Estimating Hospital Quality with Quasi-experimental Data*

Peter Hull†

February 2018

Abstract

Non-random sorting can bias observational measures of institutional quality and distort quality-based polices. I develop alternative quasi-experimental approaches to quality estimation that accommodate nonlinear causal effects, institutional specialization, and unobserved selection-on-gains. I use this framework to compute empirical Bayes posteriors of the quality of 4,821 U.S. hospitals, combining estimates from ambulance referral quasi-experiments with predictions from observational risk-adjustment models. Higher-spending, higher-volume, and privately-owned hospitals are of higher quality, and most healthcare markets exhibit positive Roy selection-on-gains. I then simulate Medicare reimbursement and consumer guidance programs based on different hospital quality measures. Higher-spending providers tend to see moderately larger performance-linked subsidies when quality posteriors replace conventional rankings, while teaching hospitals are reimbursed relatively less. Admissions policy simulations highlight limitations of consumer guidance programs in settings with unobserved Roy selection: redirecting patients to top-ranked hospitals may worsen expected survival when based on observational rankings, while quasi-experimental rankings appear to generate modest gains.

*I thank Alberto Abadie, Nikhil Agarwal, Isaiah Andrews, Josh Angrist, Amitabh Chandra, David Cutler, Dave Deming, Joe Doyle, Amy Finkelstein, Matt Gentzkow, Jon Gruber, Nick Hagerty, Nathan Hendren, Max Kasy, Larry Katz, Pat Kline, Jack Liebersohn, Bentley MacLeod, Rachael Meager, Magne Mogstad, Yusuke Narita, Parag Pathak, Bryan Perry, Jesse Shapiro, Doug Staiger, Chris Walters, Glen Weyl, and seminar participants at Brown, Columbia, Harvard, IIES, Microsoft, MIT, NBER, FRBNY, Penn State, Princeton, Stanford, U Chicago, UC Berkeley, U Washington, UW Madison, and Yale for their many helpful comments and suggestions. I am especially thankful to the Doyle et al. (2015) research team for sharing their code for the ambulance instruments and to EMS professionals Ben Artin, Mark Millet, Laura Segal, Julia Taylor, and Kevin Wickersham for their institutional wisdom. I gratefully acknowledge funding from the National Institute on Aging (#T32-AG000186) and the Spencer Foundation (#201600065); all views are my own. Keywords: hospital quality, instrumental variables, Roy selection. JEL C26, C36, I11, I18, L15

†The University of Chicago and Microsoft Research. Email: hull@uchicago.edu; website: http://peterhull.net

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 come to play an important policy role, particularly in the regulation of education and healthcare systems. Hospitals with low risk-adjusted mortality rates, for example, tend to be rewarded with higher Medicare reimbursements, while providers with poor survival outcomes may be flagged publicly as low-performers. These quality-based policies, in turn, have been found to shape both hospital incentives and patient admission decisions (Norton et al., 2017; Gupta, 2017; Dranove and Sfekas, 2008; Chandra et al., 2015).

To date, performance-based regulation has exclusively relied on observational quality estimators, such as value-added models (VAMs) in education and risk-adjustment models (RAMs) in health. These models leverage strong selection-on-observables assumptions – that a patient's choice of hospital, for example, is as-good-as-random conditional on a set of observed covariates. When hospital sorting is instead correlated with unobserved determinants of patient health, hospital RAMs and the supervisory policies leveraging them are prone to systematic bias. RAM-based admission guidance may itself be a source of this selection bias, to the extent it encourages the selection of more appropriate hospitals, as may be other factors like the medical expertise of a patient's ambulance driver or the non-random co-location of high-quality providers and differentially-healthy patients.

As in other settings, instrumental variable (IV) methods in principle offer a solution to this sort of selection bias. In practice, researchers hoping to exploit plausibly quasi-experimental variation in institutional choice face several methodological challenges. Linear IV methods, including those developed by Angrist et al. (2017) 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 on all potential patients.[2] This rules out both institutional comparative advantage and selection-on-gains, forces that are likely important in many settings including healthcare (Chandra and Staiger, 2007; Chandra and Staiger, 2017). Constant effect restrictions are also largely inappropriate for modeling binary outcomes, including the 30-day survival indicators underlying hospital RAMs.

[1] See, for example, Chetty et al. (2014b), Angrist et al. (2017), Chetty and Hendren (2017), Hoxby (2017), and Card et al. (2013) for recent estimates of the 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 additional assumptions (Behaghel et al., 2013; Kirkebøen et al., 2016; Hull, 2015; Blackwell, 2017). Even in these cases, LATE-based quality measures are likely to be 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 treatment effects on the treated, is also possible.

This paper develops a flexible approach for measuring institutional quality with quasi-experimental data that, in contrast to linear IV methods, allows for nonlinear causal response functions and unobserved selection on institutional specialization. Conventional nonlinear IV estimators are likelihood-based in a way that can be computationally intractable or requiring parametric assumptions which are difficult to assess or interpret in practice. Simply estimating a nonlinear "first stage" for institutional sorting requires fitting a high-dimensional multinomial choice model, the practical difficulties of which are well known (Hausman and Wise, 1978; McFadden, 1989; McColloch and Rossi, 1994; Berry et al., 1995). In related work, Geweke et al. (2003) use Markov-chain Monte Carlo techniques to simplify the estimation of hospital quality for 114 Los Angeles County providers, leveraging differential distance instruments, a multinomial probit model of hospital selection, and a probit specification for the short-term mortality outcomes of elderly pneumonia patients. To evaluate this likelihood, Geweke et al. (2003) specify independent priors for each of the model's 268 free parameters and make several additional functional form restrictions and calibrations, some of which appear to be rejected by the data. Characterizing the role played by such parameterizations, versus the potentially-exogenous variation of the instruments, can prove quite difficult.

Rather than fitting a fully-specified likelihood to individual microdata, my approach matches a sparse set of moments from a multi-dimensional Roy (1951) selection model to quantities identified under quasi-random instrument assignment. This yields a tractable framework for estimating institutional effectiveness that is fully non-parametric given sufficiently-rich instrument variation. When quasi-experimental data is more limited, I propose and build intuition for using distributional restrictions on the model's latent variables in order to extrapolate from observed quasi-experimental quantities to the structural parameters of interest. A minimum distance procedure easily implements this semi-parametric approach, even for large-scale problems.

I next use this framework to estimate the quality of a large set of U.S. hospitals from a nationally-representative sample of emergency Medicare patients. Leveraging quasi-experimental variation in the referral preferences of a patient's assigned ambulance company, I fit a multivariate probit model for hospital admissions and 30-day survival that allows for, but does not impose, Roy selection on hospital specialization. Doyle et al. (2015) first propose ambulance company instruments as a more credible alternative to distance-based identification strategies, which may be biased by non-random hospital and patient co-location (Hadley and Cunningham, 2004). While Doyle et al. (2015) and, more recently, Doyle et al. (2017a) use ambulance referral variation to instrument in linear models for the average Medicare spending and quality rating of a patient's hospital, respectively, my nonlinear approach instruments a patient's hospital directly. This allows ambulance referral variation in affect patient outcomes through any hospital-specific characteristic, observed or unobserved, rather than

through spending or other quality proxies.

This first estimation step yields a set of noisy quality estimates for 2,082 hospitals with sufficient quasi-experimental data, representing 85% of admissions. As in other recent explorations of quasi-experimental institutional effects (Angrist et al., 2017, Chetty and Hendren, 2017, Finkelstein et al., 2017), I then use these estimates to fit a hierarchical linear model relating true hospital quality to conventional RAM quality predictions. This model is used to compute empirical Bayes posteriors for the full sample of hospitals which optimally combine quasi-experimental and observational estimates in a classical bias-variance tradeoff, reducing overall mean squared prediction error.

Analyses of the quality posteriors reveal several dimensions of hospital quality and patient sorting. Higher-volume hospitals and those that spend more per Medicare patient produce better average survival outcomes over a given patient population, while government-run hospitals are of systematically lower-quality. Specifically, I estimate that moving a typical patient to a hospital with a one standard deviation higher log volume or log average spending increases her expected 30-day survival probability by 0.2 and 0.1 percentage points, respectively, while moving her to a government-run hospital from one that is privately-owned decreases expected survival by 0.4 percentage points. These results are qualitatively similar to earlier observational and quasi-experimental studies (Foster et al., 2013; Chandra et al., 2015; Doyle et al., 2015), and I find similar patterns in analyses of conventional RAM predictions.

Turning to sorting, I find that hospital quality and selection bias are negatively correlated, suggesting that better hospitals tend to serve sicker patients on average. Consequently, conventional rankings of hospitals and rankings based on quality posteriors are highly correlated despite pervasive bias. I moreover find robust evidence of hospital comparative advantage and positive Roy selection, with patients being admitted to more appropriate hospitals on average. Only 5-15% of the survival rate benefit of this selection-on-gains appears due to patients choosing hospitals that are closer to them or that tend to serve observably-similar individuals, implying both that hospitals specialize on unobservables and that patients and ambulance companies at least somewhat sort on this specialization.

Finally, I quantify the economic importance of non-random hospital selection by simulating quality-based Medicare reimbursement and patient guidance policies. I find that counterfactually ranking hospitals by their quality posteriors changes the types of providers subsidized by a Value-Based Purchasing program relatively little, though higher-spending and non-teaching hospitals tend to see moderately higher reimbursements. In simulations of quality-based admission policies, I find that a typical patient has a 3.2 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 3.6 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, which makes current patient choices better than random. Moving a random patient from her *selected* hospital to the the highest-RAM provider is found to decrease her expected survival by around 0.2 percentage points; only with quasi-experimental reductions in average selection bias do quality-based redirection policies improve survival outcomes from this benchmark, by around 0.2-1.1 percentage points. This exercise highlights a general issue for performance-based guidance policies in the presence of unobserved institutional specialization, as well as the potential value of new econometric methods to accommodate treatment effect heterogeneity.

I organize the remainder of this paper as follows: the next section develops a moment-based approach for estimating institutional quality with instrumental variables and discusses both non- and semi-parametric quality identification. Section 3 outlines the institutional setting for hospital quality and describes the Medicare analysis sample and estimation procedure. Lastly, 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

We observe an outcome Y_i for a set of individuals i selecting one of many possible institutions $j = 1, \ldots, J$. A set of indicators D_{ij}, collected in the vector D_i, index this choice. 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 her admission with $Y_i = 0$ otherwise. 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 their quality, defined here as $q_j = E[Y_{ij}]$.[3] Quality represents the expected outcome from sending a representative individual in the population to institution j, so that institutional quality comparisons avoid any bias from non-random sorting that leads to correlation between Y_{ij} and D_{ij}. The difference in average selected and potential outcomes,

[3] I choose this definition of quality as it is appears closest to the one-dimensional representative summary of hospital performance, or "value," referenced by policymakers (e.g. DHHS (2015)). As shown below, other moments of the potential outcome distribution are identified by my semi-parametric framework; I leave for future work the analysis of using these for quality-based hospital regulation.

$E[Y_{ij}|D_{ij}=1]-E[Y_{ij}]$, quantifies 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, 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 institutions are chosen to maximize these subjective 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, $(Y_i,D_i',Z_i')'$, with a generalized multi-dimensional Roy (1951) selection model (e.g., 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, however, the model does not limit the possibility of either institutional comparative advantage or endogenous selection on potential outcomes: the causal institutional effects $Y_{ij}-Y_{ik}$ need not be constant across individuals i, and potential outcomes may be correlated with the latent utilities governing institutional choice. This allows for the possibility of "essential heterogeneity," in the language of Heckman et al. (2006).

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

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

Quasi-random instrument assignment ensures that while realized institutional choices may be correlated with potential outcomes, there is variation in conditionally-exogenous factors $Z_{i\ell}$ that can affect sorting through the frontier of latent payoffs, $U_{ij}(Z_i)$. My approach leverages this variation with knowledge or first-step 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 $p_\ell(X_i)$ is bounded away from zero for each ℓ. All individuals in the population of interest thus face a non-zero risk of assignment to each of the L instrument values. I return to how the analysis sample is constructed to plausibly satisfy this condition in section 3.

2.2 Non-parametric Identification

Quasi-experimental instrument assignment is sufficient for non-parametric estimation of certain causal moments of the model's latent variables, Y_{ij} and $U_{ij}(z_\ell)$. Namely, the following lemma 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") along with the means of any function of potential treatment-j outcomes for such individuals (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 if the instrument is unconditionally independent of potential outcomes, as in a randomized control trial, we may take X_i to be empty so that the propensity scores $p_\ell(X_i)$ are constant; the right-hand side of equations (3) and (4) would then reduce to $E[D_{ij}|Z_{i\ell} = 1]$ and $E[f(Y_i)|D_{ij} = 1, Z_{i\ell} = 1]$, respectively. More generally, these formula use the non-parametrically identified propensity scores to appropriately re-weight the data in order to mimic this idealized experimental setting. Even with high-dimensional confounders in X_i, the scalar propensity score adjustment is sufficient for identifying choice probabilities and mean selected outcomes (Rosenbaum and Rubin, 1983), simplifying subsequent analyses.

Without further parameterizations of the model, equations (3) and (4) are enough to estimate institutional quality in settings with sufficiently rich quasi-experimental data. Intuitively, by varying ℓ and setting $f(Y_i) = Y_i$ we may 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 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 that place more weight on instrument values z_ℓ with the highest estimated choice probabilities.

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

Proposition 1 (*Non-parametric 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 $(Y_i, D_i', Z_i', X_i')'$,

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

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

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 positive. □

The local linear regression estimator $\hat{q}_j$ is consistent for institution j's quality when $G_j' Z_i$, the institution-specific 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 developed in Lewbel (2007) will also be consistent for q_j under a somewhat stronger support condition than the one used in Proposition 1.[4] 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 regressors, bandwidths, and weights is left for future research.

2.3 Semi-parametric Identification

Further restrictions on the selection model can substitute for rich quasi-experimental data when institutional quality is not non-parametrically identified. Intuitively, parameterizations of the joint distribution of latent variables Y_{ij} and $U_{ij}(z_\ell)$ in terms of a low-dimensional vector θ_0 also render the identified moments in Lemma 1 known functions of θ_0. Quasi-experimental variation in the moments that is sufficient to uniquely pin down the elements of θ_0 relevant to $E[Y_{ij}]$ would then identify quality without the large support condition of Proposition 1. Unlike conventional nonlinear IV estimators, this approach does not require specifying and maximizing a complete likelihood for the

[4] 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 $[p, 1]$.

individual microdata. Rather, quality is estimated by computationally-tractable minimum distance procedure applied to a set of estimated causal moments.[5] I first motivate and describe this approach with a generic parameterization of the latent variables; the next subsection then formally establishes identification of quality in a particular class of parameterizations for binary potential outcomes, which I later use to estimate hospital quality.

For some known distribution function $F(\cdot)$, suppose

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

so that the various choice probabilities and mean selected outcomes identified under Assumption 1 are also known functions of the unknown $\theta_0 \in \mathbb{R}^K$. Let $m(\cdot)$ be a vector collecting some subset of these functions and let $\hat{m}$ 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 set of propensity score estimators $\hat{p}_\ell(\cdot)$. Under mild regularity conditions (see, e.g., Hirano et al. (2003)), we have $\sqrt{N}(\hat{m}-m(\theta_0)) \Rightarrow N(0,Q)$, where Q is an identified asymptotic variance matrix. Next, suppose we may partition the structural parameter vector as $\theta_0 = [\tilde{\theta}_0, \bar{\theta}_0]$ where for any $\theta = [\tilde{\theta}, \bar{\theta}]$ satisfying $m(\theta) = p\lim \hat{m}$ we have $\tilde{\theta} = \tilde{\theta}_0$. A consistent minimum distance estimator of the identified subvector of θ_0 is then given by the corresponding elements of

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

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

$$\sqrt{N}(\hat{\tilde{\theta}} - \tilde{\theta}_0) \Rightarrow N\left(0, (M'AM)^{-1}M'AQAM(M'AM)^{-1}\right) \tag{8}$$

where $M = \frac{\partial m(\theta)}{\partial \tilde{\theta}}|_{\theta_0}$ and $\hat{A} \xrightarrow{p} A$.[6] Therefore if, given (6), $E[Y_{ij}] = \tilde{F}_j(\tilde{\theta}_0)$ for known $\tilde{F}_j(\cdot)$, we may form a consistent and asymptotically normal quality estimate $\hat{q}_j = \tilde{F}_j(\hat{\tilde{\theta}})$ based on the first-step quasi-experimental moment estimates.

