Estimating Hospital Quality with Quasi-Experimental Data*

Peter Hull†

January 2020

Abstract

Unobserved selection can bias observational estimates of institutional quality, while conventional instrumental variables (IV) quality estimators can impose strong or intractable restrictions on effect heterogeneity. I develop an alternative quasi-experimental approach that allows for unobserved Roy selection-on-gains. Key potential outcome moments are first estimated non-parametrically for each institution. These reduced-form data are then extrapolated, via structural restrictions, to estimate institutional quality. I show how this approach generalizes the implicit extrapolations of conventional IV estimators in simple experimental data, and propose a flexible family of extrapolations for settings with binary outcomes. I use this approach to estimate U.S. hospital quality from 30-day patient mortality outcomes and quasi-random ambulance company assignment within markets. I combine the estimates with conventional risk-adjustment methods to obtain hybrid empirical Bayes quality posteriors. I find that higher-spending, higher-volume, and privately owned hospitals tend to be of higher quality, and that most markets exhibit positive Roy selection. Higher-spending and non-teaching hospitals see modest increases in performance-linked subsidies when quality posteriors replace observational rankings. Policy simulations highlight limitations of report card rankings in settings with positive Roy selection.

*I thank Alberto Abadie, Nikhil Agarwal, Isaiah Andrews, Josh Angrist, Kirill Borusyak, Amitabh Chandra, David Cutler, Dave Deming, Joe Doyle, Amy Finkelstein, Matt Gentzkow, Gautam Gowrisankaran, Jon Gruber, Nick Hagerty, Nathan Hendren, Max Kasy, Larry Katz, Pat Kline, Jack Liebersohn, Bentley MacLeod, Rachael Meager, Magne Mogstad, Yusuke Narita, Aviv Nevo, Parag Pathak, Bryan Perry, Jesse Shapiro, Doug Staiger, Chris Walters, Glen Weyl, various seminar participants, and four anonymous referees for their many helpful comments and suggestions. I am especially thankful to the Doyle et al. (2015) research team for sharing replication code, and to Ben Artin, Mark Millet, Laura Segal, Julia Taylor, and Kevin Wickersham for sharing institutional knowledge. This work was funded in part by the National Institute on Aging and the Spencer Foundation; all policy views are my own.

Keywords: hospital quality, instrumental variables, Roy selection, empirical Bayes. *JEL:* C36, I11, I18

†University of Chicago Department of Economics and NBER. Contact: hull@uchicago.edu

1 Introduction

Outcome-based measures of institutional quality draw interest in many settings, from school and teacher value-added to the lasting effects of residential, educational, and occupational choices.[1] In the United States, institutional quality rankings have also come to play a central role in policymaking, particularly in the healthcare setting. Hospitals with low risk-adjusted 30-day mortality, for example, tend to be rewarded with increased Medicare reimbursement rates, while hospital "report cards" publicly flag institutions with poor health outcomes. Such policies can shape both provider and patient incentives, with many important and far-reaching consequences (Gupta, 2017; Norton et al., 2018; Dranove and Sfekas, 2008; Pope, 2009; Chandra et al., 2016).

A primary concern with institutional quality rankings is selection bias. To date, U.S. policymakers have exclusively relied on observational estimators, such as linear value-added models (VAMs) for teachers and nonlinear risk-adjustment models (RAMs) for hospitals. These estimators recover institutional quality under a standard selection-on-observables assumption: institutional selection is as-good-as-random, given observable characteristics. When selection is instead correlated with unobserved potential outcomes, observational quality estimates and associated policies can be biased. In the healthcare setting, selection bias might arise from quality-based policymaking itself if healthier patients respond more or less to public hospital report cards (Oster, Forthcoming). Other potential sources of bias in hospital RAMs include the medical expertise of referring doctors, the preferences of ambulance company drivers, and the non-random distance between hospitals and patients of different unobserved health (e.g., Hadley and Cunningham (2004)).

Instrumental variable (IV) quality estimators address selection bias by combining upstream quasi-experimental variation with certain structural restrictions. For example, Gowrisankaran and Town (1999) and Angrist et al. (2017) use a linear IV framework to estimate hospital and school effects, respectively, while Geweke et al. (2003) estimate a nonlinear IV model of hospital quality from patient mortality outcomes. The linear approach accomodates selection-on-unobservables by an assumption of constant treatment effects, as linear IV with multiple treatments (here, institutions) is otherwise difficult to causally interpret (Behaghel et al., 2013; Kirkebøen et al., 2016; Hull, 2017). Though tractable, the constant effects assumption rules out institutional comparative advantage and Roy selection on heterogeneous gains – a dynamic thought to be especially important in the healthcare setting (Chandra and Staiger, 2007; 2017). Constant effects is moreover incompatible with limited dependent outcomes, such as the 30-day survival indicators used in hospital RAMs. Nonlinear IV quality estimators address both shortcomings by imposing alternative parametric assumptions, such as the joint normality of a latent outcome index and latent patient utility from different hospital

[1] See, e.g., Chetty et al. (2014b), Chetty and Hendren (2018), Hoxby (2018), and Card et al. (2013) for recent analyses of the effects of different teachers, neighborhoods, colleges, and firms.

choices. These estimators can, however, be computationally challenging or intractable, while restricting effect heterogeneity in ways that, relative to the linear constant effects assumption, can be difficult to interpret or assess in practice.

This paper develops and applies a new approach to IV quality estimation that tractably accommodates unobserved Roy selection-on-gains by separating the roles of quasi-experimental instrument variation on one hand and structural assumptions on the other. I first show how in simple experimental settings (i.e., with unconfounded instrument assignment), conventional linear and nonlinear IV quality estimators can be viewed as enacting different parametric extrapolations of the same non-parametric estimates of certain reduced-form moments. The moments capture the extent of positive or negative selection into each institution, and the nature of their extrapolation is given by the structural assumptions, such as constant effects or joint normality of latent indices. This simple unifying interpretation, however, is lost in quasi-experimental data with conditionally random instrument assignment. Conventional IV estimators leverage additional restrictions on how treatment effect heterogeneity and selection depend on the various quasi-experimental design controls. These additional restrictions can also give rise to computational complexity in the case of nonlinear IV.

The new approach recovers the simple moment-extrapolation logic in quasi-experimental data by a two-step estimation procedure. In the first step, the key reduced-form moments are non-parametrically estimated by inverse (instrument) propensity score weighting. Given sufficiently rich quasi-experimental variation these estimates can be non-parametrically extrapolated, achieving quality identification "at infinity" in the sense of Heckman (1990). In settings with more limited quasi-experimental data, parametric moment extrapolations are derived from structural restrictions on potential outcome selection and applied by a minimum distance estimator. This way of separating the non-parametric estimation of reduced-form moments from the optimization of small-scale nonlinear objectives simplifies computation, relative to conventional likelihood-based estimators. The two-step approach also facilitates the interpretation of the minimum distance estimates and exploration of their sensitivity to alternative structural assumptions, holding fixed available quasi-experimental variation.

This general approach has many connections to the literature. The quasi-experimental quality estimation agenda is formally one of estimating average treatment effects (ATEs) of multiple unordered treatments, given conditionally unconfounded assignment to a discrete instrument. Discreteness necessitates an alternative to the local instrumental variables approach of Heckman et al. (2008), while both the interest in ATEs and the multiplicity of treatments require departures from standard approaches to local average treatment effect estimation (e.g., Abadie (2003)). The solution is similar to recent approaches by Brinch et al. (2017) and Mogstad et al. (2018), who show how shape restrictions on marginal treatment response functions can, in the binary treatment case, be used to identify policy-relevant parameters from reduced-form moments. I propose moment extrapolations

arising implicitly from structural assumptions on potential outcome distributions which are more appropriate for settings with limited dependent variables and multiple treatments. Importantly, this approach does not restrict the joint distribution of latent variables and observed characteristics, in contrast with the conditional moment extrapolation of Brinch et al. (2017) which may be impractical or undesirable in complex quasi-experimental designs. Rather, similar to the marginal structural model approach of Robins et al. (2000), I use inverse propensity score weighting (Hirano et al., 2003) to impose unconditional restrictions on effect heterogeneity and selection, enacting the kinds of unconditional moment extrapolations that would be feasible in the experimental ideal.

To apply this general approach, I develop a family of potential outcome parameterizations appropriate for measuring an institution's quality by its effect on binary outcomes (such as 30-day mortality). These models specify an elliptical dependence between latent potential outcome indices and the utility indices rationalizing institutional selection, while imposing a first-stage monotonicity assumption (Imbens and Angrist, 1994; Heckman and Pinto, 2018). I show that each model in this class is without observational loss for the marginal distribution of potential outcomes and institutional substitution patterns, while allowing for more flexible patterns of institutional comparative advantage and Roy selection than traditional linear and nonlinear IV models. In particular, the models do not impose constant or monotone treatment effects, allowing some individuals to benefit from institutions that are less appropriate for others. This flexibility notwithstanding, I show that quality is identified in this family by a relatively small number of reduced-form moments, with estimation imposing intuitive extrapolations of this quasi-experimental variation.

I use this framework to estimate the quality of U.S. hospitals from a nationally representative sample of Medicare patients admitted with a nondeferrable (emergency) condition. Exploiting the quasi-random assignment of patients to ambulance companies with different referral preferences, I fit Gaussian models of hospital selection and 30-day patient survival to estimate the relative quality of hospitals within local healthcare markets. Doyle et al. (2015) first use ambulance company instruments in linear IV regressions of patient mortality on the average spending or observable RAM prediction of a patient's hospital. I instead use ambulance assignment variation to instrument for hospital choice directly, leveraging a weaker exclusion restriction that allows quality to differ across hospitals with the same annual spending or observational quality proxies.

Applying the method yields a set of local quality estimates for 2,082 U.S. hospitals in 968 hospital service areas (HSAs) with sufficient quasi-experimental variation. Overall, these estimates are positively but imperfectly correlated with both observational RAM and linear IV estimates, indicating unobserved selection on both levels and gains. Individual hospital estimates, however, tend to be noisy. For a more fine-grained analysis of hospital quality and sorting, I next combine the quasi-experimental estimates with conventional RAM quality predictions in an empirical Bayes framework (Angrist et al., 2017; Chetty and Hendren, 2018; Finkelstein et al., 2017). This yields a set of quality

posteriors for all U.S. hospitals which optimally trade off variance and bias across the two estimates.

An analysis of the quality posteriors reveals several robust patterns in hospital quality and selection. First, I find that higher-volume and higher-spending hospitals tend to produce better survival outcomes for patients in their HSAs, while government-run hospitals are of systematically lower quality. Moving a typical emergency patient to a nearby hospital with a one standard deviation higher log patient volume or log average spending increases her 30-day survival probability by 0.2 and 0.1 percentage points (pp), respectively. Moving her instead to a local government-run hospital from one that is privately owned decreases expected survival by 0.6pp. These results are qualitatively similar to findings in earlier observational and quasi-experimental studies (Foster et al., 2013; Chandra et al., 2016; Doyle et al., 2015), and are also reflected in conventional RAM predictions.

Second, I find strong evidence of hospital comparative advantage and positive Roy selection, with patients tending to be admitted to more appropriate hospitals. A relatively small share of this selection appears due to patients choosing hospitals that are closer to them or that tend to serve observably similar individuals. This implies both that hospitals specialize on patient unobservables and that either patients or ambulance company drivers are partially aware of institutional comparative advantage. This positive Roy selection-on-gains appears to coexist with negative selection-on-levels, as better hospitals tend to serve sicker patients on average. As a result, hospital quality and RAM selection bias are negatively correlated, and there is a strong positive correlation between conventional hospital rankings and rankings based on quality posteriors (despite the bias).

I conclude by showing how non-random hospital selection impacts quality-based payment and admission guidance policies. Replacing conventional hospital rankings with rankings based on quality posteriors has little impact on the types of providers receiving higher Medicare reimbursement rates, though higher-spending and non-teaching hospitals tend to see moderately higher payments. In simulations of report card admission policies, I find that a typical patient has a 3.2pp higher survival rate when choosing hospitals on the basis of RAM predictions rather than selecting at random. Admission to hospitals with the highest quality posteriors in a local market yields larger survival rate improvements, in the range of 3.6-4.7pp. Nevertheless, the scope for health gains from such policies is limited by the extent of positive Roy selection, which makes prevailing patient choices better than random. Moving a typical patient from her hospital of choice to the highest-RAM provider in her region *decreases* expected survival by 0.6pp, and the gains from redirecting patients to hospitals with the highest average quality are similarly dampened (-0.2-0.6pp). This result highlights a general issue with report card policies admission in settings with meaningful unobserved Roy selection-on-gains.

The remainder of the paper is organized as follows. Section 2 presents the general econometric setting and results. Section 3 presents the institutional setting and data. Section 4 discusses the findings on hospital quality, patient sorting, and policy simulations. Section 5 concludes.

2 Econometric Framework

2.1 Setting and Motivation

We suppose an econometrician observes an *iid* sample of data $\mathcal{Y}_i = (Y_i, D_i', Z_i', X_i')'$ for N individuals, where Y_i is an outcome of individual i and D_i consists of indicators D_{ij} indexing her choice across J institutions. For example, in the healthcare setting, $D_{ij} = 1$ indicates patient i's admission to hospital j, while Y_i measures her post-admission mortality. Alternatively, in an education setting, Y_i may denote the test score of student i following enrollment in a school j. Before selecting an institution, each individual is assigned to a discrete instrument Z_i, with elements $Z_{i\ell}$ indicating her assignment to one of L groups. For example, $Z_{i\ell} = 1$ in the healthcare application indicates that ambulance company ℓ was dispatched to individual i; in the education application of Angrist et al. (2017), the $Z_{i\ell}$ correspond to admission offers for student i to attend school ℓ. The vector X_i contains a set of controls, such as demographics or other characteristics determined before $(Z_i', D_i', Y_i)'$.

We assume that institutional choices and outcomes given the instrument are generated from vectors of latent *iid* variables $\mathcal{U}_i = ((Y_{ij}, (D_{ij\ell})_{\ell=1}^L)_{j=1}^J)'$. The Y_{ij} denote potential outcomes of indidvidual i if she were to select each institution j; realized outcomes are then given by

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

The $D_{ij\ell}$ similarly allow instrument assignment to affect institutional choice D_i: with $D_{ij\ell} = 1$ indicating that individual i would select institution j if assigned to instrument value ℓ,

$$D_{ij} = \sum\nolimits_\ell D_{ij\ell} Z_{i\ell}. \tag{2}$$

Absent further restrictions, this reduced-form model allows for arbitrary institutional comparative advantage and endogenous selection: the institutional treatment effects $Y_{ij} - Y_{ik}$ may be heterogeneous across individuals and correlated with the preferences governing $D_{ij\ell}$.[2] For instance, the individuals most likely to select an institution may see systematically higher or lower gains from doing so, generating positive or negative selection in the sense of Roy (1951).[3]

The econometrician is tasked with using observations of $\mathcal{Y}_i$ to estimate institutional quality $q_j = E[Y_{ij}]$. Quality is here defined as the average potential outcome of institution j in a population, formalizing the unidimensional outcome-based summary of average performance, or "value" often targeted by policymakers (e.g., DHHS (2015) in the healthcare setting). Quality comparisons reflect average institutional treatment effects, $q_j - q_k = E[Y_{ij} - Y_{ik}]$, and the quality estimation agenda

[2] By not indexing potential outcomes by ℓ the model imposes an implicit exclusion restriction, that the instrument only affects outcomes through the choice of institution. I return to this assumption in the context of the application.

[3] In general, the $D_{ij\ell}$ should be understood as capturing the joint decision of all relevant agents when i is assigned the ℓth instrument value, with Roy selection interpreted accordingly. In the hospital setting, for example, it is not possible to distinguish between admission decisions made by the patient from those made by the ambulance company.

is formally one of estimating intercepts in a sample selection or treatment effects model (Heckman, 1990; Andrews and Schafgans, 1998; Lewbel, 2007), with discrete instruments $Z_{i\ell}$ and multiple unordered treatments D_{ij}.[4]

The econometrician's challenge is to use conditional variation in potential outcomes Y_{ij}, observed among self-selected individuals with $D_{ij} = 1$, to recover the unconditional potential outcome mean q_j. When institutional choice is as-good-as-random, in the sense of $Y_{ij} \perp\!\!\!\perp D_{ij}$, this problem has a trivial solution: the conditional and unconditional distributions of Y_{ij} coincide, so quality is revealed by the average outcomes of each institution: $E[Y_i|D_{ij} = 1] = E[Y_{ij}|D_{ij} = 1] = E[Y_{ij}] = q_j$. Obervational quality estimators extend this logic by a combination of parametric and selection-on-observables assumptions.[5] More generally, differences between $E[Y_i \mid D_{ij} = 1]$ and q_j reflect selection bias due to non-random sorting across institutions which causes D_{ij} and Y_{ij} to be correlated.

The intuition for this paper's quasi-experimental approach is that in settings with non-random sorting, variation in average potential outcomes of self-selected individuals across different as-good-as-random instrument assignments can be structured and extrapolated to eliminate selection bias. Suppose, for example, that for a given institution j the econometrician finds that the instrument values ℓ which induce more individuals to choose j tend to yield lower average outcomes there. That is, suppose j's mean selected outcomes $E[Y_{ij} \mid D_{ij\ell} = 1]$ are lower for the ℓ which generate higher choice probabilities $Pr(D_{ij\ell} = 1)$. We might infer from this that there is positive selection into the institution, in the sense that the individuals least likely to choose j tend to have lower outcomes there. This finding would suggest the econometrician should apply a negative extrapolation from the mean selected outcomes, appropriate to remove the positive selection bias in each $E[Y_{ij} \mid D_{ij\ell} = 1] - q_j$. Indeed, quality $q_j = E[Y_{ij}]$ may be thought as the limit of an implicit function linking mean selected outcomes to potential choice probabilities, as the latter tends towards one. Quality is identified by appropriately extending the points of the function that are revealed by the instrument to this limit.

I next show how conventional linear and nonlinear IV approaches to quality estimation can be understood as performing such mean selected outcome extrapolations, in simple experimental settings with fully random instrument assignment. I then show how this logic can be applied to more complex quasi-experimental designs, where conventional IV estimators may become intractable or impose strong or opaque restrictions on the extent of Roy selection on institutional comparative advantage. I conclude this section by discussing a particular class of extrapolations, appropriate for estimating quality from binary outcomes while allowing for flexible Roy selection-on-gains.

[4]Heckman et al. (2008) discuss identification of average and local average treatment effects (LATEs) from local instrumental variable functions in multiple treatment settings. In general these functions are not non-parametrically identified given a fixed set of discrete quasi-experimental assignments Z_i. While not the focus of this paper, other functions of $\mathcal{U}_i$ such as LATEs may be estimated by extending the semi-parametric approach.

[5]For example if treatment effects are constant, $Y_{ij} = q_j + \varepsilon_i$, and each D_{ij} is uncorrelated with the population residual from projecting ε_i on X_i, then an ordinary least squares regression of Y_i on D_i and X_i identifies quality.

2.2 IV Quality Estimation in Experimental Data

When instrument assignment is as-good-as-random, as when determined by a simple experiment, different IV quality estimators can be understood as enacting different extrapolations of a common set of non-parametric mean selected outcome estimates. These estimates are given by the sample analogues of $E[Y_i D_{ij} Z_{i\ell}]/E[D_{ij} Z_{i\ell}]$, which identify the mean selected outcomes $E[Y_{ij} \mid D_{ij\ell} = 1]$ under random instrument assignment ($\mathcal{U}_i \perp\!\!\!\perp Z_i$).[6] The degree of extrapolation from these moments is generally given by a common set of choice probability estimates: the sample analogues of $E[D_{ij} Z_{i\ell}]/E[Z_{i\ell}]$. The nature of extrapolation arises implicitly from different structural assumptions on the latent variables in $\mathcal{U}_i$, restricting heterogeneity in causal effects and non-random selection.

Linear IV quality estimators, like those used by Gowrisankaran and Town (1999) and Angrist et al. (2017), naturally apply a linear extrapolation to mean selected outcome estimates.[7] These estimators arise from a constant treatment effects restriction, as linear IV with multiple unordered treatments and unrestricted effect heterogeneity typically fails to recover interpretable causal parameters (Behaghel et al., 2013; Kirkebøen et al., 2016).[8] Formally, if (*i*) $Y_{ij} = q_j + \varepsilon_i$ for $E[\varepsilon_i] = 0$, such that every individual faces the same outcome gain $Y_{ij} - Y_{ik} = q_j - q_k$ from switching between any two institutions, (*ii*) $\mathcal{U}_i \perp\!\!\!\perp Z_i$, and (*iii*) the usual rank condition holds, then a linear IV regression of Y_i on all but one D_{ij} indicator, instrumented by all but one $Z_{i\ell}$ indicator, identifies a set of $J-1$ quality comparisons. Appendix B.1 shows that the implied estimates of q_j can be written as a weighted average of mean selected outcome estimates $\hat{H}_{j\ell} = \left(\frac{1}{N}\sum_i Y_i D_{ij} Z_{i\ell}\right) / \left(\frac{1}{N}\sum_i D_{ij} Z_{i\ell}\right)$, with non-convex weights given by a set of choice probability estimates $\hat{G}_{j\ell} = \left(\frac{1}{N}\sum_i D_{ij} Z_{i\ell}\right) / \left(\frac{1}{N}\sum_i Z_{i\ell}\right)$.

To illustrate this result, consider the simple case of two institutions and two instrument values. Appendix B.1 shows that the linear IV quality estimate of institution 1 can be written

$$\hat{q}_1^{LIV} = \hat{\omega}\hat{H}_{11} + (1-\hat{\omega})\hat{H}_{12} + \hat{c} \tag{3}$$

where $\hat{\omega}$ is a function of institution 1's choice probability estimates and $\hat{c}$ is a linear function of the other institution's mean selected outcome estimates. Here $\hat{\omega} < 0$ or $\hat{\omega} > 1$, so the first two terms of (3) represent a non-convex average of institution 1's estimated mean selected outcomes. Specifically,

[6] That is, $E[Y_i D_{ij} Z_{i\ell}] = E[Y_{ij} \mid D_{ij\ell} = 1, Z_{i\ell} = 1]E[D_{ij} Z_{i\ell}] = E[Y_{ij} \mid D_{ij\ell} = 1]E[D_{ij} Z_{i\ell}]$, where the first equality holds by (1)-(2) and the second by experimental instrument assignment. Identification of choice probabilities similarly follows from the fact that $E[D_{ij} Z_{i\ell}] = Pr(D_{ij\ell} = 1)E[Z_{i\ell}]$.

[7] Gowrisankaran and Town (1999) and Geweke et al. (2003) use hospital distance instruments to estimate hospital quality. For the purposes of this discussion, one could imagine Z_i indicating random assignment of patients to a set of potential locations. Angrist et al. (2017) estimate school quality in a constant effects model with fewer assignment instruments $Z_{i\ell}$ than schools; here one could imagine interacting assignment instruments with student observables to overcome this underidentification challenge.

[8] Even in the two-institution case, where linear IV may estimate a local average treatment effect, linear IV quality estimates may only be of policy interest when treatment effect heterogeneity is limited. Appendix B.6 illustrates this in a stylized model of quality regulation in which hospitals may "cream-skim" patients with certain characteristics, and where such gaming incentives are stronger when linear IV quality rankings are used to rank hospitals with comparative advantage. This dynamic reinforces the importance of quality estimators that allow for such heterogeneity.

$\hat{\omega} > 1$ when $\hat{G}_{11} > \hat{G}_{12}$, or when the share of individuals selecting institution 1 is found to be higher when $Z_{i1} = 1$ than when $Z_{i2} = 1$. In this case, (3) enacts a linear extrapolation of institution 1's mean selected outcome estimates, in the direction of $\hat{H}_{11}$: all else equal, increases in $\hat{H}_{11} - \hat{H}_{12}$ lead to increased quality estimates $\hat{q}_1^{LIV}$.

Nonlinear IV quality estimators can be similarly characterized in simple experimental settings. These estimators typically replace the constant effects assumption with a likelihood for the observed data $\mathcal{Y}_i$, which by (1)-(2) and $\mathcal{U}_i \perp\!\!\!\perp Z_i$ can be understood as restricting $\mathcal{U}_i$. This likelihood might allow for certain forms of treatment effect heterogeneity, while also being more appropriate than linear IV when Y_i has limited support. Consider, for example, a bivariate probit model of binary potential outcomes in the two-institution case:

$$Y_{ij} = \mathbf{1}[\beta_j + \varepsilon_{ij} > 0] \tag{4}$$

$$D_{ij} = \mathbf{1}[\pi_{j\ell} + \eta_{ij} > \pi_{k\ell} + \eta_{ik}, \forall k \neq j], \tag{5}$$

where $J = 2$ and $(\varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) \sim N(0, \Sigma)$. This model allows for flexible treatment effect heterogeneity and non-random selection on gains. In particular, it generalizes the usual single-index bivariate probit parameterization, which by restricting $\varepsilon_{ij} = \varepsilon_i$ imposes rank-invariance of the potential outcomes: $Pr(Y_{ij} > Y_{ik}) = 1$ for all $j > k$ (without loss). Geweke et al. (2003) use a multivariate version of the single-index probit model in their estimation of hospital quality, as discussed below.

