex, race, and home ZIP code); diagnoses and procedures from previous inpatient and outpatient claims ("comorbidities"); the identity of, ZIP code location of, and procedures performed by a patient's assigned ambulance company; the identity and location of the hospital; and subsequent mortality. As in Card et al. (2009), I restrict the sample to patients admitted for a "non-deferrable" primary condition, i.e. those with a weekend admissions rate close to 2/7ths. These are the same conditions used by Doyle et al. (2015) and are listed in the notes to Table 1. I also follow standard CMS risk-adjustment methodology in attributing outcomes to a patient's first hospital admission in 2010-2012, ignoring all subsequent transfers or readmissions. Finally, I divide the national sample of patients, ambulances, and hospitals into hospital service areas (HSAs), which are sets of ZIP codes defined by the

[11] Unlike in some RAMs, I am not able to include Veterans Affairs facilities in this analysis.

Dartmouth Atlas of Health Care as narrow regions where patients receive most of their emergency care. I use HSAs to delineate local emergency care markets, within which it is plausible that ambulance company propensity scores have full support. As Appendix Table A3 illustrates, I obtain similar findings throughout with hospital referral regions (HRRs). A data appendix describes the sample construction in detail.

Table 1 summarizes the distribution of diagnoses, ambulances, hospitals, HSAs, and 30-day survival probabilities. Hospital RAMs were first developed to measure quality by the mortality of Medicare patients with circulatory and respiratory conditions, such as acute myocardial infarction, heart failure, and pneumonia, though often with the stated goal of extending the methods to a broader patient population (Krumholz et al., 2006).[12] Panel A of Table 1 shows that circulatory and respiratory diagnoses make up 42% of nondeferrable admissions in my sample, with the remainder split between digestive (7%), injury (18%), and all other conditions (34%).

Each patient in the analysis sample was assigned to one of $9,590$ ambulance companies and admitted to one of $4,821$ hospitals.[13] Panel B of Table 1 reports that the distribution of within-HSA hospital counts is highly skewed, with around half (2,464) of all hospitals operating in their own single-hospital market. Since the ambulance design leverages within-market admissions variation, my quality analysis focuses on local comparisons for the other 2,357 hospitals in 695 multi-hospital HSAs. Column 5 of Table 1 summarizes average 30-day patient survival, which is the usual outcome of mortality RAMs. Around 83% of patients survive the first 30 days following their emergency admission, with survival rates as low as 78% for patients with respiratory conditions and as high as 93% for those with injuries. Panel B shows that average survival does not seem to vary much by the number of available hospitals.

I first use this sample to obtain a set of observational RAM quality predictions, following standard CMS risk-adjustment methodology. These specify an additively-separable latent index model for 30-day survival:

$$Y_{ij} = \mathbf{1}[\alpha_j + \epsilon_i \geq 0], \tag{18}$$

where

$$\epsilon_i = \gamma' W_i - \nu_i \tag{19}$$

for a set of observed risk-adjusters W_i. Thus in a conventional RAM

$$Y_i = \mathbf{1}[\alpha' D_i + \gamma' W_i \geq \nu_i], \tag{20}$$

where α collects the quality indices α_j. Identification of the RAM parameters α and γ follows from a selection-on-observables assumption that hospital choice is independent of latent health conditional on the

[12] A related quality measurement effort models patient readmissions. Since a patient who dies at a low-quality hospital cannot be readmitted, more involved assumptions are required to causally attribute variation in these outcomes to hospital performance; I leave this issue for future work.

[13] 41% of Medicare patients hospitalized for a nondeferrable condition in 2010-2012 were admitted by an ambulance company; these and other comparisons are reported in columns 1 and 2 of Appendix Table A1 and discussed in the data appendix.

included controls, $\nu_i \perp\!\!\!\perp D_i \mid W_i$. Following YNHHSC/CORE (2013), I parameterize η_i by an independent logit distribution and obtain quality predictions $\hat{\alpha}_j$ by estimating logit regressions of 30-day survival on hospital random effects and patient age, sex, and diagnosis and comorbidity indicators; the data appendix details the RAM estimation procedure.

Observational RAMs in my sample leave unexplained most of the national variation in survival outcomes. This is illustrated in Figure 2, which plots the ratio of residual to total 30-day survival variance in five diagnosis-specific RAMs. Only around 7% of circulatory and respiratory survival variance is due to a patient's hospital, admitting diagnosis, and year of admission. The reduction is smaller for digestive conditions and injuries, and larger, around 14%, for other diagnoses in the analysis sample. Patient demographics and comorbidities account for an additional 4% of circulatory and respiratory survival variance, with similarly modest declines for the other diagnosis categories.

If the significant residual survival determinants are exogenous to the hospital selection process, predictions from these RAMs may still provide unbiased measures of hospital quality. However, to the extent survival variance may be further reduced by observable admission determinants, such as a patient's assigned ambulance company, observational RAMs are likely to be biased. The econometric appendix formalizes this argument and develops instrument-based tests for nonlinear RAM unbiasedness that extend earlier methods for validating linear education VAMs (Kane and Staiger, 2008; Chetty et al., 2014a; Deming, 2014; Angrist et al., 2016). These tests, summarized in Appendix Table A2, decisively reject the null of selection-on-observables ($p < 0.001$), suggesting scope for bias in the observational RAMs. Motivated by these findings, I next describe the implementation of the semi-parametric IV techniques that I use to quantify and characterize hospital selection bias and quality.

3.2 Estimation

I use the identification result in Proposition 3 to semi-parametrically estimate the quality of 1,041 hospitals operating in one of 626 multi-hospital HSAs with at least 25 patients in the analysis sample and sufficient quasi-experimental admissions variation. Doyle et al. (2015) first propose that in regions served by multiple ambulance companies, centralized policies of rotational and simultaneous 911 dispatch generate plausibly-exogenous company assignment, while the subsequent expression of non-random ambulance preferences can systematically affect the admissions of otherwise identical patients. Table 2 explores both of these claims by comparing individuals in the same ZIP code who are assigned to different ambulance companies likely to refer to hospitals with high and low RAM predictions. Specifically, I compute the distance between each ambulance company's office and each nearby hospital using the provider ZIP codes contained in Medicare claims, and label companies as likely to deliver patients to a low- or high-ranked provider if their closest hospital is in the first or fourth quartile of RAM quality predictions in the HSA. I then regress patient characteristics on either these group indicators (with group means reported in columns 1 and 2) or the ambulance company's closest hospital's predicted RAM itself (with the coefficient reported in column 4),

along with a full set of ZIP code fixed effects in the subsample of 254,101 admissions in multi-hospital HSAs.

Table 2 shows that patients assigned to ambulance companies based close to a high-ranked hospital see significantly increased RAM-predicted hospital quality, despite appearing identical to other patients in terms of their demographics, the location of their emergency, and their admitting diagnosis (panel A), as well as a host of comorbidity indicators describing their medical history (panel B). This balance of observable characteristics validates the quasi-random assignment of ambulance company indicators $Z_{i\ell}$, conditional on patient location X_i (Assumption 1). Ambulance assignment also appears balanced across a set of ambulance services performed pre-hospitalization (such as distance traveled in excess of the hospital ZIP code distance, whether the patient was assigned paramedics, or whether intravenous medication was delivered en route), a fact documented in panel C of Table 2. This supports the exclusion of ambulance-based instruments from potential survival outcomes Y_{ij}, allowing for interpretation of reduced-form ambulance effects on mortality outcomes by way of first-stage admission effects (a weaker restriction than in Doyle et al. (2015), where ambulances can only affect outcomes by changing the treatment intensity of a patient's provider). The p-value for a joint test of balance on assignment to ambulances based close to high- vs. low-RAM hospitals, across all 32 covariates in panels A, B, and C, is 0.89.[14]

As in Doyle et al. (2015) I leverage a first-stage monotonicity restriction, namely that differences in ambulance referral patterns do not systematically vary by patient characteristics (Assumption 2). Although not directly testable, Doyle et al. (2015) provide anecdotal support for monotone referral from their interviews with emergency care technicians – differences in referral patterns across ambulance companies appear to be driven by institutional and personal relationships with hospitals, rather than by patient heterogeneity. This is especially plausible in the relatively homogenous sample of emergency Medicare patients studied here. Differential treatment of uninsured patients by profit-driven ambulance companies, for example, is not a concern for this population.

My own interviews with current and former emergency medical staff across the U.S. support the alternative-specific model used in Proposition 3 as appropriate for ambulance assignment instruments: when differentially redirecting patients, ambulance companies seem to prefer returning to the hospital based closest to their offices in order to minimize excess travel time and maximize local availability.[15] The estimation strategy given by Proposition 3 is also attractive in practice as the analysis sample contains many hospital-ambulance combinations with relatively few non-zero observations of $D_{ij}Z_{i\ell}$, which may lead to unreliable choice prob-

[14]Similarly, Doyle et al. (2015) find no relationship between their ambulance-based instrument and a patient's probability of emergency room admission conditional on ZIP code; see their Figure A1. They likewise validate instrument balance in their analysis sample (see their Tables 1 and A3) and report anecdotal evidence for Assumption 1 from a 30-city survey of dispatch policies. My interviews with ambulance technicians in Connecticut, Massachusetts, Nevada, Philadelphia, Washington, and Wyoming further corroborate the assumption of quasi-random assignment. Note that the findings in Sanghavi et al. (2015) that advanced life support (ALS) services lead to higher cardiac arrest mortality are not at odds with my framework, since most ambulance companies provide both ALS and basic life support services and preference variation across companies is unlikely to be correlated with ALS availability.

[15]This appears especially true for ambulances owned by municipal and local fire departments, which are often the only local emergency transport provider and thus have a strong preference to return when dispatched outside of their home ZIP code.

ability and quality estimates from Proposition 2. I thus use the closest-hospital mapping from Table 2 to partition instrument vectors to alternative-specific subvectors and use only the largest ambulance company in each Z_{ij} to estimate π_{jk} for $k \neq j$. Table A3 shows qualitatively similar results when Z_{ij} instead comprises the ambulance companies that most-often refer patients to hospital j in the universe of 2010-2012 Medicare claims (excluding observations in the analysis sample).[16]

My estimates of hospital choice probabilities and mean selected survival outcomes are based on a flexible probit specification for ambulance company propensity scores $p_\ell(X_i)$ that model the latent risk of assignment by a cubic polynomial in company-patient distance:

$$E[Z_{i\ell}|X_i] = \Phi\left(\delta_{0\ell} + \delta_{1\ell} d_\ell(X_i) + \delta_{2\ell} d_\ell(X_i)^2 + \delta_{3\ell} d_\ell(X_i)^3\right), \tag{21}$$

where $d_\ell(x)$ denotes the distance between ambulance company ℓ's institutional address and a patient located in ZIP code x. Minimum distance quality estimates correct for first-step error in approximating these conditional expectations. For robustness I also include the vector of RAM controls W_i in the propensity scores of my benchmark specification, though, consistent with Assumption 1, Table A3 demonstrates that all results are essentially unchanged when these are excluded from the probit model.[17] This table also illustrates robustness to the health and utility probit specification (Assumption 3), with similar conclusions drawn from a fatter-tailed multivariate Student's $t(2)$ distribution that yields quality identification under the same assumptions as in the normal case. Quality is only identified by Proposition 3 for hospitals with $L_j \geq J(h(j))$ ambulance companies in their instrument subvectors Z_{ij}, where $J(h)$ is the hospital count of HSA h and $h(j)$ indexes hospital j's HSA; for these I use only the $J(h(j))$ largest companies in order to keep the model just-identified and reduce the scope for finite sample bias from many-weak IV identification.[18]

Figure 3 summarizes the available quasi-experimental data by plotting the joint distribution of differences in estimated hospital choice probabilities and mean selected outcomes for each of the 1,041 hospitals with enough ambulance company instruments to identify their quality. These differences are taken over the two ambulance companies generating the highest choice probability gap for each hospital; the marginal x-axis distribution thus summarizes the maximal variation in institutional choice generated by the instruments. The average choice probability difference is 0.4, with 43% of hospitals seeing a higher estimated choice probability. The average associated mean selected outcome difference is negative, and increasingly so as the first stage gap grows. As in the bivariate probit example in section 2.3, this suggests most hospitals in the sample see positive selection bias, which the generalized Roy model later confirms.

The solid blue curve in Figure 4 plots the distribution of the 1,041 minimum distance estimates of hospital

[16] Judgments based on ambulance company size are also made on the basis of this larger disjoint sample.

[17] In some small samples where maximum likelihood estimates of equation (21) fail to converge, RAM controls and higher-order distance terms are sequentially dropped until convergence is achieved.

[18] See Cattaneo et al. (2016) for discussion of many-weak bias in estimating generalized Roy models. Appendix Figure A1 plots the distribution of minimum distance first stage F-statistics that test equality of choice probabilities for each hospital against quality estimate standard errors. As expected, the hospitals with lower first stage F-statistics tend to have higher quality standard errors; less weight will be placed on these estimates in the empirical Bayes procedure.

quality indices, $\beta_j = \Phi^{-1}(q_j)$. Due to the HSA-stratified estimation procedure, the wide dispersion in these estimates reflects both causal (within-HSA) differences in potential survival outcomes for the same patient population and variation in average patient health across different HSAs, along with estimation error. I next outline an empirical Bayes procedure to account for these different variance components and produce more accurate posterior predictions of hospital quality.

3.3 Posteriors

Under assumptions 1-3 we obtain, for a subset of hospitals j with sufficient quasi-experimental data, minimum distance estimates $\hat{\beta}_j$ that are noisy but consistent measures of the true hospital quality indices β_j. At the same time, we observe a full set of observational RAM predictions $\hat{\alpha}_j$ from equation (20), which are likely positively, but not perfectly, correlated with quality due to the sorting bias detected in section 3.1. Following Morris (1983) and Raudenbush and Byrk (1986), I next estimate a hierarchical linear model (HLM) to link these two quality measures.[19] This is

$$\hat{\beta}_j = \kappa + \lambda\hat{\alpha}_j + \mu_{h(j)} + \upsilon_j + \iota_j, \tag{22}$$

where $\kappa + \lambda E[\hat{\alpha}_j] = E[\beta_j]$ is the average hospital quality index, $\mu_{h(j)}$ is a random effect for the HSA of hospital j, υ_j is the residual true quality index of hospital j, and ι_j is a mean-zero estimation error term. The HSA random effects, assumed to be identically normally-distributed with mean zero and variance σ^2, capture between-HSA variation in unmeasured quality, while within-HSA variation in residual quality indices $\upsilon_j \sim N(0, \phi^2)$ reflect causal differences not accounted for by observational RAMs. Subject to the usual first-order asymptotic approximation, the estimation error term ι_j can also be modeled as normally-distributed, with a known covariance structure. Consistent estimation of the HLM's hyperparameters κ, λ, σ, and ϕ comes from an ordinary least squares (OLS) regression of quality index estimates $\hat{\beta}_j$ on RAM predictions $\hat{\alpha}_j$, while efficient estimates leverage a feasible generalized least squares (FGLS) procedure that uses first-step estimates of σ and ϕ and the covariance of ι_j to iteratively solve for the hyperparameters by weighted least squares. The econometric appendix describes these procedures in more detail.

Table 3 reports OLS and FGLS hyperparameter estimates of equation (22), where for ease of interpretation the standard deviation of $\hat{\alpha}_j$ has been normalized to one. Column 1 shows that minimum distance quality estimates are indeed correlated with observational RAM predictions, though the OLS estimate of $\hat{\lambda} = 0.11$ is far from statistically significant due to the relative imprecision of the equal-weighted regression. Using the OLS residual variance estimates of $\hat{\sigma} = 0.88$ and $\hat{\phi} = 0.23$ to compute inverse-variance weighted FGLS estimates in column 3 dramatically increases precision: the standard error of $\hat{\lambda}$ falls from 0.16 to 0.04 without much change in the coefficient estimate. Iterating this procedure to convergence yields modest additional precision gains in column 4, and a Hausman (1978) test of the random-effects specification relative

[19] McClellan and Staiger (1999) also use a HLM to combine multiple hospital quality measures.

to a model with HSA fixed effects (reported in column 2) returns a p-value of 0.79. Overall, the HLM's decomposition suggests that around 90% of the national variation in quality indices β_j is found between HSAs, with only 20% of the remaining within-HSA variation explained by observational RAM predictions and 80% left unexplained.