Minimum distance quality estimation is likely to be tractable, even as the numbers of institutions J, instrument values L, and controls in X_i grow 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 in θ_0 and may be separately approximated by standard techniques, such as the method of nonparametric sieves (Ge-

[5] One may make a link between this approach and classic notion of "indirect inference" (Gourieroux et al., 1993), with parameterizations of the potential outcome distribution acting as an auxiliary model for the complete likelihood.

[6] As usual, an efficient choice of $\hat{A}$ is $\hat{Q}^{-1}$ for some consistent variance estimator $\hat{Q} \xrightarrow{p} Q$, in which case $\sqrt{N}(\hat{\tilde{\theta}}-\tilde{\theta}_0) \Rightarrow N\left(0, (M'Q^{-1}M)^{-1}\right)$. Note that with Q identified without knowledge of the structural θ_0, 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 $\hat{M} = \frac{\partial m(\theta)}{\partial \tilde{\theta}}|_{\hat{\theta}}$.

man and Hwang, 1982) or doubly-robust estimators in settings with high-dimensional data (Belloni et al., 2014). Given $\hat{m}$, evaluating the minimum distance objective function requires computing at most $((D+1)J-1)L$ nonlinear functions of the structural parameters, where D is the dimension of the outcome function $f(\cdot)$.[7] Importantly, these functions do not depend on the data so that, unlike with conventional likelihood-based IV estimators, the difficulty of the nonlinear computation does not increase with the number of observations. For some specifications, including several within the class considered in the next subsection, $m(\theta)$ will take a form that is straightforward to evaluate by standard statistical software. Simulation methods can solve more exotic parameterizations; again the fact that the simulated objects are non-stochastic simplifies this relative to classical applications of simulated minimum distance (McFadden, 1989, Pakes and Pollard, 1989).

The separation of quasi-experimental moments in $\hat{m}$ from the structural assumptions underlying $m(\theta)$ also helps to clarify the role of the latter in semi-parametric quality estimation. The minimum distance approach uses a low-dimensional parameterization of the latent variable distribution in order to extrapolate from a discrete set of non-parametric instrumental variable moments to the structural parameters of interest. Figure 1 illustrates an example of this extrapolation. The vertical and horizontal values of each point give an estimated mean selected outcome and choice probability, respectively, of a fixed institution over a range of observed instrument values. Given choice probabilities that are arbitrarily close to one, quality would be revealed non-parametrically by the corresponding mean selected outcomes, as in Proposition 1. Alternatively, structural assumptions like equation (6) implicitly specify a family of curves that are assumed to generate the population quasi-experimental moments. The minimum distance procedure searches within this family for the curve that best fits the sample moments, accounting for first-step estimation error, and produces a quality estimate equal to curve-of-best-fit's horizontal intercept at one. When the latent variable model is overidentified, in the sense that only a subset of estimable quasi-experimental moments uniquely pin down the true curve, the quality of curve fit will moreover inform an omnibus specification test statistic equal to the efficiently-maximized objective function (7). Thus for any given problem, the roles played by the natural experiment and parametric restrictions can be readily visualized, and alternative restrictions can be imposed without recomputing the sparse causal moments generated by the experiment.[8]

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

[8] This graphical intuition is similar to that of the approach of Brinch et al. (2017), who parameterize conditional marginal treatment effect curves in the binary treatment case; here both the extrapolation and number of instruments needed for identification in the multiple-treatment Roy model are given implicitly by the choice of $F(\cdot)$ and do not depend on the distribution of quasi-experimental controls except through the set of non-parametric instrument propensity scores. As a result, it does not requires estimating the functions $E[Y_i|D_{ij}=1, Z_{i\ell}=1, X_i=x]$ and $E[D_{ij}|Z_{i\ell}=1, X_i=x]$ for each j, ℓ, and value in x in the support of the quasi-experimental controls X_i, as in Brinch

2.4 An Elliptical Parameterization for Binary Potential Outcomes

The minimum distance approach requires a researcher to specify appropriate structural restrictions, trading off flexibility (a $F(\cdot)$ that accommodates a wide variety of potential treatment effects and institutional selection patterns) and tractability (a quality-relevant parameter subvector $\tilde{\theta}_0$ that is uniquely pinned down by a small set of quasi-experimental moments). Here I propose one class of latent variable parameterizations that are likely to strike such a balance, in settings with binary outcomes. These specify a multivariate elliptical copula for the dependence of institutional selection on potential outcomes, with minimal restrictions on marginal substitution patterns. I show that quality is identified in these models provided a researcher has access to as many distinct instrument values as institutions.

The model starts with a latent index parameterization of binary potential outcomes. Suppose

$$Y_{ij} = \mathbf{1}[h_{ij} \geq 0]. \tag{9}$$

for unobserved h_{ij}. For example in the hospital application h_{ij} may denote the latent health of emergency patient i upon admission to hospital j, with patients surviving the first 30 days after admission ($Y_{ij} = 1$) when their health is above some threshold, here normalized to zero. With the vector h_i collecting the J health indices, the observed outcome equation (1) becomes

$$Y_i = \mathbf{1}[h_i' D_i \geq 0]. \tag{10}$$

Note here that 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 ($h_{ij} > h_{ik}$), while for others such a move may result in worse health outcomes.

To structure institutional selection patterns I first impose 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*): $\forall \ell,j,m$, $Pr(U_{ij}(z_\ell) \geq U_{ij}(z_m)) = 1$ or $Pr(U_{ij}(z_\ell) < U_{ij}(z_m)) = 1$.

This assumption restricts the instrument to only monotonically affect institutional selection, in the sense that a change from $Z_i = z_m$ to $Z_i = z_\ell$ that makes selection of institution j strictly more likely for any mass of individuals i cannot make selection of institution j strictly less likely for any other mass of individuals. In the healthcare application, emergency patients are referred to hospitals

et al. (2017), which can be infeasible or difficult in practice and complicate finite-sample inference (Robins and Ritov (1997); Angrist and Hahn (2004); Hirano et al. (2003); Cattaneo et al. (2017).

by ambulance, with $Z_{i\ell}$ indicating the quasi-experimental assignment of ambulance company ℓ to patient i. As suggested by Doyle et al. (2015), differences in ambulance referral preferences may generate variation in hospital admissions given this assignment. Assumption 2 then requires that, to the extent any two ambulance companies have different preferences for referring to a hospital j, they are fixed over different subpopulations of patients. I discuss the appropriateness of this restriction in the next section.

Assumption 2 is valuable in that it significantly reduces the dimensionality of possible latent variable parameterizations. As shown by Heckman and Pinto (2017), monotonicity over multiple unordered treatments is equivalent to an additively-separable model of latent utility,

$$U_{ij}(z_\ell) = \pi_{j\ell} + \eta_{ij}, \tag{11}$$

so that, for example, $Cov(Y_{ij}, U_{ij}(z_\ell)) = Cov(Y_{ij}, U_{ij}(z_m)) = Cov(Y_{ij}, \eta_{ij})$ for any ℓ, m. With the vector π_j collecting the $\pi_{j\ell}$, the generic selection equation (2) can then be written

$$D_{ij} = \mathbf{1}[\pi_j' Z_i + \eta_{ij} \geq \pi_k' Z_i + \eta_{ik}, \forall k] \tag{12}$$

A final parametric assumption defines $F(\cdot)$, along with equations (9) and (11): that the set of latent outcome and utility indices are distributed continuously and joint-elliptically in the population of individuals

Assumption 3 (*Ellipticity*): Given equations (9) and (11), and with the vector η_i collecting the set of η_{ij}, the density of $(h_i', \eta_i')'$ satisfies, for some known positive function $g(\cdot)$,

$$f_{(h_i', \eta_i')'}(t) \propto |S|^{-1/2} g((t-s)' S^{-1} (t-s)), \tag{13}$$

where s is a vector of length $2J$ and S is a $2J \times 2J$ positive-definite matrix.

The set of elliptical distributions satisfying Assumption 3 is quite large, including the familiar multivariate normal (where $g(u) = \exp(-u/2)$), logistic ($g(u) = \frac{\exp(-u)}{(1+\exp(-u))^2}$), and Student's t distributions ($g(u) = (1 + \frac{u}{m})^{-(n+m)/2}$ for some $n, m \in \mathbb{N}^+$), along with less-commonly encountered distributions such as the symmetric multivariate stable, Laplacian, and general hyperbolic distributions. Moreover the restriction on *marginal* selection patterns imposed by Assumptions 2 and 3 is generally without observational loss: as shown in the econometric appendix, any set of positive choice probabilities can be rationalized by equation (12) with elliptical η_i. This flexibility notwithstanding, the following result shows that each elliptical joint distribution of latent variables provides a tractable model for the dependence between potential outcomes and selection that yields quality identification with a relatively small number of quasi-experimental moments:

Proposition 2 (*Semi-parametric quality identification*): Suppose Assumptions 1-3 hold. Then if $L \geq J$ and the $J \times 1$ vectors of choice probabilities are unique across instrument values, quality is identified from a consistent estimator of all quasi-experimental moments.

Proof: See the econometric appendix.

Although a large number of parameters underlie the elliptical model - including the $J \times L$ utility shifters $\pi_{j\ell}$ and the $2J + J(2J+1)$ elements of the s vector and S matrix in Assumption 3, the proof to Proposition 2 shows that all quasi-experimental moments can, under Assumptions 2 and 3, be written as a function of only $J^2 + (J-1)L$ parameter combinations, J of which determine institutional quality. These parameter combinations make up the relevant structural subvector $\tilde{\theta}_0$. Since with Bernoulli outcomes we have at most $(2J-1)L$ linearly-independent moments identified under Assumption 1, this subvector is identified by the minimum distance estimator provided there are $L \geq J$ non-redundant instrument values. That is, as with linear IV models, minimum distance quality identification requires as many instruments $Z_{i\ell}$ as endogenous variables D_{ij}.

To build further intuition for this result, consider a setting with $J = 2$ institutions and a multivariate normal specification for the latent outcome and utility indices:

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

$$D_{i1} = \mathbf{1}[\pi_1' Z_i + \eta_{i1} \geq \pi_2' Z_i + \eta_{i2}] \tag{15}$$

$$D_{i2} = \mathbf{1}[\pi_2' Z_i + \eta_{i2} \geq \pi_1' Z_i + \eta_{i1}], \tag{16}$$

where $(h_{i1}, h_{i2}, \eta_{i1}, \eta_{i2})' \sim N(\mu, \Sigma)$. With L instrument values, there are a total of $2L+14$ underlying parameters in π_1, π_2, μ, and Σ. Nevertheless, each of the $2L$ choice probabilities and mean selected outcomes for institution 1 can be written, respectively,

$$\begin{aligned} Pr(\pi_{1\ell} - \pi_{2\ell} \geq \eta_{i2} - \eta_{i1}) &= \Phi\left(\frac{\pi_{1\ell} - \pi_{2\ell} - (\mu_{\eta 2} - \mu_{\eta 1})}{\sqrt{\sigma^2_{\eta 2} + \sigma^2_{\eta 1} - 2\sigma_{\eta 1 \eta 2}}}\right) \\ &\equiv \Phi(\tilde{\pi}_{1\ell}) \end{aligned} \tag{17}$$

and

$$Pr(h_{i1} \geq 0 \mid \pi_{1\ell} - \pi_{2\ell} \geq \eta_{i2} - \eta_{i1}) = \Phi\left(\frac{\mu_{h1}}{\sigma_{h1}}, \tilde{\pi}_{1\ell}; \rho_1\right) / \Phi(\tilde{\pi}_{1\ell}), \tag{18}$$

which are functions of only $2+L$ combinations of these parameters (here $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\Phi(\cdot, \cdot; \cdot)$ is the standard bivariate normal cumulative distribution function, and $\rho_1 = Corr(h_1, \eta_{i1} - \eta_{i2})$). The order condition for identifying these parameter combi-

nations is thus $2L \geq 2+L$, or $L \geq J$, while the rank condition holds if $\tilde{\pi}_{1\ell} \neq \tilde{\pi}_{1m}$ for $\ell \neq m$. Finally, note that, institution 1's quality is a known function of the identified parameter combinations, since $q_1 = E[Y_{i1}] = Pr(h_{i1} \geq 0) = \Phi(\mu_{h1}/\sigma_{h1})$. This same logic holds for identifying the quality of institution 2, and generalizes for the case of $J > 2$ institutions and other elliptical distributions.

3 Estimating Hospital Quality

3.1 Data and Observational RAMs

I use the preceding econometric framework to estimate the quality of U.S. hospitals according to their effects on short-term patient mortality. Typical hospital RAMs are based on three-year windows of emergency Medicare claims (YNHHSC/CORE, 2013); correspondingly, I draw a sample of 405,172 Medicare fee-for-service beneficiaries, brought to an acute-care hospital by an ambulance for one of 29 emergency conditions in 2010-2012.[9] Observations come from a nationally-representative 20% random 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, admitting condition, 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 patient's hospital; and her subsequent short-term mortality. As in Card et al. (2009), I restrict this sample to patients admitted for a "nondeferrable" primary condition, i.e. those averaging 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 into groups of patients residing in the same hospital service area (HSAs). HSAs are sets of ZIP codes defined by the Dartmouth Atlas of Health Care as regions where patients receive most of their emergency care. For my purposes, HSAs delineate local emergency care markets, within which it is plausible that ambulance company propensity scores (defined below) have full support. A data appendix describes the sample construction in detail.

Table 1 summarizes the distribution of patient conditions, ambulances, hospitals, HSAs, and 30-day survival probabilities in the analysis sample. Hospital RAMs were first developed to measure quality by the short-run mortality of Medicare patients with circulatory and respiratory conditions, such as acute myocardial infarction, heart failure, and pneumonia, though often with an eye towards

[9] Due to data limitations, I am not able to include Veterans Affairs hospitals in this analysis.

extending the measures to a broader patient population (Krumholz et al., 2006).[10] My focus on emergency conditions is driven both by the availability of quasi-experimental ambulance company referral variation and by the traditional focus on patients with an elevated risk of mortality.[11] Panel A of Table 1 shows that circulatory and respiratory conditions make up 42% of emergency admissions in my sample, with the remainder split between digestive (7%), injury (18%), and other (34%) conditions.

Each patient in the analysis sample was referred by one of 9,590 ambulance companies to one of 4,821 hospitals. Panel B of Table 1 shows that the distribution of within-HSA hospital counts is skewed, with around half of all hospitals operating in their own single-hospital market. These markets tend to be small, with an average of only 61 admitted patient observations. In contrast, the remaining 695 multi-hospital HSAs n the sample average 366 admitted patient observations and represent 63% of the analysis sample. Since the ambulance referral design leverages within-market differences in company assignment, causal quality analyses will tend to focus on local comparisons between the 2,357 hospitals in a multi-hospital HSA. Nevertheless, as on average roughly 30% of patients are referred to hospitals outside of their HSA, even in single-hospital HSAs the quasi-experimental variation has scope to detect and correct for non-random sorting. I discuss how this market outside option is treated in the estimation strategy below.

The last column of Table 1 describes average 30-day patient survival, which is the typical horizon for 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 the average survival rate does not vary much by the HSA hospital count. The 25th and 75th percentiles of hospital-specific 30-day survival rates (not reported in the table) are 0.8 and 0.92, respectively.