In simple experimental data, maximum likelihood quality estimates based on (4)–(5) are functions of the same first-step mean selected outcome and choice probability estimates which linear IV uses for extrapolation. This result follows from an elementary observation, shown in Appendix B.2, that when $\mathcal{U}_i \perp\!\!\!\perp Z_i$ and Y_i is discrete, any maximum likelihood quality estimate equivalently solves a minimum distance problem – one that matches non-parametric estimates of choice probabilities $Pr(D_{ij\ell} = 1)$ and conditional outcome probabilities $Pr(Y_{ij} = y \mid D_{ij\ell} = 1)$ to the corresponding moments of the model.[9] When Y_i is binary, these moment estimates are the same $\hat{G}_{j\ell}$ and $\hat{H}_{j\ell}$ as in the linear IV case. Different likelihood specifications for $\mathcal{U}_i$ imply different mappings between the population moments and institutional quality, rendering maximum likelihood estimates $\hat{q}_j^{MLE}$ as different implicit functions of these moment estimates.

This minimum distance interpretation can be used to characterize how nonlinear IV estimates impose different extrapolations of common mean selected outcome estimates. In the bivariate probit example, one can write the $2L$ moments for each institution j as $E[Y_{ij} \mid D_{ij\ell} = 1] = h_j(\pi_{j\ell})$ and $Pr(D_{ij\ell} = 1) = \Phi(\pi_{j\ell})$, where $h_j(\cdot)$ is a monotone function known up to two j-specific parameters

[9]Here it is assumed the observed data likelihood is differentiable and maximized on the interior of the parameter space, so that the maximum likelihood esitmates satisfy its first-order conditions. See Kline and Walters (2019) for a similar equivalence in the case of a "Heckit" likelihood, in which Y_i is normally distributed and $J = 2$.

and $\Phi(\cdot)$ is the standard normal cumulative distribution function.[10] When the model is just-identified (i.e., $L = 2$), the maximum likelihood procedure fits these moments exactly, finding parameters such that $\hat{h}_j(\cdot)$ satisfies $\hat{H}_{j\ell} = \hat{h}_j(\Phi^{-1}(\hat{G}_{j\ell}))$ for $\ell = 1, 2$. Mean selected outcomes in this model are monotone in the associated choice probabilities, $E[Y_{ij} \mid D_{ij\ell} = 1] = h_j(\Phi^{-1}(Pr(D_{ij\ell} = 1)))$, and quality is the intercept of this function at one: $q_j = \lim_{p\to 1} h_j(\Phi^{-1}(p))$. Correspondingly, the maximum likelihood quality estimate $\hat{q}_j^{MLE}$ can be understood as a monotone extrapolation of the two mean selected outcome estimates along the fitted curve $\hat{h}_j(\Phi^{-1}(p))$ to $p = 1$. As with the linear IV example, this extrapolation is in the direction of the instrument value with a higher choice probability estimate, so $sgn(\hat{q}_j^{MLE} - \hat{H}_{j1}) = sgn(\hat{H}_{j1} - \hat{H}_{j2})$ when $\hat{G}_{j1} > \hat{G}_{j2}$.[11] Unlike the linear IV extrapolation, however, the nonlinear extension of mean selected outcomes arises from the structure of equations (4)–(5) ensuring that $\hat{q}_j$ stays within the logical bounds of $(0, 1)$.

Panel A of Figure 1 illustrates this geometric interpretation of nonlinear IV quality estimation in the bivariate probit model, plotting example mean selected outcome and choice probability estimates for a given institution j. Maximum likelihood estimation minimizes a weighted distance between these estimates and the model-implied moments, resulting in estimates of the mean selected outcome function $h_j(\cdot)$ and choice probability parameters $\pi_{j\ell}$. The resulting $\hat{h}_j(\cdot)$ can be visualized as a weighted curve-of-best-fit through the non-parametric estimates, with a vertical intercept at one equal to $\hat{q}_j^{MLE}$. Panel B gives an analogous visualization of maximum likelihood estimation of an expanded model with three institutions, maintaining the assumption of joint-normality between the ε_{ij} and η_{ij}. Here the non-parametric moment estimates are plotted in three dimensions, because $E[Y_{ij} \mid D_{ij\ell} = 1] = h_j(\pi_{1\ell}, \pi_{2\ell})$ is now a function of two choice probability parameters.[12] The geometric interpretation of $\hat{q}_j^{MLE}$ remains, with the maximum likelihood estimator finding a weighted surface-of-best-fit through the mean selected outcome estimates and yielding an extrapolation that estimates $q_j = \lim_{\pi_1\to\infty} h_j(\pi_1, \pi_2)$ in the lower-right corner of the figure.

Plots like Figure 1 are useful not only for visualizing the role of structural restrictions in experimental IV quality estimation, but also for clarifying the usefulness of such restrictions in general.

---

[10] Namely, $h_j(\pi) = \int_{-\infty}^{\pi} \Phi\left(\frac{\beta_j + \rho_j v}{\sqrt{1-\rho_j^2}}\right) \frac{\phi(v)}{\Phi(\pi)} dv$, where $\rho_j = Corr(\varepsilon_{ij}, \eta_{ij})$ and without loss $\pi_{\neg j\ell} = \eta_{i\neg j} = 0$ and $Var(\varepsilon_{ij}) = Var(\eta_{ij}) = 1$. Here $h_j'(\pi)$ is proportional to $\Phi\left(\frac{\beta_j + \rho_j \pi}{\sqrt{1-\rho_j^2}}\right) - h_j(\pi) = Pr(\beta_j + \varepsilon_{ij} > 0 \mid \pi = -\eta_{ij}) - Pr(\beta_j + \varepsilon_{ij} > 0 \mid \pi > -\eta_{ij})$, the sign of which only depends on ρ_j.

[11] Note that this model satisfies monotonicity in the sense of Imbens and Angrist (1994). When $\pi_{j1} > \pi_{j2}$, $\hat{H}_{j2}$ estimates the mean Y_{ij} among treatment j "always-takers," while $\hat{H}_{j1}$ estimates a weighted average of "complier" and always-taker outcomes. Thus, $\hat{q}_j^{MLE}$ can be seen as an extrapolation based on the difference in outcomes between these two behavioral groups, to the subpopulation of "never-takers" for whom Y_{ij} is unobserved.

[12] For example, $h_1(\pi_1, \pi_2) = \int_{-\infty}^{\pi_1} \int_{-\infty}^{\pi_1 - \pi_2} \int_{\infty}^{\beta_j} \frac{\tilde{\phi}(\varepsilon, \nu, u; \rho_{j1}, \rho_{j2})}{\tilde{\Phi}(\pi_1 - \pi_2, \pi_1)} d\varepsilon, d\nu, du$, where $\rho_{j1} = Corr(\varepsilon_{i1}, \eta_{i1} - \eta_{i2})$; $\rho_{j2} = Corr(\varepsilon_{i1}, \eta_{i1})$; $\tilde{\phi}(\cdot; \rho_1, \rho_2)$ denotes the density of a mean-zero, unit-variance trivariate normal with the first element correlated by ρ_1 and ρ_2 with the second and third elements, which are themselves correlated by 1/2; and $\tilde{\Phi}(\cdot)$ is the cumulative density of the latter two variables. This follows by normalizing $\pi_{3\ell} = \eta_{i3} = 0$, $Var(\varepsilon_{ij}) = Var(\eta_{ij}) = 1$, and $Cov(\eta_{i1}, \eta_{i2}) = 0$. That the latter is without loss is shown in the proof to Proposition 2 (Appendix B.5), below.

With more instrument values, and thus more points in Figure 1A, we might imagine relaxing the parametric structure of (4)–(5) to fit a more flexible curve to the non-parametric points. In particular, if we were to observe more points near the maximal choice probability of one as the sample grows, we might imagine non-parametrically estimating the quality intercepts by local curve-fitting. This approach, formalized below, substitutes structural restrictions with a particular richness of quasi-experimental data, achieving quality identification "at infinity" in the sense of Heckman (1990).

The simple geometric interpretation of conventional nonlinear IV quality estimators nevertheless tends to become obscured outside of the experimental ideal. Consider the single-index multivariate probit model that Geweke et al. (2003) use to estimate hospital quality:

$$Y_i = \mathbf{1}[\textstyle\sum_j D_{ij}\beta_j + X_i'\gamma + \varepsilon_i > 0] \tag{6}$$

$$D_{ij} = \mathbf{1}[\textstyle\sum_\ell Z_{i\ell}\pi_{j\ell} + X_i'\kappa_j + \eta_{ij} > \sum_\ell Z_{i\ell}\pi_{k\ell} + X_i'\kappa_k + \eta_{ik}, \forall k \neq j], \tag{7}$$

where $(\varepsilon_i, \eta_i')' \mid X_i, Z_i \sim N(0, \Sigma)$ and X_i again denotes observed controls.[13] Equations (6)–(7) follow from a restriction on $\mathcal{U}_i$ in which $Y_{ij} = \mathbf{1}[\beta_j + X_i'\gamma + \varepsilon_i > 0]$ and $D_{ij\ell} = \mathbf{1}[\pi_{j\ell} + X_i'\kappa_j + \eta_{ij} > \pi_{k\ell} + X_i'\kappa_k + \eta_{ik}, \forall k \neq j]$. As mentioned, the former single-index restriction implies that conditional on X_i there is no comparative advantage: if $Y_{ij} > Y_{ik}$ for some patients i and some hospitals j and k, then hospital j is weakly better for all patients in the population. The latter first-stage restriction similarly implies the instrument has at most a monotone effect on treatment choice given X_i. Combining these restrictions with the quasi-experimental assumption that $(\varepsilon_i, \eta_i')' \perp\!\!\!\perp (X_i', Z_i')'$ weakens the previous experimental requirement that $\mathcal{U}_i \perp\!\!\!\perp Z_i$: the instrument need now only be as-good-as-randomly assigned conditional on X_i. This weaker assumption can be essential in practice: ambulance company assignment may only be plausibly exogenous among observably similar individuals from the same location (Doyle et al., 2015), while centralized school admission offers may only be random given student preferences and priorities (Abdulkadiroglu et al., 2017).

This conventional accommodation of quasi-experimental instrument assignment has several related drawbacks. First, as with linear IV, equations like (6)–(7) impose strong restrictions on unobserved treatment effect heterogeneity and non-random selection: though unobserved outcome heterogeneity ε_i is potentially correlated with unobservable determinants of institutional choice η_i, both are assumed to be independent of the non-experimental variation in X_i. Second, the addition of controls complicates the minimum distance interpretation of $\hat{q}_j^{MLE}$ and thus makes it difficult to characterize maximum likelihood estimates as simple extrapolations of quasi-experimental variation. Third, in practice it may be difficult or even intractable to obtain such estimates as the number of institutions, instruments, or controls in X_i grow moderately large – just estimating the first-stage

[13] In practice, Geweke et al. (2003) further restrict the instruments to be institution-specific, with $\pi_{j\ell} = 0$ for $\ell \neq j$, and impose sphericality among the latent selection indices.

equation (7) then requires fitting a high-dimensional multinomial choice model, the practical difficulties of which are well-known (Hausman and Wise, 1978; McColloch and Rossi, 1994).[14] Intuitively, computational difficulty stems from the fact that the data likelihood is no longer a function of a small number of non-parametric moment estimates; one must in general integrate over the unobservables in (6)–(7) for each observation separately. Controlling for observables flexibly with various transformations of X_i may further introduce an incidental parameters problem, yielding inconsistent maximum likelihood estimates.

I next discuss an alternative approach to quasi-experimental quality estimation that relaxes the constant effects assumption of linear models while overcoming each of these issues with nonlinear IV estimation. The key insight is that even when Z_i is only unconfounded conditionally on a (potentially rich) control vector X_i, one may still estimate the mean selected outcomes and choice probabilities used by experimental IV quality estimators. By directly applying different structural restrictions on $\mathcal{U}_i$, rather than specifying a likelihood for the observed data $\mathcal{Y}_i$, one may leverage the same kinds of intuitive and tractable extrapolations of reduced-form variation, despite the more complex quasi-experimental design.

2.3 Quasi-Experimental Quality Estimation

The quasi-experimental setting is one in which instrument assignment is as-good-as-random with respect to an individual's vector of potential outcomes and choices $\mathcal{U}_i$, given the control vector X_i:

Assumption 1 (*Quasi-experimental instrument assignment*): $\mathcal{U}_i \perp\!\!\!\perp Z_i \mid X_i$.

Under such conditional unconfoundedness, identification of mean selected outcomes and choice probabilities follows from the instrument propensity scores $p_\ell(x) = Pr(Z_{i\ell} = 1 \mid X_i = x)$. Provided $Pr(p_\ell(X_i) > 0) = 1$,

$$E[Y_{ij}|D_{ij\ell} = 1] = \frac{E[Y_i D_{ij} Z_{i\ell}/p_\ell(X_i)]}{E[D_{ij} Z_{i\ell}/p_\ell(X_i)]}, \tag{8}$$

and

$$Pr(D_{ij\ell} = 1) = E\left[D_{ij} Z_{i\ell}/p_\ell(X_i)\right]. \tag{9}$$

Here $Pr(p_\ell(X_i) > 0) = 1$ ensures that each individual is subject to some risk of assignment to each instrument value, so that the reduced-form moments (8)–(9) capture potential outcomes and treatments in the full population of interest.[15]

[14] This complexity is the basis for the Bayesian inference approach of Geweke et al. (2003), which specifies a prior distribution for the parameters of (6)–(7) and further restrictions.

[15] Proofs of (8)–(9) follow similarly to the proofs in footnote 6. Note that $p_\ell(x)$ is constant in the experimental setting, simplifying both population and sample analogues of (8)–(9) to the previous expressions.

In this way, the instrument propensity scores reduce the complexity of quasi-experimental designs. In some settings, like with the school quality estimation of Angrist et al. (2017), the instruments are determined by a known stochastic mechanism such that the $p_\ell(\cdot)$ are either known or can be simulated to arbitrary precision (e.g., Abdulkadiroglu et al. (2017)). More generally, propensity scores may be non-parametrically estimated from observations of $(Z_i', X_i')'$. Sample analogues of equations (8)–(9) based on these estimates reveal patterns in institutional selection and outcomes without restricting the structural relationship between $\mathcal{U}_i$ and X_i.

Per the intuitive discussion of Figure 1, the reduced-form variation may be enough to non-parametrically estimate quality from rich quasi-experimental data. To formalize this approach, consider a hierarchical data-generating process in which the choice probabilities $G_{j\ell}$ and mean selected outcomes $H_{j\ell}$ for $\ell = 1, \ldots, L$ are themselves drawn *iid* from a j-specific distribution. In the bivariate probit model (4)–(5), for example, this hierarchical structure would arise when the first-stage parameters $\pi_{j\ell}$ are *iid* random coefficients. Without restricting the conditional expectation $E[H_{j\ell} \mid G_{j\ell} = p] = h_j(p)$, quality is again given by its intercept at one: $q_j = \lim_{p \to 1} h_j(p)$. It follows that consistent non-parametric quality estimation is possible given sufficient choice probability variation near one and consistent estimates of the reduced-form moments. For instance, one might estimate the quality of institution j by the constant term in a local linear regression of $\hat{H}_{j\ell}$ on $1 - \hat{G}_{j\ell}$, for $\ell = 1, \ldots, L$. Appendix B.3 derives formal conditions for the consistency of this approach, building on Heckman (1990), Heckman et al. (2008), and Andrews and Schafgans (1998); in short, it requires the number of observed instrument values to grow with the sample, a positive choice probability density at one, and uniformly consistent $(\hat{H}_{j\ell}, \hat{G}_{j\ell})$. In the hospital application, one might thus obtain a good non-parametric estimate of quality given many ambulance companies, each serving many patients, with some companies willing to take virtually all patients to a given hospital j.[16]

As with conventional linear and nonlinear IV, structural restrictions can bridge the gap posed by more limited quasi-experimental variation. To formalize this approach, consider a restriction of the $\mathcal{U}_i \sim F(u; \theta_0)$ for a distribution function $F(\cdot)$ and unknown parameter vector θ_0. The parameter vector may be finite-dimensional, as in the bivariate probit model (4)–(5), or infinite-dimensional; the latter nests more limited restrictions like constant effects. For a given j, we partition $\theta_0 = (\bar{\theta}_0', \tilde{\theta}_0')'$ where $\bar{\theta}_0$ is finite dimensional, has a leading element of q_j, and is sufficient to parameterize some fixed subset of mean selected outcome and choice probabilities. That is, letting $M(\theta)$ be a function with rows $E_\theta[Y_{ik} \mid D_{ik\ell} = 1]$ and $E_\theta[D_{ik\ell}]$ for some (k, ℓ), where $E_\theta[\cdot]$ denotes expectations taken under the model $\mathcal{U}_i \sim F(u; \theta)$, the partition is such that for all $\bar{\theta} \in \bar{\Theta}$ and $\tilde{\theta}_a, \tilde{\theta}_b \in \tilde{\Theta}$ in some

[16] While unrealistic in the hospital application, Arnold et al. (2020) show how this approach can be used to estimate the average treatment effect of bail release from quasi-experimental judge assignment in NYC. This setting involves over 250 judge assignments, each comprising over 100 cases, with many choice probabilities estimated close to one.

parameter spaces containing $\bar{\theta}_0$ and $\tilde{\theta}_0$, $M((\bar{\theta}', \tilde{\theta}_a')') = M((\bar{\theta}', \tilde{\theta}_b')')$.[17] We then say that quality is identified under the following condition:

Assumption 2 (*Quality Identification*): For all $\theta = (\bar{\theta}', \tilde{\theta}')'$ with $\bar{\theta} \in \bar{\Theta}$ and $\tilde{\theta} \in \tilde{\Theta}$, if $M(\theta) = M(\theta_0)$, then $\bar{\theta} = \bar{\theta}_0$.

Intuitively, Assumption 2 holds when the structure placed on potential outcome and choice heterogeneity by $\mathcal{U}_i \sim F(u;\theta)$ is sufficient to narrow to a singleton the set of parameter subvectors $\bar{\theta}_0$ which rationalize variation in mean selected outcomes and choice probabilities in $M(\theta_0)$. The next subsection discusses a family of restrictions, generalizing the bivariate probit model (4)–(5), which is shown to satisfy this high-level condition.

Assumption 2 motivates a minimum distance quality estimator that matches quasi-experimental moment estimates to functions of the parameters implied by the distributional restriction:

$$\hat{q}_j^{MD} = \arg_1 \min_{\theta} \left(\hat{M} - M(\theta)\right)' \hat{A} \left(\hat{M} - M(\theta)\right) \tag{10}$$

for some $\hat{A} \xrightarrow{p} A$ (both symmetric positive-definite), where $\hat{M}$ estimates $M(\theta_0)$. I consider $\hat{M}$ with rows given by sample analogues of equations (8)–(9), of the form

$$\hat{H}_{k\ell} = \frac{\frac{1}{N}\sum_{i=1}^N Y_i D_{ik} Z_{i\ell}/\hat{p}_\ell(X_i)}{\frac{1}{N}\sum_{i=1}^N D_{ik} Z_{i\ell}/\hat{p}_\ell(X_i)} \tag{11}$$

and

$$\hat{G}_{k\ell} = \frac{1}{N}\sum_{i=1}^N \frac{D_{ik} Z_{i\ell}}{\hat{p}_\ell(X_i)}, \tag{12}$$

where $\hat{p}_\ell(\cdot)$ is a first-step estimate of the propensity score $p_\ell(\cdot)$. Following Hirano et al. (2003), I take these as given by the method of sieves (Geman and Hwang, 1982), using a series logit estimator.[18]

Appendix B.4 establishes the consistency and asymptotic normality of the semi-parametric quality estimator $\hat{q}_j^{MD}$ under Assumptions 1-2 and additional regularity conditions. Specifically, it proves the following result.

Proposition 1: Suppose Assumptions 1-2 hold, $M(\theta)$ is continuously differentiable in $\bar{\theta}$ over compact $\bar{\Theta}$ with a bounded Jacobian, and Assumptions B1-B4 in Appendix B.4 hold. Then

[17] More generally, we may allow $M(\theta)$ to contain moments of the form $E_\theta[f(Y_{ij}, X_i) \mid D_{ij\ell} = 1]$, for some measurable function $f(\cdot)$. Under Assumption 1 these elements of $M(\theta_0)$ are then identified by $\frac{E[f(Y_i,X_i)D_{ij}Z_{i\ell}/p_\ell(X_i)]}{E[D_{ij}Z_{i\ell}/p_\ell(X_i)]}$. I restrict to the case of $f(y,x) = y$ for simplicity of exposition and because these moments are sufficient for estimating quality both under the constant effects restriction of linear IV and in the elliptical parameterizations considered in the next section.

[18] The series logit approach is attractive for estimating conditional expectations of the binary $Z_{i\ell}$. See Abadie (2003) for an alternative approach based on the linear approximation theory of Newey (1994). When X_i is high-dimensional, the propensity scores moments may be estimated by the double machine learning approach of Chernozhukov et al. (2017). Extensions of Proposition 1 follow with appropriate regularity conditions replacing Assumptions B1-B4.

$\hat{q}_j^{MD} \xrightarrow{p} q_j$ and

$$\sqrt{N}(\hat{q}_j^{MD} - q) \Rightarrow N\left(0, ((\nabla' A \nabla)^{-1} \nabla' A V A \nabla (\nabla' A \nabla)^{-1})_{(1,1)}\right), \tag{13}$$

where $\sqrt{N}(\hat{M} - M(\theta_0)) \Rightarrow N(0, V)$ and ∇ denotes the Jacobian of $M(\theta)$ at θ_0. There moreover exists a consistent estimator of the asymptotic variance of $\hat{q}_j^{MD}$.

The appendix proof combines the non-parametric approximation theory of Hirano et al. (2003) with classic results on minimum distance estimation (Newey and McFadden, 1994). The appendix regularity conditions B1-B4 are adapted from the former; they importantly restrict the support of the propensity scores to be bounded above zero (permitting identification of each quasi-experimental moment in the full population of interest), restrict the rate at which terms are included in the series approximation of $p_\ell(x)$, and impose other smoothness, support, and rate conditions.

Equation (13) shows that given a smooth moment function $M(\theta)$, the asymptotic variance of $\hat{q}_j^{MD}$ is a function of the limiting moment-weighting matrix A, the asymptotic variance of the moment vector V, and the Jacobian ∇ of the function linking the reduced-form moments to the identified structural parameters $\bar{\theta}$. Appendix B.4 also derives a consistent estimator for V, which accounts for the first-step error from estimating the $p_\ell(\cdot)$; a consistent estimate of the asymptotic variance of $\hat{q}_j^{MD}$ is given by combining this estimate with $\hat{A}$ and the Jacobian evaluated at the $\hat{\theta}$ solving (10).

In simple experimental data, the minimum distance quality estimator may coincide with conventional IV estimators. That is, when the control vector X_i is constant, the elements of $\hat{M}$ will reduce to the non-parametric moment estimates discussed in Section 2.2; appropriate choices of restrictions $F(\cdot)$ and weight matrices $\hat{A}$ may then, per Appendices B.1–B.2, equate $\hat{q}_j^{MD}$ with different linear or nonlinear IV estimates. In general, however, $\hat{q}_j^{MD}$ imposes different restrictions than conventional estimators by not modeling the dependence of potential outcomes and selection patterns on the design controls. Rather, the controls are used to non-parametrically reweight the observed data via the propensity score estimates, allowing partial restrictions on the latent variable distribution to be imposed directly. As with the analyses of equation (3) or Figure 1, these restrictions can be viewed and evaluated in terms of their implied extrapolation of the quasi-experimental moments.