I use these estimates to generate empirical Bayes posterior predictions of hospital quality that, as in Angrist et al. (2015) and Chetty and Hendren (2015), shrink asymptotically-unbiased but noisy quasi-experimental estimates of institutional quality towards precise, but likely biased, observational predictions. The random-effects structure of equation (22) further allows the vector of estimates for each HSA to be jointly shrunk towards a HSA-specific mean, thereby accounting for the high local correlation in hospital quality found in Table 3 by $\hat{\sigma} > 0$. In particular, the posterior mean and variance of a HSA's quality indices given vectors of its RAM predictions $\hat{\alpha}_h$ and minimum distance estimates $\hat{\beta}_h$ are

$$E[\beta_h|\hat{\alpha}_h, \hat{\beta}_h] = \Omega_h\hat{\beta}_h + (I_{J(h)} - \Omega_h)(\kappa + \lambda\hat{\alpha}_h) \tag{23}$$

$$Var(\beta_h|\hat{\alpha}_h, \hat{\beta}_h) = (I_{J(h)} - \Omega_h)(\phi^2 I_{J(h)} + \sigma^2), \tag{24}$$

where Ω_h is a weighting matrix given by the variance hyperparameters and Ξ_h, the variance-covariance matrix of estimation error:

$$\Omega_h = (\phi^2 I_{J(h)} + \sigma^2)(\phi^2 I_{J(h)} + \sigma^2 + \Xi_h)^{-1}. \tag{25}$$

Without HSA-level random effects ($\sigma = 0$) and correlated estimation error across hospitals serving the same HSA population (so that Ξ_h is diagonal), these formulas yield the usual empirical Bayes procedure seen in Morris (1983), applied hospital-by-hospital. When additionally $\lambda = 0$, so that observational RAM predictions do not reveal anything about true hospital quality, the minimum distance estimates are shrunk towards the grand mean κ in proportion to one-minus the quality signal-to-noise ratio, as with the simplest empirical Bayes procedures. Given the posterior mean and variance of hospital j's quality index β_j, posterior mean hospital quality is given by

$$\begin{aligned} E[q_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] &= E[\Phi(\beta_j)|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] \\ &= \Phi\left(\frac{E[\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}]}{\sqrt{1 + Var(\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)})}}\right) \end{aligned} \tag{26}$$

since β_j is normally-distributed conditional on $\hat{\alpha}_{h(j)}$ and $\hat{\beta}_{h(j)}$.[20]

I construct hospital quality posteriors using these formulas and the iterated FGLS estimates of the hyperparameters κ, λ, σ, and ϕ.[21] The dashed red line in Figure 4 shows the distribution of quality

[20] If $x \sim N(m, v)$, $E[\Phi(x)] = Pr(y - x < 0)$ for independent $y \sim N(0, 1)$. Thus $E[\Phi(x)] = \Phi(-E[y - x]/\sqrt{Var(y - x)}) = \Phi(m/\sqrt{1 + v})$.

[21] As usual with empirical Bayes procedures, I treat hyperparameter estimates as known when constructing posteriors. The high degree of precision in Table 3's iterated FGLS estimates justifies this simplification in my setting.

index posteriors for the 1,041 hospitals with first-step estimates (Appendix Figure A2 instead plots the full distribution of quality posteriors). As expected, the posterior mean distribution is tighter than the estimate distribution, reflecting empirical Bayes shrinkage and theoretically-improved mean squared prediction error. The posterior mean distribution is also more symmetric, as equation (23) downweights the heteroskedastic distribution of estimation error ι_j. The dotted green line in Figure 4 shows the distribution of posterior within-HSA quality indices $\kappa + \lambda\hat{\alpha}_j + \upsilon_j$, which is narrower still.

Importantly, equation (22) also produces posterior quality predictions for hospitals without a first-step quality estimate due to insufficient quasi-experimental data. In the 69 HSAs without any minimum distance estimates (mostly two hospital HSAs with fewer than 25 admissions), the posterior quality index is simply the HLM fitted values $\hat{\kappa} + \hat{\lambda}\hat{\alpha}_h$, which uses the population relationship between observational RAM and hospital quality to extrapolate to underidentified regions. In the other 626 HSAs these predictions are then shrunk toward the HSA-average quality estimate due to the HLM's random-effects structure. This extrapolation is valid when equation (22) describes the relationship between quality indices and observational RAM across all hospitals, whether or not they have enough quasi-experimental variation. Appendix Tables A1 and A4 show that the average characteristics of patients and hospitals across these two groups are quite similar, while Table A3 shows that all main results continue to hold or are strengthened when the HLM includes interactions with the HSA's hospital count, which is the main driver of minimum distance estimate availability and the only observable characteristic that meaningfully varies across the columns of Table A4. I next discuss these findings in detail.

4 Results

The hyperparameter estimates in Table 3 indicate significant within-HSA variation in true hospital quality that is positively, but only partially, correlated with observational RAM predictions. I next use the 2,357 empirical Bayes posterior mean predictions of hospital quality from 695 multi-hospital HSAs to characterize this variation as well as the non-random patient sorting that causes observational and quasi-experimental quality estimates to diverge. I then quantify the significance of this selection bias in two quality-based policies currently in place in U.S. healthcare markets.

4.1 Hospital Quality and Patient Sorting

Within-HSA comparisons of quality $E[Y_{ij}] = q_j$ reflect average causal effects of moving a representative patient across different local hospital types. I quantify these effects by regressing various hospital characteristics on a quality measure and HSA fixed effects in the set of multi-hospital HSAs. The characteristics include indicators for a hospital's ownership structure (either private non-profit, private for-profit, or government owned); an indicator for whether it is a teaching hospital; log average hospital spending on emergency Medicare patients; log emergency Medicare patient volume; and log bed capacity. Correlations with posterior

quality are reported in the first row of Table 4, while the second row regresses hospital characteristics on posterior quality indices $\beta_j = \Phi^{-1}(q_j)$. For comparison purposes, the last two rows report coefficients from regressions on two existing quality measures, conventional RAM predictions and observed hospital survival, and all regressors are normalized to standard deviation units.

The first two rows of Table 4 show that moving patients to providers with higher posterior quality and quality indices tends to place them in hospitals that spend more on emergency Medicare patients, have a larger HSA market share, and are less likely to be government-run. I do not find a statistically-significant difference in the probability of admission to for-profit vs. non-profit hospitals, nor any significant correlation with teaching status or bed capacity, though the associated standard errors are sometimes large. With a quality posterior standard deviation of around 12 percentage points, the estimates in the first row of Table 4 imply that moving a random patient to a hospital with a one percentage point higher potential 30-day survival rate reduces the chances of admission in a government-run provider by 0.7 percentage points and places the patient in a hospital with 1.9% higher emergency Medicare spending and 4.3% higher volume, on average. A supplementary results appendix section analyzes additional quality dimensions and finds significant within-hospital correlation in quality posteriors across time and by admitting conditions, positive correlations between quality posteriors and measurable inputs (in particular average staff salary), and increases in average quality following a hospital merger.

The findings in the first row of Table 4 are broadly consistent with previously documented correlates of observational quality measures, including in Sloan et al. (2001), Silber et al. (2010), Foster et al. (2013), Doyle et al. (2015), and Chandra et al. (2015).[22] Moreover, the third row of Table 4 shows similarly signed coefficients from each hospital characteristic regression on RAM predictions, though the strength of the relationship is attenuated with the more-biased quality measure. Hospitals with quality posteriors (RAM predictions) one standard deviation above the HSA mean are 8% (2%) less likely to be government owned, spend 23% (7%) more per Medicare patient, and have a 50% (16%) larger Medicare market share, on average.

This attenuation suggests a *negative* correlation between true hospital quality and the residual selection bias of observational RAMs: better hospitals appear to attract relatively sicker patients, thereby reducing the observed relationship between, say, average spending and mortality. Indeed, the fourth row of Table 4 shows no statistically-significant correlation between the most biased quality proxy, observed survival $E[Y_{ij}|D_{ij} = 1]$, and any of the spending, volume, or ownership structure measures found to correlate with the quality posteriors.[23] The negative quality-bias correlation is more broadly illustrated in Figure 5, which plots observed survival against quality posteriors net of their HSA means. Points above the dashed 45 degree line represent hospitals with relatively higher selection bias, $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$, while points below are

[22]The instrumented quality measures used by McClellan and Staiger (2000) and Geweke et al. (2003) also show small and rarely significant differences between for-profit and non-profit hospitals.

[23]For consistency I also shrink observed survival rates towards their grand mean in proportion to one minus the signal-to-noise ratio, though all results are virtually unchanged by this empirical Bayes procedure.

less positively selected than average. The figure shows that hospitals with relatively higher quality posteriors – those to the right of the origin – tend to fall below the 45 degree line and thus be less positively selected. Overall, I find a within-HSA correlation of quality and bias posteriors of -0.83.

The generalized Roy (1951) framework underlying these estimates provides another way to characterize selection: the extent to which patient sorting exploits comparative advantage by admitting at more appropriate hospitals (i.e., selection-on-gains). To explore this, Figure 6 plots the distribution of volume-weighted average selection bias posteriors for all multi-hospital HSAs. In a constant effects framework, HSA-average bias equals zero by construction; in contrast, the wide distribution in Figure 6 suggests a large degree of comparative advantage across emergency healthcare providers. Moreover, most HSAs (86%) appear to have positive average selection bias. In these markets, a typical patient is more likely to survive at the selected hospital than at a hospital picked at random from the market, thus implying that patients benefit from positive Roy selection. Only 15 HSAs (2%) have an average bias posterior of less than -10 percentage points, while the average bias posterior in 440 HSAs (63%) exceeds 10 percentage points.

This finding does not appear to be driven by hospitals specializing in treating different emergency conditions: the shares of positively-selected HSAs in models, described in the supplementary appendix, that estimate quality separately by diagnostic category all exceed 85%. Nor does the result appear driven by the normality assumption, as Table A3 shows a similar 80% of HSAs have positive average selection bias when a Student's $t(2)$ distribution is used. Recall that the non-parametric estimates plotted in Figure 3 also suggest pervasive positive selection bias.

A more plausible driver of match-specific quality is differential hospital distance, since individuals suffering from an acute emergency may only survive if brought to the closest available emergency room. Table 5 examines the extent to which distance explains selection-on-gains by estimating the average selection bias that would be found if patients were not more likely to attend hospitals close to them. Virtually all HSAs have a negative volume-weighted "distance bias," $E[d_{ij}|D_{ij} = 1] - E[d_{ij}]$, where d_{ij} denotes the ZIP code distance between patient i and hospital j. The mean of this measure across the 695 multi-hospital HSAs is -0.91 miles. However, there is also considerable variation, with patients in some regions sorting to hospitals no more than 0.1 miles closer to them than a provider picked at random from the HSA.

Panel A of Table 5 regresses HSA-level survival bias on flexible polynomials in HSA-level distance bias and indeed finds a strong correlation. Nevertheless, the constant in even the most flexible cubic regression in column 3, representing average outcome selection bias in a HSA given zero selection-on-distance, remains significantly positive at 15 percentage points. Panel B reports non-parametric estimates of this quantity by directly computing mean selection bias in HSAs with relatively little distance bias. Even in the 39 regions where average distance bias is above -0.01 miles, patients are still around 9 percentage points more likely to survive at their chosen hospitals then via random admission (74% of these HSAs have positive average bias posteriors). Thus differential hospital distance appears to explain some, but not all, of the Roy selection

shown in Figure 6.[24] Accommodating unobservable hospital comparative advantage and selection-on-gains with the heterogenous-effects multivariate probit specification – features ruled out by other models such as the linear IV specification of Angrist et al. (2015) or the fixed-coefficient probits of conventional RAMs and Geweke et al. (2003) – is therefore empirically important in this setting.[25]

4.2 Policy Consequences of RAM Bias

Non-random patient sorting generates a sizable distribution of posterior selection bias, with a within-HSA standard deviation of 2.8 percentage points. Although conventional risk-adjustment appears to offset some of this bias, quality posteriors and RAM predictions often disagree, with a within-HSA correlation of 0.68.[26] Around 19% of hospitals (131) with the best quality posteriors in each multi-hospital HSA are ranked differently by RAM, while a similar 20% of HSAs (138) see disagreements on the worst local hospital. Nevertheless, it is difficult to gauge the economic importance of RAM bias from these statistics alone – as shown in other settings, policy decisions based on biased quality rankings may still generate large social gains (Angrist et al., 2015). Furthermore, the negative correlation found in Figure 5 means that policies that reward or punish hospitals according to observational RAM rankings are most likely to understate true quality differentials, as in Table 4. To better assess the economic implications of RAM bias, I next simulate these policies directly.

Medicare Reimbursement

I first consider how payments from Medicare's Value-Based Purchasing (VBP) program would differ if hospital ranks were based on quality posteriors instead of RAMs. VBP was launched in 2013 with the goal of incentivizing hospitals with quality-linked Medicare reimbursement adjustments in a budget-neutral way (DHHS/CMS, 2015). Along with clinical process-of-care measures and patient surveys, risk-adjusted mortality became a part of a "total performance score" (TPS) assigned to each hospital receiving Medicare reimbursement payments in fiscal year 2014. CMS withheld 1.25% of each participating hospital's FY2014 diagnosis-related group (DRG) payment, redistributing around $1.1 billion of total withholdings by a linear TPS schedule. Currently, VBP affects only a small share of a hospital's reimbursements; in FY2014, the average VBP penalty was a 0.26 percentage points and the average bonus was a 0.24 percentage points (Conway, 2013). Nevertheless, the program has proved quite controversial as the withholding rate has steadily increased, reaching to 2% in 2016 (Pear, 2014), and as CMS recently announced new plans to tie 90% of

[24] I find similarly reduced average selection bias within diagnosis categories, with the largest for circulatory and injury conditions.

[25] The EMS staff I interviewed were very receptive to the possibility of comparative advantage and selection on salient unobserved local factors: many hospitals have specialized services such as trauma centers or advanced CT scanners, for example, that are essential for some but not all patients. Ambulance company EMTs and paramedics seem well-poised to exploit these gains; in some states like Massachusetts there are explicit "Point of Entry" guidelines formalizing this institutional knowledge.

[26] For comparison, Angrist et al. (2015) find a correlation between conventional middle school value-added predictions and quasi-experimental quality posteriors of 0.85-0.93 in Boston.

all traditional Medicare payments to quality programs like VBP by 2018 (DHHS, 2015). In recent work Norton et al. (2016) show that hospitals indeed respond to the program's seemingly modest incentives, with providers facing higher marginal VBP returns improving their TPS components in subsequent years, while Gupta (2016) finds large incentive effects from the hospital readmissions reduction program, another recently-introduced quality-based reimbursement policy.

I replicate the FY2014 VBP payment schedule to simulate payment adjustments under alternative hospital rankings. Total performance scores combine "achievement points," which are based on hospital quality estimates in the most recent period, and "improvement points," which are based on a hospital's gain relative to a previous period. In FY2014, CMS computed points from hospital risk-standardized mortality rates, defined with the notation of equation (20) as

$$RSMR_j = \frac{1 - \sum_{i:D_{ij}=1} F_\nu(\hat{\alpha}_j + \hat{\gamma}' W_i)}{1 - \sum_{i:D_{ij}=1} F_\nu(\bar{\alpha} + \hat{\gamma}' W_i)} (1 - \bar{Y}), \tag{27}$$

where F_ν is the distribution of the observational RAM error term ν_i, $\hat{\gamma}$ is an estimate of the RAM parameter γ, $\bar{\alpha}$ is the mean RAM prediction $\hat{\alpha}_j$, and $1-\bar{Y}$ is the average mortality rate in the sample. In practice, risk-standardized survival rates, $1 - RSMR_j$, correlate strongly with observational RAM predictions ($\rho = 0.98$).