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

$$Y_{ij} = \mathbf{1}[\alpha_j + \epsilon_i \geq 0], \tag{19}$$

[10] A related quality-based regulatory effort models Medicare patient readmissions (Gupta, 2017; Doyle et al. (2017a)). Since a patient who dies at a low-quality hospital cannot be readmitted, more complex assumptions would be required to causally attribute variation in these outcome to hospital quality. I leave this issue for future work.

[11] 41% of Medicare patients hospitalized for a nondeferrable condition in 2010-2012 were admitted by an ambulance company; comparisons of the analysis sample with this broader group of emergency admissions are reported in columns 1 and 2 of Appendix Table A1 and discussed in the data appendix.

where

$$\epsilon_i = \gamma' W_i - \nu_i \tag{20}$$

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

$$Y_i = \mathbf{1}[\alpha' D_i + \gamma' W_i \geq \nu_i], \tag{21}$$

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 included controls, $\nu_i \perp\!\!\!\perp D_i \mid W_i$. Following YNHHSC/CORE (2013), I parameterize ν_i by an independent logit distribution and obtain quality predictions $\hat{\alpha}_j$ by estimating condition-specific regressions of 30-day survival on hospital effects and patient age, sex, and 17 comorbidity indicators; the data appendix describes this estimation procedure in more detail.

Observational RAMs leave unexplained most of the sample variation in patient mortality. This is illustrated in Figure 2, which plots the ratio of residual to total 30-day survival variance in three condition-specific RAMs. Only around 7% of circulatory and respiratory survival variance is explained by a patient's hospital, admitting condition, and year of admission in the most basic RAM1 specification. This reduction is somewhat smaller for digestive conditions and injuries, and around 14% for other conditions. Relatively more sophisticated specifications, which add patient demographics (RAM2) and both demographics and patient comorbidities (RAM3), account for an additional 4% of circulatory and respiratory survival variance, with similarly modest declines in the other condition categories.

If the residual determinants of patient mortality are completely exogenous to the hospital selection process, predictions from observational RAMs may still provide unbiased measures of hospital quality. However, to the extent survival variance may be further reduced by observable admission correlates, such as a patient's referring ambulance company, observational RAMs are likely to be biased. The econometric appendix formalizes this argument and develops instrument-based tests for RAM unbiasedness, extending earlier methods for validating linear education VAMs (Kane and Staiger, 2008; Chetty et al., 2014a; Deming, 2014; Angrist et al., 2016; Angrist et al., 2017). These results of these tests, summarized in Appendix Table A2, decisively reject the RAM assumption of selection-on-observables, with a p-value of less than 0.001.[12] This suggests pervaisive selection bias in the observational RAMs, as well as scope for leveraging the yet-unused variation in ambulance

[12] I verify that these rejections are not driven by my benchmark RAM specifications by replicating and testing the official 2013 CMS models for AMI, pneumonia, and heart failure patients. RAM bias tests continue to forcefully reject the selection-on-observables null for these models; see the appendices for details.

company referral patterns to both characterize and reduce this bias. I next describe how I make use of this variation.

3.2 Quasi-experimental Estimation

I apply the identification result in Proposition 3 to semi-parametrically estimate the quality of 2,082 hospitals with sufficient quasi-experimental admissions variation. Doyle et al. (2015) and Doyle et al. (2017b) first propose that in regions served by multiple ambulance companies, centralized policies of rotational and simultaneous dispatch generate plausibly-exogenous ambulance company assignment, while subsequent expression of non-random ambulance preferences systematically affect the admissions of otherwise identical patients. Table 2 explores both of these claims empirically by comparing individuals in the analysis sample who live in the same ZIP code but who are assigned to different ambulance companies likely to refer to hospitals with higher and lower RAM predictions. Specifically, I compute the ZIP code distance between each ambulance company's office and each nearby hospital, and label companies as likely to deliver patients to a low- or high-ranked hospital if their closest hospital is in the first or fourth quartile of predictions, in the HSA, from the fullest RAM3 specification. I then regress patient characteristics on either these group indicators (with group means reported in columns 1 and 2) or on the ambulance company's closest hospital's predicted RAM itself (with the regression coefficient reported in column 4), along with a full set of ZIP code fixed effects. P-values for the test that patient characteristics are not systematically correlated with these dimensions of ambulance company heterogeneity are reported in columns 3 and 5.

Patients assigned to ambulance companies located 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 condition (panel A), as well as a host of comorbidity indicators describing their medical history (panel B). This balance of observable characteristics is consistent with the quasi-random assignment of ambulance company indicators $Z_{i\ell}$, conditional on patient location X_i (Assumption 1). Panel C of Table 2 further suggests that ambulance assignment is 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. 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 hospital admission. This is a weaker exclusion restriction than in Doyle et al. (2015), Doyle et al. (2017a), and Doyle et al. (2017b), where ambulances are assumed to only affect outcomes via the treatment intensity or observational RAM of a patient's provider. P-values for the joint test of

balance across all 32 covariates in panels A, B, and C, is 0.98 in column 3 and 0.88 in column 5.[13]

Despite this balance, the first row of Table 2 suggests ambulances do vary systematically in their referral preferences, with patients assigned to ambulances close to highly-ranked hospitals more likely to see a higher-ranked hospital admission. I structure this variation with a first-stage monotonicity restriction (Assumption 2), implying here that differences in referral patterns do not systematically vary by patient heterogeneity. Doyle et al. (2015) defend a similar assumption of monotone referral based on a survey of emergency care technicians, finding that differences in referral patterns across ambulance companies are largely driven by institutional and personal relationships with hospitals rather than patient characteristics. Monotonicity 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 likely a concern. My own interviews with emergency medical staff suggest differential hospital distance may play a particularly strong role in determining referral patterns: ambulance companies may tend towards referring patients to the hospital based closest to their offices in order to minimize excess travel time and maximize local availability.[14]

I derive first-step estimates of hospital choice probabilities and mean selected survival outcomes from a flexible probit specification for ambulance company propensity scores $p_\ell(X_i)$. These model the latent risk of assignment by a cubic polynomial in company-patient distance:

$$E[Z_{i\ell}|X_i] = \Phi\left(\delta_{0\ell} + \sum_k \left(\delta_{1\ell k} d_k(X_i) + \delta_{2\ell k} d_k(X_i)^2 + \delta_{3\ell k} d_k(X_i)^3\right)\right), \tag{22}$$

where k indexes the set of ambulance company instruments and $d_k(x)$ denotes the distance between ambulance company k's home office and a patient located in ZIP code x. Minimum distance quality estimates correct for first-step error in estimating these scores. For robustness I also control for the vector of RAM controls W_i though, consistent with Assumption 1, results are essentially unchanged when these are excluded (see Appendix Table A6).[15]

My benchmark specification uses a multivariate normal model latent health and utility (Assump-

[13]Doyle et al. (2015) 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. They also show that there is no relationship between their ambulance-based instrument and a patient's probability of emergency room admission conditional on ZIP code; see their Figure A1. I find that patient observables are also balanced by assignment to ambulance companies that tend to refer to high- vs. low-survival (instead of RAM prediction) hospitals, with overall joint p-values of 0.81 and 0.93 (instead of 0.98 and 0.88).

[14]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 to their home ZIP code.

[15]Given the sometimes large number of controls and the sometimes small estimation samples, maximum likelihood estimates of equation (22) occasionally fail to converge, I sequentially drop the RAM controls and higher-order distance terms in these samples until convergence is achieved.

tion 3), to form minimum distance estimates of hospital quality.[16] As shown in Appendix Table A6, however, similar results are obtained throughout with a fatter-tailed multivariate Student's $t(2)$ distribution. Quality is identified in both models within HSAs with $L \geq J$ ambulance companies, where J is the number of institutional alternatives. Although most hospitals in the analysis sample are in single-hospital HSAs, patients are sometimes referred outside of their home HSA, yielding scope for non-random selection bias even in small markets. I therefore include the market outside option as a distinct treatment, so that J equals one plus the number of hospitals located in a patients HSA. Finally, to keep the model just-identified and reduce the scope for finite sample bias, I use only the J largest companies for estimation and restrict the analysis to hospitals with at least 25 observed admissions.[17]

Figure 3 summarizes the full set of quasi-experimental estimates by plotting the joint distribution of differences in estimated hospital choice probabilities and mean selected outcomes for each of the 2,082 hospitals with enough quasi-experimental variation. These differences are taken over the two ambulance instruments with the largest and smallest estimated choice probabilities for each hospital; the marginal x-axis distribution thus summarizes one dimension of first-stage variation in institutional choice. The average choice probability difference is 0.33, with more than two-thirds of hospitals seeing a choice probability difference of at least 0.2. This suggests significant first-stage admissions variation throughout the sample. Nevertheless, only 11% of hospitals have a maximal estimated choice probability of at least 0.9, with only 5% exceeding 0.95. The quasi-experimental data are thus likely not rich enough for the non-parametric approach of section 2.3, justifying the extrapolative structure of joint-normality.

Figure 3 also shows that the average associated mean selected outcome difference is negative (-2 percentage points), and grows more negative with the difference in choice probabilities: a simple regression of mean selected outcome gaps on choice probability differences yields a coefficient of -0.06 (0.01), indicated by the solid line in the figure. Intuitively, a negative relationship between mean selected outcomes and choice probabilities tends to suggest that patients are positively selected; that is, average hospital survival tends to decline when patients less likely to select the hospital are admitted. This intuition is exact in the multivariate normal case with $J = 2$, since mean selected

[16]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.

[17]See Cattaneo et al. (2017) for a recent discussion of bias in estimating generalized Roy models with high-dimensional instruments or controls. Appendix Figure A1 plots the distribution of minimum distance first stage F-statistics, testing the 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. The median first-stage F-statistic here is 10.1.

outcomes are then everywhere monotone in choice probabilities.[18] Here it suggests we are likely to find positive Roy selection at the HSA level, which subsequent analyses confirm.

The dashed red curve in Figure 4 plots the distribution of the 2,082 minimum distance estimates of hospital quality indices, defined as $\beta_j = \Phi^{-1}(q_j)$. The mean and standard deviation of these estimates are both around 0.8, with most mass concentrated between $\Phi(0) = 0.5$ and $\Phi(2) \approx 0.98$. Due to the HSA-stratified estimation procedure, this wide dispersion 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 typically non-ignorable estimation error. I next outline an empirical Bayes procedure to account for these different variance components and produce more accurate predictions of hospital quality that combine quasi-experimental and RAM-based estimates.

3.3 Quality Posteriors

Under Assumptions 1-3 I 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, I observe a full set of observational RAM predictions $\hat{\alpha}_j$ from estimates of equation (21) that are likely to be 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 estimate a hierarchical linear model (HLM) to link these two quality measures.[19] This is

$$\begin{aligned} \hat{\beta}_j &= \beta_j + \iota_j \\ &= \kappa + \lambda\hat{\alpha}_j + \mu_{h(j)} + \upsilon_j + \iota_j, \end{aligned} \tag{23}$$

where $\kappa + \lambda E[\hat{\alpha}_j] = E[\beta_j]$ is the average hospital quality index (per Figure 4, around 0.8), $\mu_{h(j)}$ is a random effect for the HSA $h(j)$ of hospital j, υ_j is the residual quality index of hospital j, and ι_j is a mean-zero estimation error term. The HSA random effects, assumed to be 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

[18] In terms of section 2.4's notation, $Pr(h_{i1} \geq 0 \mid \tilde{\pi}_{1\ell} \geq \tilde{\eta}_i) = \int_{-\infty}^{\tilde{\pi}_{1\ell}} \Phi\left((\mu_{h1}/\sigma_{h1} - \rho_1 t)/\sqrt{1-\rho_1^2}\right) \frac{\phi(t)}{\Phi(\tilde{\pi}_{1\ell})} dt$ and the sign of the derivative of this mean selected outcome with respect to $\tilde{\pi}_{1\ell}$ can be shown to be the same as the sign of $-\rho_1 = Corr(h_1, \tilde{\eta}_i)$. With $\tilde{\pi}_{1\ell}$ everywhere increasing in the corresponding choice probability, mean selected outcomes are thus decreasing if $\rho_1 > 0$, or if patients with higher utility for hospital 1 tend to have better health outcomes there.

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

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 maximum likelihood (MLE) procedure that effectively weights observations with differing degrees of minimum distance estimation error.

Table 3 reports OLS and MLE estimates of equation (23), where for ease of interpretation $\hat{\alpha}_j$ is normalized to be of zero mean and unit standard deviation and $\hat{\beta}_j$ is also de-meaned. Columns 1 and 2 regress quality index estimates on RAM1 predictions which, as in Figure 2, only control for patient diagnosis and year of admission, while columns 3 and 4 includes patient demographics in the RAM2 specification. Columns 5 and 6 report OLS and MLE hyperparameter estimates using the richest RAM3 specification, which also controls for patient comorbidities. The predictions from all three models are found to be predictive of the quality index estimates, with $\hat{\lambda} = 0.12 - 0.13$, though the OLS estimates are far from statistically significant due to the relative imprecision of the equal-weighted regression. Efficiently-weighted MLE estimates reduce the standard error on $\hat{\lambda}$ from around 0.12 to around 0.01 without much change in the estimates. Consistent with the graphical evidence in Figure 2 and the formal tests in Figure 3, including demographics and comorbidities barely increases the correlation of RAM with true quality. Overall, the HLM's variance decomposition suggests that around 50% of the national variation in quality indices β_j is found between HSAs, with 30% explained by observational RAM predictions and 20% left unexplained within HSAs. The significantly positive relationship between observational measures and true hospital quality is consistent with the linear IV estimates of Doyle et al. (2017a).

I take column 6 of Table 3 as my preferred estimate of equation (23), though I also check sensitivity to other HLM specifications. Column 7, for example, includes the predictions of all three RAMs simultaneously, showing that the most sophisticated RAM3 predictions remains highly predictive conditional on the other two. Columns 8 and 9, in turn, test for nonlinearities in the relationship between observational RAM and quality indices by including a cubic polynomial in RAM3 predictions and interactions with the HSA hospital counts. Estimated coefficients on the added terms are small and not statistically significant at conventional levels while the overall residual variance falls little, suggesting that the parsimonious specification of (23) is a reasonable approximation to the true conditional expectation.

I use the HLM hyperparameter estimates to generate empirical Bayes posterior predictions of hospital quality that, as in Angrist et al. (2017), Chetty and Hendren (2017), and Finkelstein et al. (2017), shrink consistent but noisy quasi-experimental estimates of institutional quality towards precise, but likely biased, observational predictions. The random-effects structure of equation (23)

further allows the vector of estimates for each HSA to be jointly shrunk towards a HSA-specific mean, thereby accounting for the local correlation in hospital quality found in Table 3. Specifically, 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

$$E[\beta_h|\hat{\alpha}_h, \hat{\beta}_h] = \Omega_h\hat{\beta}_h + (I_{J(h)} - \Omega_h)(\kappa + \lambda\hat{\alpha}_h) \tag{24}$$

$$Var(\beta_h|\hat{\alpha}_h, \hat{\beta}_h) = (I_{J(h)} - \Omega_h)(\phi^2 I_{J(h)} + \sigma^2), \tag{25}$$

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

$$\Omega_h = (\phi^2 I_{J(h)} + \sigma^2)(\phi^2 I_{J(h)} + \sigma^2 + \Xi_h)^{-1}. \tag{26}$$

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 in proportion to one-minus the quality index 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 then given by

$$\begin{aligned} E[q_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] &= E[\Phi(\beta_j)|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] \\ &= \Phi\left(\frac{E[\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}]}{\sqrt{1 + Var(\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)})}}\right) \end{aligned} \tag{27}$$

since β_j is normally-distributed conditional on $\hat{\alpha}_{h(j)}$ and $\hat{\beta}_{h(j)}$ by equation (23).[20]

I construct hospital quality posteriors using these formulas and the MLE hyperparameter estimates in column 6 of Table 3. The solid blue line in Figure 4 shows the resulting distribution of quality index posteriors for the 2,082 hospitals with first-step estimates, while Appendix Figure A2 plots the full distribution of quality posteriors and observed survival rates. As expected, the posterior mean distribution is much tighter than the estimate distribution, reflecting empirical Bayes shrinkage and theoretically-improved mean squared prediction error. One-minus the mean quality index signal-to-noise ratio, which gives a rough measure of the typical shrinkage factor, is 0.77 with a standard deviation of 0.23. Correspondingly, the empirical Bayes procedure reduces the standard deviation of quality index predictions significantly, from 0.79 to 0.14.