This semi-parametric approach is also likely to be more computationally tractable than conventional nonlinear IV estimation, particularly in settings with many institutions J, instrument values L, and controls in X_i. Each element of the $\hat{M}$ vector is computed from sample moments involving at most $L-1$ linearly independent propensity score estimates, which do not depend on the structural parameters and can be computed separately. Given $\hat{M}$, evaluating the minimum distance objective requires computing at most $(J-1)L$ nonlinear functions of each candidate parameter vector $\bar{\theta}$.[19] Importantly, these are not direct functions of the data, so the difficulty of nonlinear computation

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

does not increase with the sample size N. For some restrictions of $\mathcal{U}_i$, including those used below, $M(\theta)$ takes a form that is straightforward to evaluate by standard statistical software. More generally, $M(\theta)$ can be evaluated to arbitrary precision by simulation, whenever it is possible to sample from $F(\cdot)$.[20] This ease of computation also makes it straightforward to verify the sensitivity of quality estimates to different structural extrapolations in the second step, holding the first-step reduced-form moment estimates fixed.

Applying this semi-parametric approach requires specifying a structural restriction on $\mathcal{U}_i$. An ideal restriction balances the trade off between flexibility (a $F(\cdot)$ that accommodates a wide variety of institutional treatment effects and unobserved selection patterns) and tractability (a $M(\cdot)$ that satisfies Assumption 2). I next discuss a family of parameterizations, extending the earlier bivariate probit model, which is likely to strike such a balance when quality is measured with binary outcomes.

2.4 Elliptical Restrictions for Binary Potential Outcomes

I consider restrictions on $\mathcal{U}_i$ in which binary potential outcomes and institutional choice indicators are linked by a multivariate elliptical copula. Such models allow for a flexible degree of treatment effect heterogeneity and unobserved selection; in particular, they do not impose constant or monotone treatment effects and allow individuals to sort on such heterogeneity. Despite this flexibility, quality is identified in these models when the number of instrument values is at least as large as the number of institutions and the instruments generate unique choice probabilities. The family of models generalizes the bivariate probit example (4)–(5), with the resulting minimum distance quality estimates sharing its geometric interpretation of simple mean selected outcome extrapolations.

Common to this family is a first-stage monotonicity restriction, as with the identification of local average treatment effects and related parameters (Imbens and Angrist, 1994; Heckman et al., 2006):

Assumption 3 (*Monotonicity*): $\forall j$, ℓ,m, $Pr(D_{ij\ell} \geq D_{ijm}) = 1$ or $Pr(D_{ij\ell} \leq D_{ijm}) = 1$.

Under Assumption 3, a change in assignment from $Z_{i\ell} = 1$ to $Z_{im} = 1$ which makes the selection of institution j strictly more likely for any mass of individuals cannot make the selection of institution j strictly less likely for any other mass of individuals. When $Z_{i\ell}$ indicates the quasi-experimental assignment of ambulance company ℓ to patient i, for example, Assumption 3 restricts heterogeneity in the referral preferences of different ambulance companies. Particularly, it requires differences in these preferences to be fixed over patients with different potential outcomes and hospital preferences. I discuss the appropriateness of this restriction in the next section, in the context of the application.

A consequence of first-stage monotonicity is the significant reduction in the dimensionality of any model for $\mathcal{U}_i$. In general, an internally consistent model for institutional choice (imposing

[20] Estimates of $AVar(\hat{q}_j^{MD})$ are also simple to compute, as shown in Appendix B.4.

$\sum_j D_{ij\ell} = 1$ for each i and ℓ) is given by the maximization of some latent utility indices $u_{ij\ell}$:

$$D_{ij\ell} = \mathbf{1}[u_{ij\ell} > u_{ik\ell}, \forall k \neq j]. \tag{14}$$

Since monotonicity over multiple unordered treatments is observationally equivalent to assuming additively separable indices $u_{ij\ell} = \pi_{j\ell} + \eta_{ij}$ (Heckman and Pinto, 2018), modeling the distribution of $\mathcal{U}_i$ under Assumption 3 reduces to modeling the much lower-dimensional vector $((Y_{ij}, \eta_{ij})_{j=1}^J)'$.

A natural parameterization of binary potential outcomes also involves latent indices, h_{ij}:

$$Y_{ij} = \mathbf{1}[h_{ij} > 0]. \tag{15}$$

In the hospital application we might interpret h_{ij} as the latent health of 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, normalized to zero. As with the bivariate probit example, and in contrast to Geweke et al. (2003), here we do not impose that $h_{ij} = \beta_j + \varepsilon_i$. Relaxing such separability allows a flexible degree of institutional comparative advantage: a move from hospital k to hospital j may improve the health of some patients while worsening the health of others.[21]

A final parametric restriction defines $F(\cdot)$: that the latent utility indices are distributed continuously and joint-elliptically with the latent outcome index of institution j:

Assumption 4 (*Ellipticity*): Equation (14) holds for all (k, ℓ), and (15) holds for a given j. The density of $\mathcal{V}_i = ((((u_{ik\ell})_{k=1}^J)_{\ell=1}^L), h_{ij})$ is, for a known nonnegative $g(\cdot)$, proportional to

$$|S|^{-1/2} g((\mathcal{V}_i - s)' S^- (\mathcal{V}_i - s)), \tag{16}$$

for some s and positive semi-definite matrix S, where S^- gives the generalized inverse of S.

A large set of elliptical distributions satisfy this assumption, including the multivariate normal ($g(u) = \exp(-u/2)$), logistic ($g(u) = \frac{\exp(-u)}{(1+\exp(-u))^2}$), and Student's t ($g(u) = (1 + \frac{u}{m})^{-(n+m)/2}$ for some $n, m \in \mathbb{N}^+$). A useful feature of this family is that imposing Assumptions 3-4 is generally without observational loss for the marginal distributions of potential outcomes and institutional substitution patterns. As shown in the proof to the following proposition, any set of positive choice probabilities across institutions can be rationalized by any such $g(\cdot)$, and the marginal distribution of Y_{ij} is clearly unrestricted by assuming a marginally elliptical h_{ij} provided $q_j \in (0, 1)$.

An elliptical joint distribution of h_{ij} and, under monotonicity, $(\eta_{i1}, \ldots, \eta_{iJ})'$ nevertheless yields a tractable model for the dependence between heterogeneous potential outcomes and institutional choice, in that it implies a low-dimensional parameterization of quality, mean selected outcomes, and choice probabilities. This $\bar{\theta}_0$ is furthermore identified by a relatively small number of moments,

[21] Importantly, equations (14)–(15) and Assumption 8 do not rule out single-index outcome heterogeneity: the health indices across different institutions j and k may be perfectly correlated, such that h_{ij} has a representation of $\beta_j + \varepsilon_i$.

as formalized in the following result:

Proposition 2: Suppose Assumptions 3-4 hold for some j, such that $u_{ij\ell} = \pi_{j\ell} + \eta_{ij}$, and suppose the submatrix of S corresponding to $(\eta', h_{ij})'$ is positive definite. Then if $L \geq J$ and the choice probability vectors $G_\ell = (Pr(D_{i1\ell} = 1), \ldots, Pr(D_{i,J-1,\ell} = 1))'$ are unique across ℓ, Assumption 2 holds with $M(\theta_0) = ((G_\ell, H_{j\ell})_{\ell=1}^L)'$.

The proof to Proposition 2, given in Appendix B.5, shows that while a large number of parameters enter θ_0 under Assumptions 3-4 (including the $J \times L$ utility shifters $\pi_{j\ell}$ and the $(1+J) + \frac{(1+J)(2+J)}{2}$ relevant elements of s and S), the $L(J-1)+L$ quasi-experimental moments in $M(\theta_0)$ are uniquely given by the quality of institution j and another $(L+1)(J-1)$ elements of a transformed parameter vector $\bar{\theta}_0$. With $L \geq J$ the order condition for identification is thus satisfied, while the rank condition is shown to hold when each instrument induces a unique vector of choice probabilities. Thus, as with linear IV, identification follows from having as many instruments $Z_{i\ell}$ as treatments D_{ij}.

To illustrate this result, consider the elliptical generalization of the bivariate probit model (4)-(5). Here there are two outcome coefficients β_j, $2L$ selection coefficients $\pi_{j\ell}$, and 10 shape parameters in S. Nevertheless, the $2L$ choice probabilities and mean selected outcome in institution 1's $M(\theta_0)$ moment vector simplify to known functions of only $2 + L$ parameters:

$$Pr(\pi_{1\ell} + \eta_{i1} > \pi_{2\ell} + \eta_{i2}) = Pr(\bar{\pi}_{1\ell} > u_i) \tag{17}$$

$$Pr(\beta_1 + \varepsilon_{i1} > 0 \mid \pi_{1\ell} + \eta_{i1} > \pi_{2\ell} + \eta_{i2}) = Pr(Q^{-1}(q_1) > \rho_1 u_i + \sqrt{1-\rho_1^2} v_i \mid \bar{\pi}_{1\ell} > u_i) \tag{18}$$

where $(u_i, v_i)'$ are spherically distributed random variables with density generator $g(\cdot)$, $Q^{-1}(\cdot)$ is a known inverse cumulative distribution function, $\bar{\pi}_{1\ell} = \frac{\pi_{1\ell} - \pi_{2\ell}}{\sqrt{S_{\eta 2} + S_{\eta 1} - 2S_{\eta 1 \eta 2}}}$, and $\rho_1 = \frac{S_{\varepsilon 1 \eta 1} - S_{\varepsilon 1 \eta 2}}{\sqrt{S_{\varepsilon 1}(S_{\eta 2} + S_{\eta 1} - 2S_{\eta 1 \eta 2})}}$. Proposition 2 shows that this representation generalizes when $J > 2$: each additional institution generates an additional parameter like ρ_1, capturing the relationship between outcomes and institutional selection, and one additional set of first stage parameters analogous to the $\bar{\pi}_{1\ell}$. The latter are identified by the choice probability vectors, with the former given from the relationship between choice probabilities and institution 1's mean selected outcomes.

3 Estimating Hospital Quality

3.1 Data and Observational RAMs

To apply the quasi-experimental approach to hospital quality estimation, I construct a nationally representative sample of elderly (over 65) Medicare patients referred to a hospital by ambulance

for an emergency condition in 2010-2012.[22] The data are drawn from a 20% random sample of administrative fee-for-service 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 subsequent patient mortality. I augment these data with definitions of hospital service areas (HSAs) from the Dartmouth Atlas of Health Care and various hospital characteristics from other CMS sources. In defining admissions, I follow the standard practice in CMS risk-adjustment of attributing outcomes to a patient's initial ("index") acute-care hospital, ignoring all subsequent transfers or readmissions. I restrict attention to emergency admissions, both because these are most affected by ambulance company referral variation and because they produce the most variation in short-run mortality.[23] Following Card et al. (2009) and Doyle et al. (2015), these nondeferrable conditions have an average weekend admissions rate close to 2/7. Appendix A describes the sample construction in detail.

Table 1 summarizes the resulting sample of 405,172 patients, who were transported by one of 9,590 ambulance companies to one of 4,821 hospitals for one of 29 emergency conditions. 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 (AMI), heart failiure, or pneumonia, though with the goal of extending such models to a broader patient population (Krumholz et al., 2006).[24] Panel A of Table 1 shows that circulatory and respiratory conditions make up 42% of all emergency admissions, with the remainder split between digestive conditions like gastroenteritis (7%), injuries like hip fracture (18%), and other conditions (34%). The final column summarizes mortality over 30 days after admission, which is the typical horizon for mortality RAMs. The overall 30-day survival rate is 83%, with survival rates as low as 78% for patients with respiratory conditions and as high as 93% for those with injuries.

Panel B of Table 1 summarizes the distribution of patients, ambulances, hospitals, and 30-day mortality across hospital markets (HSAs) of different sizes. As discussed below, I exploit local quasi-experimental variation in ambulance company assignment to construct within-HSA estimates of hospital quality. Table 1 shows that the number of possible within-market comparisons has a skewed distribution, with around half of all hospitals operating in their own single-hospital HSA.

[22] Three years is the typical window for conventional hospital RAMs (YNHHSC/CORE, 2013), and 2012 is the latest year of data available to me. I explore time-series patterns in hospital quality in Section 4.

[23] 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 Appendix A.

[24] 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, a more complex model would be required to causally attribute variation in these outcome to hospital quality. I leave this important issue for future work.

These markets tend to be small, with an average of only 61 admitted patient observations. In contrast, the remaining 695 multi-hospital HSAs average 366 observations and represent two-thirds of all admissions. The quality analysis will tend to focus on these larger markets. Nevertheless, as on average roughly 30% of patients are referred to hospitals outside of their HSA, even in single-hospital HSAs quasi-experimental referral variation will have scope to correct for unobserved patient sorting.

I first use the analysis sample to construct observational hospital quality predictions, following standard CMS risk-adjustment methodology (e.g., YNHHSC/CORE, 2013). The observational estimates come from an additively separable latent index model for potential 30-day survival:

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

where $\alpha_j - \alpha_k$ denotes the constant effect of hospital j versus k on patient health. The residual variation in patient health is further decomposed into an observable and unobservable component: $\varepsilon_i = \gamma' W_i - \nu_i$, where W_i is a vector of *risk-adjusters* (demographics and comorbidities) and $\nu_i \mid W_i$ is assumed to follow a standard Gumbel distribution. Thus, in these conventional RAMs, a patient's observed mortality is given by

$$Y_i = \mathbf{1}[\textstyle\sum_j \alpha_j D_{ij} + \gamma' W_i \geq \nu_i]. \tag{20}$$

Identification of the parameters of (20) follows from a selection-on-observables assumption: hospital choice is unrelated to the unobserved component of health, conditional on the risk-adjustors ($\nu_i \perp\!\!\!\perp D_i \mid W_i$). I use this assumption to estimate condition-specific RAM quality indices α_{jc} in each of the five samples summarized in Panel A of Table 1, including patient age, sex, and 17 comorbidity indicators in the risk-adjuster vector W_i. To obtain an overall RAM quality index, I then average these coefficients by the overall patient condition mix. Appendix A describes this estimation procedure in more detail.

RAM estimates suggest a significant amount of heterogeneity in hospital quality. With an estimated standard deviation in quality indices of around 0.04, a hospital one standard deviation better in terms of its impact on latent patient health is predicted to increase 30-day survival by roughly 0.5 percentage points. At the same time, the observational RAMs appear to leave most mortality variation unexplained. This result 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 the variance in survival following a circulatory or respiratory diagnosis is explained by a patient's hospital, admitting condition, and admission year in a very basic specification (which I label "RAM1"). This reduction is somewhat smaller for digestive conditions and injuries, and around twice as large (14%) for other conditions. Relatively more sophisticated specifications which add patient demographics ("RAM2") and comorbidities ("RAM3") account for an additional 4% of circulatory and respiratory survival variance, with similarly modest declines elsewhere.

If the residual determinants of mortality are idiosyncratic with respect to hospital selection, predictions from observational RAMs may provide reliable measures of hospital quality. However, to the extent the variance in 30-day survival outcomes can be further reduced by observable admission correlates, such as a patient's ambulance company, the selection-on-observables assumption is likely to be violated and the observational RAMs are likely to be biased. Appendix B.7 formalizes this argument by developing instrument-based tests of nonlinear RAM validity, extending earlier methods for validating linear VAMs in education settings (Kane and Staiger, 2008; Chetty et al., 2014a; Angrist et al., 2016). Applying these tests to the hospital setting yields a decisive rejection of the null (see Appendix Table A2).[25] This finding suggests pervasive selection bias in the observational RAMs, as well as scope for leveraging the yet-unused variation in ambulance company referral patterns. I next describe this variation and how it is used.

3.2 Quasi-Experimental Estimation

Doyle et al. (2015) first note that centralized policies of rotational and simultaneous dispatch may generate plausibly exogenous ambulance company assignment. This quasi-experimental assignment may further affect the hospital admissions of otherwise identical emergency patients, via the expression of company referral preferences. Doyle et al. (2015) and Doyle et al. (2017a,b) use this variation to instrument for the average spending or quality rating of a patient's hospital, in linear IV regressions of mortality or readmission outcomes. I first apply this variation to the semi-parametric hospital quality estimation framework in Sections 2.3 and 2.4. The next section discusses how I then combine this quasi-experimental data with the foregoing observational quality measures.

Table 2 establishes the plausibility of quasi-random ambulance company assignment by comparing patients in the analysis sample who live in the same ZIP code but who are assigned to ambulance companies which are likely to refer to different hospitals. Specifically, I compute the ZIP code distance between each ambulance company's main office and each nearby hospital, labeling companies as likely to deliver patients to a high- or low-ranked hospital if their closest hospital is in the first or fourth quartile of observational RAM predictions in the HSA. I then regress patient characteristics on either these two group indicators (with group means reported in columns 1 and 2 of Table 2) or on the assigned company's predicted hospital RAM itself (with the regression coefficient reported in column 4), along with a full set of ZIP code fixed effects. Individual and joint p-values for tests that patient characteristics are not systematically correlated with these dimensions of ambulance company heterogeneity are reported in columns 3 and 5 of the table.

The first row of Table 2 shows that patients who are assigned to ambulance companies based near

[25] I verify in Appendix B.7 that these tests also reject more narrow replications of the official 2013 CMS models for patients admitted for AMI, pneumonia, and heart failure.

a high-ranked hospital are indeed more likely to be referred to a high-ranked hospital. Nevertheless, panel A shows that such patients are statistically indistinguishable from other patients in terms of their demographics, the location of their emergency, and their admitting condition, while panel B shows that patients are balanced on a host of comorbidity indicators describing their pre-emergency medical history. These findings are 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 shows that company assignment is balanced with respect to a set of ambulance services performed after assignment but before hospitalization, such as the application of intravenous medication, miles traveled in excess of the most direct route to the hospitals, or whether the patient was treated by paramedics. This finding supports the implicit exclusion restriction of using ambulance company indicators as instruments for hospital admissions: if companies are not similarly trained or skilled, systematic differences in medical treatment en route to a hospital may directly affect patient survival outcomes. It is worth emphasizing that this is a weaker exclusion restriction than in Doyle et al (2015; 2017a; 2017b), where company assignment is assumed to only affect outcomes via the average spending or quality rating of a patient's hospital. This stronger assumption could be violated if hospital quality differs in ways that are not adequately captured by these spending or observational quality ratings.[26]

In bringing the Doyle et al. (2015) instruments to the previous econometric framework, it is important to recognize that ambulance-induced referral variation is inherently local: companies are only assigned within service areas, and patients are rarely transported long distances in an emergency. This institutional feature has two implications. First, the instrument propensity scores $p_\ell(\cdot)$, which here give the probability of assignment to each ambulance company given patient location, are only plausibly bounded away from zero within service areas. Thus, the preceding econometric framework should be thought to apply within HSAs, with J denoting the number of hospitals serving each local healthcare market. Second, differences in hospital quality estimated from ambulance-induced referral variation within HSAs need not be causally interpreted across different markets. While the foregoing framework permits the ranking of local alternatives and characterization of local selection patterns, additional assumptions or quasi-experimental variation would be needed to compare the quality and selection of distant hospitals.

I first use the ambulance company instruments to estimate a set of ambulance company propensity scores by series logit estimation for a large number of HSAs with moderate quasi-experimental

[26]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 in Table 2).

variation. Specifically, within each HSA with at least 50 patients over the three sample years, I separately approximate a set of instrument propensity score functions $Pr(Z_{i\ell} = 1 \mid X_i = x)$ for each ambulance company assigned to at least 10 patients. Here, again, $Z_{i\ell}$ indicates assignment to ambulance company ℓ, while the control vector X_i contains the distance of patient i to the home office of each ambulance company in the HSA. The series logit estimator uses a power series of order three to flexibly control for patient location through X_i. For robustness, I also control linearly for the vector of RAM controls W_i though, consistent with Assumption 1 and the balance in Table 2, the following results are essentially unchanged when these are excluded (see Appendix Table A6).[27] To keep the subsequent minimum distance procedure just-identified and reduce the scope for finite sample bias, I consider only the J largest companies in an HSA (recalling the order condition for identification in the elliptical model), where J is the number of hospitals in an HSA serving at least 25 patients and ambulance volume is estimated in the raw sample of emergency claims.[28]

With these restrictions, I apply equations (11) and (12) to estimate a set of mean selected outcomes and choice probabilities for 2,082 hospitals operating in 968 HSAs. Figure A2 shows that 1,677 of these hospitals operate in a HSA with at least two active hospitals, with the remainder located in a single-hospital HSA (recall that since some patients are referred outside of their HSA, such hospitals are also candidates for quasi-experimental quality estimation). While the regions represented by the quasi-experimental data tend to be larger and with more hospitals – Figure A2 shows that most excluded hospitals operate in single-hospital HSAs, for example – Appendix Table A1 shows that their patient and insitutional observables are broadly representative of the full analysis sample.

Figure 3 summarizes the reduced-form variation by plotting the joint distribution of differences in mean selected outcomes estimates $\hat{H}_{j\ell}$ and choice probability estimates $\hat{G}_{j\ell}$ for each of the 2,082 hospitals j in the quasi-experimental sample. Specifically, it shows a binned scatter plot of $\hat{H}_{j,jMax} - \hat{H}_{j,jMin}$ against $\hat{G}_{j,jMax} - \hat{G}_{j,jMin}$, where $jMax$ and $jMin$ denote the ambulance companies with the largest and smallest $\hat{G}_{j\ell}$ for each j. The distribution of points along the horizontal axis, representing percentiles of the choice probability estimate differences $\hat{G}_{j,jMax} - \hat{G}_{j,jMin}$, thus summarizes one dimension of first-stage variation in hospital choice. The corresponding distribution of points along the vertical axis shows how the average survival outcomes of admitted patients at a hospital tend

[27] Given the sometimes large number of controls and the sometimes small estimation samples, maximum likelihood estimates of propensities scores occasionally fail to converge with the full series logit estimator. In these cases, which represent less than 5% of the sample, I sequentially drop the RAM controls and higher-order distance terms until convergence is achieved. As a further regularization, I trim estimated propensity scores above 0.95 or below 0.05; results are qualitatively unchanged in adjustments to both procedures.

[28] Appendix Figure A1 plots the distribution of propensity score estimates. While many individuals are predicted to have a low probability of assignment to certain ambulance companies, there is a wide distribution of such scores with significant mass everywhere between zero and one. In contrast, the red dashed line in this figure plots the distribution of propensity score estimates that are obtained when larger hospital referral regions (HRRs) are instead used to define local markets. In this case, virtually all individuals have propensity scores near zero, with considerable mass at the trimmed lower bound of 0.05. This skewness supports the use of more narrow HSAs in the analysis.

to change as more patients are quasi-experimentally referred by ambulance. The joint distribution of points shows how this change relates to the ambulance-induced change in hospital selection.

The average difference in estimated choice probabilities is 0.34, indicated by the dashed vertical line in Figure 3, with more than two-thirds of hospitals having a $\hat{G}_{j,jMax} - \hat{G}_{j,jMin}$ of at least 0.2. This result suggests significant first-stage admissions variation throughout the sample, arising from meaningful heterogeneity in ambulance company referral preferences.[29] At the same time, only 11% of hospitals have a maximal estimated choice probability (i.e. $\hat{G}_{j,jMax}$) 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 discussed in Section 2.3, justifying the additional structure developed in Section 2.4.

Figure 3 also shows that the average estimated mean selected outcome difference is negative (-1.2 percentage points), and increasingly so with larger differences in estimated choice probabilities. A simple regression of $\hat{H}_{j,jMax} - \hat{H}_{j,jMin}$ on $\hat{G}_{j,jMax} - \hat{G}_{j,jMin}$, indicated by the solid line in the figure, yields a coefficient of -0.03 with a robust standard error of 0.01. Both of these reduced-form facts suggest positive patient selection in hospital admissions. To see this, consider a representative hospital for which $\hat{H}_{j,jMax} - \hat{H}_{j,jMin} < 0$. This inequality suggests that $E[Y_{ij} \mid D_{i,jMax} = 1] < E[Y_{ij} \mid D_{i,jMin} = 1]$, or that the patients referred to hospital j by the ambulance company with a higher referral rate ($jMax$) tend to have lower survival rates than the patients referred by the more selective ambulance company ($jMin$). This pattern in turn suggests that the patients most likely to be referred to hospital j (those for which $D_{i,jMin} = 1$) tend to have higher potential outcomes at j compared to the patients who would only be referred to j when assigned the less selective ambulance company (those for which $D_{i,jMin} < D_{i,jMax}$), as in Figure 1. Furthermore, all else equal, one expects under positive selection that $E[Y_{ij} \mid D_{i,jMax} = 1] - E[Y_{ij} \mid D_{i,jMin} = 1]$ is more negative given a larger difference in selectivity $Pr(D_{i,jMax} = 1) - Pr(D_{i,jMin} = 1)$. This pattern is indeed suggested by the correlation in Figure 3.