These rates are converted to points by a coarse schedule, with the greater of achievement and improvement points constituting a hospital's outcome domain score. In FY2014 outcome scores made up 25% of a hospital's TPS. Hospitals were refunded none of their DRG withholdings if they scored the minimum level across all three quality domains and linearly accrued payments with higher TPSs. In simulating the distribution of FY2014 payments I hold the non-outcome domains and FY2014 DRG totals fixed, generating benchmark outcome achievement points from the estimated 2010-2012 RAM and computing improvement points from the gain in a hospital's risk-standardized mortality rate between 2007-2009 and 2010-2012. I then compare simulated VBP reimbursement adjustment rates with those that would be produced with posteriors of the within-HSA component of hospital quality, $\kappa + \lambda\hat{\alpha}_j + \nu_j$, rather than $1 - RSMR_j$. The data appendix describes the construction of simulated payments in more detail.

The results of this simulation are summarized in Table 6. Column 1 reports, for different hospital types, the percentage point change in the relative value-based purchasing adjustment from incorporating quasi-experimental data, compared with the prevailing RAM-based adjustment. Column 2 contains this benchmark adjustment, while column 3 reports the implied percentage change in relative VBP adjustments. The results indicate that when using quality posteriors, non-profit and teaching hospitals would see an average of 8.7% and 14.9% higher VBP adjustments, respectively, while government-run hospitals would have their relative VBP adjustment lowered by 8.5%. Table 5 also suggests that higher-volume and higher-capacity hospitals would see their VBP payments raised, though the coefficient on log average spending is not statistically significant. As in Table 4 and Figure 5, the estimates in Table 5 show the residual bias in conventional RAM rankings tends to attenuate quality-based VBP differentials rather than changing the

types of hospitals that are generally rewarded by performance-linked subsidies.

The magnitudes of changes in column 3 of Table 6 are modest, reflecting both the low weight of the outcome domain (25%) and the coarseness of achievement and improvement point schedules. As columns 4-6 show, eliminating the contributions of process-of-care measures and patient surveys magnifies the average change in relative adjustment rates for non-profit, government-run, and teaching hospitals to 44.3%, -45.6%, and 70.2%, respectively. VBP adjustments for relatively higher-volume and higher-capacity hospitals similarly increase, and higher spending hospitals begin to see both higher benchmark reimbursement adjustments and increased payments for quality. Although the policies represented by these columns are far from current VBP practice, together the simulation results suggest bias in observational RAMs has significant capacity to affect performance-based hospital incentive schemes, especially as outcome-based measures become more important. Nevertheless, reducing bias in performance rankings primarily rewards benchmark-subsidized hospitals further and intensifies existing incentive margins, at least along observable dimensions.

Patient Guidance

Along with hospital incentives, supervisory quality rankings have begun to shape patient admission decisions. The federal Hospital Compare website, launched in 2005 to help consumers make informed decisions about their inpatient options, reports multiple hospital performance measures, including observational RAM predictions starting in 2008. At the same time a growing number of private organizations, including the U.S. News and World Report, Consumer Reports, and the Joint Commission, have developed competing hospital "report cards" with alternative risk-adjustment measures. Although patients increasingly consult such rankings (Rice, 2014), and research shows that higher-ranked hospitals tend to see increased future emergency patient market shares (Chandra et al., 2015), there is little evidence on how quality-based admissions may affect patient survival.

The hyperparameter estimates in Table 3 suggest that redirecting a typical patient from a random hospital to the provider with the highest RAM ranking likely increases her expected 30-day survival, and that decisions based on less-biased quality posteriors should generate even better average health outcomes. At the same time, the significant degree of positive selection bias shown in Figure 6 suggests these gains may be offset by the fact that a typical patient's admissions is better than random: on average, patients already see large survival gains from selecting more appropriate hospitals.

I quantify these effects by simulating 250 realizations of quality indices β_j from the iterated FGLS estimates of κ, λ, σ, and ϕ, holding the distribution of observational RAM predictions fixed. I then draw estimation error components ι_j and construct simulated quality estimates and posteriors. From these data, I compute the average 30-day survival rates for a typical patient admitted to a random hospital within her HSA, the local hospital with the highest survival rate, or the local hospital ranked best by either RAM predictions or quality posteriors. While abstracting away from various general equilibrium effects or capacity constraints, these estimates give a rough sense of the relative public health value of guiding patient admissions

by various supervisory quality rankings.

Results of this exercise are plotted in Figure 7. Selection bias notwithstanding, an emergency patient sent to the lowest-mortality local hospital is on average 0.9 percentage points more likely to survive their first 30 days after admission, relative to the random admissions benchmark. Using a conventional RAM for admissions further increases the policy's health effect, to 2.8 percentage points. This reduction in 30-day emergency condition mortality is quite large in the historical context: among Medicare patients admitted for pneumonia, for example, Ruhnke et al. (2011) estimate an average mortality decline due to technological advances of around 3.4 percentage points between 1987 and 2005.

Incorporating quasi-experimental data leads to larger survival gains from report card admission policies, though this improvement is limited by imprecision in minimum distance quality estimates. The last two bars in Figure 7 depict the range of possible improvements, from a feasible admission policy with the actual estimation error level found in my sample to an infeasible regime in which all choice probabilities and mean selected outcomes used to construct minimum distance quality estimates are assumed to be known without error. Sending patients to hospitals with the highest quality posteriors leads to incremental 30-day survival rate gains of between 0.5 and 1.7 percentage points, or 18-60% of the 2.8 percentage point gain from RAM-based admission policies. This suggests using less-biased hospital rankings to guide admissions would deliver meaningful partial-equilibrium health returns, particularly when rankings are estimated on larger administrative datasets or by more efficient semi-parametric methods.

At the same time, the simulation results in Figure 7 highlight the inherent limitation of supervisory quality-based admission policies applied to settings with significant institutional comparative advantage and positive Roy selection. Moving a patient from the selected (rather than a randomly-chosen) hospital to the local hospital with the highest average quality actually *decreases* expected survival by 11 percentage points. Consumer guidance policies that make average emergency care patients more likely to select high-ranked hospitals (in circumstances where their ambulance operator gives them the choice), as well as policies that close or limit the growth of low-ranked providers, may therefore undermine the prevailing health benefits of hospital selection-on-gains and have unintended negative consequences for average patient health.

5 Conclusions

Policymakers in many settings now rely on outcome-based quality measures to incentivize institutions and inform consumers, despite concerns that existing observational methods only partially offset bias from non-random institutional choice. This paper develops a flexible framework for quantifying institutional performance and selection bias with quasi-experimental data. Quality in these models can be non-parametrically estimated from rich instrument variation, while distributional restrictions may substitute for constant effects to extrapolate from narrower quasi-experimental designs. Unlike previous likelihood-based estimation methods, a tractable minimum distance procedure implements this semi-parametric approach. Moreover, the

models estimated here allow for both institutional comparative advantage and Roy-style selection-on-gains, two important features previously lacking in both linear and nonlinear IV frameworks.

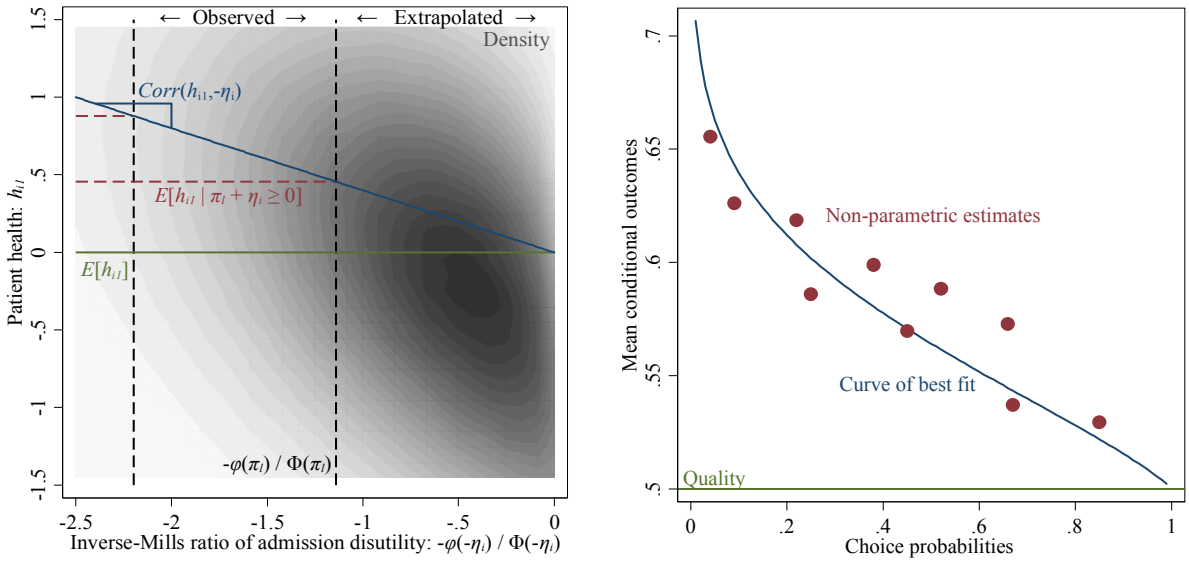
These features are highly relevant in emergency healthcare. I both find a large degree of match-specific hospital quality and that most markets exhibit positive Roy selection, with patients admitting to more appropriate hospitals on average. This non-random sorting generates pervasive selection bias, with a negative quality-bias correlation obscuring important relationships with hospital ownership structure, patient volume, and average spending. Observational risk-adjustment methods remove some of this bias, generating survival gains in simulations of ranking-based guidance policies, while quasi-experimental quality posteriors can further improve the targeting of both Medicare reimbursement and patient guidance programs.

Ultimately, more work is needed to characterize the ways in which these policies may shape long-run hospital quality supply and demand. As long as biased quality measures are used to structure the Value-Based Purchasing program, providers may find ways to "game the system," boosting their payments without improving actual performance. While the simulations in section 4 show that most observable hospital characteristics currently rewarded by VBP are only further subsidized by policies based on less-biased quality posteriors, there may remain various hospital-controlled unobservables that correlate with RAM rankings but not true quality. Detecting VBP "gaming" may become easier as the scope of performance-linked healthcare reimbursement and the strength of incentives grow.

The simulations also raise new questions about the efficacy of demand-side interventions, including the large and growing set of hospital report cards currently consulted by patients. With constant causal effects, the finding that higher-ranked hospitals tend to attract more emergency patients in the future, as in Chandra et al. (2015), has unambiguously positive implications for public health. Accounting for the significant extent of selection on match-specific quality, however, requires a more nuanced analysis. On one hand, report cards may cause patients to update weak or incorrect priors on their most appropriate hospital and induce the selection of providers with high average quality, thus increasing patients' chances of survival. However, widely-known rankings may also disrupt prevailing beneficial selection patterns, to the extent they also influence patients with better private information. Understanding the ways in which hospital performance measures actually affect admission decisions and characterizing the optimal design of public quality signals in settings with Roy selection are two important goals raised by the heterogeneous-effects framework.

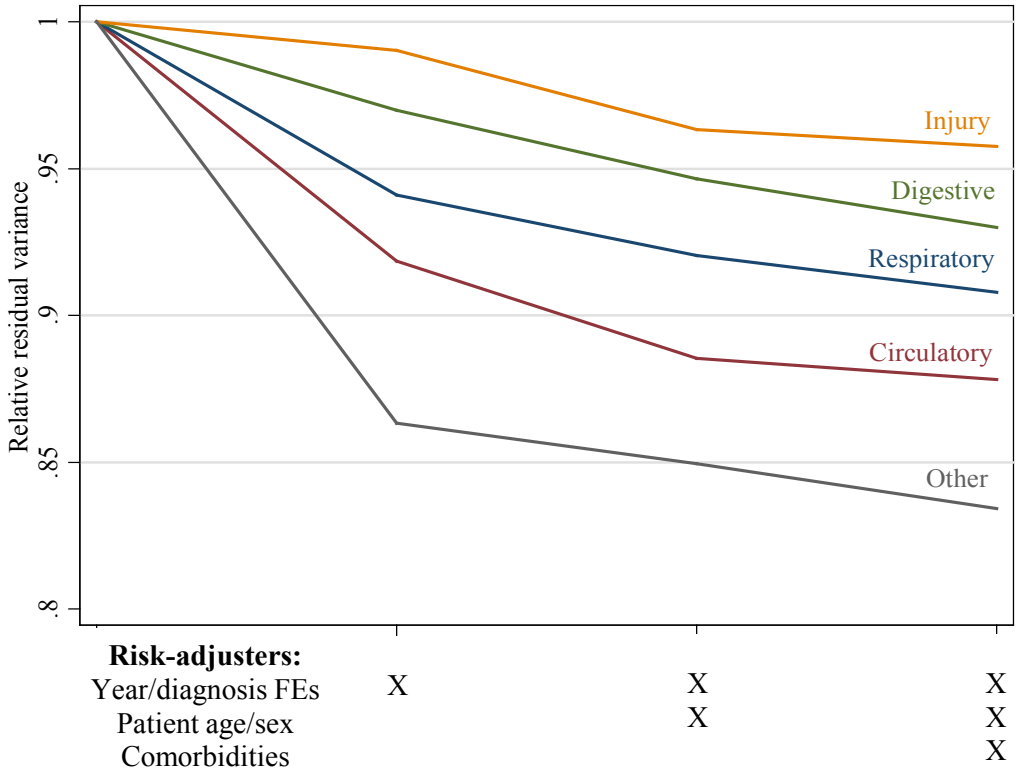
Figure 1: Quality identification and estimation in a bivariate probit model

A. Identification ($L = 2$) B. Estimation ($L > 2$)



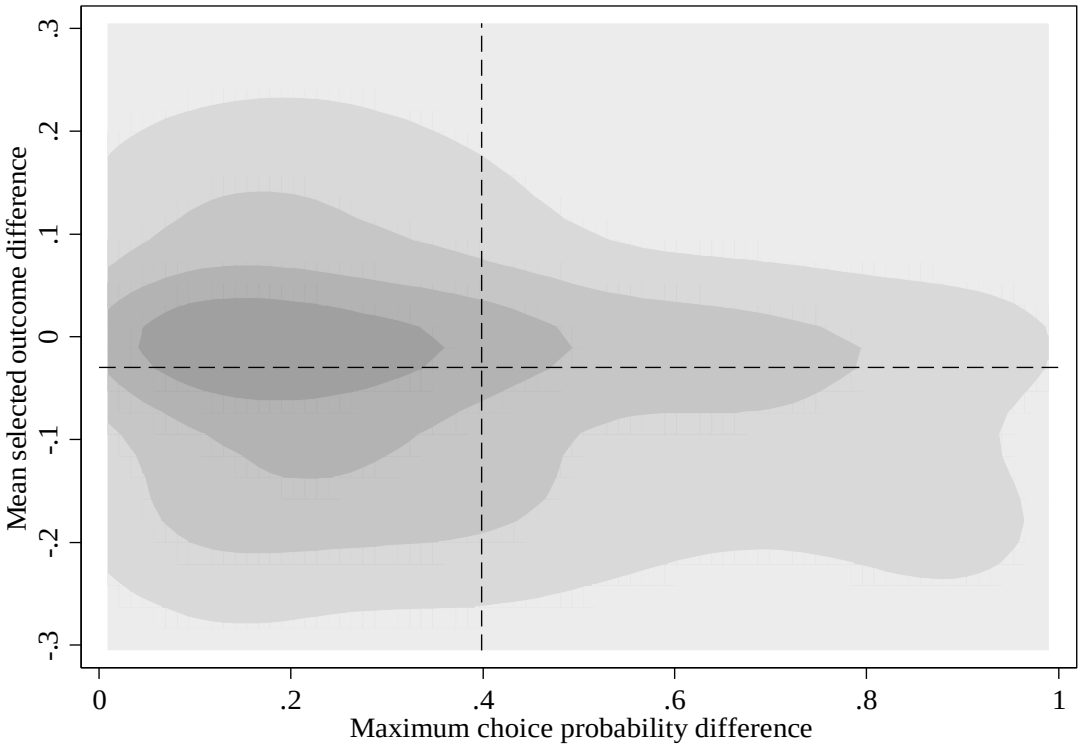
Notes: Panel A shows the probability density function of potential patient health and the inverse Mills ratio of latent admission disutility for a hospital with positive selection bias and joint-normal health and utility. The vertical dashed lines indicate inverse Mills transformations of the first-stage preference parameters for two different ambulance companies, while the horizontal dashed lines indicate the average health of patients that would be admitted to the hospital by each company. The downward-tilted line meets zero on the x-axis at the hospital's population average health ($\beta_1 = 0$) on the y-axis, and its slope ($-\rho_1 = -0.4$) represents the population correlation of health and disutility. Panel B shows estimated mean conditional outcomes (survival probabilities) for patients admitted by a set of ambulance companies against the associated choice probability from the same model. The curve of best fit equals the population survival probability ($\Phi(\beta_1) = 0.5$), that is, the hospital's quality, when the choice probability equals one.