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

Importantly, equation (23) also produces posterior quality predictions for hospitals without a first-step quality estimate due to small size or insufficient ambulance instruments. These are the HLM fitted values, plotted by a dotted green distribution curve in Figure 4, which uses the population relationship between observational RAM and hospital quality to extrapolate to underidentified regions. This extrapolation is valid when equation (23) describes the relationship between quality indices and observational RAM across all hospitals, whether or not they have enough quasi-experimental data to be included in estimation. Appendix Table A1 shows that the average characteristics of patients and hospitals in HSAs with and without minimum distance estimates are quite similar, while Table A6 shows that all main results continue to hold or are strengthened when the HLM includes interactions with the HSA's hospital count, the main driver of minimum distance estimate availability. I next describe these main results 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 empirical Bayes posterior predictions of hospital quality 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 selection bias in two quality-based policies currently in place in U.S. healthcare markets.

4.1 Correlates of Hospital Quality

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 quality posteriors and other outcome measures on hospital characteristics and HSA fixed effects, within the set of 695 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 annual spending on the hospital's emergency Medicare patients over 2010-2012; and log emergency Medicare patient volume over the same period. As shown in Table A1, most hospitals in the sample (61%) operate as private non-profits, with 18% and 21% registered as for-profit and government-run hospitals respectively. 22% of providers are categorized as teaching hospitals, and average log patient spending and volume come to 9.3 and 3.5.

Columns 1-3 of Table 4 summarize regressions of observed hospital survival rates, while columns 4-6, 7-9, and 10-12 regress hospital RAM predictions, minimum distance quality index estimates, and

quality posteriors, all normalized to standard deviation units for comparability.[21] While the survival rate coefficient estimates are small and mostly insignificant, a consistent pattern appears from the less-biased quality measures in columns 4-12: government-run hospitals tend to be of statistically-significantly lower quality on average (panel A), while higher-spending and higher-volume hospitals tend to be of higher quality (panel B). I do not find statistically-significant differences between for-profit and non-profit hospitals, nor any significant correlation with teaching status, though the associated standard errors are sometimes large. Notably, the minimum distance quality estimates generate very similar correlations as the RAM predictions and quality posteriors (which are partly influenced by the RAM), though they are much less precise.

Rescaling columns 10-12, the estimates suggest that moving a typical patient from a government-run hospital to a local privately-owned hospital increases her expected 30-day survival by 0.4 percentage points on average, while moving patients between providers with a one standard deviation difference in log volume or log average spending affects short-run mortality by around 0.2 and 0.1 percentage points, respectively. Log spending and volume are positively correlated in this sample ($\rho = 0.44$), however, and column 12 shows that there is no significant effect on expected mortality from moving patients to hospitals with different spending levels holding volume fixed. Moving representative patients from non-teaching to teaching hospitals is predicted to decrease expected 30-day survival by 0.4 percentage points, though again this estimate is not statistically significant.

Qualitatively, the findings in 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] This is perhaps unsurprising, as similar conclusions can be drawn just from the observational RAM estimates in columns 4-6. As a quantitative benchmark, Ruhnke et al. (2011) estimate an average decline in the 30-day mortality of Medicare patients with pneumonia of around 3.4 percentage points over the nearly 30 years between 1987 and 2005 due to technological advances. The predicted survival effects above are significant in this historical context.

Appendix Tables A3 and A4 explore other dimensions of hospital quality, including its correlation over time, across different emergency conditions, and with measured quality inputs such as average staff salary, the use of electronic records or case management, and the breadth of accreditation or types of imaging technologies. Quality indices are estimated to be quite persistent, with a typical

[21] For consistency, here and throughout I use empirical Bayes posterior survival rates that shrink observed rates towards the grand mean in proportion to one-minus the signal to noise ratio. This matters little for all results, as the typical observed survival rate is quite precisely estimated.

[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.

three-year lag correlation of around 0.7 and slightly decaying autocorrelations over longer horizons. There is also a significantly positive correlation of the quality-of-care across patients with different conditions, especially respiratory and circulatory diagnoses ($\rho = 0.64$). Table A4 moreover shows that all measured inputs appear significantly correlated with quality posteriors, with staff salary remaining the most predictive in a multivariate regression with all inputs and hospital volume. It is worth emphasizing that these estimates, as well as those in Table 4, are causal in a limited sense, as they do not reveal the effect of a hospital changing its ownership structure, increasing its volume, or paying its staff more. Nevertheless, these supplementary analyses support the claim that the quality posteriors capture intuitive and persistent aspects of true hospital productivity.

4.2 Patient Sorting and Selection Bias

Comparing quality posteriors $E[Y_{ij}]$ and observed survival rates $E[Y_{ij}|D_{ij} = 1]$ captures outcome-relevant sorting patterns and the degree of selection bias affecting observational hospital quality measures. Figure 5 shows that quality and survival are positively though imperfectly correlated within multi-hospital HSAs, with $\rho = 0.64$. Moreover, the figure displays a clear pattern of negative correlation between quality and selection bias, defined again as $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$. Points above the dashed 45 degree line represent hospitals with relatively higher selection bias, which tend to be those of lower relative quality, while points below the solid line less positively selected than average while tending to be of relatively higher quality. Overall, I find a rank correlation of quality and bias posteriors of -0.81. This negative correlation suggests that hospitals that are of higher quality tend to treat sicker-than-average patients, and implies that reductions in selection bias may not dramatically alter the performance rankings of hospitals, despite a meaningful distribution of bias with a standard deviation of 2.8 percentage points. Indeed, RAM predictions and quality posteriors in this sample have a correlation of around 0.9. Interestingly Angrist et al. (2017) also find a negative correlation between school quality and selection bias in their study of Boston middle schools, along with a correlation between conventional VAM predictions and quasi-experimental quality posteriors of 0.85.

Unlike in the constant effects model of Angrist et al. (2017), however, the generalized Roy (1951) framework deployed here provides another way to characterize institutional selection: the extent to which patient sorting exploits hospital comparative advantage by favoring more appropriate hospitals (i.e., selection-on-gains). To explore this, Figure 6 plots the distribution of volume-weighted average selection bias posteriors for the set of 695 multi-hospital HSAs. Under constant treatment effects, HSA-level selection bias equals zero by construction: any volume-preserving transfer of patients across hospitals trades off equal gains and losses across different subpopulations. The

wide distribution in bias posteriors plotted in Figure 6, however, suggests a large degree of hospital specialization and selection on match-specific quality. Furthermore, the vast majority of HSAs (88%) appear to have positive average selection bias, and the average HSA-level selection bias posterior is 3.6 percentage points. This suggests that a typical patient is more likely to survive at her selected hospital than at a hospital picked at random from most healthcare markets; that is, patients appear to benefit significantly from positive selection-on-gains.

What explains this striking pattern of positive Roy selection? Differential hospital distance is among the most obvious candidate, since individuals suffering from an acute emergency may only survive if brought to the closest available emergency room, a hospital characteristic that varies trivially over patients. On the other hand, although distance is an obvious predictor of hospital admissions, it is found to be a weak one empirically. Chandra et al. (2015), for example, find that only half of emergency patients in their sample are admitted to their nearest hospital. Similarly, in my sample there is significant variation in hospital "distance bias," defined as $E[d_{ij}|D_{ij} = 1] - E[d_{ij}]$, where d_{ij} denotes the ZIP code distance between patient i and hospital j. The volume-weighted HSA-level mean of this measure is -0.91 across the 695 multi-hospital HSAs, reflecting that on average patients are admitted to hospitals around 0.91 miles closer to them than the typical hospital in their market. However, this average varies widely, with patients in some regions sorting to hospitals no more than 0.1 miles closer to them than a provider picked at random.

Table 5 uses the HSA-level survival and distance bias measures to examine the extent to which selection-on-distance explains the overall estimated 3.6 percentage-point survival benefit from selection-on-gains. Panel A regresses survival bias on flexible polynomials in distance bias and indeed finds a strong correlation, with a marginal "effect" of around a 0.1-0.25 percentage point decline in average HSA-level selection bias per mile of increased average HSA-level distance bias. Nevertheless, the constant in even the most flexible cubic regression in column 3, representing average outcome selection bias in a HSA with zero selection-on-distance, remains a significantly positive 3.4 percentage points. This suggests selection-on-distance explains only around 5% of the benefit from Roy selection. 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 3 percentage points more likely to survive at their chosen hospitals then with random admission (and 87% of these HSAs have positive average bias posteriors). Thus selection on hospital distance appears to explain only a small fraction of hospital Roy selection.[23]

[23]Appendix Table A6 also shows that these findings are also not driven by the tail behavior of the multivariate

Similar analyses in Appendix Table A5 suggest this is also the case for other observables, including patient diagnosis, demographics, and pre-existing conditions. Average HSA-level selection bias controlling for the HSA-level observable "distance" between admitted patients and average patients in the HSA range from 3.1 to 3.5 percentage points, suggesting observables explain at most 13% of the estimated benefit of positive Roy selection (here distance for non-geographic observables is computed by the Mahalanobis metric). Together, these results indeed suggest that hospitals specialize in ways that are unobserved to the econometrician (consistent with Chandra and Staiger (2007) and Chandra and Staiger (2017)) but that are observed and selected on by ambulance companies and patients. Accommodating these unobservables with the heterogenous-effects multivariate probit specification – which are ruled out by other models such as the linear IV specification of Angrist et al. (2017) or the fixed-coefficient probits of conventional RAMs and Geweke et al. (2003) – thus appears empirically important in this setting.

4.3 Policy Consequences of Selection Bias

Non-random hospital choice generates a sizable distribution of posterior selection bias, with a overall 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: 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 Angrist et al. (2017), policy decisions based on biased quality rankings may still generate large social gains. Furthermore, the overall negative quality-bias correlation means that policies that reward or punish hospitals according to observational RAM rankings are most likely to understate true quality differentials rather than drastically change the types of hospitals that are subsidized. 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 CMS' hospital ranks were based on the estimated quality posteriors. 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

normal distribution assumed for latent health and utility. A multivariate Student's $t(2)$ specification finds that 84% of HSAs exhibit positive Roy selection, with an average HSA-level bias posterior of 3.3 percentage points.

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. This round of VBP payments affected only a small share of a hospital's reimbursements: the average VBP repayment was 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 has announced plans to tie 90% of all traditional Medicare payments to quality programs like VBP by 2018 (DHHS, 2015; Gupta, 2017). In recent work Norton et al. (2017) show that hospitals indeed respond to the program's seemingly modest incentives, with providers that face higher marginal VBP returns seen to improve their respective TPS components in subsequent years.

I first replicate the FY2014 VBP payment schedule to simulate reimbursements adjustments based on conventional RAM 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. Following the CMS methodology, I compute points from hospital risk-standardized mortality rates, defined with the notation of equation (21) as

$$RSMR_j = \frac{1 - \sum_{i:D_{ij}=1} F_\nu(\hat{\alpha}_j + \hat{\gamma}' W_i)}{1 - \sum_{i:D_{ij}=1} F_\nu(\bar{\alpha} + \hat{\gamma}' W_i)} (1 - \bar{Y}), \tag{28}$$

where F_ν is the logit cumulative distribution function, $\hat{\gamma}$ 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 repaid none of their DRG withholdings if they scored the minimum level across all three quality domains and linearly accrued repayments with increases in the TPS. I hold the non-outcome domains and DRG totals fixed at their true FY2014 values, 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 repayment rates with those that would be produced with TPS based solely on the hospital quality posteriors. The data appendix describes the construction of simulated repayments in more detail.

The resulting distribution of differences in simulated VBP repayment rates is shown in Figure

7. Ten percent of hospitals see no difference in VBP reimbursement when quality posteriors are used in place of the current rankings; the average change is -0.1 percentage points, with a standard deviation of 0.81, and the distribution is right-skewed. Most hospitals (74%), moreover, see a repayment rate change of less than one percentage point in magnitude (recall that the withholding rate is 1.25%). Regression estimates in Table 6, moreover, show that both the benchmark VBP rates and the simulated changes from incorporating the quasi-experimental data are not correlated with most of hospital observables. Non-profit, for-profit, government-run, and teaching hospitals are all subsidized roughly the same, though the difference of -0.07 percentage points for government-run hospitals is statistically significant. Similarly while higher-spending and lower-volume hospitals tend to be subsidized more by conventional VBP, in practical terms these subsidizes are quite small. Reimbursing on the basis of quality posteriors leads similarly-small differentials, though teaching hospitals are now slightly penalized and higher-spending and volume hospitals moderately subsidized. Overall, the relatively modest extent of changes to quality-based reimbursement are consistent with the negative quality-bias correlation; as with the RAM predictions and quality posteriors, benchmark VBP and quality-based achievement points are highly correlated with $\rho \approx 0.9$. The scope for selection bias to materially change the distribution of policy winners and losers thus appears limited.

Patient Guidance

Along with hospital incentives, supervisory quality rankings have begun to shape patient admission decisions. The federal Hospital Compare website was launched in 2005 to help consumers make informed decisions about their inpatient options via 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 observational 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, particularly as they become more influential for patients and ambulance companies and as they may affect, either directly or indirectly other sorts of regulatory policies influencing hospital popularity or capacity.

The hyperparameter estimates in Table 3 suggest that redirecting a typical patient from a random hospital to the provider with the highest RAM ranking in her HSA is likely to increase 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 and pervasive extent of positive Roy selection-on-gains in Figure 6 suggests these gains may be at least partly 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, so that policies that succeed in redirecting individuals to the best-on-average hospital may actually do harm.

I quantify the net effect of these competing forces by simulating 500 sets of quality indices β_j from the HLM estimates of Table 3, column 6, holding the set of observational RAM predictions fixed. I next draw estimation error components ι_j among the hospitals with minimum distance quality estimates, based on the empirical distribution of minimum distance estimation error, and construct simulated estimates and posteriors as in the full sample. From these data, I compute the average 30-day survival rates for a typical patient admitted to either a random hospital within her HSA, the hospital in her HSA with the highest survival rate, the local hospital ranked best by one of the RAM models summarized in Figure 2, or the hospital with the highest quality posterior. While abstracting away from general equilibrium effects and capacity constraints, these estimates give an expected public health value measure of using various supervisory quality rankings to redirect a representative patient's admissions.

As expected, each ranking-based admission policy increases expected survival relative to a benchmark of random admissions. An emergency patient sent to the lowest-mortality local hospital is on average 2.1 percentage points more likely to survive her first 30 days after admission, while basing admissions on even the most naive RAM1 model increases this to 3.1 percentage points. As we've seen before in Figures 2, Table 3, and Appendix Table A2, adding patient demographics and comorbidities to the RAM specification only incrementally reduces bias, bringing the expected gain from the richest RAM3 model to 3.2 percentage points. Incorporating quasi-experimental data improves things further still, with a representative patient admitting according to the quality posteriors seeing a 3.6 percentage point gain in expected survival, relative to random admissions. These gains are on the order of the 30-year technological decline in pneumonia mortality rates estimated by Ruhnke et al. (2011).