To formally link this reduced-form variation to structural concepts of selection and quality, I apply the minimum distance estimator (10), again separately by HSA, using the elliptical model from Section 2.4. This model imposes first-stage monotonicity (Assumption 3), implying that differences in referral patterns do not systematically vary by unobserved patient heterogeneity. Doyle et al. (2015) use a similar monotonicity condition via 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.[30] Monotonicity is especially

[29] Appendix Figure A3 plots the distribution of first-stage F-statistics, testing the equality of choice probabilities across different ambulances for each hospital, against average minimum distance quality estimate standard errors. The median first-stage F-statistic is 10.1. Hospitals with lower first stage F-statistics tend to have higher quality standard errors; less weight will be placed on these estimates in the subsequent empirical Bayes procedure.

[30] A likely determinant of referral patterns is differential hospital distance: ambulance companies may prefer to send patients to the hospital based closest to their offices in order to minimize travel time and maximize local availability.

plausible in the relatively homogenous sample of emergency Medicare patients; differential treatment of uninsured patients by profit-driven ambulance companies, for example, is not a concern.

My benchmark specification uses a multivariate normal model of latent health and utility that satisfies Assumption 4. As shown in Appendix Table A6, however, similar results are obtained throughout with a fatter-tailed multivariate Student's $t(2)$ distribution. Proposition 2 shows that quality is identified in both models within HSAs with $L \geq J$ ambulance companies, where J is the number of institutional alternatives. Here J is the number of hospitals serving at least 25 patients over the three-year window, plus one for an combined fallback alternative for patients referred outside of the HSA or to a small hospital. With the model just-identified, quality estimates are obtained without specifying a weight matrix in equation (10), by solving closed-form expressions for the model-implied moment functions $M(\theta)$.

The dashed red curve in Figure 4 plots the distribution of the 2,082 minimum distance estimates of hospital quality index coefficients, $\beta_j = E[h_{ij}]$, where h_{ij} is the potential health index defined in (15). The average $\hat{\beta}_j$ is around 0.8, implying a typical hospital quality of $E[Y_{ij}] = \Phi(0.8) \approx 0.79$. That this quality average is below the typical hospital's observed admission rate of 0.83 is consistent with the reduced-form evidence of positive selection in Figure 3. When mean selected outcomes are declining in choice probabilities the unselected mean potential outcome of a hospital will be below the observed average outcome $E[Y_{ij} \mid D_{ij} = 1]$.[31]

Around this mean, Figure 4 displays wide variation in estimated hospital quality with a standard deviation in $\hat{\beta}_j$ of 0.8. 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 estimation error. I next develop a procedure to account for these different sources of variation and conduct a more fine-grained analysis of hospital quality and selection. This empirical Bayes approach combines quasi-experimental and RAM-based estimates to obtain more accurate quality posteriors.

3.3 Quality Posteriors

The minimum distance procedure yields, for a subset of hospitals j with sufficient quasi-experimental data, 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 (20) that are likely to be positively, but not perfectly, correlated with quality due to the sorting bias

[31] Appendix Figure A4 plots the distribution of minimum quality distance estimates themselves, along with a comparison distribution of linear IV quality estimates obtained by applying the mean selected outcome extrapolations discussed in Section 2.2 to the same reduced-form moments. Around 30 percent of linear IV estimates are outside of the logical bounds of zero and one (which the minimum distance procedure enforces), with 5th and 95th percentiles of -0.81 and 2.1. This finding reinforces the fact that the constant effects assumption underlying linear IV is inappropriate for binary mortality outcomes. Nevertheless, the two sets of quality estimates are positively correlated ($\rho = 0.23$).

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. This model specifies β_j and $\hat{\alpha}_j$ as being joint-normally distributed across hospitals, with unobserved quality differences allowed to be clustered within hospital markets. That is

$$\hat{\beta}_j = \beta_j + \iota_j = \kappa + \lambda\hat{\alpha}_j + \mu_{h(j)} + \upsilon_j + \iota_j, \tag{21}$$

where $\kappa + \lambda E[\hat{\alpha}_j]$ is the average hospital quality index (per Figure 4, around 0.8), $\mu_{h(j)}$ is a random qualitiy index shifter 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 asymptotic approximation in Proposition 1, the estimation error term ι_j can also be modeled as normally distributed, with a known covariance structure. Consistent estimation of the HLM's hyperparameters κ, λ, σ, and ϕ comes from an ordinary least squares (OLS) regression of quality index estimates $\hat{\beta}_j$ on RAM predictions $\hat{\alpha}_j$, while efficient estimates leverage a maximum likelihood (MLE) procedure that effectively weights the minimum distance estimates inversely to their variance.[32]

Table 3 reports OLS and MLE estimates of equation (21), where for ease of interpretation $\hat{\alpha}_j$ has been normalized to be of unit standard deviation. 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 additionally controls for patient comorbidities.

The observational measures from all three RAMs are predictive of the quality index estimates, with λ estimated between 0.11 and 0.13. The OLS estimates of these coefficients 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 point estimates, suggesting that the simple HLM specification (21) adequately captures the relationship between observational and true quality estimates across hospitals with different degrees of estimation error. Consistent with the graphical evidence in Figure 2 and the formal tests in Figure 3, including demographics and comorbidities barely increases the predictive

[32]Note that, as in Angrist et al. (2017), Chetty and Hendren (2018), and Finkelstein et al. (2017), the assumption of jointly normal quality and observational quality predictions places high-level restrictions on the underlying distribution of potential outcomes and utility that, while not inconsistent with Assumptions 1–4 (which implicitly condition on the β_j), may be difficult to formally "microfound." As usual when the assumption of normal υ_j and ι_j is violated, the posterior means (22) maintain an interpretation of best linear unbiased predictors of the true hospital quality indices β_j given the observational RAM prediction (allowing for HSA-clustered quality indices). See Bonhomme and Weidner (2019) for a recent discussion of misspecification in empirical Bayes modeling.

reliability of observational RAMs. OLS and MLE estimates of the residual variance parameters are similarly stable, with $\hat{\sigma} \approx 0.16$ and $\hat{\phi} \approx 0.17$. Together, these suggest around 35% of the national variation in quality indices β_j is found between HSAs, with 23% explained by observational RAM predictions and 41% left unexplained within HSAs. That the observational measures are predictive of true hospital quality (with a within-HSA R^2 of 0.36) is broadly consistent with the linear IV analysis of Doyle et al. (2017a).

I take column 6 of Table 3 as my preferred estimate of equation (21), though the remaining columns explore richer HLM specifications. Column 7, for example, includes the predictions of all three RAMs simultaneously, showing that the less sophisticated RAM1 and RAM2 estimates are insignificant predictors of quality once the most sophisticated RAM3 estimates are included. 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 additional regressors are small and not statistically significant while the residual variance falls little. This finding again suggests that the parsimonious HLM specification (21) is a reasonable approximation to the true conditional expectation.

I use the estimated HLM to generate empirical Bayes posterior predictions of hospital quality that, as in Angrist et al. (2017), Chetty and Hendren (2018), 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 (21) further allows the vector of estimates for each HSA to be jointly shrunk towards a random 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 given by

$$E[\beta_h|\hat{\alpha}_h, \hat{\beta}_h] = \Omega_h\hat{\beta}_h + (I_{J(h)} - \Omega_h)(\kappa + \lambda\hat{\alpha}_h) \tag{22}$$

$$Var(\beta_h|\hat{\alpha}_h, \hat{\beta}_h) = (I_{J(h)} - \Omega_h)(\phi^2 I_{J(h)} + \sigma^2), \tag{23}$$

where $\Omega_h = (\phi^2 I_{J(h)} + \sigma^2)(\phi^2 I_{J(h)} + \sigma^2 + \Xi_h)^{-1}$ is a weight matrix given by the variance hyperparameters and Ξ_h, the variance-covariance matrix of estimation error. Without HSA-level random effects ($\sigma = 0$) and correlated estimation error across hospitals serving the same HSA population (so that Ξ_h is diagonal), equations (22)–(23) coincide with the usual empirical Bayes procedure studied by 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. I construct empirical Bayes estimates of posterior mean hospital quality by plugging the MLE hyperparameter estimates in column 6 of Table 3 into

the formula

$$\hat{q}_j = E[q_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] = E[\Phi(\beta_j)|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] = \Phi\left(\frac{E[\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}]}{\sqrt{1 + Var(\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)})}}\right), \tag{24}$$

where the second line follows from the fact that β_j is normally distributed given $\hat{\alpha}_{h(j)}$ and $\hat{\beta}_{h(j)}$.[33]

The solid blue line in Figure 4 shows the resulting distribution of quality index posteriors for the 2,082 hospitals with first-step estimates. As expected, the posterior mean distribution is tighter than the estimate distribution, reflecting empirical Bayes shrinkage.[34] Appendix Figure A5 instead plots the full distribution of quality posteriors and observed survival rates. As expected from the reduced-form variation in Figure 3 and the above discussion of selection-on-gains, average hospital quality is lower than the average survival probability, by around 3 percentage points.

Importantly, equation (24) also produces posterior quality predictions for hospitals without a first-step quality estimate due to their small size or insufficient quasi-experimental variation. 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 (21) 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 the first-step 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 (consistent with the estimates in column 9 of Table 3). I next describe these substantive findings in detail.

4 Results

The hyperparameter estimates in Table 3 indicate meaningful within-HSA variation in hospital quality that is positively, but only partially, correlated with observational RAM predictions. The reduced-form variation in Figure 3 and hierarchical model predictions in Figure 4 also suggest positive Roy selection-on-gains. I next use the hospital quality posteriors to conduct more nuanced analyses of both quality and sorting. I then use simulations to quantify the significance of selection bias in two quality-based policies currently in place in U.S. healthcare markets.

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

[34] 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 0.8 to 0.14.

4.1 Hospital Quality

Within-HSA variation in quality $E[Y_{ij}]$ reflects average causal effects of moving a representative patient across different hospitals. I quantify these effects by regressing quality posteriors and other measures on hospital characteristics and HSA fixed effects, within the set of 695 multi-hospital HSAs in the sample. 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, respectively. 22% are categorized as teaching hospitals.

Columns 1-3 of Table 4 report coefficients from regressions of observed 30-day survival rates, while columns 4–6, 7–9, and 10–12 regress the hospital RAM predictions $\hat{\alpha}_j$, minimum distance quality index estimates $\hat{\beta}_j$, and quality posteriors $\hat{q}_j$, respectively, all normalized to standard deviation units for comparability across column.[35] While the survival rate coefficients in columns 1–3 are small and mostly insignificant, a consistent pattern emerges from the less-biased quality measures in columns 4–12: government-run hospitals tend to be of 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 similar correlations as the RAM predictions and quality posteriors, though the coefficients in Table 4 are much less precise.

Quantitatively, the quality posterior coefficients in columns 10–12 of Table 4 suggest large average effects from moving patients between government-run and private hospitals and hospitals with different spending and volume levels. With an outcome standard deviation of 3.8 percentage points (pp), government-run hospitals are predicted to be on average 0.6pp lower quality. With a standard deviation in log spending (volume) of 1.3 (1.8), the coefficients suggest that moving patients between providers with typical differences affects short-run mortality by around 0.1pp (0.2pp). Log spending and volume are positively correlated in this sample ($\rho = 0.44$); however, column 12 shows that there is no significant effect on expected mortality from moving patients to hospitals with different spending levels conditional on volume. Moving representative patients from non-teaching to teaching hospitals is predicted to decrease 30-day survival 0.4pp, though again this estimate is not statistically significant.[36] Qualitatively, the findings in Table 4 are consistent with previously

[35] Bonhomme and Weidner (2019) discuss the posterior average effect interpretation of regressions of empirical Bayes posterior means. Posterior survival rates are obtaind from a standard empirical Bayes procedure that shrinks observed rates towards the grand mean in proportion to one-minus the signal to noise ratio.

[36] 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

documented patterns in 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. (2016).[37] This result is perhaps unsurprising as similar conclusions can be drawn from the regressions of observational RAM predictions, in columns 4–6 of Table 4.

Appendix Table A3 explores other predictors of hospital quality, including average staff salary, the use of electronic records or case management, and the breadth of accreditation or types of imaging technologies. Coefficients on all of these input measures are statistically significant, with staff salary remaining the most predictive in a multivariate regression that includes all regressors. Column 7 of this table further shows that these patterns are not driven by the robust relationship between hospital volume and quality. It is worth emphasizing that these estimates, as well as those in Table 4, are not causal in the sense of revealing the effect of a hospital changing its ownership structure, increasing its volume, or paying its staff more. At the same time, the robust correlations between quality posteriors and various input measures suggest that the former indeed captures intuitive aspects of true hospital productivity.

Appendix Table A4 summarizes other dimensions of hospital quality heterogeneity, over time and across patients with different emergency conditions. Specifically, Columns 1–3 report autocorrelations (adjusted for estimation error) of hospital quality indices, RAM predictions, and 30-day survival rates over three-year windows from 2001 by applying the same estimation procedure as in the main analysis sample of 2010–12. The typical first-order autocorrelation in quality indices is around 0.65, with the long correlation between quality indices in the main sample and in 2001–03 remaining quite high ($\rho = 0.47$). These exceed the estimated persistence of observational RAM predictions (panel B), while being similar to the persistence in raw 30-day survival rates (panel C). Columns 4–7 of Table A4 report correlations of these quality measures across samples of patients admitted for different emergency conditions; to achieve large enough samples for each condition catagory in Table 1, these are computed over the full 2001–2012 sample. Panel A shows large positive correlations in hospital quality indices, particularly between the sample of respiratory conditions and both samples of circulatory ($\rho = 0.62$) and digestive ($\rho = 0.74$) conditions. Panels B and C report similar but smaller correlations of observational RAM predictions and 30-day survival rates.

4.2 Patient Sorting and Selection Bias

Differences in a hospital's quality $E[Y_{ij}]$ and survival rate $E[Y_{ij}|D_{ij} = 1]$ reveal non-random selection on potential mortality outcomes. Figure 5 plots quality and survival rate posteriors within the 695 multi-hospital HSAs in the sample, adjusting for HSA fixed effects. These two measures are

predicted survival effects above are therefore quite significant in this historical context.

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

positively, but imperfectly correlated ($\rho = 0.62$). The figure further documents a negative correlation between hospital quality and selection bias $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$. Points above the solid 45-degree line in the figure represent hospitals with higher bias, which tend to be those of lower relative quality; similarly points below the line are less positively selected while also tending to be of relatively higher quality. Overall, the within-HSA correlation of hospital quality and bias posteriors is ρ =-0.85.

The negative relationship between selection bias and quality suggests that hospitals that are of higher quality tend to treat sicker-than-average patients. It also suggests that reductions in selection bias may not alter observational rankings of hospitals, despite a large distribution of bias (the within-HSA standard deviation of bias in this sample is 2.8pp). Such bias corrections will tend to reduce the rankings of low-quality hospitals and increase the rankings of high-quality hospitals, increasing the overall variance in the observational measure but inducing less change in the relative rankings. Indeed, observational RAM predictions and quality posteriors in this sample have a high rank correlation of 0.91.[38]

Along with selection on average potential mortality, the quality estimation framework also allows for an analysis of heterogeneous Roy selection-on-gains. To explore this, Figure 6 plots the distribution of volume-weighted average selection bias posteriors across the 695 multi-hospital HSAs in the sample. Absent Roy selection, for example when hospital treatment effects are constant, the market-average $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$ must equal zero: any volume-preserving transfer of patients across hospitals trades off equal gains and losses across different individuals. A positive (negative) market-average selection bias indicates that a typical patient is more (less) likely to survive at her selected hospital than at a hospital picked at random, suggesting positive (negative) sorting on hospital comparative advantage. Figure 6 shows that most HSAs (91%) exhibit such positive Roy selection; the average HSA-level selection bias posterior is 3.8pp. This substantive finding also reinforces the importance of the nonlinear minimum distance procedure; the constant effects assumption underlying conventional linear IV estimation rules out this type of sorting.

What explains hospital comparative advantage and Roy selection? Differential hospital distance is an obvious candidate, since individuals with acute conditions may only survive if brought to the closest available emergency room – a hospital quality which inherently varies over patients. On the other hand, though different distance is a clear driver of hospital admissions, it is a surprisingly weak predictor empirically. Chandra et al. (2016), 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

[38] Angrist et al. (2017) also find a negative correlation between school quality and student 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.

this measure is -0.91 across the 695 multi-hospital HSAs in my sample, reflecting that patients are admitted to hospitals that are on average 0.91 miles closer to them than the typical hospital in the market. However, this market-average distance bias 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 variation in the market-average selection bias plotted in Figure 6 and the market-average distance bias measure to examine the extent to which selection-on-distance explains selection-on-gains. Panel A reports the overall average selection bias of 3.8pp illustrated in Figure 6. Panel B reports estimates from regressions of HSA-level selection bias on polynomials in average distance bias. There is indeed a strong association, with a marginal 0.13–0.35pp 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 polynomial regression in column 3, representing average outcome selection bias in a HSA with zero selection-on-distance, remains a significantly positive 3.6pp. Panel C of Table 5 reports non-parametric estimates by directly computing average selection bias in HSAs with relatively small average distance bias. Even in the 39 regions where average distance bias is above -0.01 miles, patients are still around 3.3pp more likely to survive at their chosen hospitals than with random admission (87% of these HSAs have positive average bias posteriors). Together, these results suggest that the tendency for patients to select hospitals that are closer to them explains a small fraction of the overall benefit from positive Roy selection.[39]

Appendix Table A5 summarizes parallel analyses of patient sorting on other observables: diagnoses, demographics, and comorbidities. Columns 3–5 of the table again report constants from regressions of average market-level selection bias on a cubic polynomial of the average "distance" between admitted patients in a hospital and average patients in the HSA (here computed by the Mahalanobis metric). These estimates of selection-adjusted average bias range from 3.4pp to 3.7pp, suggesting a small role for observables in driving the unadjusted average 3.8pp. Including all observables in column 6 reduces the average bias only somewhat more, to 3.3pp. Together, these results suggest that hospitals specialize in ways that are unobserved by the econometrician (consistent with Chandra and Staiger (2007; 2017) but that are observed and selected on by ambulance companies and patients. The results also reinforce the importance of accommodating such unobserved selection-on-gains with the more flexible minimum distance approach.

4.3 Policy Consequences

Bias in observational hospital quality measures can distort the incentives of both providers and patients through federal reimbursement and report card policies. To quantify the potential for such

[39] Appendix Table A6 shows that these findings are also not driven by the tail behavior of the multivariate normal distribution assumed for latent health and utility. A multivariate Student's $t(2)$ specification finds an average HSA-level bias posterior of 3.4pp.

effects, as well as the potential improvement from using quasi-experimental data, I next simulate two leading U.S. policies. The first Value-Based Purchasing (VBP) policy redistributes Medicare reimbursements via incentive payments to hospitals ranked higher on mortality RAMs. The second Hospital Compare policy attempts to guide patient care decisions by publicizing local RAM rankings.

Medicare Reimbursement

VBP was launched in 2013, with the goal of redistributing a small share of Medicare reimbursement funds towards higher-performing hospitals (DHHS/CMS, 2015). Along with clinical process-of-care measures and patient surveys, risk-adjusted mortality became a part of the "total performance score" (TPS) used for this redistribution in fiscal year (FY) 2014. In that year, CMS withheld 1.25% of each participating hospital's diagnosis-related group (DRG) payment and redistributed the resulting $1.1 billion pool by a linear TPS schedule. VBP payments affect only a small share of a hospital's reimbursements: around half of participating hospitals saw a payment change of between -0.2 and 0.2 percent in FY2014 (Conway, 2013). Nevertheless, the program has proved quite controversial as the withholding rate has steadily increased, to 2% in 2016 (Pear, 2014). Norton et al. (2018) find evidence that hospitals respond to the program's incentives in these early years.

To gauge the role of selection bias in VBP, I first simulate benchmark FY2014 repayment rates with the observational RAM rankings informing total performance scores. I then recompute counterfactual repayments with performance scores based on the hospital quality posteriors. Performance scores combine "achievement points," which are based on data from the most recent period (2010–2012, in my sample), and "improvement points" based on a hospital's gain relative to a previous period (2007–2009). In the benchmark simulation, I follow the current methodology of computing points from risk-standardized mortality rates and combining performance scores with the earliest available scores from non-outcome domains from FY2014 to compute each hospital's TPS. In the counterfactual simulation, I recompute TPS based solely on achievement and improvement as measured by the hospital quality posteriors. Appendix A describes these two simulations in more detail.

Table 6 summarizes how benchmark, counterfactual, and simulated VBP repayment rate changes correlate with different hospital observables via multivariate regression. Teaching hospitals are slightly penalized by incorporating quasi-experimental data, with an average change in repayment rates of -0.12pp. Higher-spending hospitals also see a change in reimbursements, with an average 0.4pp increase in repayment rates associated with each percentage point increase in average spending (relative to -0.2pp in the observational benchmark). Higher-volume, non-profit, for-profit, and government-run hospitals are subsidized similarly in the counterfactual policy regime as in the observational benchmark. Appendix Figure A6 further shows that most hospitals see relatively small changes in the counterfactual regime: ten percent see no difference in VBP reimbursement when quality posteriors replace benchmark rankings, while 75% see an change of less than 1pp.

These simulation results suggest a limited scope for selection bias to affect the set of VBP winners and losers. This finding partly reflects the negative "selection-on-levels" documented in Figure 5, where hospitals of higher quality tend to treat sicker-than-average patients. A negative correlation of quality and bias will tend to cause less-biased quality rankings to generally agree with more naive measures. Indeed, counterfactual quality-based achievement points are highly correlated with the RAM-based scores determining benchmark VBP payments ($\rho = 0.84$).

Patient Guidance

The federal Hospital Compare website was launched in 2005 with a variety of hospital performance measures, including rankings derived from 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 – had developed competing hospital "report cards" with alternative observational risk-adjustment measures. Patients increasingly consult such rankings (Rice, 2014), and higher-ranked hospitals tend to see increased future emergency patient market shares (Chandra et al., 2016). Less is known about how report card guidance may affect patient health, particularly as such guidance becomes more influential and may increasingly affect admission decisions either directly (by influencing patients or referring providers) or indirectly (from regulatory policies influencing hospital popularity or capacity).

The hyperparameter estimates in Table 3 suggest that redirecting a typical patient from a randomly selected hospital to the provider with the highest RAM ranking in her HSA is likely to increase her 30-day survival probability, and that decisions based on less-biased quality posteriors will 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 effects will be partly offset by the fact that a typical patient's admission choice is better-than-random. On average, patients already see large survival gains from selecting more appropriate hospitals, suggesting that policies which succeed in redirecting individuals to the best-on-average hospital may actually do harm.

I quantify the net effect of such policies by simulating 500 sets of quality indices β_j from the HLM estimates of Table 3, column 6, given the set of observational RAM predictions. I then 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 the expected health value of using various supervisory quality rankings to redirect a representative patient's admissions.

Figure 7 shows that directing patients to top-ranked hospitals increases 30-day survival rates relative to a benchmark where uninformed patients pick hospitals at random within their HSAs. Patients who are redirected to the lowest mortality hospital in their HSA are on average 2.1pp more likely to survive their first 30 days after admission, while basing admissions on even the most naive RAM1 model increases this gain to 3.1pp. Adding patient demographics and comorbidities to the RAM specification only incrementally reduces bias (consistent with Figure 2 and Table 3), bringing the expected gain from the richest RAM3 model to 3.2pp. Incorporating quasi-experimental data improves the policy further still, with a representative patient seeing a 3.6pp gain in expected survival relative to random admissions. These gains are on the order of the 30-year technological decline in pneumonia mortality rates documented by Ruhnke et al. (2011). The final column of Figure 7 shows the maximal gain from quasi-experimental variation, in which all of the quasi-experimental moments used to construct the minimum distance quality estimates are known without error. This policy leads to a survival rate gain of around 4.5pp, relative to uninformed hospital choice.

Existing Roy selection-on-gains dominates all but the last ranking-based admission polices, however. The average gain in 30-day survival due to unobserved selection is found in Table 5 to be 3.8pp. Relative to this benchmark, redirecting a random patient from her preferred hospital to the one with the highest RAM3 rank reduces survival by 0.6pp. Quality posterior-based redirection closes this gap significantly (to -0.2pp), but not entirely. Only in the infeasible quasi-experimental policy, in which the reduced-form moments are known without error, is a ranking-based admission policy found to increase 30-day survival (by 0.7pp) relative to existing Roy selection-on-gains.

Broadly, these simulation results suggest that using less-biased quasi-experimental rankings to guide hospital sorting can deliver meaningful health benefits for patients who would otherwise make uninformed choices or would use observational hospital report card rankings. At the same time, report card policies that make informed 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 undermine the health benefits of prevailing hospital sorting. This result highlights a general limitation of unidimensional report card rankings in settings with positive Roy selection-on-gains: without knowing which types of patients respond to report card rankings, it is difficult to ensure that such rankings effectively boost patient outcomes.