Figure 2: Residual survival variance in observational RAMs



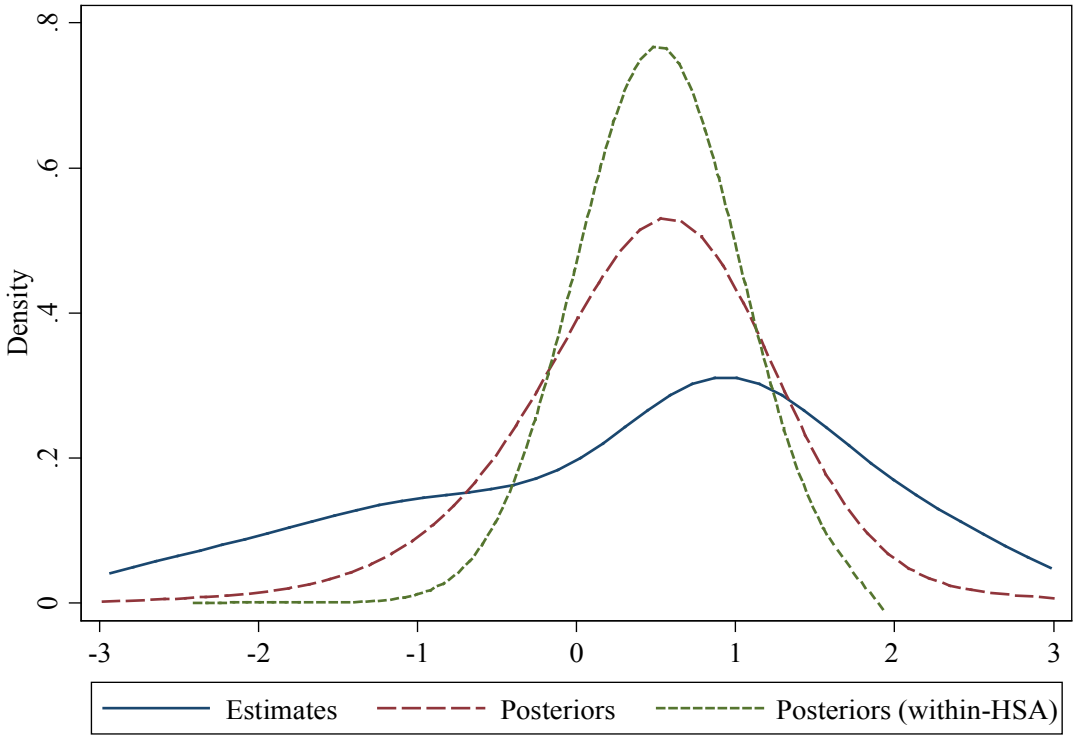
Notes: This figure plots the variance of risk-adjusted 30-day survival relative to the unadjusted survival variance for three risk-adjustment models, estimated separately by diagnosis category. See Table 1 for a description of each diagnosis category, Table 2 for a list of included comorbidities, and the data appendix for a description of the RAM estimation procedure.

Figure 3: The joint distribution of ambulance effects on hospital choice and patient survival



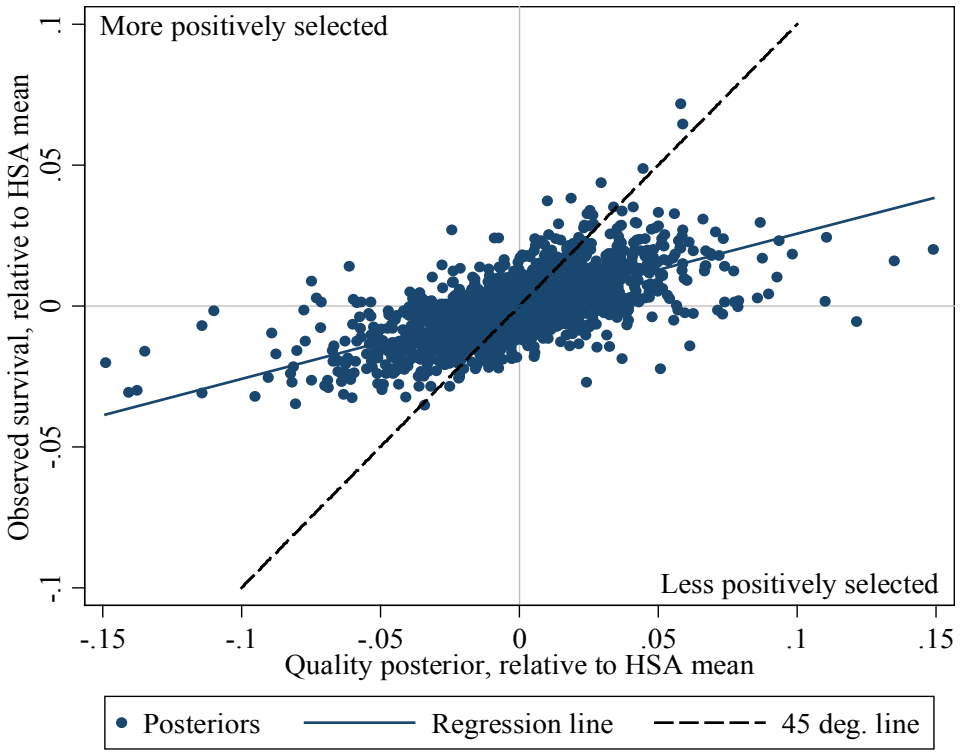
Notes: This figure plots a Gaussian kernel density estimate of the joint distribution of estimated mean selected outcome differences and estimated choice probability differences for 1,041 hospitals with minimum distance quality estimates. Differences are taken across the two ambulance companies with the maximal estimated choice probability difference for each hospital and estimate causal effects of differential ambulance company assignment on hospital choice and 30-day survival for admitted patients. The vertical and horizontal bandwidths used to estimate this distribution are 0.05 and 0.1. Dashed lines indicate sample means.

Figure 4: The distribution of hospital quality index estimates and posteriors



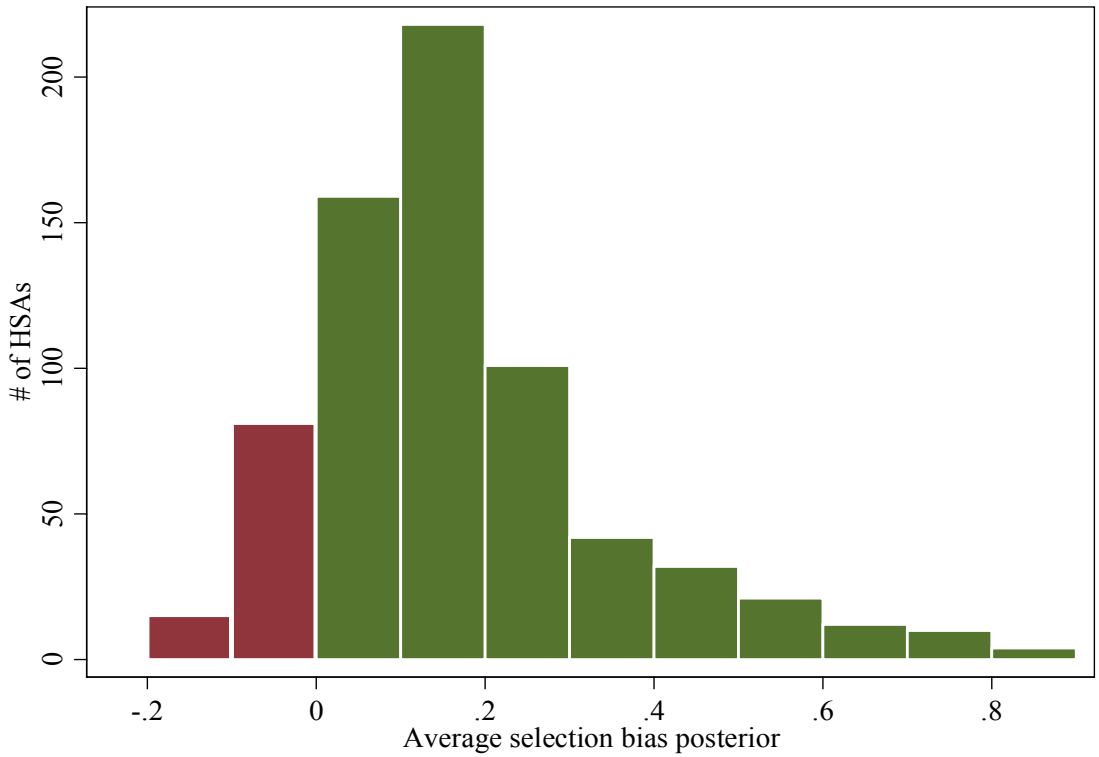
Notes: This figure plots Gaussian kernel density estimates of the distribution of minimum distance hospital quality index estimates and empirical Bayes posteriors of both the overall and within-HSA quality indices. The sample includes 1,041 hospitals operating in 626 multi-hospital HSAs with a first-step quality estimate. The bandwidth used to estimate each distribution is 0.5.

Figure 5: Within-HSA variation in hospital quality and selection bias



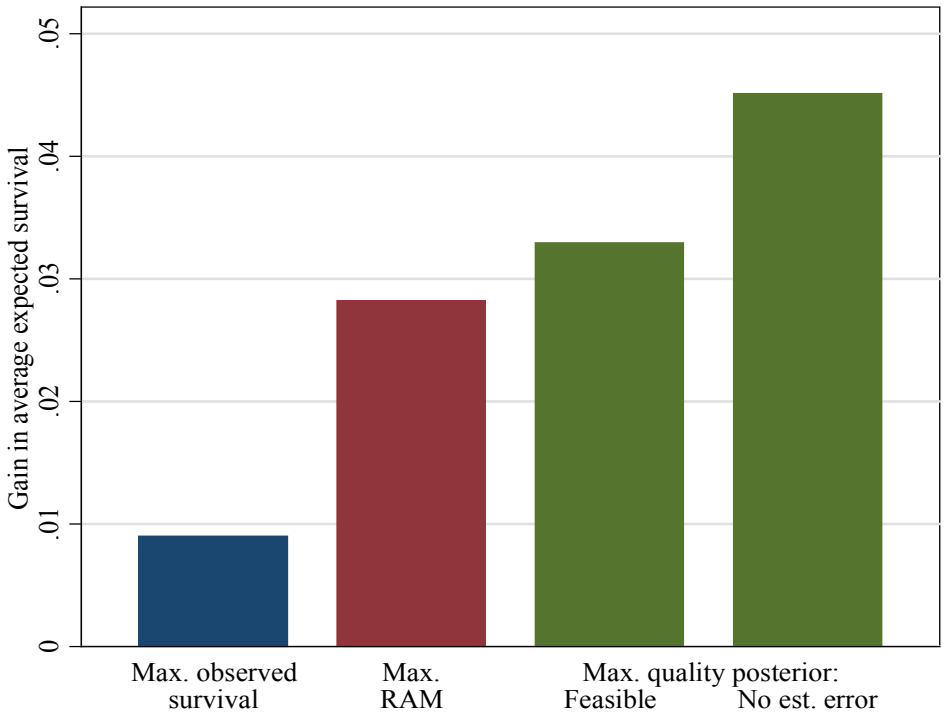
Notes: This figure plots posterior hospital survival rates against posterior quality, both net of their HSA means. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs. Points above the dashed 45-degree line represent hospitals that are relatively more positively selected within their HSA, while hospitals below the 45-degree line are relatively less positively selected.

Figure 6: The distribution of HSA-average selection bias



Notes: This figure plots the distribution of volume-weighted average posterior selection bias across 695 multi-hospital HSAs. HSAs with negative selection bias would see higher average 30-day survival if patients were randomly allocated to hospitals, while a positively-selected HSA would have a lower survival rate under random admissions.

Figure 7: Survival gains from selecting a top-ranked hospital, relative to random admissions



Notes: This figure plots simulated gains in average expected survival for a random patient sent to the highest-ranked hospital in her HSA, relative to a random admission, according to the hospital's 30-day survival rate, observational RAM prediction, or quality posterior (with and without estimation error). The sample consists of 2,357 hospitals operating in 695 multi-hospital HSAs. Estimates are from 250 draws of the hierarchical model described in the text.

Table 1: The analysis sample

	Diagnoses (1)	Patients (2)	Ambulances (3)	Hospitals (4)	HSAs (5)	30-day survival (6)
Full sample	29	405,173	9,590	4,821	3,159	0.833
			A. By diagnosis category			
Circulatory	5	89,077	7,578	3,879	2,777	0.807
Respiratory	4	81,021	7,432	4,224	2,980	0.781
Digestive	6	26,359	5,244	3,323	2,354	0.902
Injury	8	71,616	7,396	3,634	2,561	0.931
All other	6	137,100	8,064	4,441	2,997	0.815
			B. By HSA hospital count			
One	29	151,072	6,756	2,464	2,464	0.831
Two	29	84,634	3,578	800	400	0.837
Three	29	44,399	2,302	396	132	0.835
Four	29	24,398	1,227	212	53	0.829
Five or more	29	100,670	3,775	949	110	0.832

Notes: This table summarizes the distribution of diagnoses, ambulances, hospitals, and 30-day survival in the sample of Medicare FFS patients admitted for one of 29 nondeferrable diagnoses in 2010-2012. Circulatory diagnoses include acute myocardial infarction, intracerebral hemorrhage, occlusion and stenosis of the precerebral artery, occlusion of cerebral arteries, and transient cerebral ischemia. Respiratory diagnoses include pneumonia due to solids and liquids, pneumonia (organism unspecified), other bacterial pneumonia, and other diseases of the lung. Digestive diagnoses include diseases of the esophagus, gastric ulcer, duodenal ulcers, vascular insufficiency of the intestine, intestinal obstruction without mention of hernia, and other/unspecified noninfectious gastroenteritis and colitis. Injury diagnoses include fracture of the ribs, sternum, larynx, and trachea; fracture of the pelvis; fracture of the neck or femur; fracture of the tibia and fibula; fracture of the ankle; poisoning by angesicis; antipyretics, and antirheumatics; poisoning by psychotropic agents; and other/unspecified injury. All other diagnoses include septicemia; malignant neoplasm of the trachea, bronchus, and lung; secondary malignant neoplasm of respiratory and digestive systems; other disorders of the urethra and urinary tract; disorders of muscle, ligament, and fascia; and general symptoms.