Nevertheless, Roy selection dominates many of these polices. In this sample of quasi-experimental HSAs, the average gain in 30-day survival from prevailing selection patterns is 3.4 percentage points. Admission policy effects net of this gain are plotted in Figure 8, revealing that all but the policies based on quality posteriors lead on average to reduced expected survival rates. In particular, redirecting a random patient from her preferred hospital to the one with the highest RAM3 rank is found to reduce survival by -0.2 percentage points. Quality posterior-based redirection still increases expected survival, though by a diminished 0.2 percentage points. For completeness, I also simulate

gains from an infeasible policy in which all of the quasi-experimental moments used to construct minimum distance quality estimates are assumed to be known without error. Redirecting patients based on true quality has significant expected survival effects of roughly 1.1 percentage points. Thus, using less-biased hospital rankings to guide admissions may deliver meaningful partial-equilibrium health benefits, particularly when rankings are estimated on larger administrative datasets or by more efficient quasi-experimental methods. At the same time, policies that make patients or ambulance companies more likely to select high-ranked hospitals, as well as policies that close or limit the growth of low-ranked providers, may also undermine the prevailing health benefits of hospital selection-on-gains and have unintended negative consequences on patient mortality.

5 Conclusions

Policymakers in many settings increasingly 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. Instead, this paper develops a flexible framework for measuring institutional productivity and selection bias with quasi-experimental data. Quality can be non-parametrically estimated from rich instrument variation, while distributional restrictions may substitute for traditional constant effects assumptions in order to extrapolate from narrower quasi-experiments. Rather than relying on costly likelihood-based estimation, a tractable minimum distance procedure implements this semi-parametric approach. I establish identification of a large class of elliptical models for binary potential outcomes which flexibly allow for both institutional comparative advantage and Roy-style selection-on-gains, features lacking in existing estimation frameworks.

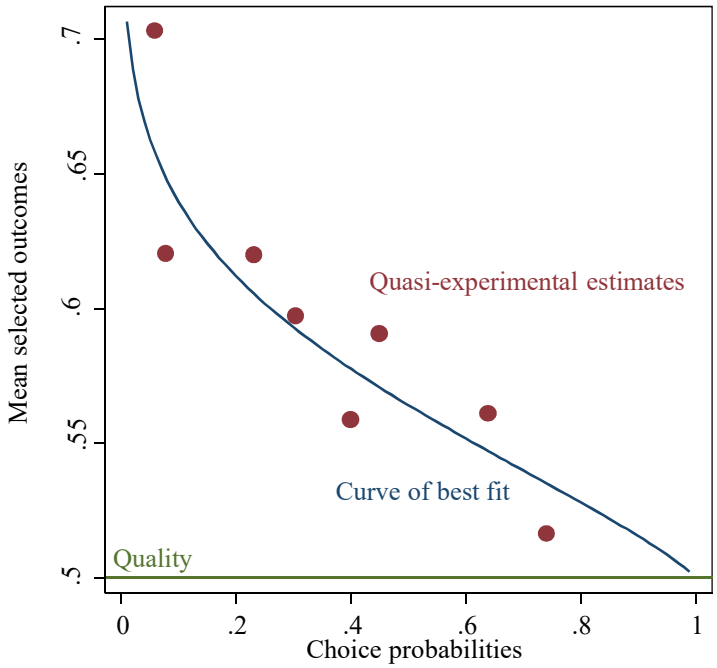
These features appear to be highly relevant in the application to hospital quality. I find a large degree of hospital specialization and unobserved positive Roy selection-on-gains, with typical patients in most markets benefiting from being admitted to more appropriate hospitals. This non-random sorting across hospitals generates pervasive selection bias, though this bias tends to be negatively correlated with true hospital quality. Higher spending, higher volume, and privately owned hospitals appear to be of higher average quality, and only a small share of the Roy selection appears correlated with observables such as differential hospital distance. Observational risk-adjustment methods remove some selection bias, and policies based on less-biased quasi-experimental quality rankings for regulatory policy may lead to modest survival gains for patients without significantly altering the distribution of performance-linked Medicare reimbursements.

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 are at best weakly correlated with bias in VBP rankings, there may remain various hospital-controlled unobservables that correlate with 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, while basing quality on "upstream" instrumental variable variation rather than admissions itself may reduce the ability of hospitals to game.

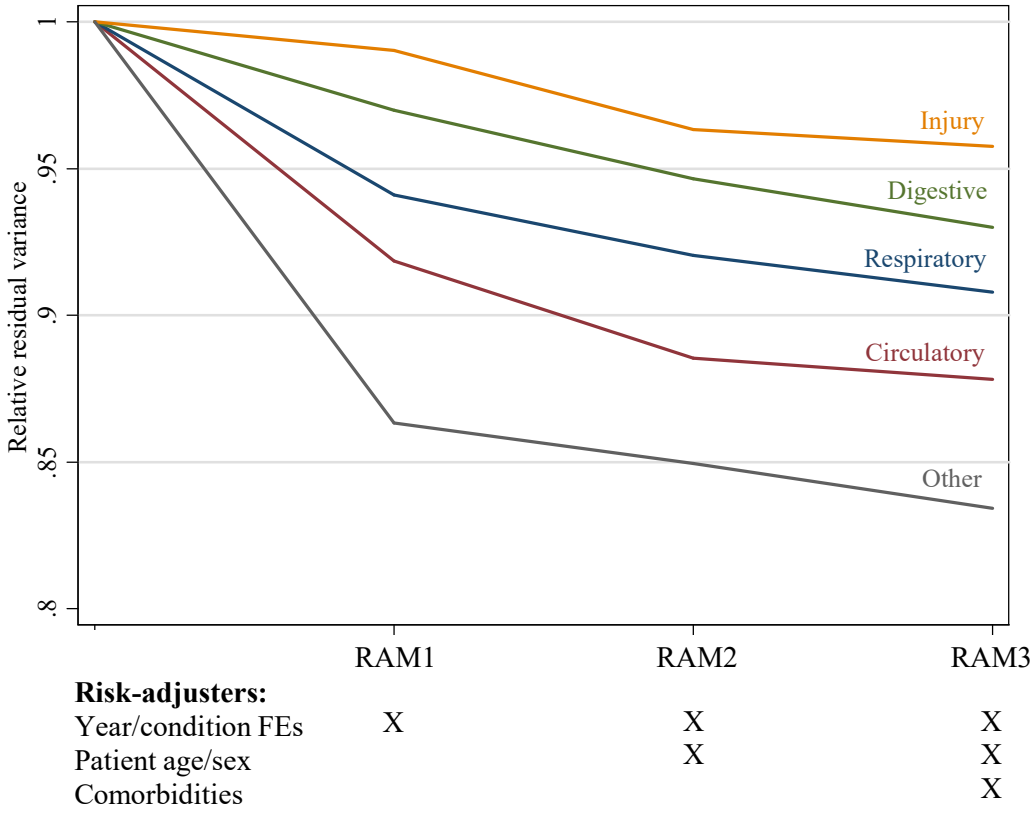
The policy simulations also raise new questions about the efficacy of demand-side quality interventions, including the large and growing set of hospital report cards currently accessible to patients. Without significant hospital specialization, the finding that higher-ranked hospitals tend to attract 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 approach. 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 quality measures actually affect admission decisions and characterizing the optimal design of public quality signals in settings with meaningful Roy selection are two important goals raised by the heterogeneous-effects approach.

Figure 1: Semi-parametric Quality Estimation



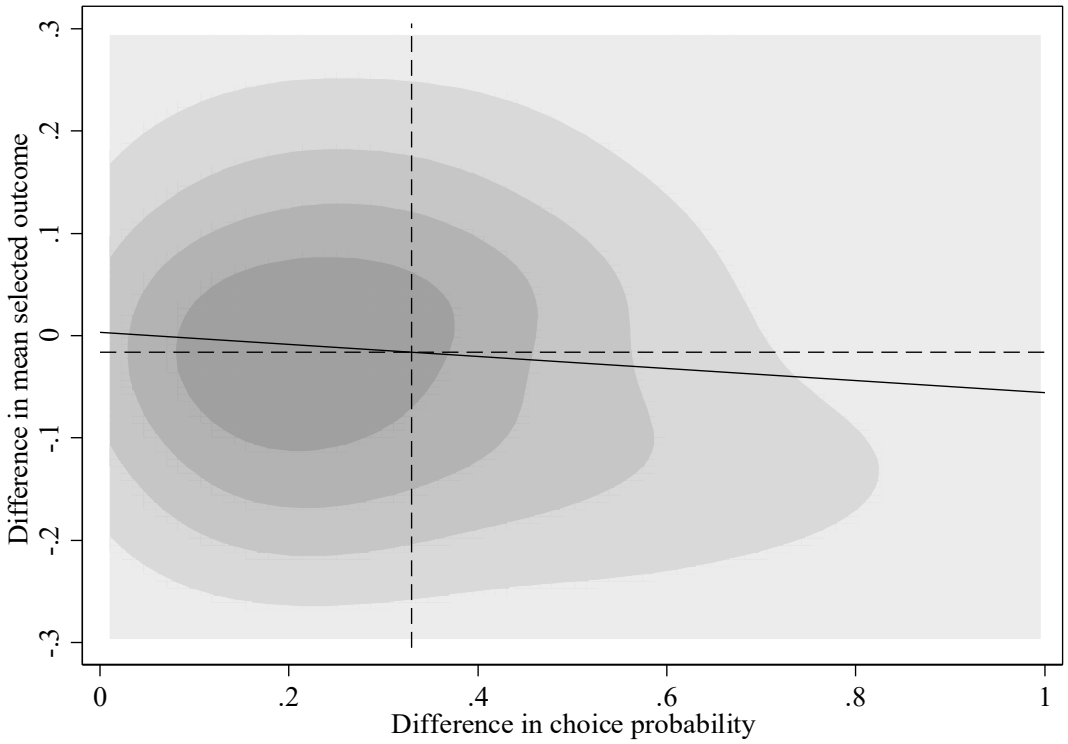
Notes: This figure illustrates mean selected outcome and choice probability estimates for a single institution across eight instrument values, simulated from the two-institution multivariate normal model described in Section 2.4. The minimum distance curve of best fit intersects the institution's quality at a choice probability of one.

Figure 2: Residual survival variance in observational RAMs



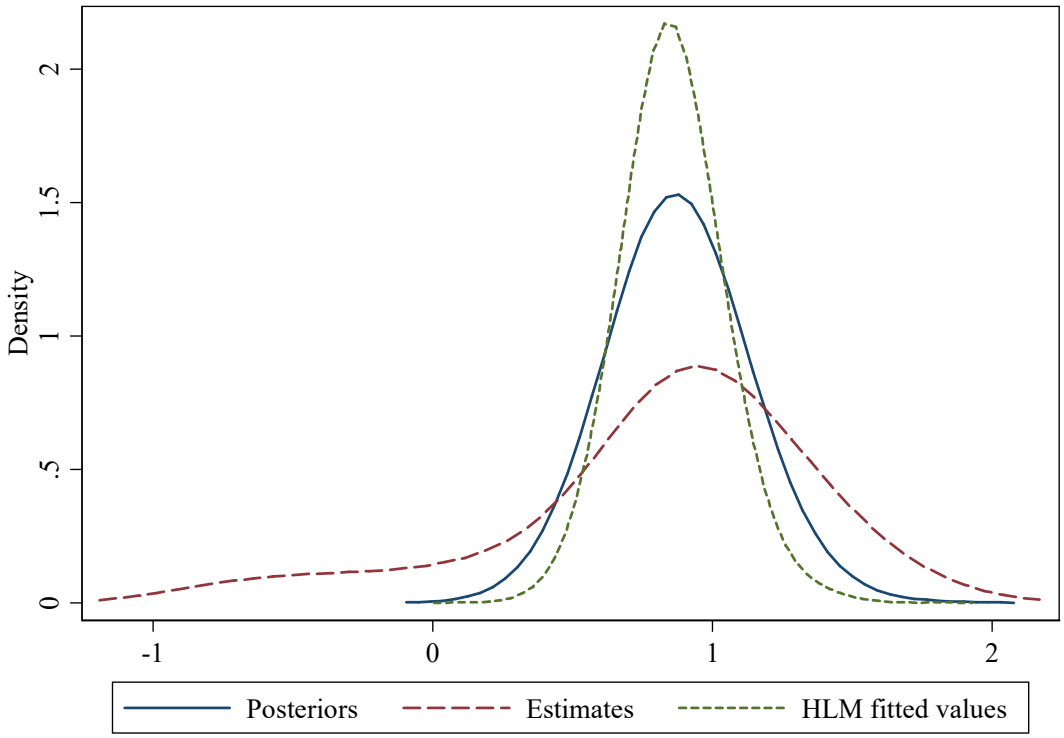
Notes: This figure plots the variance of risk-adjusted 30-day survival relative to the unadjusted survival variance for three risk-adjustment models, estimated separately by condition. See Table 1 for a description of each condition category, Table 2 for a list of included comorbidities, and the data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

Figure 3: The joint distribution of quasi-experimental hospital moments



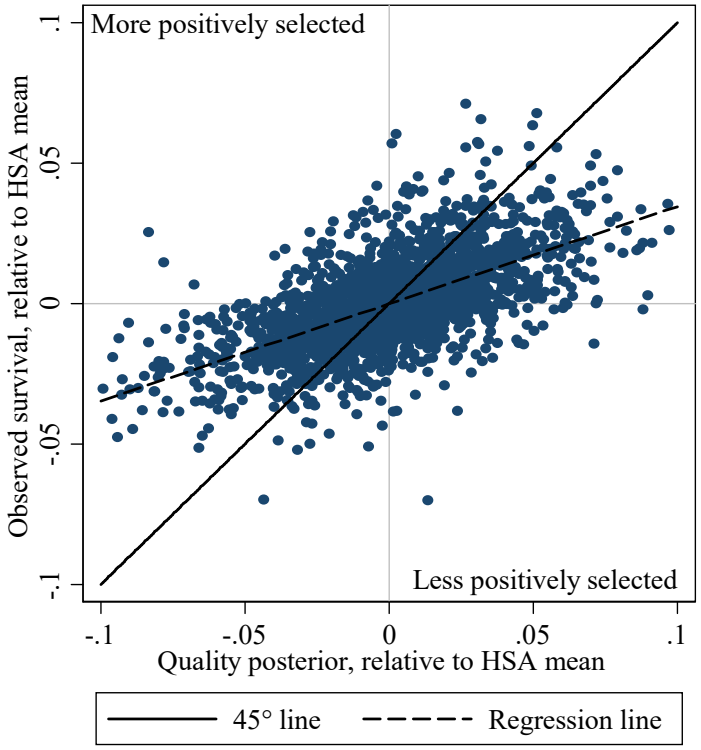
Notes: This figure plots a Gaussian kernel density estimate of the joint distribution of estimated mean selected outcome differences and estimated choice probability differences for the 2,082 hospitals with minimum distance quality estimates. Differences are taken across the two ambulance companies with the largest and smallest estimated choice probability for each hospital. The vertical and horizontal bandwidths used to estimate this distribution are 0.05 and 0.1, respectively. Dashed lines indicate sample means, and the solid line indicates the regression of mean selected outcome differences on choice probability differences.