5 Conclusions

Policymakers in many settings increasingly rely on outcome-based quality measures, despite concerns of selection bias in conventional quality estimators. This paper develops a flexible framework for measuring institutional quality and selection with quasi-experimental data. Reduced-form moments capturing institutional selection patterns are first estimated non-parametrically for each institution.

Structural assumptions are then used to extrapolate the differential selection revealed by the quasi-experiment. This two-step minimum distance approach is computationally tractable while allowing for more general patterns of Roy selection-on-gains than both conventional linear IV and previous nonlinear IV quality estimators. I establish quality identification in a broad class of elliptical models for binary potential outcomes, which enact intuitive extrapolations of the reduced-form moments.

IV methods are, of course, not a panacea to institutional quality estimation. For one, valid instruments for institutional choice are not always available. The recent rise of quasi-experimental "examiner" assignment in various settings, such as ambulance companies to patients (Doyle et al., 2015), judges to criminal defendants (Kling, 2006), or administrators to public benefit applicants (Maestas et al., 2013), may however make the methods in this paper broadly applicable for estimating the relative effects of various institutions and policies. A second concern is that IV estimation tends to produce more imprecise quality predictions than standard observational methods, despite a possible reduction in bias. Here the empirical Bayes approach of combining noisy but unbiased IV estimates with precise but potentially confounded observational estimates presents an attractive hybrid path. Third, the IV approach is not without structural assumptions and is generally more involved than simple observational quality estimation based on a selection-on-observables restriction. While this paper's proposal of applying structural assumptions directly on a set of non-parametric reduced-form moments can address both the computational burden and possible opacity of alternative methods, applying inappropriate structure in either method will generally lead to biased IV estimates. Robustness checks which vary the structural extrapolation of a fixed set of reduced-form estimates can be useful for gauging the potential for such misspecification.

Applying the method to hospital quality estimation, I find a large degree of hospital specialization and positive selection-on-gains: patients in most markets benefit significantly from being admitted to more appropriate hospitals, on average. Non-random sorting across hospitals generates pervasive selection bias, though this bias tends to be negatively correlated with true hospital quality such that observational quality rankings are not overly unreliable. Higher-spending, higher-volume, and privately owned hospitals are of higher average quality, and only a small share of the overall Roy selection is explained by selection-on-observables such as differential hospital distance. Admission policies based on observational or quasi-experimental quality report cards lead to significant mortality reductions for uninformed patients, but not when patients benefit from existing Roy selection-on-gains.

Ultimately, more work is needed to characterize the ways in which outcome-based policies shape hospital quality. As long as biased quality measures are used to adjust federal reimbursement, providers may find ways to "game the system" by boosting VBP scores without improving quality. While the simulations in Section 4 show that most observable hospital characteristics are uncorrelated with bias in VBP reimbursement rates, there may be various hospital-controlled unobservables

that correlate with VBP 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 (as discussed in see Appendix B.6).

The report card policy simulations also raise new questions about the efficacy of demand-side quality policies. Without hospital specialization, the finding that higher-ranked hospitals tend to attract more patients (as in Chandra et al. (2016)) has unambiguously positive implications for public health. Accounting for significant selection-on-gains, however, requires a more nuanced approach. The central question becomes whether the patients who respond to public quality rankings are those benefitting from hospital comparative advantage or whether the marginal patient is relatively uninformed, such that her default hospital is not more appropriate than the highest-ranked hospital in her HSA. Understanding the ways in which hospital performance measures actually affect admission decisions and characterizing the optimal design of public quality signals in settings with Roy selection are two important goals raised by the heterogeneous-effects framework.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Reseponse Models," *Journal of Econometrics*, 113, 231–263.

Abdulkadiroglu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2017): "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 85, 1373–1432.

Andrews, D. W. K. and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65, 497–517.

Angrist, J., P. Hull, P. Pathak, and C. Walters (2016): "Interpreting Tests of School VAM Validity," *The American Economic Review: Papers & Proceedings*, 106, 388–392.

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

Arnold, D., W. Dobbie, and P. Hull (2020): "Measuring Racial Discrimination in Bail Decisions," Working Paper.

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.

Bonhomme, S. and M. Weidner (2019): "Posterior Average Effects," Working Paper.

Brinch, C. N., 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?" *The Quarterly Journal of Economics*, 124, 597–636.

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

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016): "Healthcare Exceptionalism? Performance and Allocation in the U.S. Healthcare Sector," *The American Economic Review*, 106, 2110–44.

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

——— (2017): "Identifying Sources of Inefficiency in Health Care," NBER Working Paper No. 24035.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

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

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

Chetty, R. and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *The Quarterly Journal of Economics*, 133, 1163–1228.

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.

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., J. A. Graves, and J. Gruber (2017a): "Evaluating Measures of Hospital Quality," NBER Working Paper No. 23166.

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

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

Fan, J. (1992): "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

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.

Genest, C. and J. Neslehova (2007): "A Primer on Copulas for Count Data," *Astin Bulletin*, 37, 475–515.

Geweke, J., G. Gowrisankaran, and R. J. Town (2003): "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71, 1215–1238.

Gowrisankaran, G. and R. J. Town (1999): "Estimating the Quality of Care in Hospitals Using Instrumental Variables," *Journal of Health Economics*, 18, 747–767.

Gupta, A. (2017): "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program," Becker Friedman Institute for Research in Economics 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, 1527–1546.

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

Heckman, J. (1990): "Varieties of Selection Bias," *The American Economic Review*, 80, 313–318.

Heckman, J., S. Urzúa, and E. Vytlacil (2008): "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case," *Annals of Economics and Statistics*, 91, 151–174.

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35.

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

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

Hoxby, C. (2018): "The Productivity of U.S. Postsecondary Institutions," in *Productivity in Higher Education*, NBER.

Hull, P. (2017): "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons," Working Paper.

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

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

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

Kline, P. and C. R. Walters (2019): "On Heckits, LATE, and Numerical Equivalence," *Econometrica*, 87, 677–696.

Kling, J. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876.

Krumholz, H. M., Y. Wang, J. A. Mattera, Y. Wang, L. F. Han, M. J. Ingber, S. Roman, and S.-L. T. 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, 777–806.

Maestas, N., K. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review*, 103, 1797–1829.

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

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

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619.

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

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. McFadden, Elsevier, vol. 4 of *Handbook of Econometrics*, chap. 36, 2111 – 2245.

Norton, E. C., J. Li, A. Das, and L. M. Chen (2018): "Moneyball in Medicare," *Journal of Health Economics*, 61, 259–273.

Oster, E. (Forthcoming): "Health Recommendations and Selection in Health Behaviors," .

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.

Pope, D. G. (2009): "Reacting to Rankings: Evidence from "America's Best Hospitals"," *Journal of Health Economics*, 28, 1154–1165.

Raudenbush, S. and A. S. Byrk (1986): "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59, 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. M., M. A. Hernan, and B. Brumback (2000): "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

Rotemberg, J. J. (1983): "Instrumental Variable Estimation of Misspecified Models," Working Paper 1508-83, MIT Sloan.

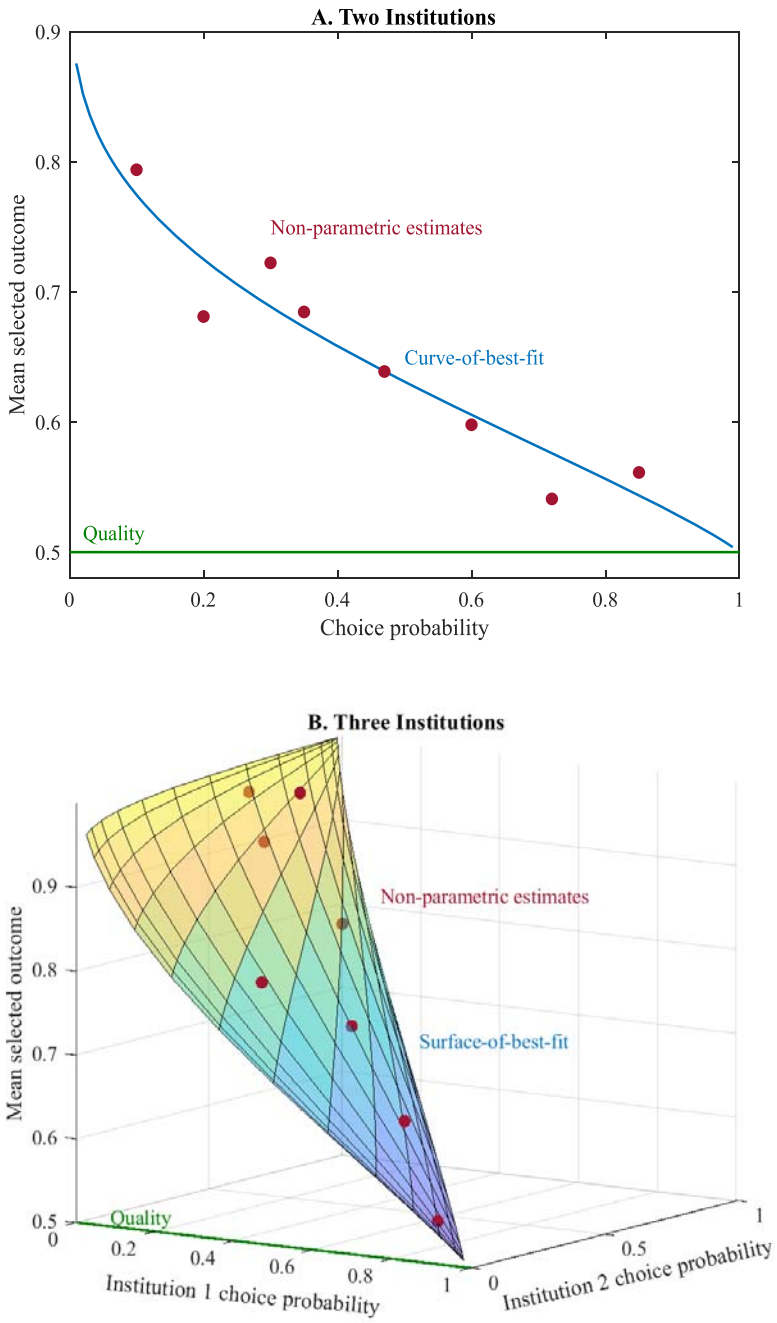
Ruhnke, G. W., M. Coca-Perraillon, B. T. Kitch, and D. M. Cutler (2011): "Marked Reduction in 30-Day Mortality Among Elderly Patients with Community-acquired Pneumonia," *American Journal of Medicine*, 124, 171–178.

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

Sloan, F. A., G. A. 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, 1–21.

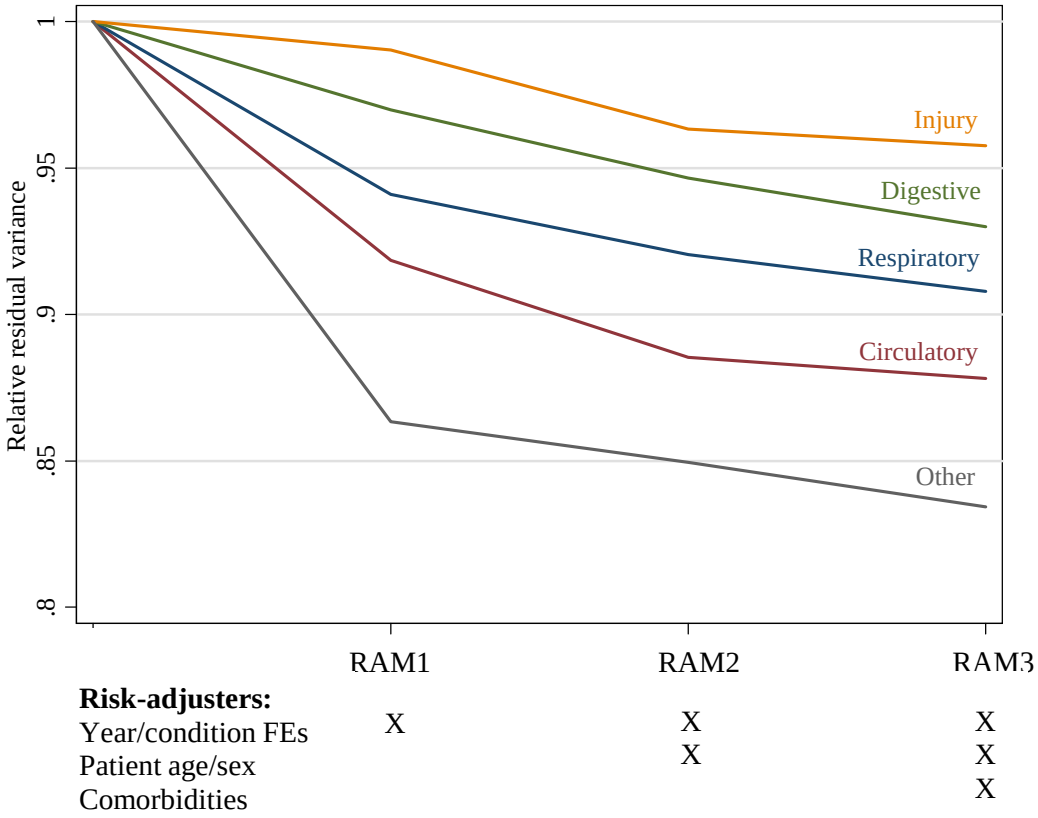
YNHHSC/CORE (2013): "2013 Measures Updates and Specifications: Acute Myocardial Infarction, Heart Failure, and Pneumonia 30-Day Risk-Standardized Mortality Measure (Version 7.0)," Available at `http://www.qualitynet.org/dcs/ContentServer?cid=1228774398696&pagename=QnetPublic%2FPage%2FQnetTier4&c=Page`.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

Figure 1: Nonlinear IV Extrapolation of Mean Selected Outcome Estimates



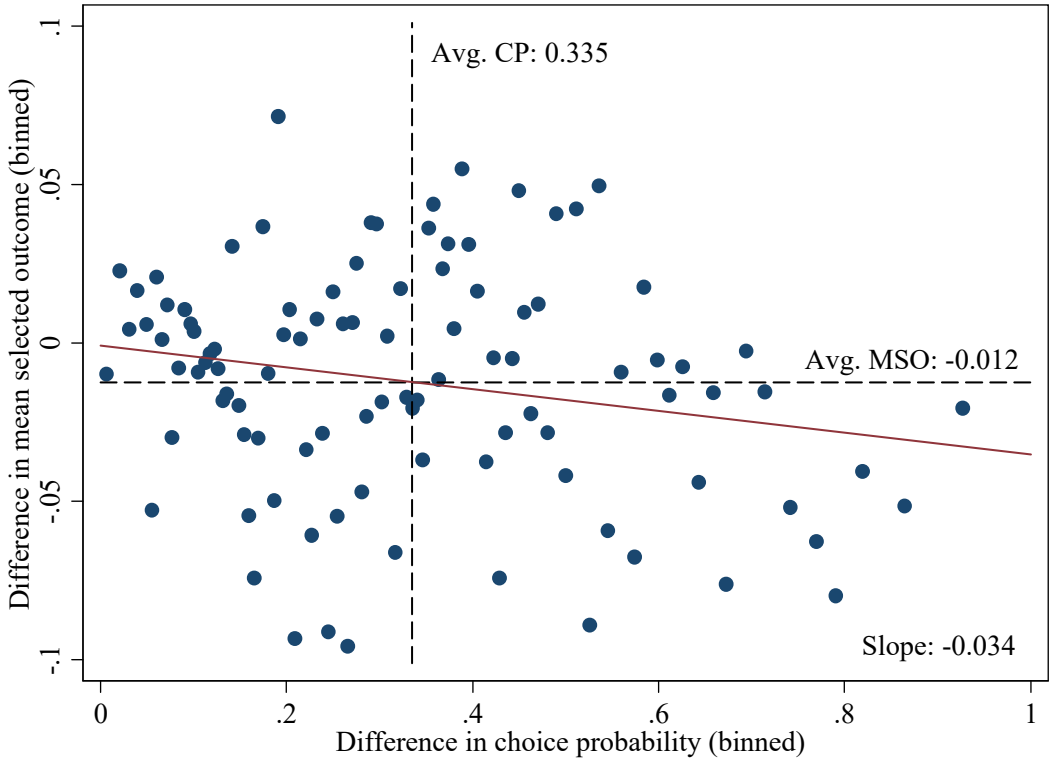
Notes: Panel A of this figure illustrates the mean selected outcome extrapolation implied by the bivariate probit model in Section 2.2 for a single institution, across eight different instrument values. Panel B simulates a trivariate probit model belonging to the family described in Section 2.4. The quality of the target institution (0.5) is indicated by a green line in both figures, and the red dots indicate non-parametric estimates of mean selected outcomes and choice probabilities. The line or curve of best fit is given by the minimum distance procedure discussed in the text.

Figure 2: Residual Survival Variation in Observational RAMs



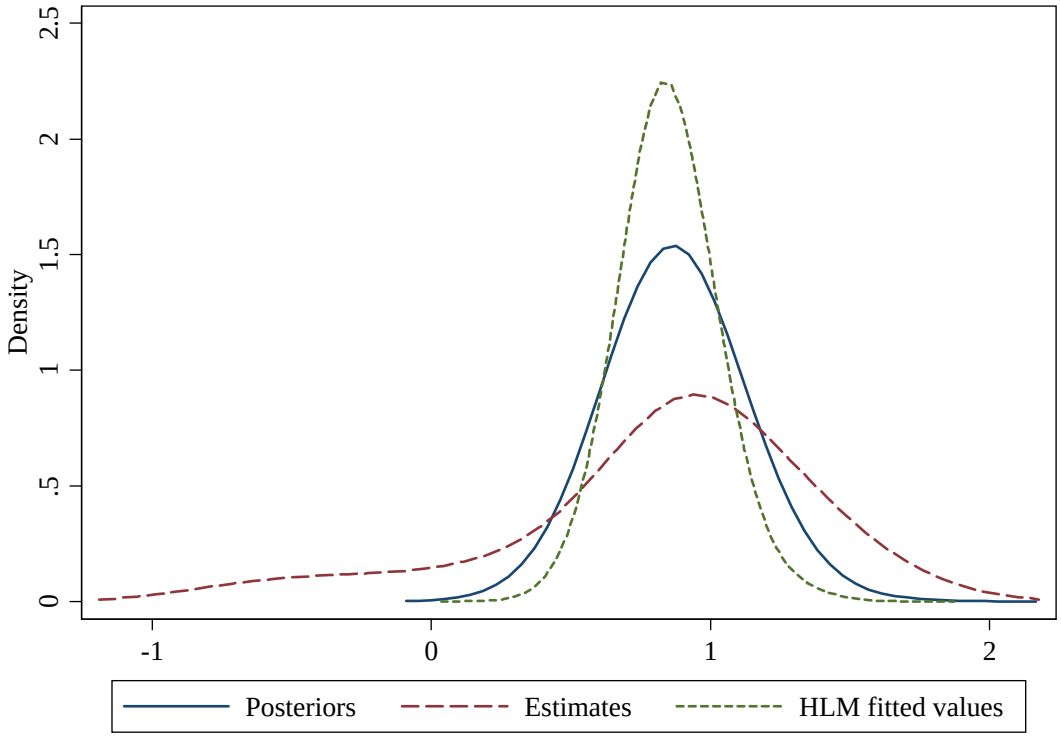
Notes: This figure plots the ratio of residual 30-day survival variation, after risk adjustment, relative to the total variance for three risk-adjustment models estimated separately by five condition catagories. See the notes to Table 1 for a list of conditions included in each catagory, Table 2 for a list of included comorbidities, and Appendix A for a description of the RAM estimation procedure.

Figure 3: Binscatter of mean selected outcome and choice probability differences



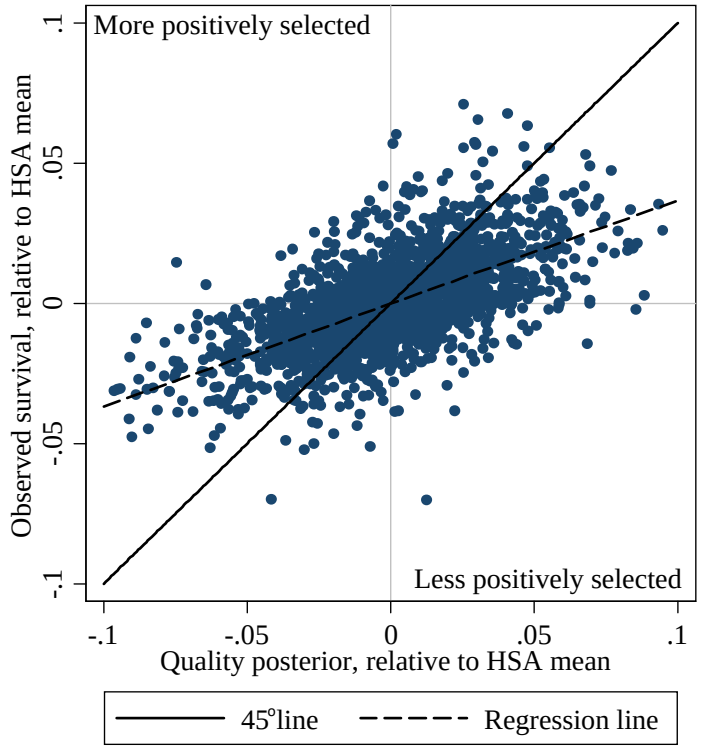
Notes: This figure average differences in mean selected outcome and choice probability estimates for the 2,082 hospitals with minimum distance quality estimates, by percentiles of choice probability differences. Differences are taken across the two ambulance companies with the largest and smallest estimated choice probability for each hospital. Dashed lines indicate sample means, and the solid line indicates an ordinary least squares fit.

Figure 4: Distributions of Hospital Quality Index Estimates and Posteriors



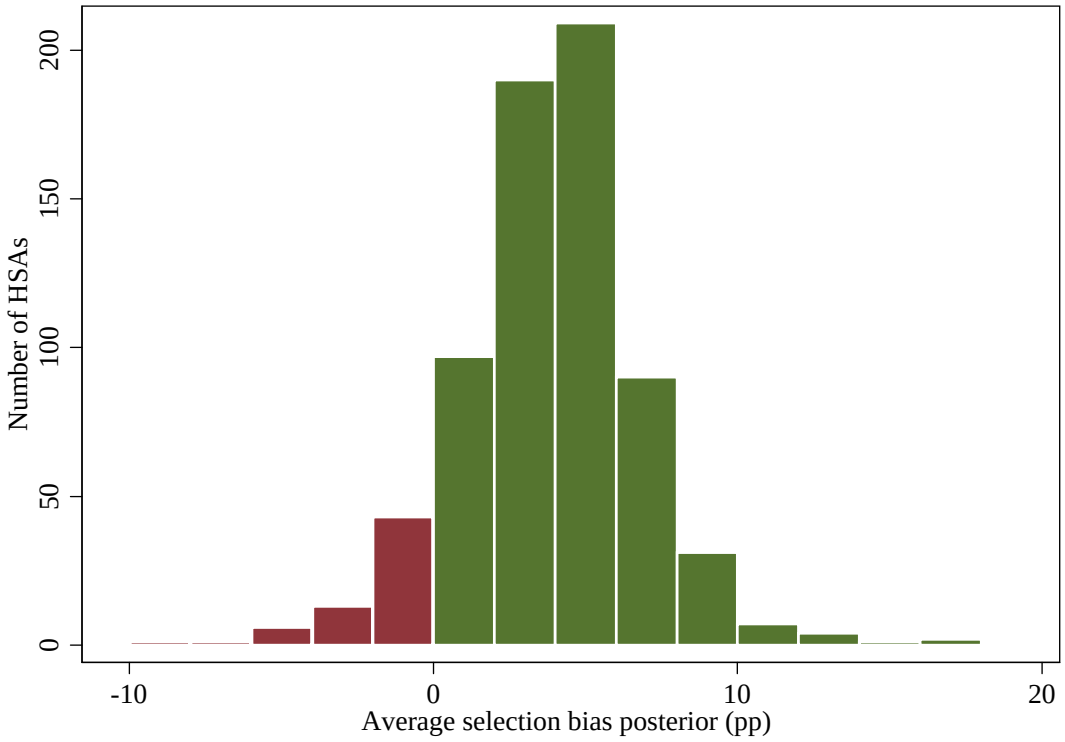
Notes: This figure plots Gaussian kernel density estimates of the distribution of minimum distance hospital quality index estimates and empirical Bayes posteriors, along with fitted values from the hierarchical linear model's projection on conventional RAM predictions. The sample includes 2,082 hospitals 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 Rates



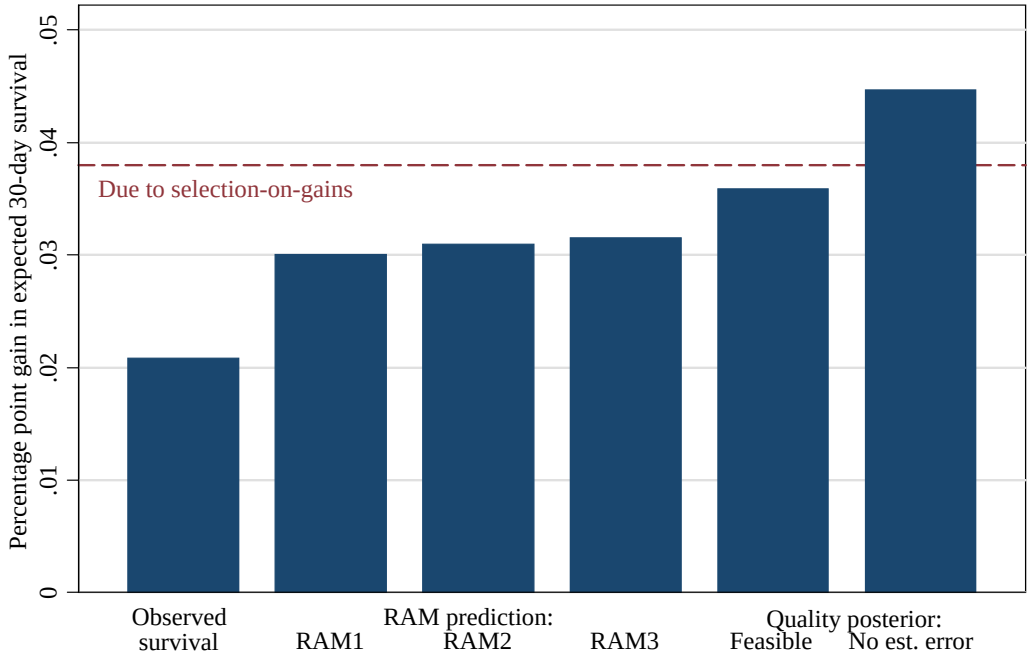
Notes: This figure plots posterior hospital 30-day survival rates against posterior hospital quality, adjusting for HSA fixed effects. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs.