Table 2: Ambulance company assignment balance

	Assigned ambulance company's closest hospital		Equality p-value	Regressions on RAM of the ambulance's closest hospital
	Low RAM	High RAM		
	(1)	(2)	(3)	(4)
RAM prediction	-0.058	0.015	<0.001	0.101 (0.010)
	A. Demographics			
Age	81.56	81.62	0.790	0.148 (0.214)
Male	0.380	0.384	0.784	-0.002 (0.013)
White	0.859	0.849	0.259	-0.004 (0.009)
Black	0.092	0.099	0.395	0.004 (0.007)
Referred from home	0.639	0.607	0.008	-0.043 (0.013)
Referred from accident	0.130	0.125	0.546	-0.009 (0.009)
Circulatory diagnosis	0.236	0.229	0.531	-0.014 (0.011)
Respiratory diagnosis	0.187	0.185	0.852	0.009 (0.010)
Digestive diagnosis	0.065	0.067	0.800	0.003 (0.006)
Injury diagnosis	0.174	0.184	0.294	0.002 (0.010)
	B. Comorbidities			
Hypertension	0.262	0.271	0.419	0.009 (0.012)
Stroke	0.011	0.012	0.708	0.002 (0.003)
Cerebrovascular disease	0.032	0.034	0.614	0.003 (0.005)
Renal failure	0.117	0.121	0.643	0.009 (0.009)
Dialysis	0.012	0.012	0.943	0.001 (0.003)
Chronic obstructive pulmonary disease	0.107	0.107	0.992	0.005 (0.008)
Pneumonia	0.052	0.055	0.682	0.003 (0.006)
Diabetes	0.120	0.133	0.126	0.012 (0.009)
Protein-calorie malnutrition	0.035	0.037	0.680	0.003 (0.005)
Dementia	0.082	0.093	0.132	0.009 (0.007)
Paralysis	0.032	0.037	0.213	0.006 (0.005)
Peripheral vascular disease	0.073	0.077	0.498	0.001 (0.007)
Metastic cancer	0.020	0.020	0.896	0.001 (0.004)
Trauma	0.057	0.058	0.820	0.003 (0.006)
Substance abuse	0.039	0.037	0.727	0.001 (0.006)
Major psychological disorder	0.029	0.031	0.672	0.001 (0.005)
Chronic liver disease	0.007	0.007	0.796	0.002 (0.002)
	C. Ambulance services			
Excess miles transported	-0.044	0.048	0.985	0.072 (0.069)
Emergency transport	0.956	0.961	0.331	0.013 (0.005)
Advanced life support	0.727	0.728	0.994	0.001 (0.012)
Intravenous fluids administered	0.009	0.008	0.869	-0.004 (0.002)
Intubation performed	<0.001	<0.001	0.228	-0.001 (0.001)
Demographics, comorbidities, and ambulance services joint p-value: 0.887				

Notes: This table compares the characteristics of patients referred by ambulance companies located close to hospitals with high and low RAM predictions, within patient ZIP codes. The sample includes 254,101 patients admitted to 2,357 hospitals in 695 multi-hospital HSAs. Columns 1 and 2 report average characteristics of patients assigned to ambulances that are closest (in terms of ZIP code centroid distance) to hospitals in the first and fourth quartiles of RAM predictions in their HSA, controlling for ZIP code fixed effects. Column 3 reports robust p-values for tests of equality across the two groups. Column 4 reports coefficients and robust standard errors from regressions on the assigned ambulance's closest hospital's RAM, controlling for ZIP code fixed effects. Excess miles transported is computed as a patient's transported miles minus the ZIP code centroid distance to a patient's hospital.

Table 3: Hierarchical linear model estimates

	OLS	HSA fixed effects	Random effects (FGLS)	
			Two-step	Iterated
	(1)	(2)	(3)	(4)
RAM coefficient (λ)	0.106	0.029	0.111	0.111
	(0.155)	(0.311)	(0.038)	(0.038)
Constant (κ)	0.222		0.502	0.502
	(0.155)		(0.043)	(0.042)
Variance components:				
Within-HSA (φ)	0.232	0.150	0.189	0.248
	(0.039)	(0.053)	(0.045)	(0.037)
Between-HSA (σ)	0.878		0.802	0.811
	(0.034)		(0.037)	(0.037)
Hausman test statistic (1 d.f.)				0.07
				[0.790]

Notes: This table reports estimated parameters of the hierarchical linear model outlined in the text. The sample consists of 1,041 minimum distance quality index estimates and RAM predictions from 626 multi-hospital HSAs. Column 1 reports OLS coefficients and variance estimates from a regression of quality index estimates on RAM predictions and a constant. Column 2 reports estimates an OLS regression with HSA fixed-effects, while columns 3 and 4 report two-step and iterated FGLS random-effect estimates. See the econometric appendix for details on this estimation procedure. The Hausman statistic tests equality of the estimates in columns 2 and 4. Standard errors, clustered by HSA, are reported in parentheses; the test's p-value is reported in brackets.

Table 4: Quality measure correlates of hospital characteristics

Regressor:	Non-profit hospital (1)	For-profit hospital (2)	Government hospital (3)	Teaching hospital (4)	Log (avg. spending) (5)	Log (volume) (6)	Log (# of beds) (7)
Quality posterior	0.055	0.024	-0.079	-0.018	0.230	0.517	-0.057
	(0.040)	(0.033)	(0.029)	(0.041)	(0.111)	(0.223)	(0.084)
Quality index posterior	0.044	0.025	-0.069	-0.018	0.161	0.388	-0.015
	(0.034)	(0.029)	(0.022)	(0.033)	(0.106)	(0.219)	(0.088)
RAM prediction	0.022	-0.002	-0.020	-0.016	0.065	0.159	-0.013
	(0.012)	(0.010)	(0.009)	(0.015)	(0.036)	(0.068)	(0.031)
Observed survival	0.003	0.009	-0.012	-0.050	0.009	-0.089	-0.137
	(0.012)	(0.011)	(0.009)	(0.013)	(0.032)	(0.058)	(0.027)

Notes: This table reports coefficients from regressions of the hospital characteristic in each column on the quality measure in each row, controlling for HSA fixed effects. All regressors are normalized to standard deviation units. Observed survival posteriors shrink observed rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The sample is 2,357 hospitals operating in 695 multi-hospital HSAs. Standard errors, clustered by HSA, are reported in parentheses.

Table 5: Estimates of average selection bias net of hospital distance bias

Regressors:	(1)	(2)	(3)
	A. Parametric		
Constant	0.165	0.159	0.154
	(0.008)	(0.008)	(0.008)
Avg. distance bias (marginal effect)	-0.009	-0.018	-0.031
	(0.003)	(0.005)	(0.007)
Polynomial:	Linear	Quadratic	Cubic
HSAs:		695	
	B. Non-parametric		
Constant	0.160	0.117	0.089
	(0.015)	(0.019)	(0.022)
Bandwidth:	1 mile	0.1 miles	0.01 miles
HSAs:	142	66	39

Notes: This table summarizes regressions of average selection bias posteriors for 695 multi-hospital HSAs. Panel A regresses bias posteriors on polynomials in HSA-average distance bias. A hospital's distance bias is the difference between its average ZIP code centroid distance to its admitted patients and its average distance to all potential patients in the HSA. Panel B reports average selection bias posteriors for HSAs with an average distance bias that falls within narrow bandwidths of zero. The constant from both sets of regressions estimates average selection bias that is not explained by the relative distance between admitted patients and their hospitals. Robust standard errors are reported in parentheses.

Table 6: Correlates of changes in value-based purchasing adjustments

	25% outcome domain weight			100% outcome domain weight		
	Percentage point change	Benchmark adjustment	Change as % of | benchmark |	Percentage point change	Benchmark adjustment	Change as % of | benchmark |
Regressor:	(1)	(2)	(3)	(4)	(5)	(6)
Non-profit hospital	0.014	0.160	8.70	0.074	0.168	44.28
	(0.005)	(0.029)		(0.026)	(0.042)	
For-profit hospital	-0.003	-0.032	-10.04	-0.022	-0.055	-39.57
	(0.007)	(0.036)		(0.032)	(0.054)	
Government hospital	-0.018	-0.217	-8.50	-0.094	-0.206	-45.57
	(0.006)	(0.035)		(0.030)	(0.049)	
Teaching hospital	0.013	0.089	14.92	0.072	0.103	70.21
	(0.006)	(0.033)		(0.029)	(0.047)	
Log (avg. spending)	-0.001	-0.137	-0.99	0.100	0.348	28.62
	(0.008)	(0.062)		(0.036)	(0.069)	
Log (volume)	0.008	0.196	4.12	0.053	0.246	21.61
	(0.002)	(0.010)		(0.011)	(0.017)	
Log (# of beds)	0.008	0.132	6.09	0.059	0.208	28.18
	(0.003)	(0.018)		(0.015)	(0.024)	

Notes: Columns 1 and 4 report coefficients from regressions of changes in simulated value-based purchasing reimbursement adjustment percentages from using quality posteriors instead of risk-standardized survival rates. Each row represents a different regression. Columns 2 and 5 report coefficients from regressions of the original adjustment percentages, while columns 3 and 6 report the changes as a percentage of absolute benchmarks. Columns 1-3 use a 25% outcome domain weight while columns 4-6 use a 100% outcome domain weight. Both simulations use FY2014 balance sheet information, withholdings, and non-quality domain scores. See the data appendix for a detailed description of the reimbursement schemes. The sample is 2,565 hospitals with balance sheet information and quality posteriors from both the 2007-2009 and 2010-2012 periods. Robust standard errors, clustered by HSA, are reported in parentheses.

References

Angrist, J. (1991): "Grouped-data Estimation and Testing in Simple Labor-Supply Models," *Journal of Econometrics*, 47(2), 243–266.

Angrist, J. and J. Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1), 1–15.

Angrist, J., P. Hull, P. Pathak, and C. Walters (2015): "Leveraging Lotteries for School Value-Added: Testing and Estimation," NBER Working Paper No. 21748.

——— (2016): "Interpreting Tests of School VAM Validity," *American Economic Review: Papers & Proceedings*, 106(5), 388–392.

Behaghel, L., B. Crépon, and M. Gurgand (2013): "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial," IZA Discussion Paper No. 7447.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Blackwell, M. (2016): "Instrumental Variable Methods for Conditional Effects and Causal Interaction in Voter Mobilization Experiments," Working Paper.

Brinch, C., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," Statistics Norway Discussion Paper No. 703.

Capps, C. (2005): "The Quality Effects of Hospital Mergers," Department of Justice Economic Analysis Group Discussion Paper No. 05-6.

Card, D., C. Dobkin, and N. Maestas (2009): "Does Medicare Save Lives?" *Quarterly Journal of Economics*, 124(2), 597–636.

Card, D., J. Heining, and P. Kline (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *Quarterly Journal of Economics*, 128, 967–1015.

Cattaneo, M., M. Jansson, and X. Ma (2016): "Marginal Treatment Effects with Many Instruments," Working Paper.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2015): "Healthcare Exceptionalism? Performance and Allocation in the U.S. Healthcare Sector," NBER Working Paper No. 21603.

Chandra, A. and D. Staiger (2007): "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, 115(1), 103–140.

Chetty, R., J. Friedman, and J. Rockoff (2014a): "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104(9), 2593–2632.

——— (2014b): "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104(9), 2633–2679.

Chetty, R. and N. Hendren (2015): "The Impacts of Neighborhoods on Intergenerational Mobility: Childhood Exposure Effects and County-Level Estimates," Working Paper.

Conway, P. (2013): "CMS Releases Latest Value-Based Purchasing Program Scorecard," Available at https://blog.cms.gov/2013/11/14/cms-releases-latest-value-based-purchasing-program-scorecard/. Last accessed October 30, 2016.

Deming, D. (2014): "Using School Choice Lotteries to Test Measures of School Effectiveness," *American Economic Review: Papers & Proceedings*, 104(5), 406–411.

D'Haultfoeuille, X. and A. Maurel (2013): "Another Look at the Identification at Infinity of Sample Selection Models," *Econometric Theory*, 29(1), 213–224.

DHHS (2015): "Better, Smarter, Healthier: In Historic Announcement, HHS Sets Clear Goals and Timeline for Shifting Medicare Reimbursements from Volume to Value," Available at http://bit.ly/1QhLv5b. Last accessed October 26, 2016.

DHHS/CMS (2015): "Hospital Value-Based Purchasing," Available at https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf. Last accessed March 20, 2016.

Doyle, J., J. Graves, J. Gruber, and S. Kleiner (2015): "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, 123(1), 170–214.

Dranove, D. and A. Sfekas (2008): "Start Spreading the News: A Structural Estimate of the Effects of New York Hospital Report Cards," *Journal of Health Economics*, 27, 1201–1207.

Dunnett, C. (1989): "Algorithm AS 251: Multivariate Normal Probability Integrals with Product Correlation Structure," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(3), 564–579.

Efron, B. and C. Morris (1973): "Stein's Estimation Rule and Its Competitors - An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117 – 130.

Ferguson, T. (1958): "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities," *Annals of Mathematical Statistics*, 29(4), 1046–1062.

Foster, D., L. Zrull, and J. Chenoweth (2013): "Hospital Performance Differences by Ownership," Truven Health Analytics. Available at http://100tophospitals.com/portals/2/assets/HOSP_12678_0513_100TopHopPerfOwnershipPaper_RB_WEB.pdf. Last accessed May 31, 2016.

Gaynor, M. and R. Town (2012): "Competition in Health Care Markets," in *Handbook of Health Economics*, ed. by M. McGuire, V. Pauly, and P. Barrow, Elsevier, vol. 2, chap. 9, 1 ed.

Geman, S. and C.-R. Hwang (1982): "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *Annals of Statistics*, 10, 401–414.

Geweke, J., G. Gowrisankaran, and R. Town (2003): "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 171(4), 1215–1238.

Grossman, M. (1972): "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy*, 82(2), 223–255.

Gupta, A. (2016): "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program," Working Paper.

Hadley, J. and P. Cunningham (2004): "Availability of Safety Net Providers and Access to Care of Uninsured Persons," *Health Services Research*, 39(5), 1527–1546.

Hausman, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46(6), 1251–1271.

Hausman, J. and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46(2), 403–426.

Heckman, J. and B. Honore (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58(5), 1121–1149.

Heckman, J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *The Review of Economics and Statistics*, 88(3), 389–432.

——— (2008): "Instrumental Variablesin Models with Multiple Outcomes: The General Unordered Case," *Annals of Economics and Statistics*, 91, 151–174.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," 71, 1161–1189.

Ho, V. and B. Hamilton (2000): "Hospital Mergers and Acquisitions: Does Market Consolidation Harm Patients?" *Journal of Health Economics*, 19, 767–791.

Hoxby, C. (2015): "Computing the Value-Added of American Postsecondary Institutions," Working Paper.

Hull, P. (2015): "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons," Working Paper.

Imbens, G. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," 62, 467–475.

Imbens, G. and D. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Kane, T. and D. Staiger (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607.

Keane, M. (1992): "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics*, 10(2), 193–200.

Kirkebøen, L., E. Leuven, and M. Mogstad (2014): "Field of Study, Earnings, and Self-Selection," NBER Working Paper 20816.

Krumholz, H., Y. Wang, J. Mattera, Y. Wang, L. F. Han, M. Ingber, S. Roman, and S.-L. Normand (2006): "An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates Among Patients With and Acute Myocardial Infarction," *Circulation*, 113, 1683–1692.

Lewbel, A. (2007): "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 141(2), 777–806.

McClellan, M. and D. Staiger (1999): "The Quality of Health Care Providers," NBER Working Paper No. 7327.

——— (2000): "Comparing Hospital Quality at For-Profit and Not-for-Profit Hospitals," in *The Changing Hospital Industry*, ed. by D. Cutler, University of Chicago Press.

McCulloch, R. and P. Rossi (1994): "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64(1), 207 – 240.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57(5), 995 – 1026.

Morris, C. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78(381), 47 – 55.

Norton, E., J. Li, A. Das, and L. Chen (2016): "Moneyball in Medicare," NBER Working Paper No. 22371.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027 – 1057.

Pear, R. (2014): "Health Law's Pay Policy is Skewed, Panel Finds," The New York Times. Available at `http://nyti.ms/1rvCy8O`. Last accessed September 1, 2015.

Raudenbush, S. and A. Byrk (1986): "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59(1), 1–17.

Rice, S. (2014): "Experts Question Hospital Raters' Methods," Modern Healthcare. Available at `http://www.modernhealthcare.com/article/20140531/MAGAZINE/305319980`. Last accessed June 1, 2016.