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



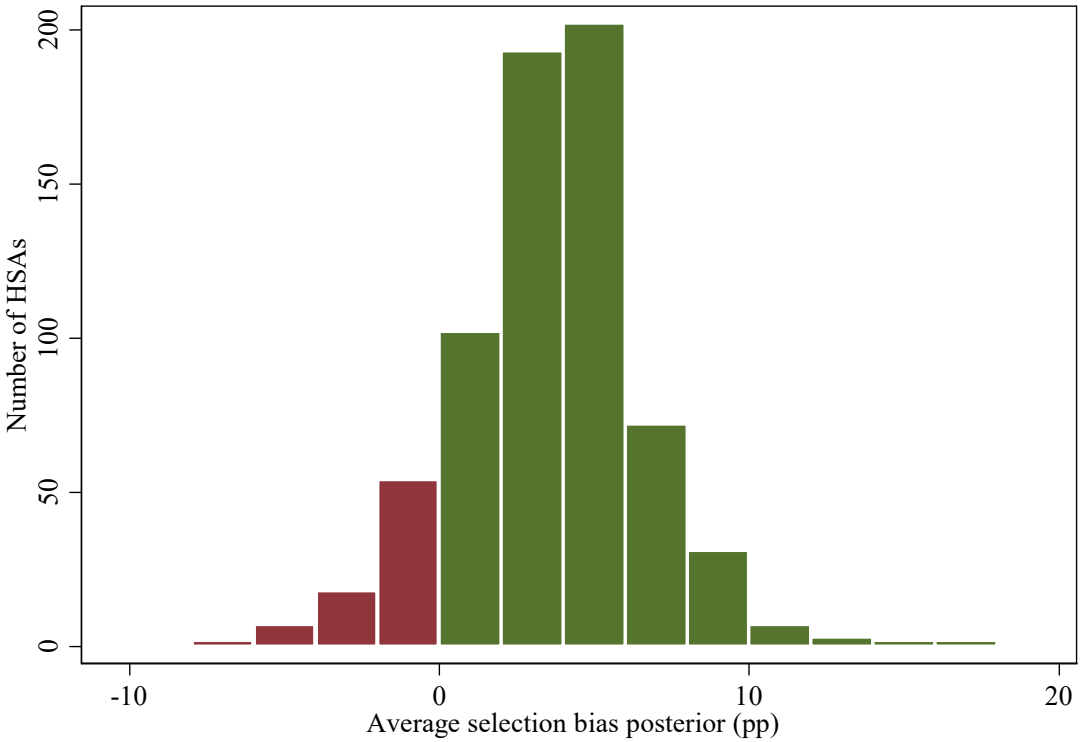
Notes: This figure plots Gaussian kernel density estimates of the distribution of minimum distance hospital quality index estimates and empirical Bayes posteriors, along with fitted values from the hierarchical linear model's projection on conventional RAM predictions. The sample includes 2,082 hospitals with a first-step quality estimate. The bandwidth used to estimate each distribution is 0.2.

Figure 5: Within-HSA variation in hospital quality and 30-day survival



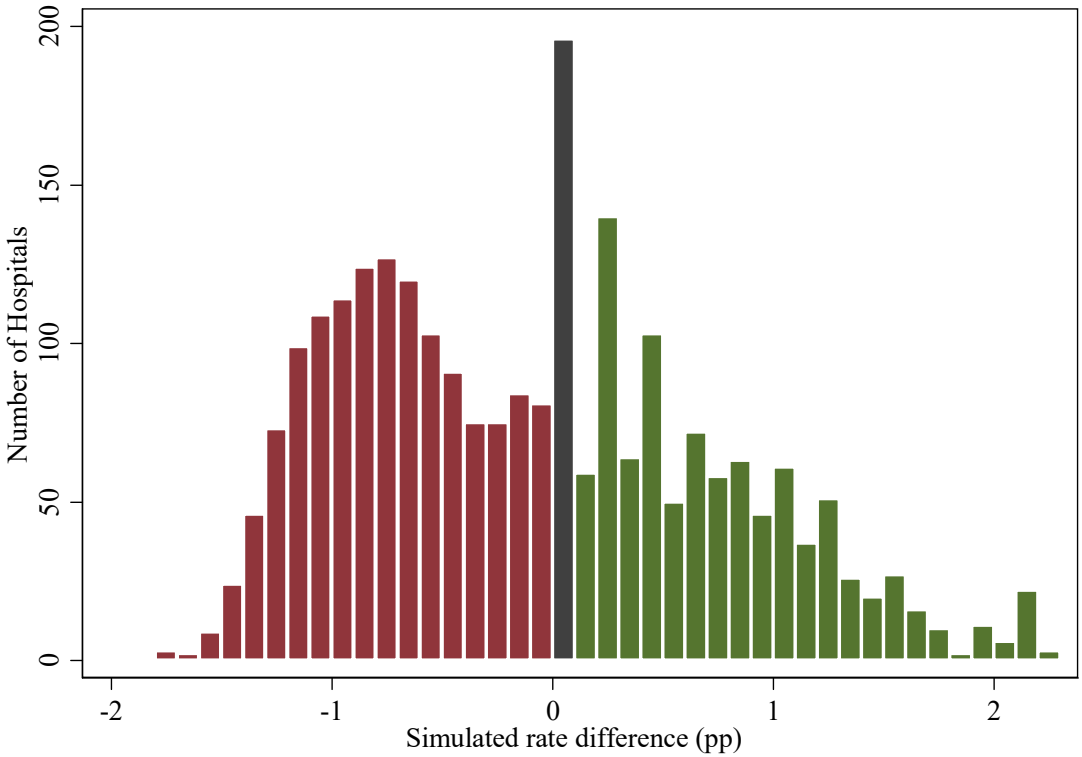
Notes: This figure plots posterior hospital 30-day survival rates against posterior hospital quality, both net of their HSA means. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs.

Figure 6: The distribution of HSA-level selection bias



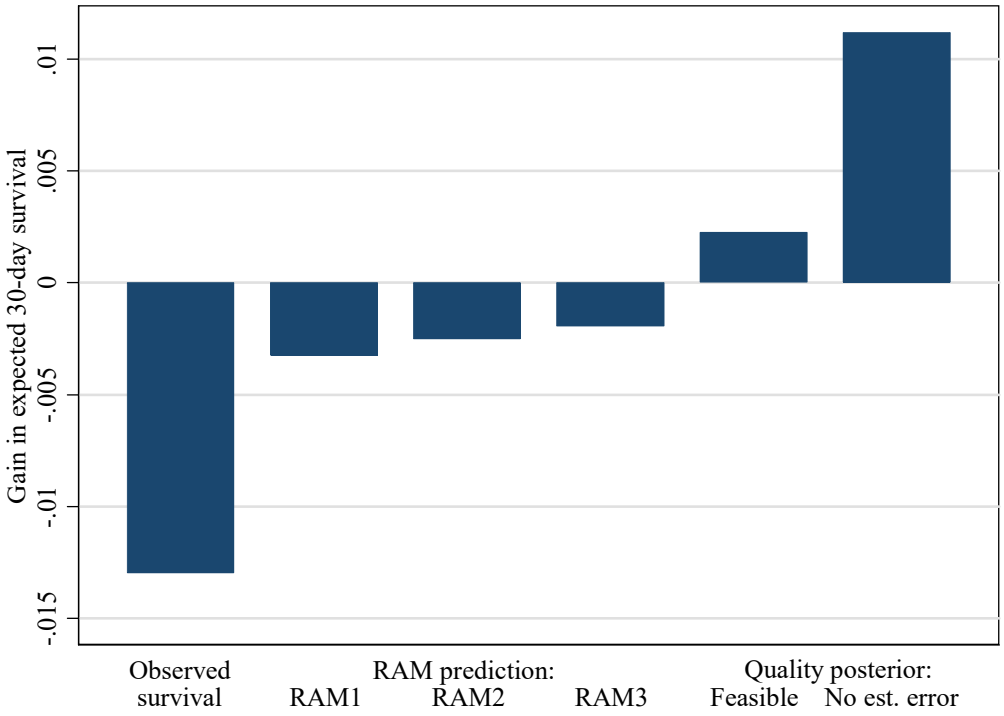
Notes: This figure plots the distribution of volume-weighted average posterior selection bias across 695 multi-hospital HSAs. HSAs with negative selection bias would see higher average 30-day survival if patients were randomly allocated to hospitals, while a positively-selected HSA would have a lower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

Figure 7: The distribution of Value-Based Purchasing repayment rate differences



Notes: This figure plots the distribution of the change in simulated VBP repayment rates when quality posteriors replace benchmark hospital rankings. See the text and data appendix for details.

Figure 8: Expected survival gains from redirecting patients to a top-ranked hospital



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 her current choice, according to either the hospital's 30-day survival rate, observational RAM prediction, or quality posterior (with or without estimation error in the quasi-experimental moments). See Table A2 for a description of the RAM specifications. Estimates are from 500 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	405,172	9,590	4,821	3,159	0.833
			A. By patient's condition			
Circulatory	5	89,076	7,576	3,879	2,777	0.807
Respiratory	4	81,021	7,434	4,224	2,980	0.781
Digestive	6	26,358	5,242	3,323	2,354	0.902
Injury	8	71,616	7,399	3,634	2,561	0.931
All other	6	137,101	8,063	4,441	2,997	0.815
			B. By patient's admitted HSA's hospital count			
One	29	151,071	6,760	2,464	2,464	0.831
Two	29	84,634	3,576	800	400	0.837
Three	29	44,399	2,303	396	132	0.835
Four	29	24,398	1,227	212	53	0.829
Five or more	29	100,670	3,781	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

	Comparison by RAM of assigned company's closest hospital			Regression on closest-hospital RAM	
	Low (1)	High (2)	p-value (3)	Coefficient (4)	p-value (5)
RAM prediction	-0.044	0.021	<0.001	0.101	<0.001
			A. Demographics		
Age	81.59	81.52	0.726	0.144	0.591
Male	0.379	0.385	0.554	-0.002	0.914
White	0.863	0.852	0.157	-0.004	0.722
Black	0.091	0.099	0.276	0.004	0.683
Referred from home	0.635	0.621	0.189	-0.042	0.010
Referred from accident	0.125	0.128	0.636	-0.009	0.408
Circulatory condition	0.234	0.236	0.851	-0.014	0.314
Respiratory condition	0.188	0.189	0.955	0.008	0.516
Digestive condition	0.064	0.066	0.711	0.003	0.688
Injury condition	0.176	0.179	0.730	0.002	0.899
Joint p-value			0.875		0.221
			B. Comorbidities		
Hypertension	0.263	0.267	0.670	0.009	0.559
Stroke	0.012	0.012	0.913	0.002	0.650
Cerebrovascular disease	0.032	0.034	0.729	0.003	0.640
Renal failure	0.118	0.118	0.988	0.009	0.411
Dialysis	0.012	0.011	0.894	0.001	0.862
COPD	0.108	0.108	0.925	0.004	0.671
Pneumonia	0.053	0.054	0.868	0.003	0.655
Diabetes	0.121	0.129	0.311	0.011	0.294
Protein-calorie malnutrition	0.035	0.037	0.734	0.003	0.585
Dementia	0.085	0.087	0.698	0.009	0.327
Paralysis	0.033	0.035	0.662	0.005	0.354
Peripheral vascular disease	0.073	0.076	0.618	0.001	0.873
Metastic cancer	0.020	0.021	0.720	0.001	0.846
Trauma	0.057	0.057	0.946	0.003	0.678
Substance abuse	0.039	0.038	0.840	0.000	0.939
Major psychological disorder	0.030	0.030	0.953	0.000	0.967
Chronic liver disease	0.007	0.007	0.871	0.002	0.580
Joint p-value			0.922		0.893
			C. Ambulance services		
Excess miles transported	-0.041	0.034	0.987	0.070	0.991
Emergency transport	0.954	0.954	0.972	0.012	0.063
Advanced life support	0.720	0.746	0.010	0.001	0.955
Intravenous fluids administered	0.008	0.007	0.644	-0.004	0.133
Intubation performed	<0.001	<0.001	0.270	-0.000	0.174
Joint p-value			0.141		0.205
Overall joint p-value			0.976		0.876

Notes: This table compares the characteristics of patients referred by ambulance companies that are located close to hospitals with high and low RAM predictions, controlling for patient ZIP code fixed effects. The sample in columns 1-3 includes 175,485 patients by companies that are closest (in terms of ZIP code centroid distance) to a hospital in either the first ("low") or fourth ("high") quartile of RAM predictions in their hospital service area. Columns 4-5 regress characteristics on the company's closest-hospital RAM in the full patient sample. Reported p-values are for the null of no difference across patients, and are based on robust standard errors. The "overall" p-values jointly test across panels A, B, and C. 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	MLE	OLS	MLE	OLS	MLE	MLE		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RAM1	0.123	0.121					0.023		
	(0.116)	(0.011)					(0.038)		
RAM2			0.123	0.125			0.015		
			(0.117)	(0.011)			(0.086)		
RAM3					0.127	0.127	0.119	0.145	0.157
					(0.117)	(0.011)	(0.059)	(0.016)	(0.027)
$(RAM3)^2$								-0.003	
								(0.008)	
$(RAM3)^3$								-0.005	
								(0.003)	
$(RAM3)\times J$									-0.021
									(0.019)
Residual std. dev.:									
Within-HSA	0.096	0.096	0.099	0.099	0.099	0.099	0.098	0.098	0.118
	(0.099)	(0.098)	(0.104)	(0.102)	(0.102)	(0.102)	(0.101)	(0.104)	(0.077)
Between-HSA	0.165	0.165	0.162	0.162	0.162	0.162	0.163	0.162	0.147
	(0.064)	(0.064)	(0.070)	(0.069)	(0.069)	(0.069)	(0.067)	(0.070)	(0.066)

Notes: This table reports estimated parameters of the hierarchical linear model outlined in the text. The sample consists of 2,082 minimum distance quality index estimates and RAM predictions. Columns 1, 3, and 5 report OLS coefficients and variance component estimates from a regression of quality index estimates on RAM predictions. Columns 2, 4, and 6 report corresponding maximum likelihood estimates. See the text for details of this model and a description of the three RAM specifications. Columns 7-9 report MLE estimates of multivariate models; in column 9 the the main effect of J, the total number of hospitals in each HSA, is estimated to be -0.038 (0.016). Standard errors, clustered by HSA, are reported in parentheses.

Table 4: Within-HSA correlates of hospital quality measures

	Observed survival			RAM prediction			Quality index estimate			Quality posterior		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
						A. Ownership structure						
For-profit	-0.019		-0.036	-0.065		-0.080	0.097		0.086	-0.063		-0.081
	(0.082)		(0.081)	(0.083)		(0.081)	(0.196)		(0.197)	(0.080)		(0.078)
Government	-0.072		-0.072	-0.178		-0.177	-0.325		-0.312	-0.154		-0.153
	(0.083)		(0.083)	(0.077)		(0.076)	(0.226)		(0.231)	(0.075)		(0.075)
Teaching		-0.103	-0.104		-0.092	-0.096		-0.121	-0.065		-0.108	-0.112
		(0.077)	(0.077)		(0.084)	(0.084)		(0.185)	(0.190)		(0.081)	(0.080)
						B. Average spending and patient volume						
Log(spending)	0.011		0.017	0.025		0.003	0.002		-0.007	0.027		0.005
	(0.012)		(0.013)	(0.013)		(0.013)	(0.336)		(0.335)	(0.013)		(0.013)
Log(volume)		-0.005	-0.011		0.039	0.038		0.068	0.068		0.032	0.030
		(0.013)	(0.014)		(0.017)	(0.018)		(0.107)	(0.108)		(0.016)	(0.017)

Notes: This table reports coefficients from regressions of the hospital quality measure in each column on the row characteristics, controlling for HSA fixed effects. All quality measures are normalized to standard deviation units; see Table A2 for a description of the RAM prediction specification (RAM3). Observed survival posteriors shrink observed rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The sample in columns (1)-(6) and (10)-(12) is 2,357 hospitals operating in 695 multi-hospital HSAs, while the sample in columns (7)-(9) include 1,145 with minimum distance quality estimates. Standard errors, clustered by HSA, are reported in parentheses.

Table 5: Average HSA-level selection bias, adjusting for selection-on-distance

	(1)	(2)	(3)
		A. No adjustment	
Avg. selection bias (pp)		3.60	
		(0.12)	
HSAs:		695	
		B. Parametric adjustment	
Avg. selection bias (pp)	3.51	3.48	3.43
	(0.13)	(0.14)	(0.14)
Marginal distance bias "effect"	-0.10	-0.15	-0.25
	(0.04)	(0.07)	(0.10)
Polynomial:	Linear	Quadratic	Cubic
HSAs:		695	
		C. Non-parametric adjustment	
Avg. selection bias (pp)	3.23	3.14	3.01
	(0.26)	(0.34)	(0.45)
Bandwidth:	1 mile	0.1 miles	0.01 miles
HSAs:	142	66	39

Notes: This table summarizes average HSA-level selection bias posteriors, expressed in percentage points of 30-day survival. Panel A reports the average across all 695 multi-hospital HSAs, while Panel B reports the constant from regressions of HSA-level 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 C reports average HSA-level selection bias posteriors for HSAs with an average distance bias that falls within the indicated bandwidth of zero. Robust standard errors are reported in parentheses.