Figure 6: Distribution of Average Selection-on-Gains across HSAs



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. HSAs with positive (negative) average selection thus exhibit positive (negative) Roy selection-on-gains

Figure 7: Average Survival Rate Gains from Redirecting Patients to Top-Ranked Hospitals



Notes: This figure plots simulated gains in average 30-day survival for a random patient sent to the highest-ranked hospital in her HSA, relative to making a random selection within the market. Top-ranked hospitals are given by 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 alternative RAM specifications. Estimates are from 500 draws of the hierarchical model described in the text.

Table 1: Hospital Quality 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 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 Market (HSA) Size			
One Hospital	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 rates in the sample of Medicare FFS patients admitted for one of 29 emergency conditions 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 analgesicis; 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: Tests of Quasi-Experimental Ambulance Company Assignment

	Comparison by RAM of the 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 mMalnutrition	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
Panels A-C 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 analysis sample. The p-values reported in columns 3 and 5 are for the null of no difference across patients, and are based on robust standard errors. 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	0.112	0.115					-0.001		
	(0.116)	(0.012)					(0.042)		
RAM2			0.123	0.122			0.027		
			(0.125)	(0.012)			(0.102)		
RAM3					0.127	0.128	0.139	0.159	0.139
					(0.117)	(0.012)	(0.089)	(0.016)	(0.030)
$(RAM3)^2$								0.018	
								(0.012)	
$(RAM3)^3$								-0.013	
								(0.031)	
$(RAM3)\times J$									-0.010
									(0.021)
Residual Std. Dev.:									
Within-HSA	0.170	0.170	0.172	0.171	0.172	0.171	0.171	0.161	0.166
	(0.053)	(0.053)	(0.057)	(0.057)	(0.057)	(0.058)	(0.057)	(0.068)	(0.095)
Between-HSA	0.165	0.165	0.161	0.162	0.159	0.159	0.156	0.150	0.178
	(0.061)	(0.060)	(0.067)	(0.066)	(0.067)	(0.056)	(0.066)	(0.078)	(0.061)

Notes: This table reports estimated parameters of the hierarchical linear model relating quality indices to observational RAM predictions. The sample consists of 2,082 hospitals with initial quasi-experimental quality estimates. 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 0.006 (0.009). Standard errors, clustered by HSA, are reported in parentheses.

Table 4: Within-HSA Predictors of Hospital Quality Measures

	30-Day Survival			RAM Predictions			Quality Index Estimates			Quality Posteriors		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
						A. Ownership and Teaching Status						
For-Profit	-0.019		-0.036	-0.065		-0.080	0.068		0.050	-0.058		-0.075
	(0.082)		(0.081)	(0.083)		(0.081)	(0.188)		(0.188)	(0.077)		(0.074)
Government	-0.072		-0.072	-0.178		-0.177	-0.350		-0.329	-0.156		-0.155
	(0.083)		(0.083)	(0.077)		(0.076)	(0.217)		(0.221)	(0.072)		(0.071)
Teaching		-0.103	-0.104		-0.092	-0.096		-0.160	-0.108		-0.107	-0.110
		(0.077)	(0.077)		(0.084)	(0.084)		(0.189)	(0.194)		(0.077)	(0.076)
						B. Average Spending and Patient Volume						
Log(Spending)	0.011		0.017	0.025		0.003	0.011		-0.011	0.021		0.005
	(0.012)		(0.013)	(0.013)		(0.013)	(0.357)		(0.356)	(0.013)		(0.013)
Log(Volume)		-0.005	-0.011		0.039	0.038		0.088	0.088		0.029	0.027
		(0.013)	(0.014)		(0.017)	(0.018)		(0.114)	(0.115)		(0.015)	(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). Survival rate posteriors shrink observed 30-day survival rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The sample in columns 1--6 and 10--12 includes 2,357 hospitals operating in 695 multi-hospital HSAs, while the sample in columns 7--9 includes the 2,082 hospitals with minimum distance quality estimates. Standard errors, clustered by HSA, are reported in parentheses.

Table 5: Average Selection-on-Gains, Adjusting for Selection-on-Distance

	(1)	(2)	(3)
		A. No Adjustment	
Avg. Selection Bias (pp)		3.83	
		(0.12)	
HSAs		695	
		B. Parametric Adjustment	
Avg. Selection Bias (pp)	3.71	3.67	3.60
	(0.13)	(0.14)	(0.14)
Distance Bias Marginal Effect	-0.13	-0.21	-0.35
	(0.04)	(0.07)	(0.11)
Polynomial	Linear	Quadratic	Cubic
HSAs		695	
		C. Non-parametric adjustment	
Avg. Selection Bias (pp)	3.49	3.39	3.32
	(0.25)	(0.34)	(0.41)
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, adjusting for average "bias" in hospital selection. 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

	Benchmark Rankings	Quality Rankings	Difference
	(1)	(2)	(3)
For-Profit	-0.013	-0.058	-0.045
	(0.028)	(0.045)	(0.048)
Government	-0.069	-0.062	0.007
	(0.028)	(0.042)	(0.045)
Teaching	0.006	-0.135	-0.140
	(0.027)	(0.042)	(0.043)
Log(Spending)	-0.237	0.167	0.404
	(0.046)	(0.058)	(0.072)
Log(Volume)	0.164	0.205	0.041
	(0.009)	(0.014)	(0.015)

Notes: This table reports coefficients from multivariate regressions of the share of total value-based purchasing withholdings that is repaid to a hospital under different quality ratings. Column 1 uses benchmark total performance scores to compute repayment rates, while column 2 uses hospital quality posteriors. Column 3 reports coefficients from regressing the change in repayment rates. VBP simulations use FY2014 balance sheet information and non-quality domain scores; see the data appendix for a description of the repayment rate simulation. 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.

A Data Appendix

I primarily follow Doyle et al. (2015) in constructing my 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 observations associated with 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 reports 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 overall.

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, suggesting a faithful replication.

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, defined in terms of equation (20) as

$$RSMR_j = \frac{1 - \sum_{i:D_{ij}=1} F_\nu(\hat{\alpha}_j + \hat{\gamma}'W_i)}{1 - \sum_{i:D_{ij}=1} F_\nu(\bar{\alpha} + \hat{\gamma}'W_i)}(1 - \bar{Y}), \tag{25}$$

where F_ν is the logit cumulative distribution function, $\hat{\gamma}$ estimates 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. The resulting risk-standardized survival rates, $1 - RSMR_j$, correlate highly with observational RAM predictions ($\rho = 0.98$). Scores for the counterfactual VBP simulation instead come from hospital quality posteriors. In both simulations, achievement scores are estimated in the main 2010–2012 analysis sample, while improvement scores come from changes between 2007–2009 and 2010–2012. Achievement scores are converted to points 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 and 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.

B Econometric Appendices

B.1 Linear IV Quality Estimation

This appendix shows that experimental linear IV quality estimates can be expressed as a weighted average of mean selected outcome estimates, with non-convex weights. The model is $Y_{ij} = q_j + \varepsilon_i$, where $E[\varepsilon_i] = 0$ and $\mathcal{U}_i \perp\!\!\!\perp Z_i$, and we without loss consider estimation of q_1 by omitting the D_{i1} treatment indicator and Z_{i1} instrument indicator. For notational simplicity we assume $L = J$, though the weighted average representation is straightforward to generalize to overidentified cases using the equivalence result in Rotemberg (1983).

Given the usual rank condition, an IV regression of Y_i on $D_{i2}, \ldots, D_{iJ}$ instrumented by $Z_{i2}, \ldots, Z_{iL}$ produces a coefficient vector estimating $\beta = (q_2 - q_1, \ldots, q_J - q_1)'$, of

$$\hat{\beta}^{LIV} = \underbrace{\begin{bmatrix} \hat{G}_{22} - \hat{G}_{21} & \ldots & \hat{G}_{J2} - \hat{G}_{J1} \\ \vdots & \ddots & \vdots \\ \hat{G}_{2L} - \hat{G}_{21} & \ldots & \hat{G}_{JL} - \hat{G}_{J1} \end{bmatrix}^{-1}}_{\equiv \hat{\Delta}^{-1}} \begin{bmatrix} \sum_j \hat{H}_{j2}\hat{G}_{j2} - \sum_j \hat{H}_{j1}\hat{G}_{j1} \\ \vdots \\ \sum_j \hat{H}_{jL}\hat{G}_{jL} - \sum_j \hat{H}_{j1}\hat{G}_{j1} \end{bmatrix}, \tag{26}$$

where $\hat{G}_{j\ell} = (\frac{1}{N}\sum_i D_{ij}Z_{i\ell})/(\frac{1}{N}\sum_i Z_{i\ell})$ and $\hat{H}_{j\ell} = (\frac{1}{N}\sum_i Y_i D_{ij} Z_{i\ell})/(\frac{1}{N}\sum_i D_{ij}Z_{i\ell})$ estimate choice probabilities and mean selected outcomes, respectively. Each row of $\hat{\Delta}$ captures the estimated first-stage effects of changing instrument assignment from $Z_{i1} = 1$ to $Z_{i\ell} = 1$ on institutional choice. The rows of the second matrix estimate the corresponding reduced-form effects on Y_i, since $\sum_j \hat{H}_{j\ell}\hat{G}_{j\ell} = (\frac{1}{N}\sum_i Y_i Z_{i\ell})/(\frac{1}{N}\sum_i Z_{i\ell})$. This can be derived from the fact that the model implies $Y_i = q_1 + \sum_{j>1}\beta_j D_{ij} + \varepsilon_i$ and that $\hat{\beta}^{LIV}$ satisfies, for each ℓ, the sample linear IV moments

$$\begin{aligned} 0 = \frac{1}{N}\sum_i \hat{\varepsilon}_i Z_{i\ell} &= \frac{1}{N}\sum_i \left(Y_i - \hat{q}_1^{LIV} - \sum_{j>1} \hat{\beta}_{j-1}^{LIV} D_{ij} \right) Z_{i\ell} \\ &= \hat{p}_\ell \left(\sum_j \hat{H}_{j\ell}\hat{G}_{j\ell} - \hat{q}_1^{LIV} - \sum_{j>1} \hat{\beta}_{j-1}^{LIV} \hat{G}_{j\ell} \right), \end{aligned} \tag{27}$$

where $\hat{p}_\ell = \frac{1}{N}\sum_i Z_{i\ell}$ and the second line uses the definitions of $\hat{G}_{j\ell}$ and $\hat{H}_{j\ell}$. Defining $\hat{\Omega}_\ell$ as the ℓth

column of $\hat{\Omega} = -\left[\hat{G}_{21} \dots \hat{G}_{J1}\right] \hat{\Delta}^{-1}$, we have by substituting (26) into (27) for $\ell = 1$ that

$$\begin{aligned}
\hat{q}_1^{LIV} &= \sum_j \hat{H}_{j1}\hat{G}_{j1} - \sum_{j>1} \hat{\beta}_{j-1}^{LIV}\hat{G}_{j1} \\
&= \left(1 - \sum_{\ell>1} \hat{\Omega}_{\ell-1}\hat{G}_{j1}\right) \sum_j \hat{H}_{j1}\hat{G}_{j1} + \sum_{\ell>1} \hat{\Omega}_{\ell-1} \sum_j \hat{H}_{j\ell}\hat{G}_{j\ell} \\
&= \sum_j \sum_\ell \hat{\omega}_{j\ell}\hat{H}_{j\ell}
\end{aligned} \tag{28}$$

where

$$\hat{\omega}_{j\ell} = \begin{cases} \left(1 - \sum_{m>1} \hat{\Omega}_{m-1}\right) \hat{G}_{j1}, & \ell = 1 \\ \hat{\Omega}_{\ell-1}\hat{G}_{j\ell}, & \ell > 1 \end{cases} \tag{29}$$

such that $\sum_j \sum_\ell \hat{\omega}_{j\ell} = \sum_j \hat{G}_{j1} = 1$. This shows that $\hat{q}_1$ is a non-convex average of the mean selected outcome estimates $\hat{H}_{j\ell}$, with weights given by the set of choice probability estimates $\hat{G}_{j\ell}$. In particular, it shows that when $L = J = 2$

$$\begin{aligned}
\hat{q}_1 &= (1 + \hat{\Delta}^{-1})\hat{G}_{11}\hat{H}_{11} - \hat{\Delta}^{-1}\hat{G}_{12}\hat{H}_{12} + (1 + \hat{\Delta}^{-1})\hat{G}_{21}\hat{H}_{21} - \hat{\Delta}^{-1}\hat{G}_{22}\hat{H}_{22} \\
&= \frac{\hat{G}_{11}(1 - \hat{G}_{12})}{\hat{G}_{11} - \hat{G}_{12}}\hat{H}_{11} + \frac{-\hat{G}_{12}(1 - \hat{G}_{11})}{\hat{G}_{11} - \hat{G}_{12}}\hat{H}_{12} + \frac{(1 - \hat{G}_{11})(1 - \hat{G}_{12})}{\hat{G}_{11} - \hat{G}_{12}}(\hat{H}_{21} - \hat{H}_{22}),
\end{aligned} \tag{30}$$

where we use the fact that $\hat{G}_{21} = 1 - \hat{G}_{11}$ and $\hat{G}_{22} = 1 - \hat{G}_{12}$. This gives equation (3) in the main text. Note that when $\hat{G}_{11} > \hat{G}_{12}$ the weight on $\hat{H}_{11}$ ($\hat{\omega}$, in the main text) is strictly above one, and also strictly decreasing in $\hat{G}_{11}$. As $\hat{G}_{11} - \hat{G}_{12}$ increases, $\hat{q}_1$ tends towards $\hat{H}_{11}$.

B.2 Minimum Distance Representation of Maximum Likelihood

This appendix shows that in the experimental setting, with $\mathcal{U}_i \perp\!\!\!\perp Z_i$, maximum likelihood quality estimates derived from discretely supported outcomes Y_i can be written as a minimum distance estimator involving mean selected outcome and choice probability estimates. For this I suppose the observed data log-likelihood is differentiable, with maximum likelihood estimates obtained in the interior of the parameter space.

Let $\mathcal{L}(\mathcal{Y}_i, \theta_0)$ denote the likelihood of the *iid* observations $\mathcal{Y}_i$ given a parameter vector θ_0 containing q_j, and define $f_c(y) = \mathbf{1}[y = y_c]$ for each observed value y_c of Y_i. The log-likelihood of

observed data is then proportional to

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ln\mathcal{L}(\mathcal{Y}_i;\theta_0) =& \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C}\ln Pr(Y_i=y_c, D_{ij}=1, Z_{i\ell}=1;\theta_0)f_c(Y_i)D_{ij}Z_{i\ell}\\ =&\sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C}\ln Pr(Y_{ij}=y_c \mid D_{ij\ell}=1, Z_{i\ell}=1;\theta_0)\hat{H}_{cj\ell}\hat{G}_{cj\ell}\hat{p}_\ell \\ &+\sum_{\ell=1}^{L}\sum_{j=1}^{J}\ln Pr(D_{ij\ell}=1 \mid Z_{i\ell}=1;\theta_0)\hat{G}_{j\ell}\hat{p}_\ell + \sum_{\ell=1}^{L}\ln Pr(Z_{i\ell}=1;\theta)\hat{p}_\ell\end{aligned} \tag{31}$$

where $\hat{H}_{cj\ell} = (\frac{1}{N}\sum_{i=1}^{N} f_c(Y_i)D_{ij}Z_{i\ell})/(\frac{1}{N}\sum_{i=1}^{N} D_{ij}Z_{i\ell})$, $\hat{G}_{j\ell} = (\frac{1}{N}\sum_{i=1}^{N} D_{ij}Z_{i\ell})/(\frac{1}{N}\sum_{i=1}^{N} Z_{i\ell})$, and $\hat{p}_\ell = \frac{1}{N}\sum_{i=1}^{N} Z_{i\ell}$.

Since $\mathcal{U}_i \perp\!\!\!\perp Z_i$, we can partition the parameter vector as $\theta_0 = (\bar{\theta}_0, \tilde{\theta}_0)$ by defining the functions $H_{cj\ell}(\bar{\theta}_0) = Pr(Y_{ij}=y_c \mid D_{ij\ell}=1;\theta_0)$, $G_{j\ell}(\bar{\theta}_0) = Pr(D_{ij\ell}=1;\theta_0)$, and $p_\ell(\tilde{\theta}_0) = Pr(Z_{i\ell}=1;\theta_0)$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ln\mathcal{L}(\mathcal{Y}_i;\theta_0) =& \sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C}\ln H_{cj\ell}(\bar{\theta}_0)\hat{H}_{cj\ell}\hat{G}_{j\ell}\hat{p}_\ell + \sum_{\ell=1}^{L}\sum_{j-1}^{J}\ln G_{j\ell}(\bar{\theta}_0)\hat{G}_{j\ell}\hat{p}_\ell + \sum_{\ell=1}^{L}\ln p_\ell(\tilde{\theta}_0)\hat{p}_\ell\\ =&\sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C-1}\ln H_{cj\ell}(\bar{\theta}_0)\hat{H}_{cj\ell}\hat{G}_{j\ell}\hat{p}_\ell + \sum_{\ell=1}^{L}\sum_{j-1}^{J-1}\ln G_{j\ell}(\bar{\theta}_0)\hat{G}_{j\ell}\hat{p}_\ell \\ &+\sum_{\ell=1}^{L}\sum_{j=1}^{J}\ln\left(1-\sum_{c=1}^{C-1}H_{cj\ell}(\bar{\theta}_0)\right)\left(1-\sum_{c=1}^{C-1}\hat{H}_{cj\ell}\right)\hat{G}_{j\ell}\hat{p}_\ell \\ &+\sum_{\ell=1}^{L}\ln\left(1-\sum_{j=1}^{J-1}G_{j\ell}(\bar{\theta}_0)\right)\left(1-\sum_{j=1}^{J-1}\hat{G}_{j\ell}\right)\hat{p}_\ell + \sum_{\ell=1}^{L}\ln p_\ell(\tilde{\theta}_0)\hat{p}_\ell.\end{aligned} \tag{32}$$

By assumption, the maximum likelihood estimate of $\bar{\theta}_0$ (which contains q_j) satisfies the first-order condition

$$\begin{aligned}0 =& \sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C-1}\left(\frac{\hat{H}_{cj\ell}\hat{G}_{j\ell}\hat{p}_\ell}{H_{cj\ell}(\hat{\theta}_{MLE})} - \frac{(1-\sum_{c'=1}^{C-1}\hat{H}_{c'j\ell})\hat{G}_{j\ell}\hat{p}_\ell}{1-\sum_{c'=1}^{C-1}H_{c'j\ell}(\hat{\theta}_{MLE})}\right)\nabla^H_{cj\ell}(\hat{\theta}_{MLE})\\ &+\sum_{\ell=1}^{L}\sum_{j=1}^{J-1}\left(\frac{\hat{G}_{j\ell}\hat{p}_\ell}{G_{j\ell}(\hat{\theta}_{MLE})} - \frac{(1-\sum_{j'=1}^{J-1}\hat{G}_{j'\ell})\hat{p}_\ell}{1-\sum_{j'=1}^{J-1}G_{j'\ell}(\hat{\theta}_{MLE})}\right)\nabla^G_{j\ell}(\hat{\theta}_{MLE})\\ =&\sum_{\ell=1}^{L}\sum_{j=1}^{J}\sum_{c=1}^{C-1}\left(\hat{H}_{cj\ell}-H_{cj\ell}(\hat{\theta}_{MLE})\right)\left(\frac{\hat{G}_{j\ell}\hat{p}_\ell\nabla^H_{cj\ell}(\hat{\theta}_{MLE})}{H_{cj\ell}(\hat{\theta}_{MLE})} + \frac{\sum_{c'=1}^{C-1}\hat{G}_{j\ell}\hat{p}_\ell\nabla^H_{c'j\ell}(\hat{\theta}_{MLE})}{1-\sum_{c'=1}^{C-1}H_{c'j\ell}(\hat{\theta}_{MLE}}\right)\\ &+\sum_{\ell=1}^{L}\sum_{j=1}^{J-1}\left(\hat{G}_{j\ell}-G_{j\ell}(\hat{\theta}_{MLE})\right)\left(\frac{\hat{p}_\ell\nabla^G_{j\ell}(\hat{\theta}_{MLE})}{G_{j\ell}(\hat{\theta}_{MLE})} + \frac{\sum_{j'=1}^{J-1}\hat{p}_\ell\nabla^G_{j'\ell}(\hat{\theta}_{MLE})}{1-\sum_{j'=1}^{J-1}G_{j'\ell}(\hat{\theta}_{MLE})}\right).\end{aligned} \tag{33}$$

where $\nabla^H_{cj\ell}(\cdot)$ and $\nabla^G_{j\ell}(\cdot)$ denote the gradients of $H_{cj\ell}(\cdot)$ and $G_{j\ell}(\cdot)$. This can be written

$$0 = \nabla(\hat{\theta}_{MLE})V(\hat{\theta}_{MLE})^{-1}\left(\hat{M} - M(\hat{\theta}_{MLE})\right), \tag{34}$$

where $\hat{M}$ is a vector stacking the $\hat{H}_{cj\ell}$ and $\hat{G}_{j\ell}$ in (33); $M(\hat{\theta}_{MLE})$ is the corresponding vector of

$H_{cj\ell}(\hat{\theta}_{MLE})$ and $G_{j\ell}(\hat{\theta}_{MLE})$; $V(\hat{\theta}_{MLE})$ is a block-diagonal matrix with blocks of the form

$$V^y_{j\ell}(\hat{\theta}_{MLE}) = \hat{p}_\ell(1-\hat{p}_\ell)\left(diag(v^y_{j\ell}(\hat{\theta}_{MLE})) - v^y_{j\ell}(\hat{\theta}_{MLE})v^y_{j\ell}(\hat{\theta}_{MLE})'\right) \tag{35}$$

$$V^d_{\ell}(\hat{\theta}_{MLE}) = \hat{p}_\ell(1-\hat{p}_\ell)\left(diag(v^d_{\ell}(\hat{\theta}_{MLE})) - v^d_{\ell}(\hat{\theta}_{MLE})v^d_{\ell}(\hat{\theta}_{MLE})'\right) \tag{36}$$

with $v^y_{j\ell}(\hat{\theta}_{MLE}) = \left[H_{1j\ell}(\hat{\theta}_{MLE}) \dots H_{C-1,j\ell}(\hat{\theta}_{MLE})\right]'$ and $v^d_{\ell}(\hat{\theta}_{MLE}) = \left[G_{j\ell}(\hat{\theta}_{MLE}) \dots G_{j\ell}(\hat{\theta}_{MLE})\right]'$; and $\nabla(\hat{\theta}_{MLE})$ has as columns the $\nabla^H_{cj\ell}(\hat{\theta}_{MLE})$ and $\nabla^G_{j\ell}(\hat{\theta}_{MLE})$. By inspection, equation (34) coincides with the first-order condition of a minimum distance estimator that uses all non-redundant mean selected outcome and choice probability estimates and a weight matrix of $V(\hat{\theta}_{MLE})^{-1}$.