Robins, J. and Y. Ritov (1997): "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine*, 16, 285–319.

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–146.

Ruhnke, G., M. Coca-Perraillon, B. Kitch, and D. Cutler (2011): "Marked Reduction in 30-Day Mortality Among Elderly Patients with Community-acquired Pneumonia," *American Journal of Medicine*, 124(2), 171–178.

Sanghavi, P., A. Jena, J. Newhouse, and A. Zaslavsky (2015): "Outcomes After Out-of-Hospital Cardiac Arrest Treated by Basic vs Advanced Life Support," *Journal of the American Medical Assocation Internal Medicine*, 175(2), 196–204.

Silber, J., P. Rosenbaum, T. Brachet, R. Ross, L. Bressler, O. Evan-Shosahn, S. Lorch, and K. Volpp (2010): "The Hospital Compare Mortality Model and the Volume-Outcome Relationship," *Health Services Research*, 45(5), 1148–1167.

Sloan, F., G. Piccone, D. Taylor, and S.-Y. Chou (2001): "Hospital Ownership and Cost and Quality of Care: Is There a Dime's Worth of Difference?" *Journal of Health Economics*, 20(1), 1–21.

YNHHSC/CORE (2013): "2013 Measures Updates and Specifications: Acute Myocardial Infarction, Heart Failure, and Pneumonia 30-Day Risk-Standardized Mortality Measure (Version 7.0)," Available at `http://www.qualitynet.org/dcs/ContentServer?cid=1228774398696&pagename=QnetPublic%2FPage%2FQnetTier4&c=Page`. Last accessed November 3, 2015.

Data Appendix

I follow Doyle et al. (2015) in constructing an analysis sample from 2010-2012 CMS claims. I first link a 20% random sample of Medicare beneficiaries that originate an ambulance company claim in the CMS Carrier file to their inpatient claims, which indicate admitting hospitals and diagnoses. The claims data also include basic demographic information on beneficiaries, including birth date, sex, race, and the ZIP code where official correspondence is sent. The data are further linked to vital statistics that record when a patient dies, thereby generating the primary 30-day survival outcome. Ambulance company data, including the company's registered ZIP code, information on miles traveled, the mode and method of transport, and any pre-hospital interventions for each claim are retained from the Carrier file. Hospital ZIP codes provided by inpatient claims and linked to hospital service areas defined by the Dartmouth Atlas. Data on hospital ownership structure (non-profit private, for-profit private, and government owned) and the total number of hospital beds come from the CMS Provider of Service files, while teaching status and total diagnosis-related group payments for FY2014 come from hospital Cost Report data. Hospital volume is computed as the total number of admitted patients in the analysis sample, while average spending includes all Medicare reimbursement paid to the hospital from the first 30 days following a patient's admission, excluding those for drugs covered under Medicare Part D due to data limitations.

Following Card et al. (2009) and others, I limit the sample to patients who were admitted by ambulance through a hospital's emergency room and receiving a primary diagnosis of one of 29 "nondeferrable" conditions wherein selection into inpatient care is unlikely to be discretionary. These are the same conditions Doyle et al. (2015) identify as having weekend admissions rates close to the 2/7ths, which would be expected given no selection, and are listed in the footnote of Table 1. As with the CMS risk-adjustment methodology established by YNHHSC/CORE (2013), I keep only a patient's first hospital admission in 2010-2012. Unlike Doyle et al. (2015), I do not drop small ZIP codes, ambulances, or hospitals, nor do I limit the sample to hospitals within 50 miles of the patient's ZIP code centroid in order to minimize endogenous sample selection concerns.

Appendix Table A1 summarizes patient demographics in the analysis sample. Around 41% of beneficiaries admitted for a nondeferrable condition in 2010-2012 (column 1) were done so via ambulance (column 2); this subsample is slightly older and more female, with somewhat higher average Medicare spending and 30-day mortality. Columns 3 and 4 of Table A1 further report demographics for patients admitted in multi-hospital HSAs and to hospitals with enough quasi-experimental data to estimate quality by the minimum distance procedure outlined in the text; these subsamples appear quite representative of the full analysis sample.

Observational RAMs are estimated as hierarchical logit regressions with normal random hospital effects, separately by each of the five diagnosis categories listed in Table 1. RAM predictions $\hat{\alpha}_j$ are the volume-weighted average posterior means of the hospital effects. The benchmark RAM specification includes diagnosis and year fixed effects, patient age and sex, and the 17 diagnosis comorbidities listed in Panel B of Table 2. Appendix Table A2 also uses estimates from replicated CMS-RAM models. For these I

follow YNHHSC/CORE (2013) as closely as possible in constructing a 20% sample from 2010-2012 inpatient claims and defining diagnosis and procedure comorbidities specific to each of their AMI, heart failure, and pneumonia risk-adjustment models. The AMI model is estimated using a sample of 107,916 patients and includes indicators for the comorbidities listed in Table 2 of YNHHSC/CORE (2013). The heart failure model uses a sample of 206,363 patients and includes the comorbidity controls listed in their Table 6. Lastly, the pneumonia specification is estimated using a sample of 205,980 patients and includes comorbidity indicators that YNHHSC/CORE (2013) list in their Table 12. Regressions of reported CMS hospital scores on those generated in my samples produce coefficients of 0.93 (AMI), 1.05 (heart failure), and 1.03 (pneumonia) with standard errors on the order of 0.04, which suggests a faithful reproduction.

To simulate payments from the CMS Value-Based Purchasing program, I replicate as closely as possible the methodology outlined in DHHS/CMS (2015). I obtain FY2014 non-outcome domain scores from the VBP website and hold them fixed throughout. Achievement and improvement scores for the outcome domain are obtained from either conventional risk-standardized survival rates, as described in the text, or from $\hat{\kappa} + \hat{\lambda}\alpha_j + E[\upsilon_j|\hat{\alpha}, \hat{\beta}]$, the posterior mean of a hospital's within-HSA quality index. I generate achievement scores from the main 2010-2012 analysis sample, while improvement scores come from changes in these measures between 2007-2009 and 2010-2012. Achievement points are awarded on a linear 0-9 scale, with zero points given to hospitals that score below the median achievement score and 9 points awarded to those scoring above the mean of hospitals in the top tenth percentile. No improvement points are assigned to hospitals with negative improvement scores but are earned linearly from positive improvement with 8 points awarded to hospitals above the mean of the top tenth percentile of improvement. A hospital's overall score for the outcome domain is the maximum of achievement and improvement points multiplied by 10, which is combined with the non-outcome domains with a weight of either 25%, 40%, or 100% to arrive at its Total Performance Score. Total VBP withholdings equal 1.25% of total hospital DRG payments in FY2014 and are fully redistributed to hospitals by a linear schedule, with hospitals scoring zero on their Total Performance Score earning back zero withholdings. VBP percentage adjustments are given by these payments divided by a hospital's withholdings.

Econometric Appendix

Proof of Lemma 1

Consider the choice probability for institution j and instrument value ℓ:

$$\begin{aligned} Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) &= E[E[\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]|X_i]] \\ &= E[E[\mathbf{1}[U_{ij}(Z_i) \geq U_{ik}(Z_i), \forall k]|Z_{i\ell} = 1, X_i]] \\ &= E[E[D_{ij}|Z_{i\ell} = 1, X_i]] \\ &= E\left[E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}|X_i\right]\right] \\ &= E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \end{aligned} \tag{28}$$

The first and fifth equalities follow from the Law of Iterated Expectations, the second holds under Assumption 1, and the third and fourth use the model for D_{ij} and definition of $p_\ell(X_i) = E[Z_{i\ell}|X_i]$. Similar logic yields equation (4), as each mean selected outcomes can be written

$$E[f(Y_{ij})|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = \frac{E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]]}{Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)}, \tag{29}$$

and, following the same steps as above,

$$E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]] = E\left[\frac{f(Y_i)D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \tag{30}$$

Combining equations (28)-(30) completes the proof. □

Multivariate Probit Choice Probabilities and Mean Selected Outcomes

Under the model given by equations (10) and (12), Assumption 3, and the necessary normalizations,

$$\begin{aligned} Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) &= Pr(\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k) \\ &= \frac{1}{(2\pi)^{(J-1)/2}|\tilde{\Sigma}_\eta|^{1/2}} \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{1\ell}} \cdots \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{J'\ell}} \exp\left(-\frac{1}{2}t'\tilde{\Sigma}_\eta^{1/2}t\right) dt, \end{aligned} \tag{31}$$

where $\tilde{\pi}_{j\ell} = \pi_{j\ell} - \pi_{\bar{j}\ell}$ for fixed $\bar{j}$, $J' = J$ if $j \neq J$ and otherwise $J' = J-1$ (i.e. the integration is taken over all $k \neq j$), and $\tilde{\Sigma}_\eta = 1 + I_{J-1}$ is the variance matrix of the vector of $\eta_{ik} - \eta_{ij}$, for $k \neq j$. By the exchangeable correlation structure of $\tilde{\Sigma}$, we can furthermore use the result in Dunnett (1989) to rewrite this as a more computationally-tractable expression involving a single integral:

$$\begin{aligned} &Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) \\ &= \frac{1}{\sqrt{\pi}} \int_0^\infty \left(\prod_{k\neq j} \Phi\left(-\sqrt{2}t - (\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell})\right) + \prod_{k\neq j} \Phi\left(\sqrt{2}t - (\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell})\right) \right) \exp\left(-t^2\right) dt. \end{aligned} \tag{32}$$

Similarly, $E[Y_{ij}|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = Pr(h_{ij} \geq 0, \tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)/Pr(\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)$ with

$$
\begin{aligned}
&Pr(h_{ij} \geq 0, \tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k) \\
&= \frac{1}{(2\pi)^{J/2}|\tilde{\Sigma}(\rho_j)|^{1/2}} \int_{-\infty}^{\beta_j} \int_{-\infty}^{\tilde{\pi}_{j\ell} - \tilde{\pi}_{1\ell}} \cdots \int_{-\infty}^{\tilde{\pi}_{j\ell} - \tilde{\pi}_{J'\ell}} \exp\left(-\frac{1}{2} t' \tilde{\Sigma}(\rho_j)^{1/2} t\right) dt, \qquad (33)
\end{aligned}
$$

where $\rho_j = (\rho_{j1}, \ldots, \rho_{jJ'})$ for $\rho_{jk} = Corr(h_{ij}\eta_{ik} - \eta_{ij})$ and

$$
\tilde{\Sigma}(\rho_j) = \begin{bmatrix} 1 & -\sqrt{2}\rho_j' \\ -\sqrt{2}\rho_j & \tilde{\Sigma}_\eta \end{bmatrix}. \qquad (34)
$$

Testing Hospital RAMs

For the general quality model given by equations (1)-(2), consider the null hypothesis

H_0 *(RAM Validity)*: $Y_{ij} = \mathbf{1}[a_j + \gamma' W_i \geq \nu_i]$, where $\nu_i \mid \left((Z_{i\ell}, (U_{ij}(z_\ell))_{j=1,\ldots,J})_{\ell=1,\ldots,L}, W_i\right) \sim F_\nu$

for some distribution F_ν. In the health context, H_0 rules out hospital comparative advantage and says that the patient sorting mechanism is independent of latent health, conditional on the controls in W_i. In particular, H_0 implies $\nu_i \perp\!\!\!\perp D_i \mid W_i$, the usual basis for consistent estimation of equation (20), and is equivalent when ignoring knife-edge cases of perfectly-offsetting dependencies between health, utility, and ambulance company assignment.

By the Law of Iterated Expectations, we have under H_0 that

$$
\begin{aligned}
E[Y_i | Z_{i\ell} = 1, D_{ij} = 1, W_i = w] &= Pr(\alpha_j + \gamma' w \geq \nu_i | Z_{i\ell} = 1, U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k, W_i = w) \\
&= F_\nu(\alpha_j + \gamma' w) \\
&= E[Y_i | D_{ij} = 1, W_i = w], \qquad (35)
\end{aligned}
$$

for any ambulance company ℓ, hospital j, and w in the support of W_i. Given a first-step estimate of the propensity score $p_\ell(D_i, W_i) = E[Z_{i\ell} | D_i, W_i]$, a non-parametric test statistic for H_0 based on equation (35) is therefore the sample analogue of

$$
\begin{aligned}
E\left[Y_i \left(\frac{Z_{i\ell} - p_\ell(D_i, W_i)}{p_\ell(D_i, W_i)(1 - p_\ell(D_i, W_i))}\right)\right] &= E\left[E\left[\frac{Y_i Z_{i\ell}}{p_\ell(D_i, W_i)} - \frac{Y_i(1 - Z_{i\ell})}{1 - p_\ell(D_i, W_i)} | D_i, W_i\right]\right] \\
&= E[E[Y_i | Z_{i\ell} = 1, D_i, X_i] - E[Y_i | Z_{i\ell} = 0, D_i, X_i] = 0 \qquad (36)
\end{aligned}
$$

where the last equality follows by H_0.

An alternative test procedure leverages knowledge of the error distribution F_ν, noting that under H_0,

$$
\begin{aligned}
E[Y_i | Z_i] &= E[E[Y_i | Z_i, D_i, W_i] | Z_i] \\
&= E[F_\nu(\alpha' D_i + \gamma' W_i) | Z_i], \qquad (37)
\end{aligned}
$$

so that $E[Y_i | Z_i] - E[F_\nu(\alpha' D_i + \gamma' W_i) | Z_i] = 0$. Given first-step coefficient estimates of the RAM parameters α and γ, this equality can be verified by a Lagrange Multiplier test statistic that checks orthogonality of the

RAM's residuals $Y_i - F_\nu(\alpha' D_i + \gamma' W_i)$ with the instrument. As when validating linear VAMs (Angrist et al., 2016), a first-order equivalent Wald test statistic uses the fact that equation (37) implies vector-equality of the coefficients μ_Y and μ_F in the regressions:

$$Y_i = \mu_Y' Z_i + e_Y \tag{38}$$

$$F_\nu(\alpha' D_i + \gamma' W_i)) = \mu_F' Z_i + e_F. \tag{39}$$

A final approach notes that equations (38) and (39) are the reduced form and first stage equations of a two-stage least squares (2SLS) procedure that uses Z_i to instrument for RAM-predicted survival in a regression of realized survival Y_i. Since $\mu_Y = \mu_F$ under H_0, this procedure should produce a 2SLS of one when the RAM is valid. As in the education setting, testing the L restrictions of the Lagrange Multiplier and Wald statistic can be viewed as combining a single degree-of-freedom test for "forecast bias" or that the 2SLS "forecast coefficient" equals one (Kane and Staiger, 2008), with the 2SLS model's $L - 1$ overidentifying restrictions.

Panel A of Appendix Table A2 reports chi-squared statistics and associated p-values for non-parametric propensity score tests, using 100 randomly-selected ambulance companies admitting at least 100 patients in the main analysis sample to simplify computation. For each observational RAM specification in Figure 2, I approximate the propensity scores $p_\ell(D_i, W_i)$ by a probit model and jointly test significance of the 100 sample analogues of equation (36), correcting inference for first-step estimation error. Adding patient demographics and comorbidity controls to W_i reduces the resulting chi-squared test statistic, with 100 degrees of freedom, from 295 in column 1 to 238 in column 3. Nevertheless, all three RAM specifications reject the null hypothesis of RAM validity with $p < 0.001$. This is similar to the rejection in column 4, which tests replicated AMI, heart failure, and pneumonia RAMs from the 2013 CMS risk-adjustment methodology (see the data appendix for details of this replication).

Panel B of Table A2 reports chi-squared statistics and associated p-values for tests of forecast bias, overidentification, and the full set of parametric restrictions given by equation (37), for the same set of 100 randomly-chosen ambulance companies. Adding demographic and comorbidity controls to the RAM brings the forecast coefficient from 1.30 to 1.09, and the latter is not statistically distinguishable from one at conventional levels. Nevertheless, p-values for tests of the 2SLS model's overidentifying restrictions (with 99 degrees of freedom) are all less than 0.001. As with the non-parametric test in panel A, joint test statistics for all forecast restrictions (again with 100 degrees of freedom) are all around 200 and produce correspondingly small p-values. Although the forecast coefficient is not statistically distinguishable from one in the CMS-RAM subsample of AMI, pneumonia, and heart attack patients, the model's overidentifying restrictions continue to drive rejections of RAM validity.