Table 6: Correlates of value-based purchasing repayment rates

	Using benchmark scores	Using quality posteriors	Difference
	(1)	(2)	(3)
For-profit	-0.013	-0.046	-0.033
	(0.028)	(0.043)	(0.046)
Government	-0.069	-0.066	0.003
	(0.028)	(0.039)	(0.042)
Teaching	0.018	-0.098	-0.116
	(0.028)	(0.045)	(0.044)
Log(spending)	-0.222	0.180	0.402
	(0.048)	(0.060)	(0.074)
Log(volume)	0.173	0.195	0.022
	(0.012)	(0.017)	(0.019)
Log(bed capacity)	-0.022	-0.015	0.007
	(0.020)	(0.030)	(0.031)

Notes: Columns 1 and 2 report coefficients from regressing the share of total value-based purchasing withholdings that is repaid to the hospital given its benchmark VBP score and its quality posterior, respectively. Column 3 reports the difference in these coefficients. VBP simulations use FY2014 balance sheet information and non-quality domain scores; see the data appendix for a description of the repayment rate construction. 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. 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 (2016): "Interpreting Tests of School VAM Validity," *American Economic Review: Papers & Proceedings*, 106(5), 388–392.

——— (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132, 871–919.

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.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

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

Blackwell, M. (2017): "Instrumental Variable Methods for Conditional Effects and Causal Interaction in Voter Mobilization Experiments," *Journal of the American Statistical Association*, 112, 590–599.

Brinch, C., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

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 (2017): "Two-Step Estimation and Inference with Possibly Many Included Covariates," Working Paper.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2015): "Healthcare Exceptionalism? Performance and Allocation in the U.S. Healthcare Sector," *American Economic Review*, 106, 2110–44.

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.

Chandra, A. and D. O. Staiger (2017): "Identifying Sources of Inefficiency in Health Care," NBER Working Paper No. 24035.

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 (2017): "The Impacts of Neighborhoods on Intergenerational Mobility: County-Level Estimates," NBER Working Paper No. 23002.

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, and J. Gruber (2017a): "Evaluating Measures of Hospital Quality," NBER Working Paper No. 14607.

——— (2017b): "Uncovering Waste in US Healthcare," *Journal of Health Economics*, 54, 25–39.

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.

Finkelstein, A., M. Gentzkow, P. Hull, and H. Williams (2017): "Adjusting Risk Adjustment – Accounting for Variation in Diagnostic Intensity," *New England Journal of Medicine*, 376, 608–610.

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.

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.

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118.

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

Gupta, A. (2017): "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program," BFI Working Paper No. 2017-07.

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. 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.

Heckman, J. J. and R. Pinto (2017): "Unordered Monotonicity," NBER Working Paper No. 23497.

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

Hoxby, C. (2017): "The Productivity of U.S. Postsecondary Institutions," in *Productivity in Higher Education*, NBER.

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," *Econometrica*, 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.

Kirkebøen, L., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings, and Self-Selection," *Quarterly Journal of Economics*, 131, 1057–1111.

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.

McColloch, 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 (2017): "Moneyball in Medicare," *Journal of Health Economics*.

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.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

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.

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 data. I first link a 20% random sample of Medicare beneficiaries originating an ambulance company claim in the CMS Carrier file to their resulting inpatient claims, which indicate admitting hospitals and diagnoses. The claims data include basic patient demographic information, including birth date, sex, race, and the ZIP code where official correspondence is sent. These data are also linked to vital statistics that record when a patient dies, allowing me to construct 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 are also retained from the Carrier file. Hospital ZIP codes provided by inpatient claims are linked to the 2010 Dartmouth Atlas hospital service area definitions. Data on hospital ownership structure (non-profit private, for-profit private, and government owned) come from the 2010-2012 CMS Provider of Service files, while teaching status and total FY2014 diagnosis-related group payments come from hospital Cost Report data. Hospital volume is computed as the total number of admitted patients observed 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 analysis sample to patients who were admitted by ambulance through a hospital's emergency room for 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 discretion, and are listed in the note below 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.

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 referred via an emergency room by an ambulance (column 2); this subsample is slightly older and more female, with somewhat higher average Medicare spending and 30-day mortality. The hospitals represented in this emergency sample are somewhat more likely to be privately owned, non-profit, and higher-volume, and more than twice as likely to be a teaching hospital. Column 3 further report characteristics for patients and hospitals in HSAs with first-step minimum distance quality estimates, which constitutes roughly 85% of the analysis sample. These subsamples appear quite representative.

Observational RAMs are estimated in this sample via hierarchical logit regressions with conditionally normal random hospital effects, separately by each of the five condition categories listed in Table 1. RAM predictions $\hat{\alpha}_j$ are the volume-weighted average posterior means of the hospital effects across conditions. The benchmark RAM3 specification includes condition and year fixed effects, patient age and sex, and indicators for the 17 comorbidities listed in Panel B of Table 2. The RAM2 specification omits comorbidity dummies, while the most basic RAM1 model includes only condition and year effects. Appendix Table A2 also uses estimates from a replicated CMS-RAM model. For these I follow YNHHSC/CORE (2013) as closely as possible in constructing an auxiliary 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.

To simulate counterfactual reimbursements from the CMS Value-Based Purchasing program, I replicate the methodology outlined in DHHS/CMS (2015). FY2014 non-outcome domain scores are drawn for each hospital in my sample from the VBP website, while achievement and improvement scores for the outcome domain are obtained from estimated risk-standardized survival rates, as described in the text. Achievement scores based on quality posteriors come from the main 2010-2012 analysis sample, while improvement scores come from changes in posteriors 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 assigned linearly from positive improvement, with 8 points awarded to hospitals above the mean of the top tenth percentile of improvement. A hospital's Total Performance Score is the maximum of achievement and improvement points multiplied by 10, which for the benchmark simulations are then combined with the non-outcome domains with a weight of 25%. 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 repayment rates are given by the share of these payments divided by a hospital's total withholdings.

Econometric Appendix

Proof of Lemma 1

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

$$\begin{aligned} Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) &= E[E[\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]|X_i]] \\ &= E[E[\mathbf{1}[U_{ij}(Z_i) \geq U_{ik}(Z_i), \forall k]|Z_{i\ell} = 1, X_i]] \\ &= E[E[D_{ij}|Z_{i\ell} = 1, X_i]] \\ &= E\left[E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}|X_i\right]\right] \\ &= E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \end{aligned} \tag{29}$$

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

$$E[f(Y_{ij})|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = \frac{E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]]}{Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)}, \tag{30}$$

and, following the same steps as above,

$$E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]] = E\left[\frac{f(Y_i)D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \tag{31}$$

Combining equations (29)-(31) completes the proof. □

Proof of Proposition 2

Let P^{J-1} denote the standard $J-1$ probability simplex and define for $j = 1, \ldots, J-1$ the functions $G_j(\cdot) : \mathbb{R}^{J-1} \to P^{J-1}$ as

$$\begin{aligned} G_j(\pi) &= K_g \int_{-\infty}^{\pi_j - \pi_{-j}} |\Sigma_j|^{-1/2} g((t-\mu_j)'\Sigma_j^{-1}(t-\mu_j))dt \\ &= K_g \int_{-\infty}^{\Sigma_j^{-1/2}(\pi_j - \pi_{-j} - \mu_j)} g(t't)dt, \end{aligned} \tag{32}$$

where $\pi_{-j} = [\pi_1, \ldots, \pi_{j-1}, \pi_{j+1}, \ldots, \pi_{J-1}, 0]'$, the vector μ_j and positive-definite matrix Σ_j are the location and shape parameters, respectively, of $[\eta_{i1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{iJ}]' - \eta_{ij}$, which is elliptically distributed with generating function $g(\cdot)$ and integration constant K_g under Assumption 3. By inspection, $G(\cdot) = [G_1(\cdot), \ldots, G_{J-1}(\cdot)]'$ is a proper mapping, in that for any sequence $\{\pi_n\}$ that escapes to infinity in $\mathbb{R}^{J-1}$ we have $\{G(\pi_n)\}$ escaping to infinity in P^{J-1} with an appropriately-

defined metric. Moreover, the Jacobian

$$J_G(\pi) = \begin{bmatrix} \frac{\partial G_1(\pi)}{\partial \pi_1} & \cdots & \frac{\partial G_1(\pi)}{\partial \pi_{J-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial G_{J-1}(\pi)}{\partial \pi_1} & \cdots & \frac{\partial G_{J-1}(\pi)}{\partial \pi_{J-1}} \end{bmatrix} \tag{33}$$

is evidently a strictly-diagonally dominant L-matrix everywhere in $\mathbb{R}^{J-1}$, so that by the Lévy-Desplanques theorem it is everywhere invertible. Thus, by Hadamard's theorem there exists a global inverse function $G^{-1}(\cdot) : P^{J-1} \to \mathbb{R}^{J-1}$. By construction this function maps each of the ℓ vectors of choice probabilities identified by Lemma 1 to a unique-up-to-a-constant vector of utility shifters $\pi^*_{j\ell}$ given any $g(\cdot)$, s, and S from Assumption 3.

Next, for $j = 1, \ldots, J$, $\ell = 1, \ldots, \overline{L} = J$, and $C^{J-1} = \{\rho \in (0,1)^{J-1} : \sum \rho_j^2 < 1\}$, let $H_{j\ell}(\cdot) : \mathbb{R} \times C^{J-1} \to (0,1)$ be given by

$$H_{j\ell}(\beta, \rho) = \frac{K_g}{G_{j\ell}} \int_{-\infty}^{\beta/\sigma_j} \int_{-\infty}^{\Sigma_j^{-1/2}(\pi^*_{j\ell} - \pi^*_{-j,\ell} - \mu_j)} \begin{vmatrix} 1 & \rho' \\ \rho & I_{J-1} \end{vmatrix}^{-1/2} g\left(\begin{bmatrix} s \\ t \end{bmatrix}' \begin{bmatrix} 1 & \rho' \\ \rho & I_{J-1} \end{bmatrix}^{-1} \begin{bmatrix} s \\ t \end{bmatrix} \right) dt ds \tag{34}$$

$$= \frac{K_g}{G_{j\ell}} \int_{-\infty}^{\frac{\beta/\sigma_j - \rho' \Sigma_j^{-1/2}(\pi^*_{j\ell} - \pi^*_{-j,\ell} - \mu_j)}{\sqrt{1 - \sum \rho_m^2}}} \int_{-\infty}^{\Sigma_j^{-1/2}(\pi^*_{j\ell} - \pi^*_{-j,\ell} - \mu_j)} g\left(\begin{bmatrix} s \\ t \end{bmatrix}' \begin{bmatrix} s \\ t \end{bmatrix} \right) dt ds, \tag{35}$$

where σ_j^2 is the shape parameters associated with h_{ij} and $G_{j\ell}$ is the choice probability for institution j and instrument value ℓ. Note that each function $H_{j\ell}(\beta, \rho)$ is invariant to the location and shape parameters of η_i, and that $H_j(\cdot) = [H_{j1}(\cdot), \ldots, H_{j\overline{L}}(\cdot)]'$ is a proper mapping. Moreover, the Jacobian

$$J_{Hj}(\beta, \rho) = \begin{bmatrix} \frac{\partial H_{j1}(\beta,\rho)}{\partial \beta} & \frac{\partial H_{j1}(\beta,\rho)}{\partial \rho_1} & \cdots & \frac{\partial H_{j1}(\beta,\rho)}{\partial \rho_{J-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial H_{j\overline{L}}(\beta,\rho)}{\partial \beta} & \frac{\partial H_{j\overline{L}}(\beta,\rho)}{\partial \rho_1} & \cdots & \frac{\partial H_{j\overline{L}}(\beta,\rho)}{\partial \rho_{J-1}} \end{bmatrix} \tag{36}$$

has elements

$$\frac{\partial H_{j\ell}}{\partial \beta}(\beta, \rho) = \frac{K_g}{\sigma G_{j\ell}} \int_{-\infty}^{\Sigma_j^{-1/2}(\pi^*_{j\ell} - \pi^*_{-j,\ell} - \mu_j)} \begin{vmatrix} 1 & \rho' \\ \rho & I_{J-1} \end{vmatrix}^{-1/2} g\left(\begin{bmatrix} \frac{\beta}{\sigma_j} \\ t \end{bmatrix}' \begin{bmatrix} 1 & \rho' \\ \rho & I_{J-1} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\beta}{\sigma_j} \\ t \end{bmatrix} \right) dt \tag{37}$$

and

$$\frac{\partial H_{j\ell}}{\partial \rho_k}(\beta, \rho) = -\left(\rho_k \frac{\sigma \rho' \Sigma_j^{-1/2}}{1 - \sum \rho_m^2} + \sigma \Sigma_{jk}^{-1/2} \right) (\pi^*_{j\ell} - \pi^*_{-j,\ell} - \mu_j) \frac{\partial H_{j\ell}}{\partial \beta}(\beta, \rho), \tag{38}$$

where $\Sigma_{jk}^{-1/2}$ is the kth row of $\Sigma_j^{-1/2}$. Thus, given sufficient variation in the choice probabilities $J_{Hj}(\cdot)$ is everywhere invertible, and again by Hadamard's theorem there exists a global inverse $H^{-1}(\cdot) : (0,1)^L \to \mathbb{R} \times C^{J-1}$. By construction this function yields a unique β_j^*/σ_j satisfying the set

of the first J population mean selected outcomes for institution j, given the set of the first J choice probabilities for all k and $g(\cdot)$. Thus $q_j = K_g \int_{-\infty}^{\beta_j^*/\sigma_j} g(s^2)ds$ is identified given consistent estimates of these moments. □

Testing Hospital RAMs

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

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

for some known distribution F_ν. In the health context, H_0 rules out hospital comparative advantage and says that the patient sorting mechanism is independent of potential survival, conditional on the controls in W_i. In particular, H_0 implies $\nu_i \perp\!\!\!\perp D_i, W_i$, the usual basis for consistent estimation of equation (21), 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 for any ℓ, j, and w,

$$\begin{aligned} E[Y_i|Z_{i\ell} = 1, D_{ij} = 1, W_i = w] &= Pr(\alpha_j + \gamma' w \geq \nu_i | Z_{i\ell} = 1, U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k, W_i = w) \\ &= F_\nu(\alpha_j + \gamma' w) \\ &= E[Y_i|D_{ij} = 1, W_i = w], \end{aligned} \tag{39}$$

That us, under selection-on-observables a patient's expected 30-day survival does not depend on the identity of her ambulance company, conditional on her admission and observables. Defining the propensity score $p_\ell(D_i, W_i) = E[Z_{i\ell}|D_i, W_i]$ and iterating expectations, we thus have under H_0

$$\begin{aligned} E\left[Y_i \left(\frac{Z_{i\ell} - p_\ell(D_i, W_i)}{p_\ell(D_i, W_i)(1 - p_\ell(D_i, W_i))}\right)\right] &= E\left[E\left[\frac{Y_i Z_{i\ell}}{p_\ell(D_i, W_i)} - \frac{Y_i(1 - Z_{i\ell})}{1 - p_\ell(D_i, W_i)} | D_i, W_i\right]\right] \\ &= E[E[Y_i|Z_{i\ell} = 1, D_i, W_i] - E[Y_i|Z_{i\ell} = 0, D_i, W_i] \\ &= 0 \end{aligned} \tag{40}$$

Given a consistent estimate of $p_\ell(D_i, W_i)$ and under appropriate regularity conditions, we may thus use the sample analogue of the left-hand side of (39) to test H_0, across each instrument value ℓ.