B.3 Non-parametric Quality Estimation

This appendix establishes the consistency of the local linear quality estimator proposed in Section 2.3. The result is essentially one of identification-at-infinity for the intercepts in a treatment effects model with many discrete instruments, building on Heckman (1990), Heckman et al. (2008), and Andrews and Schafgans (1998).

For a fixed institution j, let

$$\hat{G}_\ell = \frac{1}{N}\sum_{i=1}^{N} \frac{D_{ij}Z_{i\ell}}{\hat{p}_\ell(X_i)} \tag{37}$$

estimate the (j,ℓ)th choice probability $G_\ell = Pr(D_{ij\ell} = 1)$, and let

$$\hat{H}_\ell = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i D_{ij} Z_{i\ell}/\hat{p}_\ell(X_i)}{\frac{1}{N}\sum_{i=1}^{N} D_{ij} Z_{i\ell}/\hat{p}_\ell(X_i)}. \tag{38}$$

estimate the (j,ℓ)th mean selected outcome $H_\ell = E[Y_{ij} \mid D_{ij\ell} = 1]$. Here $\hat{p}_\ell(\cdot)$ is an estimator of the instrument propensity score $p_\ell(x)$, such as the series logit estimator discussed in Section 2.3.

Consider the quality estimator given by a local linear regression of $\hat{H}_\ell$ on $1 - \hat{G}_\ell$:

$$\hat{q}^{LL}_j = \arg_1 \min_{a,b} \sum_{\ell=1}^{L} (\hat{H}_\ell - a - (1-\hat{G}_\ell)b)^2 K_C(1-\hat{G}_\ell), \tag{39}$$

where $K_C(u) = \frac{1}{C}K(u/C)$ for some kernel function $K(\cdot)$ and bandwith C. As C increases with the sample size N, this regression places more weight on instrument values ℓ with choice probability estimates closer to one. When the quasi-experimental moment estimates are uniformly consistent, $\hat{q}_\ell$ will thus approximate the average mean selected outcomes that are least-censored by selection, and thus closest to quality. Consistency follows from observing a large number of instrument values as the sample grows, with maximal choice probabilities increasingly close to one.

Formally, consider the following assumptions:

Assumption A1 (*Data and kernel*): (G_ℓ, H_ℓ) for $\ell = 1, \dots, L$ are *iid* random variables, with $Var(H_\ell \mid G_\ell = g)$ bounded and continuous. $K(\cdot)$ is a bounded density function with

$\int_{-\infty}^{\infty} xK(x)dx = 0$ and $\int_{-\infty}^{\infty} x^4K(x)dx < \infty$. As $N \to \infty$, $C \to 0$ and $CL \to \infty$.

Assumption A2 (*Moment consistency*): $sup_\ell|\hat{G}_\ell - G_\ell| \xrightarrow{p} 0$ and $sup_\ell|\hat{H}_\ell - H_\ell| \xrightarrow{p} 0$ as $N \to \infty$.

Assumption A3 (*High choice probabilities*): The density of G_ℓ is continuous and bounded above zero at one.

Here Assumptions A1 and A2 specify a hierarchical data generating process, in which L choice probability and mean selected outcome pairs are drawn and estimated in each a sample of size N. For the bivariate probit model in Section 2.2, for example, $A1$ holds with random $\pi_{j\ell}$. The moment estimates are uniformly consistent as L grows, implying a particular richness of quasi-experimental data. For example in the experimental case of simple random assignment, where Assumption 1 holds with $X_i = 1$ and $\hat{p}_\ell(x) = p_\ell(x) = 1/L$, Assumption A2 holds when $N/L \to \infty$, with an increasing number of individuals assigned to each instrument value. Assumption A1 further imposes regularity conditions for the conditional variance and kernel functions, while requiring the bandwith C to shrink and the number of moments L to grow with the sample, such that $CL \to \infty$.

Assumption A3 restricts the quasi-experimental setting to one in which instrument-specific choice probabilities may be drawn arbitrarily close to one. In the hospital application, this means one might observe ambulance companies willing to take nearly all patients to the given hospital j, regardless of condition or distance, in large enough samples. In the bivariate probit example Assumption A3 would be satisfied for institution $j = 1$ when $\pi_{2\ell} = \eta_{i2} = 0$, $Var(\eta_{i1} \mid \pi_1) = 1$, and $\pi_{1\ell} \sim N(0, 1)$, in which case $G_{\ell 1} = \Phi(\pi_{1\ell}) \sim U(0, 1)$.

We then have the following result:

Proposition A1 (*Non-parametric identification*): $\hat{q}_j^{LL} \xrightarrow{p} q_j$ under Assumptions A1–A3.

The proof follows in two steps. First, note that Assumptions A1 and A3 imply consistency of the infeasible local linear moment regression

$$\tilde{q}_j^{LL} = \arg_1 \min_{a,b} \sum_{\ell=1}^{L} \left(H_\ell - a - (1 - G_\ell)b\right)^2 K_C(1 - G_\ell), \tag{40}$$

per, e.g., Fan (1992). In particular $\tilde{q}_j^{LL} \xrightarrow{p} E[H_j \mid G_j = 1] = E[Y_{ij}] = q_j$ by definition of H_j and G_j. Second, note that $\hat{q}_j^{LL} - \tilde{q}_j^{LL} \xrightarrow{p} 0$. This follows by A2, since the weighted regression of $\hat{H}_\ell$ on $\hat{G}_\ell$ asymptotically coincides with that of H_ℓ on G_ℓ when each corresponding data point and weight uniformly converges. Thus $\hat{q}_j^{LL}$ is consistent.

B.4 Proof of Proposition 1

This appendix proves the consistency and asymptotic normality of the semi-parametric quality estimator $\hat{q}_j^{MD}$, under the following regularity conditions:

Assumption B1 (*Controls*): The support of X_i is a Cartesian product of compact intervals, with density bounded away from zero and infinity.

Assumption B2 (*Outcomes and treatment*): For all k and ℓ such that $E[D_{ik\ell}]$ is included in $M(\theta_0)$, $E[D_{ik\ell} \mid X_i = x]$ is continuously differentiable; for all k and ℓ such that $E[Y_{ik} \mid D_{ik\ell} = 1]$ is included in $M(\theta_0)$, $E[Y_{ik}^2] < \infty$, $E[D_{ik\ell}] > 0$, and $E[Y_{ik}D_{ik\ell} \mid X_i = x]$ and $E[D_{ik\ell} \mid X_i = x]$ are both continuously differentiable.

Assumption B3 (*Bounded and smooth propensity scores*): For all ℓ such that either $E[D_{ik\ell}]$ or $E[Y_{ik} \mid D_{ik\ell} = 1]$ is included in $M(\theta_0)$, for some k, $E[Z_{i\ell} \mid X_i = x]$ is bounded away from zero and continuously differentiable of order $s_\ell \geq 7 \cdot r$, where r is the dimension of X_i.

Assumption B4 (*Series estimator*): $\hat{p}_\ell(x)$ is a series logit estimator of $E[Z_{i\ell} \mid X_i = x]$, using a power series with $K = \lfloor N^{\nu_\ell} \rfloor$ terms for $\nu_\ell \in (\frac{1}{4(s_\ell/r-1)}, \frac{1}{9})$ with s_ℓ and r as in Assumption B3.

Assumptions B1 and B2 restrict the distribution of controls and the conditional distribution of $\mathcal{U}_i$ given X_i, while Assumption B3 restricts the smoothness and support of instrument propensity scores. Assumption B4 further restricts the rate at which additional terms are added to the series approximation of $p_\ell(x)$; this depends on the number of controls in X_i and the smoothness of $p_\ell(x)$.

Applying Theorem 1 in Hirano et al. (2003), $\hat{M} \xrightarrow{p} M(\theta_0)$ under Assumption 1 and B1–B4. Furthermore $\sqrt{N}(\hat{M} - M(\theta_0)) \Rightarrow N(0, V)$, where the diagonal of V has elements of the form

$$V_{mm} = E\left[(E[D_{ik\ell} \mid X_i] - E[D_{ik\ell}])^2 + \frac{Var(D_{ik\ell} \mid X_i)}{p_\ell(X_i)}\right] \tag{41}$$

and

$$V_{mm} = E\left[(E[Y_{ik}D_{ik\ell} \mid X_i] - E[Y_{ik}D_{ik\ell}])^2 + \frac{Var(Y_{ik}D_{ik\ell} \mid X_i)}{p_\ell(X_i)}\right] / E[D_{ik\ell}]^2. \tag{42}$$

Theorem 2 in Hirano et al. (2003) further shows $\hat{V} \xrightarrow{p} V$ under Assumptions 1 and B1-B4, where $\hat{V} = (\hat{\psi} + \hat{\alpha})'(\hat{\psi} + \hat{\alpha})/N$ for the $N \times M$ matricies $\hat{\psi}$ and $\hat{\alpha}$, with elements of the form

$$\hat{\psi}_{im} = \frac{D_{ik}Z_{i\ell}}{\hat{p}_\ell(X_i)} \text{ or } \hat{\psi}_{im} = \frac{Y_i D_{ik} Z_{i\ell}/\hat{p}_\ell(X_i)}{\frac{1}{N}\sum_{i'} D_{i'k}Z_{i'\ell}/\hat{p}_\ell(X_{i'})}, \tag{43}$$

and

$$\hat{a}_{im} = -\left(\frac{1}{N}\sum_{i'=1}^{N} \frac{D_{i'k}Z_{i'\ell}}{p_\ell(X_{i'})} R_\ell(X_{i'})\right)' \left(\frac{1}{N}\sum_{i'=1}^{N} R_\ell(X_{i'})R_\ell(X_{i'})'\right)^{-1} R_\ell(X_i)(Z_{i\ell} - \hat{p}_\ell(X_i)) \tag{44}$$

or

$$\hat{a}_{im} = -\left(\frac{1}{N}\sum_{i'=1}^{N} \frac{Y_{i'}D_{i'k}Z_{i'\ell}}{p_\ell(X_{i'})} R_\ell(X_{i'})\right)' \left(\frac{1}{N}\sum_{i'=1}^{N} R_\ell(X_{i'})R_\ell(X_{i'})'\right)^{-1} \frac{R_\ell(X_i)(Z_{i\ell} - \hat{p}_\ell(X_i))}{\frac{1}{N}\sum_{i'=1}^{N} D_{i'k}Z_{i'\ell}/\hat{p}_\ell(X_{i'})}, \tag{45}$$

where $R_\ell(x)$ denotes the vector of approximating functions used to construct $\hat{p}_\ell(x)$.

Given consistency and asymptotic normality of the moment vector, it remains to apply results on classic minimum distance estimation. Namely, the consistency of $\hat{q}_j^{MD}$ follows by Theorem 2.1 in Newey and McFadden (1994): the objective in equation (10), denoted $\hat{\Lambda}(\theta)$, converges uniformly in probability to a continuous $\Lambda(\theta) = (M(\theta_0) - M(\theta))'A(M(\theta_0) - M(\theta))$ by the preceeding result and since $M(\theta)$ is Hölder continuous on the parameter space. This space is compact and $\Lambda((\bar{\theta}', \tilde{\theta}')')$ is uniquely maximized at $\bar{\theta}_0$ by Assumption 2, so $\hat{q}_j \xrightarrow{p} q_j$.

Asumptotic normality of $\hat{q}_j$ furthermore follows by applying Theorem 3.1 in Newey and McFadden (1994), since θ_0 is on the interior of its parameter space, $M(\theta)$ is continuously differentiable in this space (so $\hat{\Lambda}(\theta)$ is twice continuously differentiable and $\sup_\theta ||\frac{\partial^2}{\partial\theta^2}\hat{\Lambda}(\theta) - \nabla(\theta)|| \xrightarrow{p} 0$), $\hat{A} \xrightarrow{p} A$ for positive-definite A, and $\sqrt{N}(\hat{M} - M(\theta_0)) \Rightarrow N(0, V)$.

B.5 Proof of Proposition 2

This appendix establishes identification of the elliptical model for binary potential outcomes. Consider institution k's choice probabilities when $D_{ik\ell} = \mathbf{1}[\pi_{k\ell} + \eta_{ik} > \pi_{k'\ell} + \eta_{ik'}, \forall k' \neq k]$, per Assumption 7, with η_i elliptically distributed by the density generator $g_\eta(\cdot)$ and location and shape parameters s_η and S_η, per Assumption 8:

$$\begin{aligned} Pr(D_{ik\ell} = 1) &= Pr(\Delta_k \pi_\ell > \Delta_k \eta) \\ &= Pr\left(\Delta_k(\pi_\ell + s_\eta) > (\Delta_k S_\eta \Delta_k')^{1/2}\upsilon\right). \end{aligned} \tag{46}$$

Here Δ_k is a $(J-1)\times J$ differencing matrix such that the k'th row of $\Delta_k\pi$ is $\pi_k - \pi_{k'}'$, and $\upsilon \in \mathbb{R}^{J-1}$ is spherically distributed with a known density generator derived from $g_\eta(\cdot)$. Also consider institution j's mean selected outcomes

$$\begin{aligned} E[Y_{ij} \mid D_{ij\ell} = 1] &= Pr(h_{ij} > 0 \mid \Delta_j \pi_\ell > \Delta_k \eta) \\ &= Pr\left(s_h/\sqrt{S_h} > \rho'\upsilon + \omega\sqrt{1-\rho'\rho} \mid \Delta_j(\pi_\ell + s_\eta) > (\Delta_j S_\eta \Delta_j')^{1/2}\upsilon\right) \end{aligned} \tag{47}$$

where $\rho = (\Delta_j S_\eta \Delta_j' S_h)^{-1/2}\Delta_j S_{\eta h}$ and $(\upsilon', \omega)'$ is spherically distributed with a known density generator derived from $g(\cdot)$. From these we can see that certain parameter combinations are observationally equivalent in terms of $M(\theta_0)$: we can without loss normalize $s_\eta = \pi_{j\ell} = 0$, $S_h = 1$, and set the jth row of $S_{\eta h}$ and the jth row and column of S_η to zero. Then

$$Pr(D_{ik\ell} = 1) = Pr\left(\Delta_k(\pi_\ell', 0)' > (\Delta_k S_\eta \Delta_k')^{1/2}\upsilon\right) \tag{48}$$

$$E[Y_{ij} \mid D_{ij\ell} = 1] = Pr\left(s_h > \rho'\upsilon + \omega\sqrt{1-\rho'\rho} \mid \pi_\ell > \tilde{S}_{\eta,j}^{-1/2}\upsilon\right) \tag{49}$$

where now $\rho = -\tilde{S}_\eta^{-1/2}\tilde{S}_{\eta h}$, with $\tilde{S}_\eta$ denoting the submatrix of S_η with the jth row and column removed and $\tilde{S}_{\eta h}$ denoting the subvector of $S_{\eta h}$ with the jth row removed.

Consider next the function from $\mathbb{R}^{J-1}$ to the standard $J-1$ probability simplex P^{J-1} given by

$G(\tilde{\pi}) = (G_1(\tilde{\pi}), \ldots, G_{J-1}(\tilde{\pi}))'$, where

$$G_k(\tilde{\pi}) = Pr\left(\Delta_k(\tilde{\pi}', 0)' > (\Delta_k S_\eta \Delta_k')^{1/2} \upsilon\right) \tag{50}$$

By inspection $G(\cdot)$ is a proper mapping, in that for any sequence $\{\tilde{\pi}_n\}$ that escapes to infinity in $\mathbb{R}^{J-1}$, $\{G(\tilde{\pi}_n)\}$ escapes to infinity in P^{J-1}. Moreover, the Jacobian of $G(\cdot)$ is a strictly diagonally dominant L-matrix everywhere in $\mathbb{R}^{J-1}$, so that by the Lévy-Desplanques theorem it is everywhere invertible. By Hadamard's theorem there thus exists a global inverse $G^{-1}(\cdot) : P^{J-1} \to \mathbb{R}^{J-1}$. This has two implications. First, Assumptions 7 and 8 are generally without observational loss for the marginal distribution of $(((D_{ij\ell})_{j=1}^J)_{\ell=1}^L)$: provided the choice probability vectors G_ℓ are positive, we can rationalize them with $\pi_\ell = G^{-1}(G_\ell)$. Second, since we can do so for any S_η, and since the mean selected outcomes (49) are invariant to applying any positive-definite transformation of S_η while replacing π_ℓ with the corresponding $\pi_\ell = G^{-1}(G_\ell)$, we can without loss set $\tilde{S}_{\eta,j} = I_{J-1}$.

Quality in this model is given by $q_j = q(s_h)$, for a strictly increasing $q(\cdot)$ derived from $g(\cdot)$. We have then shown that $M(\theta_0)$ is parameterized by the $J + L(J-1) \times 1$ vector $\bar{\theta}_0 = (q_j, \rho', G_1', \ldots, G_L')'$, where the $L(J-1)$ choice probabilities are directly parameterized and the L mean selected outcomes are given by

$$E[Y_{ij} \mid D_{ij\ell} = 1] = Pr\left(q^{-1}(q_j) > \rho' \upsilon + \omega\sqrt{1 - \rho'\rho} \mid G^{-1}(G_\ell) > \upsilon\right). \tag{51}$$

It remains to show that a unique $\bar{\theta}_0$ satisfies (51) when $L \geq J$ with $G_\ell \neq G_k$ for all $\ell \neq k$. This follows from noting that the right-hand side of (51) gives, along with the $G_{j\ell}$, a J-dimensional eliptical copula with a correlation matrix of $\begin{bmatrix} 1 & \rho' \\ \rho & I_{J-1} \end{bmatrix}$, with $J-1$ parameters. There is thus a unique ρ consistent with any $J-1$ unique choice probabilities (Genest and Neslehova, 2007), with q_j given uniquely by the remaining instrument value, since (51) is strictly increasing in q_j given ρ.

B.6 Quality-Based Policies and Hospital Incentives

This appendix uses a stylized economic model to illustrate how observational and quasi-experimental quality measures may shape institutional incentives. While far from a comprehensive analysis of the various margins by which such institutions may respond to quality-based regulation, this model is useful for highlighting the role of treatment effect heterogeneity in institutional "cream-skimming," as well as formalizing the sense in which the IV quality measures developed in this paper may be less prone to such gaming.

Consider an environment with two hospitals j and k, two ambulance companies ℓ and m, and a population of patients i. Patients are differentiated by a vector of characteristics $(X_{i1}, \ldots, X_{iN})$ and potential health outcomes (Y_{ij}, Y_{ik}), which are joint-normally distributed. Without loss in what follows we assume the X_{in} are standard normal variables which are mutually uncorrelated across n.

Each patient has a normally distributed relative preference U_i for hospital j and is randomly assigned (with equal probability) to one of the two ambulance companies. The companies face relative costs C_ℓ and C_m, respectively, for referring patients to hospital j, and do so only when, e.g., $U_i > C_\ell$. With Z_i indicating assignment to company ℓ, admission to hospital j is then given by $D_i = \mathbf{1}[U_i > C_\ell Z_i + C_m(1 - Z_i)]$, and patient health outcomes are $Y_i = Y_{ij}D_i + Y_{ik}(1 - D_i)$.

A policymaker ranks and rewards hospitals by a statistic derived from the distribution of observed (Z_i, D_i, Y_i). To operationalize the notion of "cream-skimming," we assume that hospital j can affect the quality measure by altering patient utility and ambulance company costs. Specifically, the hospital can pay some cost to advertise itself more aggressively to patients with higher X_{in}, for each n, and can similarly affect C_ℓ and C_m. For simplicity here we assume hospital k is passive, hospital j's capacity $\pi \in (0, 1)$ is fixed, and that outside of hospital catering individual preferences are ideosyncratic. We then suppose

$$D_i = \mathbf{1}\left[\sum_n \tau_n X_{in} + \sqrt{1 - \sum_n \tau_n^2} V_i > \Phi^{-1}((\omega Z_i + (1 - \omega)(1 - Z_i))\pi)\right] \tag{52}$$

where $V_i \sim N(0, 1)$, independently from the X_{in} and potential outcomes, $\tau_n \in [-1, 1]$ quantifies the amount of hospital j's catering to characteristic X_{in} (with $\sum_n \tau_n^2 \leq 1$), and $\omega\pi$ (respectively, $(1 - \pi)\pi$) denotes the share of patients referred to hospital j by company ℓ (respectively, m).

Suppose first that hospitals are ranked by the observed difference in average patient outcomes. Given (52), this is

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = q_j - q_k + \sum_n \tau_n(\rho_{jn}\pi + \rho_{kn}(1 - \pi))\frac{\phi(\Phi^{-1}(\pi))}{\pi(1 - \pi)}, \tag{53}$$

where $q_j = E[Y_{ij}]$ and $q_k = E[Y_{ik}]$ denote hospital quality, $\rho_{jn} = Cov(Y_{ij}, X_{in})$ and $\rho_{kn} = Cov(Y_{ik}, X_{in})$ capture how patient characteristics covary with potential outcomes, $\phi(\cdot)$ is the standard normal probability density function, and $\Phi(\cdot)$ is the standard normal cumulative density function. From this it is clear that hospital j has incentive to increase its quality, but also to cream-skim on characteristics correlated with either Y_{ij} or Y_{ik}. The marginal benefit from such gaming (i.e. increasing τ_n) is higher for characteristics on which both hospitals specialize (i.e. n for which ρ_{jn} and ρ_{kn} are both positive or both negative). Holding the market share weighted average covariances $\rho_{jn}\pi + \rho_{kn}(1-\pi)$ fixed, gaming incentives are higher when market shares are unequal, as $\frac{\phi(\Phi^{-1}(\pi))}{\pi(1-\pi)}$ is strictly convex and minimized at $\pi = 0.5$. Iintuitively, either $E[Y_i \mid D_i = 1]$ or $E[Y_i \mid D_i = 0]$ is more strongly affected by changes in τ_n when either hospital is small. Since the ambulance companies play no role in the observational quality measure, there is no incentive to affect referral costs.

Now suppose hospitals are ranked by a conventional linear IV coefficient, specifically from regressing Y_i on D_i with Z_i as an instrument. Since the model satisfies the assumptions of Imbens

and Angrist (1994) this is a local average treatment effect, here given by

$$\frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]} = q_j - q_k + \sum_n \tau_n(\rho_{jn} - \rho_{kn}) \frac{\phi(\Phi^{-1}(\omega\pi)) - \phi(\Phi^{-1}((1-\omega)\pi))}{(2\omega - 1)\pi} \tag{54}$$

From this we see that hospital j again has incentive to cream-skim, as well as increase its quality. However which patients the hospital is likely to cater to, in terms of its comparative advantage, has reversed: the marginal benefit of increasing τ_n is higher for n with ρ_{jn} and ρ_{kn} of opposite sign, and there is no incentive to cream-skim on X_{in} that are equally correlated with potential outcomes at both institutions. Furthermore, the hospital now has incentive to affect ambulance company costs: namely to either increase or decrease ω, as $\frac{\phi(\Phi^{-1}(\omega\pi)) - \phi(\Phi^{-1}((1-\omega)\pi))}{(2\omega-1)\pi}$ is also strictly convex and minimzed at $\omega = 0.5$, for each π. Intuitively, this results in a larger share of "compliers" with higher utility for hospital j, who by the hospital's manipulation of the τ_n have more positive treatment effects than the representative patient.

Finally, note that a policymaker using the minimum distance estimator developed in 2.3 will, upon recognizing joint-normality of U_i and (Y_{ij}, Y_{ik}), recover a pure quality ranking $q_j - q_k$ regardless of the τ_n and ω. Hospital j thus only has incentive to invest in its quality in this scenario.