Estimating the Hierarchical Linear Model

Given minimum distance quality index estimates $\hat{\beta}_j$ for hospital j in HSA $h(j)$ and vectors H_j of conventional RAM predictions and a constant, an OLS procedure applied to equation (22) consistently estimates the HLM

hyperparameters $\Gamma = (\kappa, \lambda)'$, σ, and ϕ:

$$\hat{\Gamma}_0 = (H'H)^{-1}H'\hat{\beta} \tag{40}$$

$$\hat{\sigma}_0^2 = \frac{\sum_h \overline{w}_{h0}\left(\left(\overline{\hat{\beta}}_h - \overline{H}_h'\hat{\Gamma}_0\right)^2 - \overline{\Xi}_h\right)}{\sum_g \overline{w}_{g0}} \tag{41}$$

$$\hat{\phi}_0^2 = \frac{\sum_j w_{j0}\left(\left(\left(\hat{\beta}_j - \overline{\hat{\beta}}_{h(j)}\right) - \left(H_j - \overline{H}_{h(j)}\right)'\hat{\Gamma}_0\right)^2 - \tilde{\Xi}_j\right)}{\sum_k w_{k0}}, \tag{42}$$

where $\hat{\beta}$ and H collect observations of $\hat{\beta}_j$ and H_j, $\overline{\hat{\beta}}_h$ and $\overline{H}_h$ denote HSA-level averages of $\hat{\beta}_j$ and H_j, $\overline{\Xi}_h$ denotes the variance of HSA-average estimation error, $\tilde{\Xi}_j$ is the variance of hospital estimation error net of this HSA average, and $\overline{w}_{h0}$ and w_{j0} are known weights. The step-s feasible generalized least squares estimates are

$$\hat{\Gamma}_s = (H'V_s^{-1}H)^{-1}H'V_s^{-1}\hat{\beta} \tag{43}$$

$$\hat{\sigma}_s^2 = \frac{\sum_h \overline{w}_{hs}\left(\left(\overline{\hat{\beta}}_h - \overline{H}_h'\hat{\Gamma}_s\right)^2 - \overline{\Xi}_h\right)}{\sum_g \overline{w}_{gs}} \tag{44}$$

$$\hat{\phi}_s^2 = \frac{\sum_j w_{js}\left(\left(\left(\hat{\beta}_j - \overline{\hat{\beta}}_{h(j)}\right) - \left(H_j - \overline{H}_{h(j)}\right)'\hat{\Gamma}_s\right)^2 - \tilde{\Xi}_j\right)}{\sum_k w_{ks}}, \tag{45}$$

where V_s is a block-diagonal matrix with HSA blocks $V_{hs} = \hat{\phi}_{s-1}^2 I_{J(h)} + \hat{\sigma}_{s-1}^2 + \Xi_h$, and this procedure may be iterated to convergence. Many sequences of weights $\overline{w}_{hs}$ and w_{js} will yield consistent estimates of the variance hyperparameters σ^2 and ϕ^2. In practice I follow Efron and Morris (1973) in setting

$$\overline{w}_{hs} = 1/(\hat{\sigma}_{s-1}^2 + \overline{\Xi}_h)^2 \tag{46}$$

$$w_{js} = 1/(\hat{\sigma}_{s-1}^2 + \hat{\phi}_{s-1}^2 + \Xi_j)^2, \tag{47}$$

where Ξ_j is the estimated asymptotic variance of hospital-level estimation error.

In Appendix Table A5 I also use these weights to estimate covariances between two sets of minimum distance quality estimates $\hat{\beta}_{jh}^A$ and $\hat{\beta}_{jh}^B$. Since these are all generated from different datasets A and B, under independent random sampling $Cov(\hat{\beta}_{jh}^A, \hat{\beta}_{jh}^B) = Cov(\beta_{jh}^A, \beta_{jh}^B)$, and we also obtain estimates of true quality covariance. Quality correlations are then given by dividing by iterated FGLS estimates of $\sqrt{Var(\beta_{jh}^A)Var(\beta_{jh}^B)}$.

Supplementary Results

Correlation Across Time and Conditions

Appendix Table A5 reports additional correlations, computed as described in the econometric appendix, of hospital quality measures over time and across admitting diagnosis categories. For example column 1 of Panel A shows that the most naïve quality yardstick, a hospital's observed 30-day survival rate, appears to follow a white noise process over three-year windows, with signal-to-noise ratio of around 0.5. In contrast, the time-series correlation of observational RAM predictions (panel B) and hospital quality indices (panel C) tends to decline, suggesting selection bias underlies some of the permanent component in observed survival rates. The time-series correlation for quality is also somewhat lower than for RAM, with both reduced relative to the raw survival rates in panel A.

Reducing selection bias also appears to increase many of the correlations of observational quality measures across different diagnoses categories, the full matrix of which is shown in columns 2-6 of Table A5. These correlations are computed over a broader analysis sample of Medicare patients admitted in 2001-2012 to increase precision in the narrower patient groups. The correlation reduction from panels A to panel C is especially apparent for circulatory, respiratory, and other survival outcomes, though correlations within digestive and injury diagnoses are smaller in panel C. Encouragingly, all correlations in panel C are at or above 0.2, suggesting quality measures based on one subpopulation of patients may serve as rough proxies for overall hospital effectiveness.

Correlation with Measured Inputs

Appendix Table A6 presents a more fine-grained analysis of causal hospital quality variation by correlating posteriors with observed inputs: the log average staff salary, the use of electronic records and case management systems, and the numbers of earned accreditations and available imaging technologies. These measures come from 2010-2012 Annual Surveys of the American Hospital Association and are only available for two-thirds of my national hospital sample. Usefully, column 1 of Table A6 shows that quality posteriors are unrelated to this availability. To make efficient use of the limited coverage, I regress posteriors of within-HSA hospital quality indices, $\kappa + \lambda\hat{\alpha}_j + \upsilon_j$, rather than including HSA fixed effects as in Table 4, though point estimates are qualitatively similar across the two specifications.[27] For interpretability, all regressors and the quality measure in Table A6 are normalized to have a standard deviation of one.

All five input measures positively correlate with quality index posteriors, with the salary and accreditation proxies remaining statistically significant in a multivariate "horse race" regression that includes all measures (column 7). The salary-quality association appears particularly robust, holding even when the strong quality

[27] For example, the coefficient on log average staff salary in column 8 of Table A6 becomes 0.03 when HSA fixed effects are included, with a standard error of 0.11.

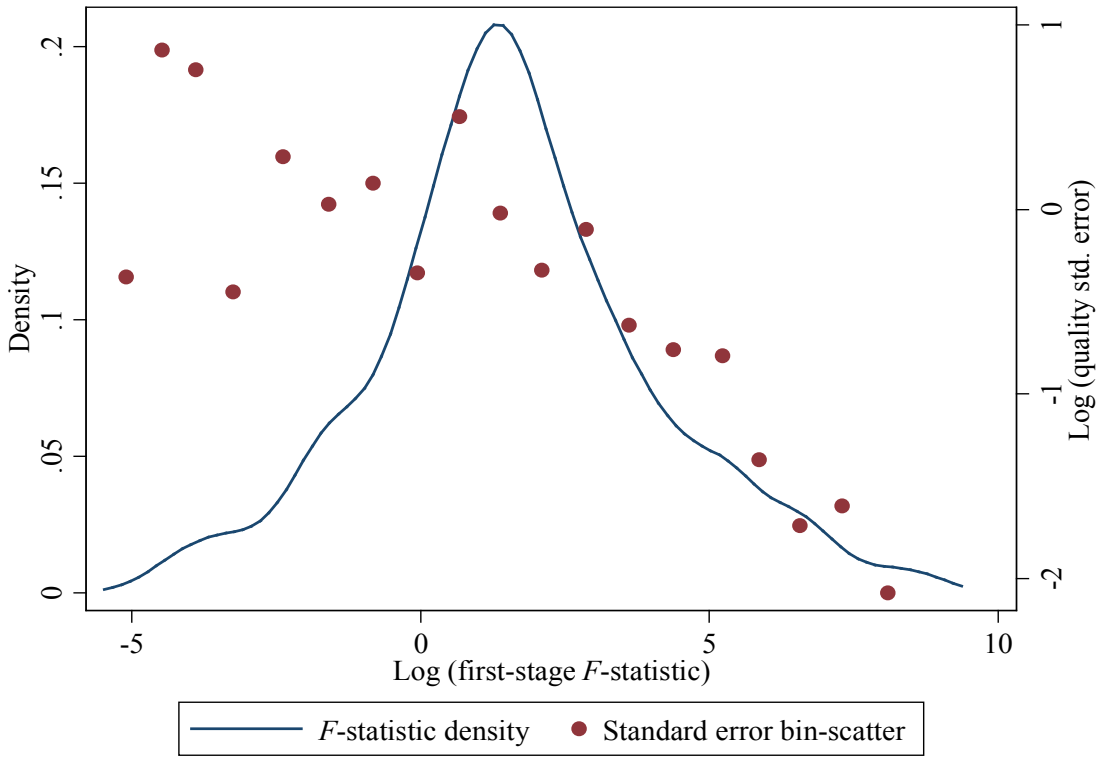
predictor of hospital volume is also included in column 8. Together these results suggest these kinds of input measures may also serve as hospital quality proxies.

Hospital Mergers

The correlations in Tables 4 and A5 need not be causal in the sense of, say, predicting the effect on hospital quality from an exogenous volume or salary increase, as omitted factors such as staff experience or doctor expertise may be correlated with both quality posteriors and observed characteristics. To explore one possible determinant of quality, I exploit the plausibly-exogenous timing of hospital mergers and estimate the average causal effect of hospital acquisition on quality posteriors. For this I obtain a list of 39 hospitals that acquired another provider between 2001 and 2012 from the American Hospital Association Summary of Changes database and recompute quality estimates, the HLM, and quality posteriors in rolling three-year lagged windows for each year in 2001-2012. I then match each acquiring hospital to a comparison group by ownership structure, teaching status, and terciles of total patient volume, average spending, and bed capacity in the year preceding the merger. The solid blue line in Appendix Figure A3, panel A, shows the trajectory of the average quality index posterior for acquiring hospitals, while the dashed red line plots the same for the set of matched comparison hospitals. That these follow roughly parallel trends up to the merger year (here normalized to zero) suggests differences in quality growth post-merger may be interpreted as causal effects in a standard difference-in-differences (DD) framework.

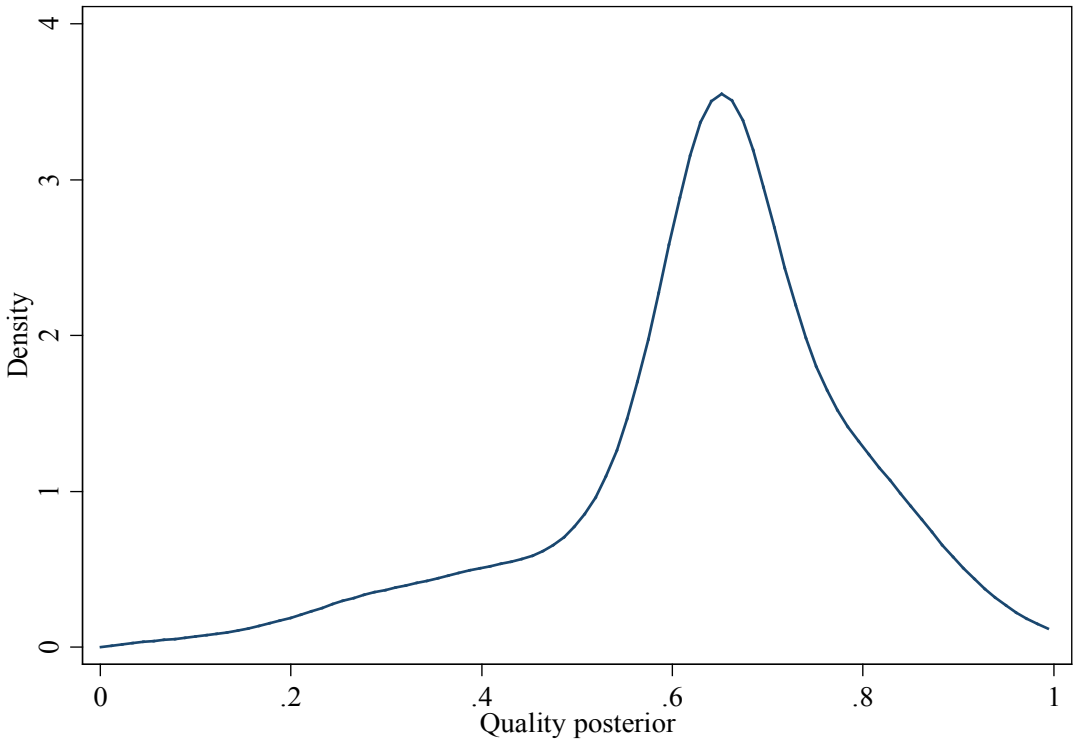
DD estimates of acquiring hospital merger effects are plotted in panel B of Figure A3 along with associated 95% confidence intervals. Pre-trend differences across the treatment and control groups are statistically indistinguishable from zero, while quality indices rise by an average of 0.16 in the first post-merger year and persist through year three. These pre-trend and effect estimates are quite robust and similar to those obtained using quality posteriors as the outcome, as shown in Appendix Table A7, with an average quality gain of around 4 percentage points. I find no significant merger effect on either observed hospital survival or observational RAM predictions, however, consistent with earlier investigations by Ho and Hamilton (2000) and Capps (2005). While a detailed analysis of hospital merger effects is outside the scope of this paper (see Gaynor and Town (2012) for a comprehensive overview of the existing literature), these results are consistent with a theory in which economies of scale in emergency healthcare production increase merging hospital quality despite offsetting anti-competitive forces. At the same time, changes in selection bias due to the new patient population post-merge appear to obscure these effects when gauged by observational measures.