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

$$\begin{aligned} E[Y_i|Z_i] &= E[E[Y_i|Z_i, D_i, W_i]|Z_i] \\ &= E[F_\nu(\alpha' D_i + \gamma' W_i)|Z_i], \end{aligned} \tag{41}$$

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 (41) implies vector-equality of the coefficients μ_Y and μ_F in the regressions:

$$Y_i = \mu_Y' Z_i + e_Y \tag{42}$$

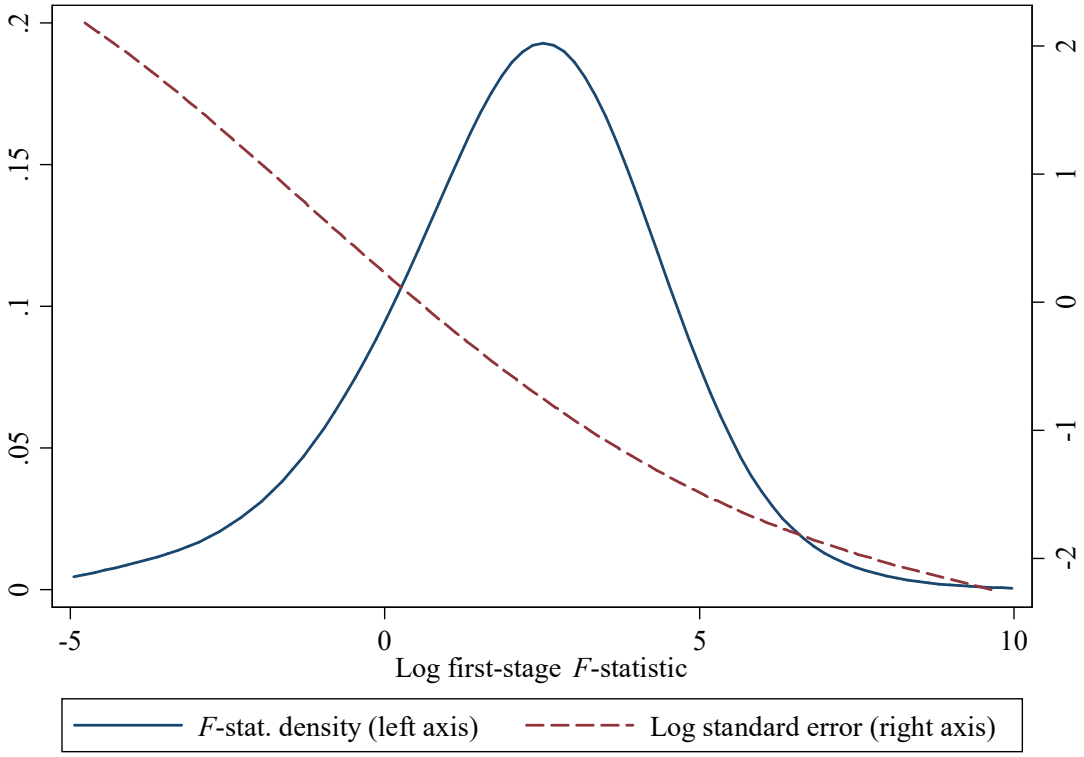
$$F_\nu(\alpha' D_i + \gamma' W_i)) = \mu_F' Z_i + e_F. \tag{43}$$

A final approach notes that equations (42) and (43) 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 $F_\nu(\alpha' D_i + \gamma' W_i)$ in a regression of realized survival Y_i. Since $\mu_Y = \mu_F$ under H_0, this procedure should produce a 2SLS coefficient 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 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 the propensity score-based 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, I approximate the propensity scores $p_\ell(D_i, W_i)$ with a probit and jointly test significance of the 100 sample analogues of the left-hand side of equation (40), correcting inference for first-step estimation error. Adding patient demographics and comorbidity controls to W_i in the RAM2 and RAM3 model reduces the resulting chi-squared test statistic somewhat from 295 in column 1 to 238 in column 3. Nevertheless, with 100 degrees of freedom, all three RAM specifications reject the null hypothesis of RAM validity ($p < 0.001$). This is similar to the rejection in column 4, which tests AMI, heart failure, and pneumonia RAMs replicating as closely as possibly the official 2013 CMS methodology (see the data appendix for details of this replication).

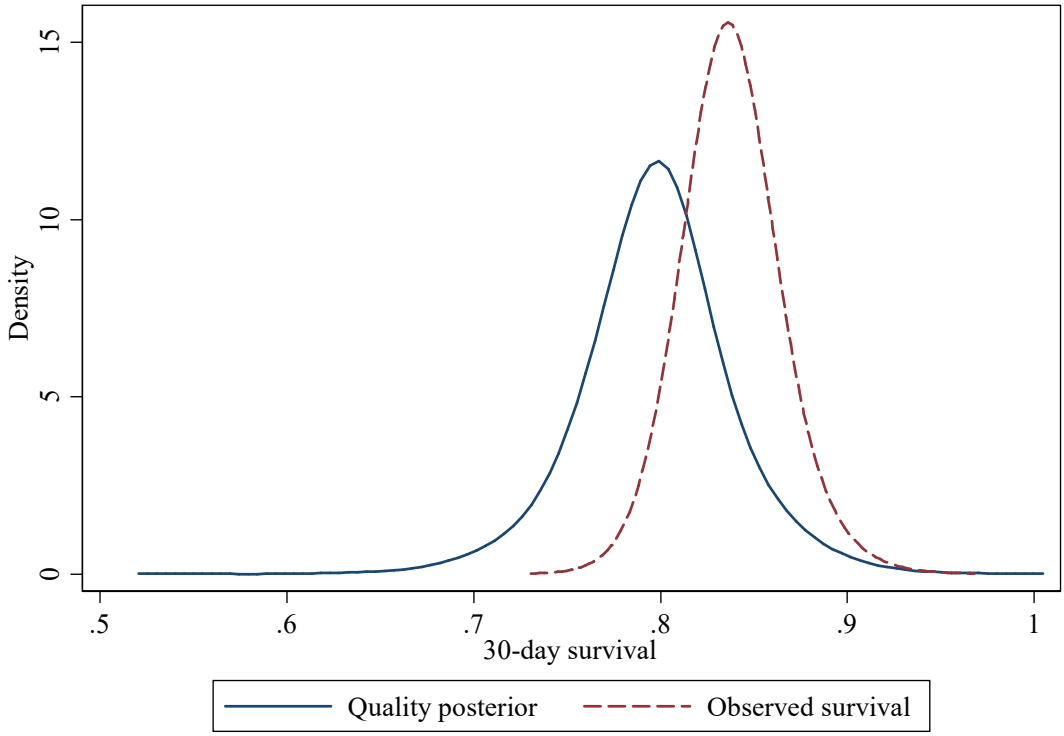
Panel B of Table A2 reports chi-squared statistics and associated p-values for tests of forecast bias, overidentification, and the full set of parametric restrictions given by equation (41), for the same set of 100 randomly-chosen ambulance companies. Adding demographic and comorbidity controls to the RAM brings the forecast coefficient from 1.30 to 1.09, and the latter is not statistically distinguishable from one. 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.

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



Notes: The solid blue line in this figure plots a Gaussian kernel density estimate of the distribution of log first-stage F-statistics for the 2,082 minimum distance hospital quality estimates. Here F-statistics test the equality of estimated choice probabilities across all ambulance company instruments for each hospital. The bandwidth used to estimate this distribution is 1. The dashed red line plots average log quality estimate standard errors for each estimate, smoothed by a quartic polynomial.

Figure A2: The distributions of hospital quality posteriors and 30-day survival rates



Notes: This figure plots Gaussian kernel density estimates of the distribution of empirical Bayes posteriors of hospital quality and hospital survival rates. The sample includes all 4,817 hospitals; the bandwidth used to estimate each distribution is 0.02.

Table A1: Patient and hospital characteristics

	All nondeferrable Medicare admissions	Analysis sample	
		All HSAs	HSAs with min. distance estimates
	(1)	(2)	(3)
		A. Patients	
30-day survival	0.875	0.833	0.834
Age	80.22	81.76	81.76
Male	0.410	0.379	0.379
White	0.873	0.875	0.881
Black	0.082	0.082	0.079
Circulatory condition	0.233	0.220	0.223
Respiratory condition	0.208	0.200	0.195
Digestive condition	0.101	0.065	0.066
Injury condition	0.118	0.177	0.178
		B. Hospitals	
RAM prediction	---	0.000	0.071
For-profit	0.200	0.155	0.179
Government	0.231	0.232	0.212
Teaching	0.216	0.223	0.215
Log(spending)	9.441	9.273	9.266
Log(volume)	4.350	3.349	3.493
Patients	998,489	405,172	346,011
Hospitals	5,162	4,817	2,839
HSAs	3,257	3,155	1,500

Notes: This table reports average patient and hospital characteristics across three samples of Medicare inpatient claims. Column 1 reflects a 20% random sample of patients admitted to a hospital in 2010-2012 for one of the 29 nondeferrable conditions listed in the notes to Table 1. Column 2 summarizes the analysis sample, described in more detail in the data appendix. Finally, column 3 reports characteristics of HSAs in the analysis sample with enough quasi-experimental data to construct minimum distance quality estimates.

Table A2: Hospital RAM bias tests

	RAM1 (1)	RAM2 (2)	RAM3 (3)	CMS-RAM (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:				
Year/condition FEs	Y	Y	Y	Y
Patient age/sex		Y	Y	Y
Comorbidities			Y	Y
Patients:		405,172		82,815

Notes: This table summarizes ambulance-based tests for bias in hospital risk-adjustment models. All RAMs are hierarchical logit models of 30-day survival, estimated separately for each condition 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 RAM1 model controls for year and diagnosis fixed effects, while the RAM2 specification includes patient age and sex and RAM3 adds all comorbidity indicators listed in Table 2. The specification in column 4 replicates the 2013 CMS 30-day risk-standardized mortality models for AMI, heart failure, and pneumonia. Tests use 100 randomly-selected ambulance companies referring at least 100 patients in the sample. 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: Correlation structure of 30-day survival, RAM predictions, and quality indices

	Over time				Across conditions			
	2010-12 (1)	2007-09 (2)	2004-06 (3)		Circulatory (4)	Respiratory (5)	Digestive (6)	Injury (7)
				A. Observed survival rates				
2007-09	0.444			Respiratory	0.232			
2004-06	0.297	0.440		Digestive	0.271	0.254		
2001-03	0.240	0.327	0.463	Injury	0.252	0.147	0.252	
				All other	0.153	0.319	0.130	0.134
				B. RAM predictions				
2007-09	0.351			Respiratory	0.261			
2004-06	0.298	0.344		Digestive	0.184	0.237		
2001-03	0.221	0.247	0.296	Injury	0.208	0.163	0.152	
				All other	0.211	0.319	0.185	0.160
				C. Hospital quality indices				
2007-09	0.669			Respiratory	0.642			
2004-06	0.598	0.669		Digestive	0.365	0.477		
2001-03	0.536	0.603	0.622	Injury	0.546	0.511	0.399	
				All other	0.641	0.763	0.306	0.476

Notes: This table reports estimated correlation coefficients for a hospital's 30-day survival rate, RAM prediction, and quality index, accounting for first-step estimation error. Columns 1-3 correlates data from the benchmark 2010-2012 analysis sample with corresponding data from 2007-2009, 2004-2006, and 2001-2003, while columns 4-7 report correlations across five patient diagnosis categories over the entire 2001-2012 period.

Table A4: Regressions of hospital quality posteriors on measured hospital inputs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log(staff salary)	0.375					0.264	0.232
	(0.068)					(0.071)	(0.071)
Uses electronic records		0.150				0.079	0.064
		(0.062)				(0.063)	(0.064)
Uses case management			0.109			-0.036	-0.073
			(0.043)			(0.047)	(0.047)
# of accreditations				0.061		0.022	0.002
				(0.015)		(0.018)	(0.021)
# of imaging technologies					0.041	0.025	0.009
					(0.007)	(0.009)	(0.011)
Log(volume)							0.056
							(0.022)

Notes: This table reports coefficients from regressions of hospital quality posteriors on measures of hospital input quality. The regressors are measured in the first year in which data are available in 2010-2012; the sample includes 3,198 hospitals with input data from the American Hospital Association. 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 A5: Average HSA-level selection bias, adjusting for selection-on-observables

	(1)	(2)	(3)	(4)	(5)	(6)
Avg. selection bias (pp)	3.60	3.43	3.20	3.46	3.48	3.13
	(0.12)	(0.14)	(0.16)	(0.16)	(0.16)	(0.20)
Selection adjustment:						
Distance		X				X
Condition			X			X
Demographics				X		X
Comorbidities					X	X

Notes: This table reports the constant, expressed in percentage points of 30-day survival, from regressions of HSA-level bias posteriors on cubic polynomials of HSA-level observable selection terms. The sample is 695 multi-hospital HSAs. A hospital's distiance selection term 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. Selection terms for each of the other covariate groups are calculated as the difference between a hospital's average Mahalanobis distance to its admitted patients and its average Mahalnobis distance to all potential patients, for observables in the group. Condition observables include a full set of indicators for the 29 diagnoses listed in the notes to Table 1. Demographic observables include patient age, sex, race, and indicators for whether a patient was referred from home or an accident. Comorbidity observables include a full set of indicators for the 17 conditions listed in Table 2. Robust standard errors are reported in parentheses.

Table A6: Robustness of the main results to alternative specifications

		Preferred specification	Robustness checks		
			Propensity scores omit RAM controls	Health/utility follows $t(2)$ distribution	HLM includes (RAM3)×J interaction
		(1)	(2)	(3)	(4)
Quality index posterior rank correlation		1.000	0.989	0.960	0.665
RAM3 coefficient in HLM		0.127	0.131	0.130	0.157
		(0.011)	(0.011)	(0.015)	(0.027)
Within-HSA quality-bias rank correlation		-0.810	-0.818	-0.825	-0.429
Within-HSA quality posterior correlates	Government	-0.154	-0.157	-0.146	-0.059
		(0.075)	(0.075)	(0.076)	(0.049)
	Log(spending)	0.027	0.021	0.032	0.011
		(0.013)	(0.013)	(0.013)	(0.006)
	Log(volume)	0.032	0.031	0.050	0.012
		(0.016)	(0.016)	(0.016)	(0.008)
Share of positively-selected HSAs		0.883	0.882	0.843	0.963
HSA-level selection bias (pp)	Average	3.60	3.68	3.34	4.15
		(0.12)	(0.12)	(0.13)	(0.21)
	Distance-adjusted avg.	3.01	2.90	2.96	3.23
		(0.45)	(0.49)	(0.47)	(0.18)
Correlates of VBP repayment rate change	Teaching	-0.111	-0.124	-0.159	-0.210
		(0.040)	(0.043)	(0.044)	(0.052)
	Log(spending)	0.407	0.415	0.382	0.361
		(0.070)	(0.074)	(0.074)	(0.078)
Expected survival gain from redirection (pp)	Max. RAM3 prediction	-0.20	-0.19	-0.21	-0.90
	Max. quality posterior	1.01	0.94	1.58	1.21

Notes: Column 1 of this table summarizes key results from the preferred analytic specification, while columns 2-4 report corresponding results from three alternative specifications. Specifically, column 2 excludes patient age, sex, and RAM comorbidities from the estimated ambulance company propensity scores, column 3 assumes patient health and utility indices are distributed by a multivariate Student's t distribution with two degrees of freedom instead of a multivariate normal, and column 4 adds the total number of hospitals in a hospital's HSA and the interaction of this number with a hospital's RAM3 prediction to the hierarchical linear model. The first row reports rank correlations of quality posteriors from each alternative specification with that of the preferred specification. The next sets of rows reports maximum likelihood estimates of the coefficient on RAM3 in the HLM as in Table 3, within-HSA rank correlations of posterior quality and bias reflected in Figure 5, within-HSA regression estimates of quality posteriors on hospital characteristics as in Table 4, the share of HSAs with positive average selection bias as in Figure 6, the average amount of HSA-level selection bias as in Table 5, the correlates of differences in VBP reimbursement rates as in Table 6, and simulated gains from rank-based admissions policies as in Figure 8. Standard errors, clustered by HSA, are reported in parentheses.