B.7 Testing Hospital RAMs

This appendix derives tests of the selection-on-observables assumption underlying observational hospital RAMs. For the general quality model given by equations (1)-(2), we consider the null

$$H_0 \ \textit{(RAM Validity)}\text{: } Y_{ij} = \mathbf{1}[f_j(W_i, \nu_i) \geq 0], \text{ where } \nu_i \mid \left((Z_{i\ell}, (D_{ij\ell})_{j=1,\ldots,J})_{\ell=1,\ldots,L}, W_i'\right)' \sim F_\nu$$

for some observable vector W_i, known function $f_j(\cdot)$ and known distribution F_ν. For example, the null model may be $Y_{ij} = \mathbf{1}[a_j + \gamma' W_i \geq \nu_i]$ with Gumbel-distributed ν_i. This is the usual model used for observational RAMs in the health context, which rules out hospital comparative advantage. More generally the selection-on-observables null H_0 assumes that any unobservable component of potential outcomes ν_i is independent of the patient sorting process (2), conditional on the risk-adjusters in W_i.[40] In particular, H_0 implies $\nu_i \perp\!\!\!\perp (D_i, W_i)$, the usual basis for consistent estimation of equation (20), and is equivalent when ignoring knife-edge cases of perfectly offsetting dependencies between (in the health context) patient health, admission decisions, and ambulance company assignment.

Each $Z_{i\ell}$ is conditionally excludable from outcomes when selection-on-observables hold. That is

[40] Note that H_0 rules out both "essential heterogeneity" and "selection bias" in the language of Heckman et al. (2006), as both tend to generate bias in observational quality estimates. For example if $Y_i = \mathbf{1}[D_i'\alpha_i > \eta_i]$ with η_i but not α_i unconditionally independent of the selection mechanism, $E[Y_{ij} \mid D_{ij} = 1]$ will generally not equal $E[Y_{ij}]$.

for any ℓ, j, and w, we have by the Law of Iterated Expectations,

$$\begin{aligned} E[Y_i|Z_{i\ell}=1, D_{ij}=1, W_i=w] &= Pr(f_j(w,\nu_i)\geq 0|Z_{i\ell}=1, D_{ij\ell}=1, W_i=w) \\ &= Pr(f_j(w,\nu_i)\geq 0|D_{ij\ell}=1, W_i=w) \\ &= E[Y_i|D_{ij}=1, W_i=w], \end{aligned} \tag{55}$$

In the context of the application, selection-on-observables implies that a patient's expected 30-day survival does not depend on the identity of her ambulance company, conditional on her hospital and risk-adjusters. Defining $p_\ell(D_i, W_i) = E[Z_{i\ell}|D_i, W_i]$, we thus have under H_0

$$\begin{aligned} E\left[Y_i\left(\frac{Z_{i\ell}-p_\ell(D_i,W_i)}{p_\ell(D_i,W_i)(1-p_\ell(D_i,W_i)}\right)\right] &= E\left[E\left[\frac{Y_iZ_{i\ell}}{p_\ell(D_i,W_i)}-\frac{Y_i(1-Z_{i\ell})}{1-p_\ell(D_i,W_i)}|D_i,W_i\right]\right] \\ &= E[E[Y_i|Z_{i\ell}=1,D_i,W_i]-E[Y_i|Z_{i\ell}=0,D_i,W_i] \\ &= 0 \end{aligned} \tag{56}$$

Given a consistent approximation to the propensity score functions $p_\ell(\cdot)$ and appropriate regularity conditions, we may thus use the sample analogue of the left-hand side of (56) to test H_0 for each instrument value ℓ. This test does not leverage knowledge of $f_\ell(\cdot)$ or $F_\nu(\cdot)$, and is in that sense non-parametric.

An alternative test leverages knowledge of the potential outcome structure, noting that by H_0,

$$\begin{aligned} E[Y_i|Z_i] &= E[E[Y_i|Z_i,D_i,W_i]|Z_i] \\ &= E\left[\sum_j D_{ij}Pr(f_j(W_i,\nu_i)\geq 0)|Z_i\right], \end{aligned} \tag{57}$$

such that in the observational RAM example, $E[Y_i|Z_i]-E[F_\nu(\alpha'D_i+\gamma'W_i)|Z_i]=0$. 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), an alternative Wald test statistic uses the fact that equation (57) implies vector-equality of the coefficients μ_Y and μ_F in the regressions:

$$Y_i = \mu_Y'Z_i + e_Y \tag{58}$$

$$F_\nu(\alpha'D_i+\gamma'W_i)) = \mu_F'Z_i + e_F. \tag{59}$$

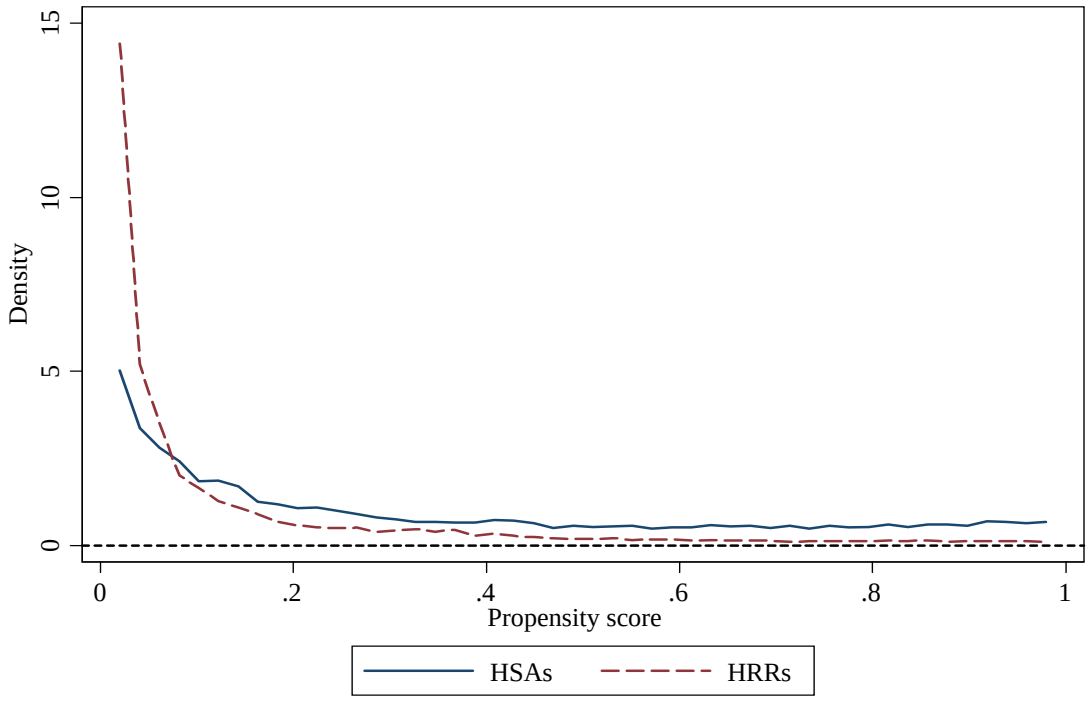
A final approach notes that equations (58) and (59) are the reduced form and first stage equations of a two-stage least squares (2SLS) procedure that uses Z_i to instrument for RAM-predicted survival $F_\nu(\alpha'D_i+\gamma'W_i)$ in a regression of realized survival Y_i. Since $\mu_Y=\mu_F$ under H_0, this procedure identifies a 2SLS coefficient of one under H_0. 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 non-parametric test, applied to 100 randomly selected ambulance companies admitting at least 100 patients in the main analysis sample. For each observational RAM specification, I approximate the propensity scores $p_\ell(D_i, W_i)$ with a flexible logit model and jointly test significance of the 100 sample analogues of the left-hand side of equation (56), correcting inference for first-step propensity score 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 that replicate as closely as possibly the 2013 CMS methodology (see Appendix A 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 (57), for the same set of 100 randomly chosen ambulance companies. Adding demographic and comorbidity controls to the RAM brings the forecast coefficient from 1.3 to 1.1, with the latter 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: Distribution of Propensity Score Estimates for Different Market Definitions



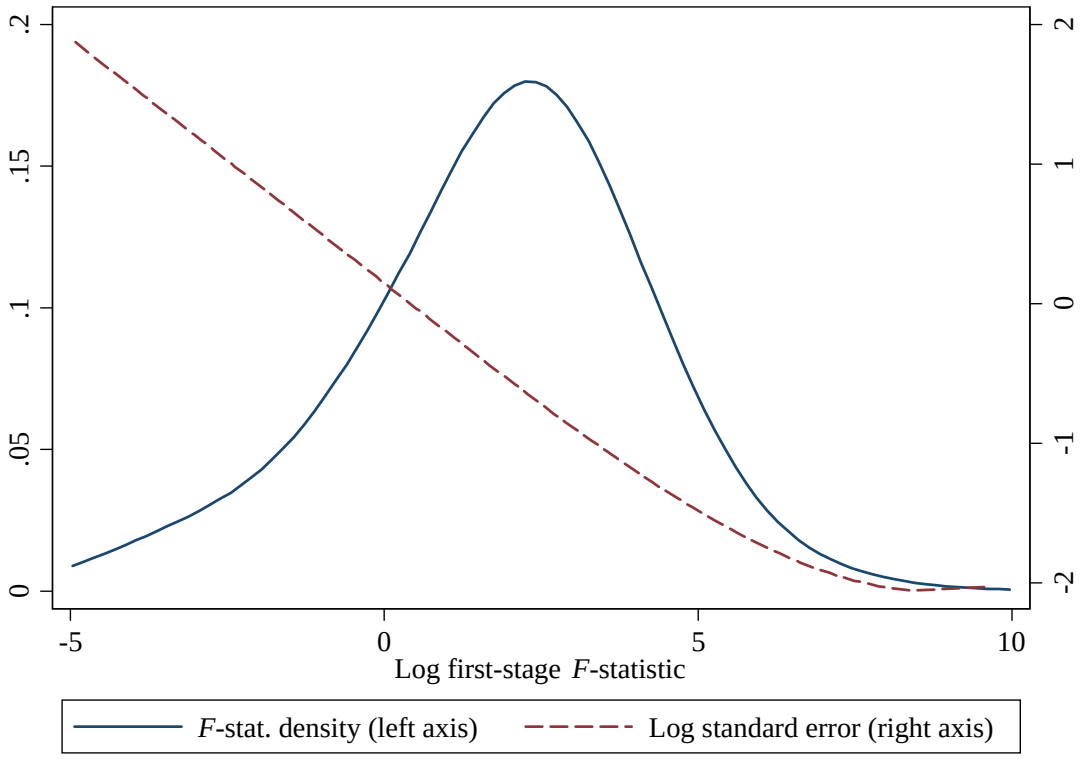
Notes: The solid blue line in this figure plots a Gaussian kernel density estimate of the distribution of ambulance company propensity score estimates used to obtain the 2,082 minimum distance hospital quality estimates. The dashed red line plots the corresponding distribution of propensity scores when hospital referral regions (HRRs) are used instead of hospital service areas (HSAs) to define local hospital markets. The bandwidth used to estimate each distribution is 0.005.

Figure A2: Hospital and HSA Samples

	Multi-Hospital HSAs	HSAs with a Single Hospital	Total
Hospitals with Sufficient Quasi-Experimental Variation	1,677 Hospitals (563 HSAs)	405 (405)	2,082 (968)
Small Hospitals	680 (132)	2,059 (2,059)	2,739 (2,191)
Total	2,357 (695)	2,464 (2,464)	4,821 (3,159)

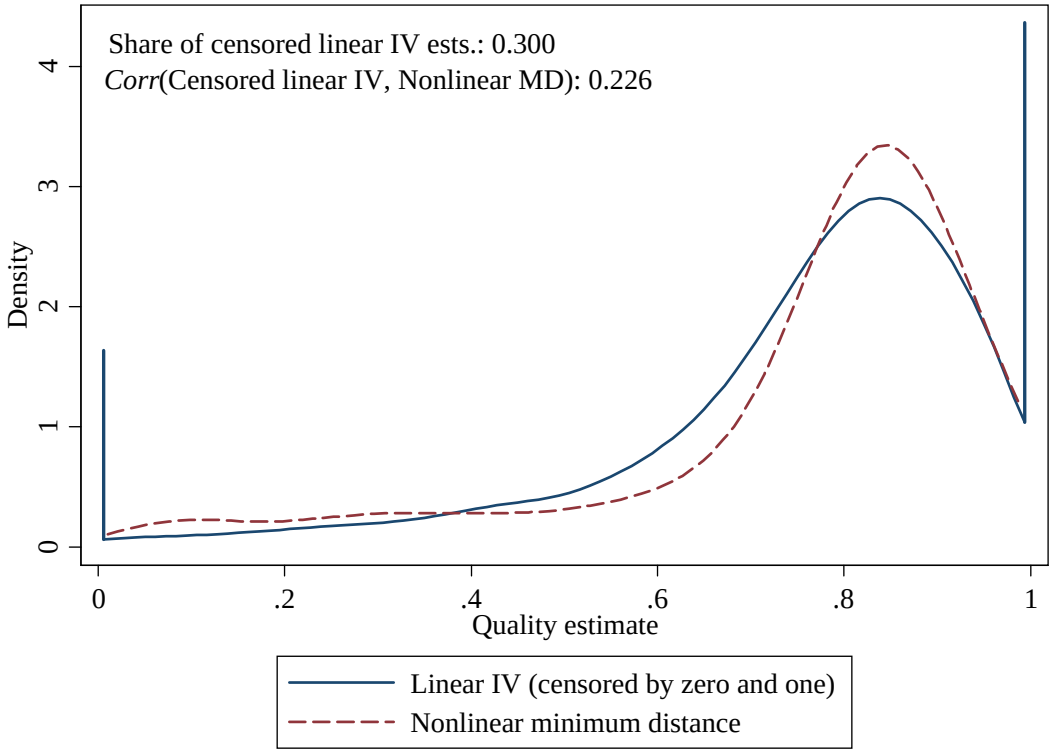
Notes: This figure summarizes the different samples of hospitals and HSAs used at different points of the analysis. The left column counts the number of hospitals (HSAs) operating in a HSA with at least one other hospital, while the right column counts the number of single-hospital HSAs and hospitals. The top row counts the number of hospitals (HSAs) for which first-step quasi-experimental estimates are obtained, while the bottom row counts hospitals which either serve fewer than 25 patients, or are located in a HSA with fewer than 50 patients or fewer ambulance companies serving at least 10 patients than the number of non-small hospitals.

Figure A3: Distribution of First-Stage F-Statistics and Quality Estimate Standard Errors



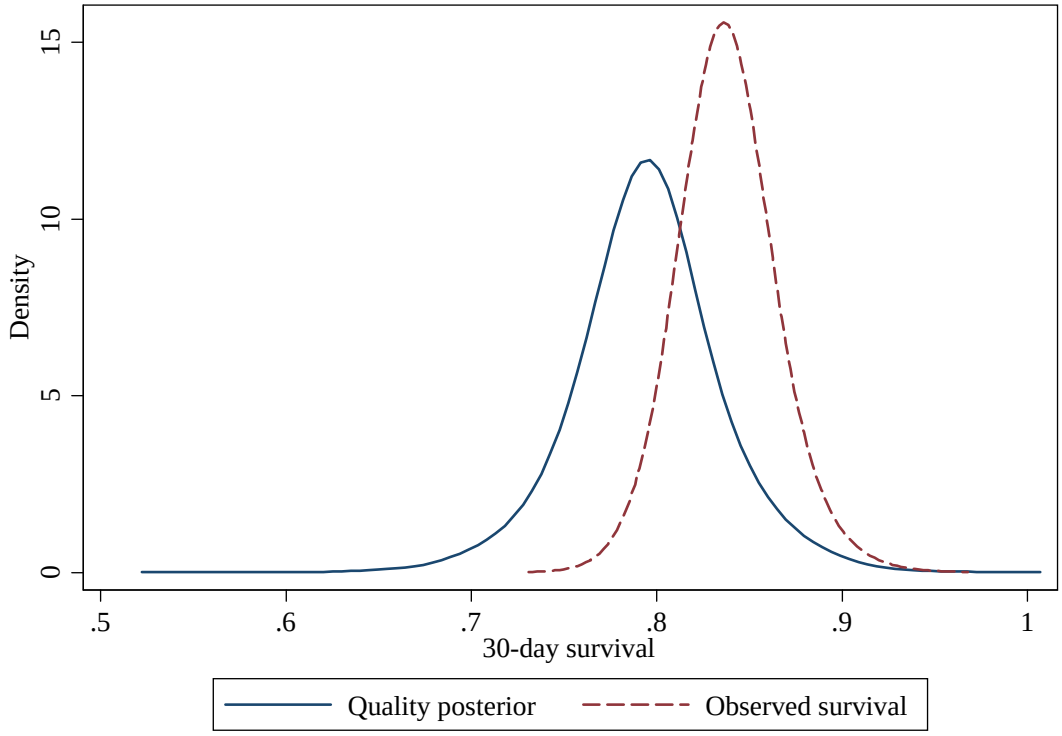
Notes: The solid blue line in this figure plots a Gaussian kernel density estimate of the distribution of log first-stage F-statistics for the 2,082 minimum distance hospital quality estimates. Here F-statistics test the equality of estimated choice probabilities across all ambulance company instruments 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 A4: Distributions of Linear IV and Minimum Distance Quality Estimates



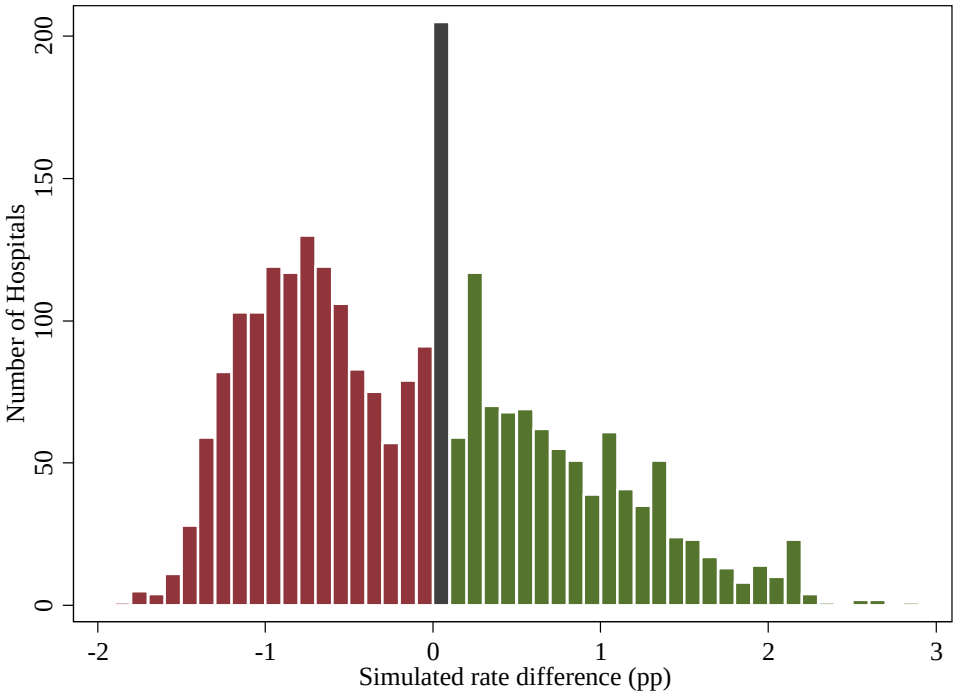
Notes: The solid blue line in this figure plots a Gaussian kernel density estimate of the distribution of linear IV quality estimates for the 2,082 hospitals with sufficient quasi-experimental variation, censored at the logical minimum and mazimum of zero and one. The dashed red line plots a corresponding density estimate of the distribution of nonlinear minimum distance quality estimates, which are by construction bounded between zero and one.

Figure A5: Distribution 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 4,821 hospitals; the bandwidth used to estimate each distribution is 0.02.

Figure A6: Distribution of Simulated Value-Based Purchasing Repayment Rate Differences



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

Table A1: Patient and Hospital Characteristics

	All Nondeferrable Medicare Admissions	Analysis Sample	
		All HSAs	HSAs with Minimum 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	4,821	2,839
HSAs	3,257	3,159	968

Notes: This table reports average patient and hospital characteristics across three samples of Medicare inpatient claims. Column 1 summarizes 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. Column 3 reports characteristics of HSAs in the analysis sample that have enough quasi-experimental data to construct minimum distance quality estimates. Note that the number of hospitals in this column is higher than the number for which first-step quasi-experimental estimates are available, since the latter excludes small hospitals (with fewer than 25 patients) that are active in these HSAs.

Table A2: Hospital RAM Bias Tests

	RAM1 (1)	RAM2 (2)	RAM3 (3)	CMS-RAM (4)
		A. Non-Parametric Test		
Test Statistic (100 d.f.)	295.37	287.78	237.52	186.42
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
		B. Forecast-Based 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]
Overidentification (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 quasi-experimental tests of observational hospital risk-adjustment model (RAM) validity. All RAMs are estimated as hierarchical logit models of 30-day survival, 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 Appendix B.7. 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: Regressions of Hospital Quality Posteriors on Input Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log(Staff Salary)	0.385					0.294	0.256
	(0.069)					(0.081)	(0.081)
Uses Electronic Records		0.173				0.109	0.092
		(0.061)				(0.064)	(0.065)
Uses Case Management Software			0.105			-0.030	-0.072
			(0.046)			(0.048)	(0.048)
# of Accreditations				0.053		0.013	-0.011
				(0.019)		(0.022)	(0.024)
# of Imaging Technologies					0.038	0.023	0.004
					(0.008)	(0.009)	(0.011)
Log(Volume)							0.064
							(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 A4: Correlations of Hospital Quality Indices, RAM Predictions, and 30-Day Survival Rates

	Over Time				Across Conditions			
	2010-12	2007-09	2004-06		Circulatory	Respiratory	Digestive	Injury
	(1)	(2)	(3)		(4)	(5)	(6)	(7)
				A. Quality Indices				
2007-09	0.572			Respiratory	0.621			
2004-06	0.489	0.687		Digestive	0.314	0.741		
2001-03	0.466	0.656	0.658	Injury	0.604	0.253	0.148	
				All Other	0.614	0.580	0.367	0.568
				B. RAM predictions				
2007-09	0.353			Respiratory	0.278			
2004-06	0.302	0.342		Digestive	0.220	0.238		
2001-03	0.221	0.249	0.294	Injury	0.223	0.168	0.160	
				All Other	0.235	0.319	0.193	0.162
				C. Survival Rates				
2007-09	0.446			Respiratory	0.254			
2004-06	0.301	0.444		Digestive	0.252	0.254		
2001-03	0.240	0.329	0.454	Injury	0.292	0.147	0.296	
				All other	0.199	0.318	0.158	0.142

Notes: This table reports estimated correlation coefficients for a hospital's 30-day survival rate, RAM prediction, and quality index. Columns 1-3 correlate 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. See Table A2 for a description of the RAM prediction specification (RAM3). Reported correlations account for estimation error by using hierarchical model estimates of variances and covariances.

Table A5: Average Selection-on-Gains, Adjusting for Selection-on-Observables

	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Selection Bias (pp)	3.83	3.60	3.42	3.63	3.73	3.28
	(0.12)	(0.14)	(0.16)	(0.16)	(0.16)	(0.19)
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 distance 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 Key Results to Alternative Specifications

			Robustness Checks		
		Baseline Specification	Propensity Scores Omit RAM Controls	Health & Utility Follow a t (2) Joint Distribution	HLM Includes (RAM3)×J Interaction
		(1)	(2)	(3)	(4)
Quality Index Posterior Rank Correlation		1.000	0.970	0.853	0.766
RAM3 Coefficient in HLM		0.128	0.125	0.146	0.139
		(0.012)	(0.010)	(0.015)	(0.030)
Within-HSA Quality-Bias Rank Correlation		-0.710	-0.811	-0.868	-0.412
Within-HSA Quality Posterior Correlates	Government	-0.156	-0.181	-0.145	-0.049
		(0.072)	(0.075)	(0.079)	(0.030)
	Log(Spending)	0.021	0.028	0.029	0.015
		(0.013)	(0.014)	(0.013)	(0.006)
	Log(Volume)	0.029	0.043	0.021	0.013
		(0.015)	(0.016)	(0.016)	(0.006)
Share of Positively-Selected HSAs		0.908	0.901	0.888	0.863
HSA-level selection bias (pp)	Average	3.83	3.88	4.35	3.25
		(0.12)	(0.11)	(0.15)	(0.15)
	Distance-Adjusted	3.32	3.48	3.36	3.08
		(0.41)	(0.42)	(0.50)	(0.34)
Correlates of VBP Repayment Rate Change	Teaching	-0.140	-0.136	-0.099	-0.174
		(0.043)	(0.043)	(0.048)	(0.061)
	Log(Spending)	0.404	0.380	0.566	0.834
		(0.072)	(0.074)	(0.083)	(0.156)
Expected Survival Gain from Redirection	Max. RAM3	-0.65	-0.47	-0.31	-0.83
	Max. Quality	0.67	0.96	1.26	0.51

Notes: Column 1 of this table summarizes key results from the baseline empirical approach, 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.