Figure A1: The distributions of first-stage *F*-statistics and quality estimate standard errors



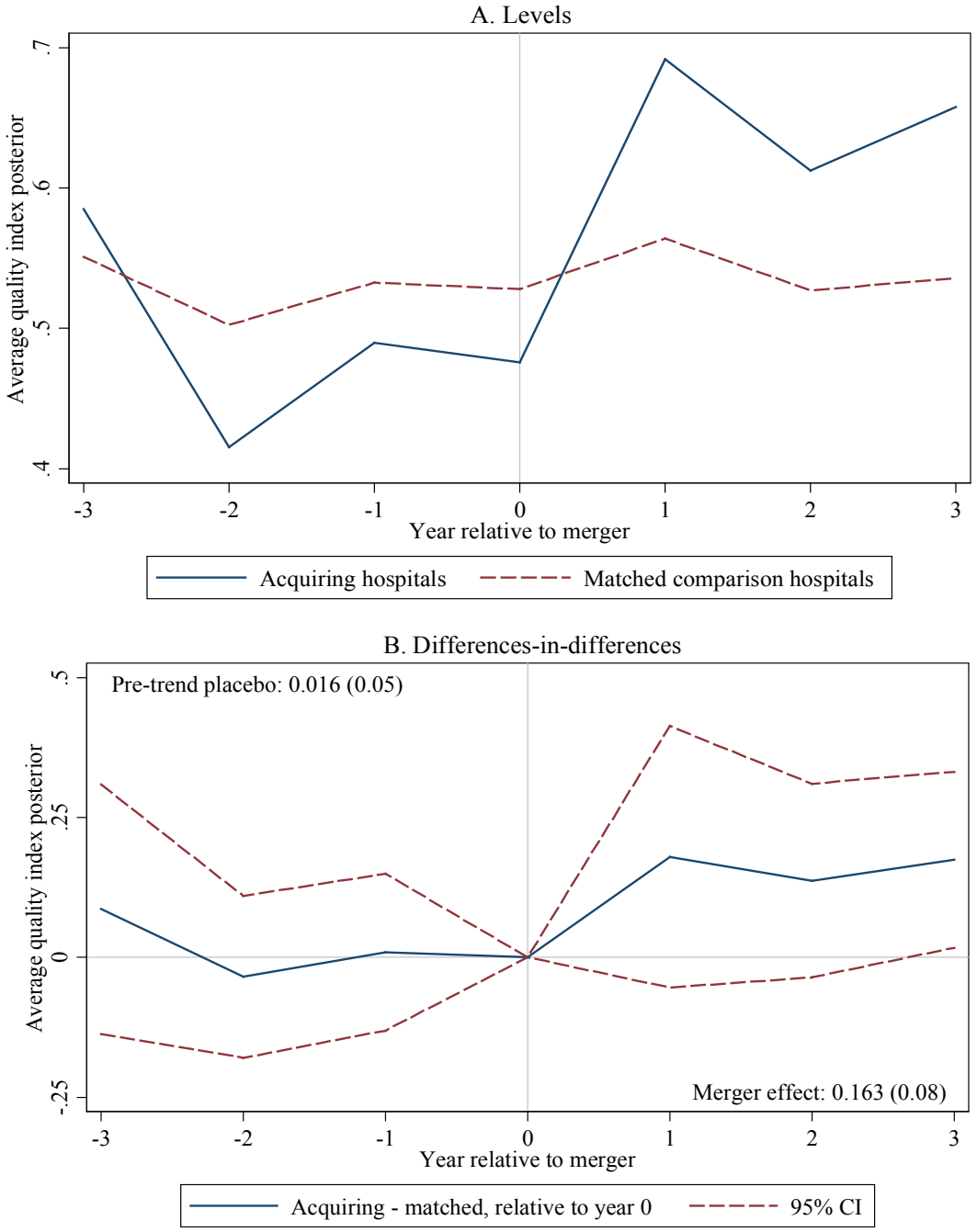
Notes: The blue line in this figure plots a Gaussian kernel density estimate of the distribution of the log of first-stage *F*-statistics for the 1,041 minimum distance quality estimates. *F*-statistics test the equality of estimated choice probabilities across all ambulance company instruments. The bandwidth used to estimate this distribution is 0.5. Red points plot the median log first-stage *F*-statistic against the median log quality estimate standard error, in 20 equal-sized bins.

Figure A2: The distribution of hospital quality posteriors in multi-hospital HSAs



Notes: This figure plots Gaussian kernel density estimates of the distribution of empirical Bayes hospital quality posteriors. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs. The bandwidth used to estimate the distribution is 0.05.

Figure A3: Estimated effects of hospital mergers on quality indices



Notes: The solid line in panel A plots average quality index posteriors for a set of 39 hospitals acquiring another hospital in a merger between 2001-2012, relative to the merger year. The dashed line shows average quality indices for a set of matched comparison hospitals that were not merged in these years. Acquiring hospitals are matched by their ownership structure (non-profit, for-profit, or government owned), teaching status, and terciles of total patient volume, average spending, and bed capacity in the year preceding the merger. Quality posteriors for each calendar year are estimated in three-year windows using the preceding two years of data and the hierarchical model described in the text. Panel B plots difference-in-differences estimates of merger effects on quality posteriors. The dashed red lines indicate 95% confidence intervals, clustered by HSA. Average effects pre- and post-merger periods are displayed along with cluster-robust standard errors

Table A1: Patient characteristics

	All nondeferrable Medicare admissions	Analysis sample		
		All hospitals	Multi-hospital HSAs	With min. dist. quality ests.
	(1)	(2)	(3)	(4)
30-day survival	0.875	0.833	0.834	0.834
Log (spending)	8.821	9.284	9.323	9.304
Age	80.22	81.76	81.58	81.61
Male	0.410	0.379	0.382	0.383
White	0.873	0.875	0.857	0.891
Black	0.082	0.082	0.093	0.073
Circulatory diagnosis	0.233	0.220	0.234	0.238
Respiratory diagnosis	0.208	0.200	0.188	0.187
Digestive diagnosis	0.101	0.065	0.066	0.065
Injury diagnosis	0.118	0.177	0.177	0.181
Patients	998,489	405,172	254,101	179,589

Notes: Column 1 of this table reports average characteristics of patients, from a 20% random sample of Medicare inpatient claims, who were admitted to a hospital in 2010-2012 for one of the 29 nondeferrable conditions listed in the notes to Table 1. Column 2 summarizes patient characteristics in the ambulance company analysis sample, while columns 3 and 4 subset this sample to patients admitted in a multi-hospital HSA and to a hospital with enough quasi-experimental data to construct minimum distance quality estimates, respectively.

Table A2: Hospital RAM bias tests

	Benchmark RAM			CMS-RAM
	(1)	(2)	(3)	(4)
		A. Propensity score test		
Test statistic (100 d.f.)	295.37	287.78	237.52	186.42
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
		B. Forecast tests		
Forecast coefficient	1.301	1.187	1.086	1.294
	(0.123)	(0.106)	(0.095)	(0.262)
Test statistics (d.f.):				
Forecast bias (1)	6.04	3.12	0.82	1.26
	[0.014]	[0.077]	[0.365]	[0.262]
Over-identification (99)	189.98	184.71	183.67	149.94
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
All restrictions (100)	201.56	192.80	189.43	171.02
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Risk-adjusters:				
Diagnosis/year FEs	Y	Y	Y	Y
Patient age/sex		Y	Y	Y
Comorbidities			Y	Y
Patients:		405,173		82,815

Notes: This table summarizes tests for bias in hospital risk-adjustment models with ambulance company instruments. All RAMs are hierarchical logit models of 30-day survival, estimated separately for each diagnosis category in Table 1. Columns 1-3 estimate RAMs in the full analysis sample, while the model in column 4 uses a nationally-representative sample of AMI, heart failure, and pneumonia Medicare patients admitted in 2010-2012. The first column includes year and diagnosis fixed effects as RAM controls, while the second adds patient age and sex, and the third includes all comorbidity indicators listed in the notes to Table 2. The specification in column 4 replicates the 2013 CMS 30-day risk-standardized mortality models. Tests use 100 randomly-selected ambulance companies referring at least 100 patients. Panel A reports test statistics for the joint significance of each company in the propensity score weighting scheme outlined in the appendix. Panel B reports forecast coefficients from 2SLS regressions of realized survival on RAM-predicted survival, instrumented by ambulance company indicators. The forecast bias test statistic is for the null hypothesis that the forecast coefficient equals 1. The full test combines forecast bias and overidentifying restrictions and is implemented by regressing RAM residuals on ambulance indicators and testing their joint significance. Propensity scores for panel A are estimated by company-specific probit models. Test statistics are robust to heteroskedasticity and account for first-step propensity score estimation error. Robust standard errors are reported in parentheses; test p-values are reported in brackets.

Table A3: Robustness of main results to alternative specifications

	Within-HSA rank corr. w/ preferred spec.	% positively-selected regions	Potential survival correlates			FY2014 VBP change (%)		Ranking survival gains (%)	
			Government hospital	Log (avg. spending)	Log (volume)	Non-profit hospital	Teaching hospital	Max. RAM prediction	Max. quality post.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Preferred specification	1.000	86.0	-0.079	0.230	0.517	8.70	14.92	3.39	3.97
			(0.029)	(0.111)	(0.223)	(3.40)	(6.78)		
Regions are HRRs (not HSAs)	0.926	99.3	-0.067	0.216	0.572	6.99	9.03	3.75	4.64
			(0.029)	(0.120)	(0.222)	(4.13)	(5.04)		
Z_{il} is in Z_{ij} if j is company l's most-referred (not closest) hospital	0.939	87.1	-0.079	0.254	0.571	8.43	9.20	3.07	3.55
			(0.034)	(0.119)	(0.232)	(5.45)	(4.95)		
No risk-adjusters in propensity scores $p_l(X_i)$	0.942	85.5	-0.082	0.298	0.699	11.05	18.64	3.44	4.13
			(0.038)	(0.153)	(0.295)	(2.80)	(5.10)		
Health/utility distributed by Student's *t(2)* (not normally)	0.959	79.6	-0.101	0.280	0.681	6.75	16.00	3.36	3.79
			(0.039)	(0.144)	(0.290)	(2.47)	(4.75)		
HLM includes J and J×RAM prediction interaction	0.681	85.6	-0.116	0.434	1.020	15.84	21.93	2.72	3.27
			(0.064)	(0.220)	(0.422)	(6.94)	(9.13)		

Notes: The first row in this table reports estimates from the hierarchical model described in the text. The second row modifies this model by using hospital referral regions rather than hospital service areas to define local emergency care markets. The third row uses referral rates rather than ambulance-hospital distance to partition ambulance company instruments, while the fourth excludes patient age, sex, and RAM comorbidities from ambulance company propensity scores. The fifth row assumes patient health and utility indices are distributed by a multivariate Student's t distribution with two degrees of freedom. The sixth row adds the total number of hospitals in a hospital's HSA and its interaction with RAM predictions to the benchmark hierarchical linear model. Column 1 reports the within-HSA rank correlation of quality posterior from each model with the preferred specification. Column 2 reports the percentage of hospital regions with positive average selection bias, and columns 3-5 report within-HSA regressions, as in Table 4. Columns 6 and 7 report regressions of the change in VBP reimbursement rates, as in Table 5, while columns 8 and 9 report simulated gains from rank-based admissions policies, as in Figure 6. Standard errors, clustered by region, are reported in parentheses.

Table A4: Hospital characteristics

	Multi-hospital HSAs		Single-hospital HSAs
	With min. dist. quality estimates	Without quality estimates	
	(1)	(2)	(3)
Observed survival	0.838	0.837	0.838
RAM prediction	0.011	0.008	0.006
Risk-adjustment index	1.977	1.972	1.992
Non-profit	0.763	0.696	0.681
For-profit	0.119	0.207	0.154
Government-owned	0.118	0.096	0.166
Teaching	0.584	0.578	0.573
Log (avg. spending)	9.600	9.693	9.496
Log (volume)	5.453	4.979	4.801
Log (# of beds)	5.916	5.802	5.159
HSA hospital count	3.654	11.174	1.000
Hospitals	1,041	1,316	2,464

Notes: This table reports average characteristics of hospitals in the analysis sample; column 1 summarizes the sample of hospitals with enough quasi-experimental data to construct minimum distance quality estimate, while columns 2-3 characterize the remaining sample. Observed survival posteriors shrink observed rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The risk-adjustment index comes from the estimated fixed portion of the observational RAM.

Table A5: Correlation structure of 30-day survival, RAM predictions, and quality indices

	Over time	Across diagnosis categories					
		Circulatory	Respiratory	Digestive	Injury	All other	
	(1)	(2)	(3)	(4)	(5)	(6)	
			A. Observed survival rates				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.529	0.179	1.000				Respiratory
2004-2006	0.517	0.394	0.204	1.000			Digestive
2001-2003	0.491	0.402	0.167	0.424	1.000		Injury
		0.259	0.168	0.334	0.471	1.000	All other
			B. RAM predictions				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.323	0.298	1.000				Respiratory
2004-2006	0.286	0.209	0.193	1.000			Digestive
2001-2003	0.238	0.211	0.182	0.157	1.000		Injury
		0.255	0.316	0.178	0.154	1.000	All other
			C. Hospital quality indices				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.292	0.385	1.000				Respiratory
2004-2006	0.204	0.413	0.279	1.000			Digestive
2001-2003	0.196	0.431	0.285	0.351	1.000		Injury
		0.319	0.430	0.191	0.222	1.000	All other

Notes: This table reports estimated correlation coefficients for a hospital's 30-day survival rate, RAM prediction, and quality index. Column 1 correlates data from the 2010-2012 analysis sample with corresponding data from 2007-2009, 2004-2006, and 2001-2003, while columns 2-6 report correlations across the five patient diagnosis categories over the entire 2001-2012 period. The sample in column 1 includes 2,357 hospitals operating in 695 multi-hospital HSAs; columns 2-6 report correlations for 5,805 hospitals in 2,021 multi-hospital HSAs. See the econometric appendix for a description of the estimation of correlations in panel C.

Table A6: Standardized regressions of residual hospital quality index posteriors on measured inputs

Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Has inputs data	0.016							
	(0.023)							
Log (average staff salary)		0.075					0.060	0.053
		(0.017)					(0.018)	(0.018)
Uses electronic records			0.050				0.043	0.040
			(0.030)				(0.032)	(0.031)
Uses case management				0.071			-0.016	-0.047
				(0.031)			(0.035)	(0.035)
# of accreditations					0.072		0.056	0.030
					(0.018)		(0.022)	(0.025)
# of imaging technologies						0.049	-0.010	-0.044
						(0.016)	(0.022)	(0.029)
Log (volume)								0.072
								(0.032)
Hospitals:	4,821				3,199			

Notes: Each column of this table reports coefficients from regressions of hospital quality index posteriors, net of HSA random effects, on hospital characteristics. All variables are normalized to standard deviation units. Column 1 regresses quality on an indicator for the availability of input data from the American Hospital Associations Annual Survey in the full sample, while columns 2-8 report coefficients from regressions in the sample of hospitals with measured inputs. Inputs are measured in the first year in which data are available in 2010-2012. Average staff salary is computed by dividing total facility payroll by full-time equivalent total personnel. Accreditations include those by The Joint Commission, recognition for one or more Accreditation Council for Graduate Medical Education accredited programs, medical school affiliation with the American Medical Association, affiliation with the National League for Nursing, accreditation by the Commission on Accreditation of Rehabilitation Facilities, membership in the Council of Teaching Hospitals of the Association of American Medical Colleges, Blue Cross contracting or participating, Medicare certification by the U.S. Department of Health and Human Services, accreditation by the Healthcare Facilities Accreditation program of the American Osteopathic Association, approval of an internship by the American Osteopathic Association, approval of a residency by the American Osteopathic Association, and DNV Healthcare accreditation. Imaging technologies include CT scanners, diagnostic radioisotope facilities, EBCT systems, full-field digital mammography, MRI machines, IMRI machines, magnetoencephalography machines, multislice spiral computed tomography scanners, PET scanners, PET/CT scanners, SPECT scanners, and ultrasounds. Standard errors, clustered by HSA, are reported in parentheses.

Table A7: Difference-in-differences estimates of hospital merger effects

	Pre-trend placebo	Merger effect	Hospitals	Obs.
	(1)	(2)	(3)	(4)
	A. Alternative quality measures			
Observed survival	-0.003	-0.001		
	(0.002)	(0.002)		
RAM prediction	-0.003	-0.009		
	(0.006)	(0.006)	456	2,225
Quality index posterior	0.016	0.163		
	(0.050)	(0.080)		
Quality posterior	-0.002	0.039		
	(0.016)	(0.023)		
	B. Quality index estimate robustness checks			
Balanced panel	0.098	0.190	196	1,226
	(0.066)	(0.098)		
Volume weighting	0.001	0.192	456	2,225
	(0.085)	(0.113)		
Quintile matching	-0.010	0.176	195	934
	(0.055)	(0.085)		
Nearest-neighbor matching	0.005	0.119	138	867
	(0.065)	(0.063)		

Notes: This table reports difference-in-differences estimates of the effect of mergers on various hospital quality measures. The sample and estimation for panel A is described in the notes to Figure 4. Panel B reports alternative estimates of the effect of mergers on hospital quality index posteriors. The first row restricts the sample to hospitals observed for three years preceding and following the merger year. The second row weights the difference-in-differences specification by a hospital's Medicare patient volume, while the third uses quintiles (rather than terciles) of total patient volume, average spending, and bed totals to match hospitals. The final row in panel B matches each merging hospital to a comparison hospital with the same ownership structure and teaching status and the shortest Mahalanobis distance in volume, spending, and bed capacity. Standard errors, clustered by HSA, are reported in parentheses.