Hospital Quality Risk Standardization via Approximate Balancing Weights*

Luke Keele† Eli Ben-Michael‡ Avi Feller§ Rachel Kelz¶ Luke Miratrix‖

May 19, 2020

Abstract

Comparing outcomes across hospitals, often to identify underperforming hospitals, is a critical task in health services research. However, naive comparisons of average outcomes, such as mortality rates, can be misleading because hospital case mixes differ—mortality rates may be lower due to more effective treatments or simply because the hospital serves a healthier population overall. Popular methods for adjusting for different case mixes, especially "indirect standardization," are prone to model misspecification, can conceal overlap concerns, and produce results that are not directly interpretable. In this paper, we develop a method of "direct standardization" where we re-weight each hospital patient population to be representative of the overall population and then compare the weighted averages across hospitals. Adapting methods from survey sampling and causal inference, we find weights that directly control for imbalance between the hospital patient mix and the target population, even across many patient attributes. We also derive principled measures of statistical precision, and use outcome modeling and Bayesian shrinkage to increase precision and account for variation in hospital size. We demonstrate these methods using claims data from Pennsylvania, Florida, and New York, estimating standardized hospital complication rates for general surgery patients. We conclude with a discussion of how to detect low performing hospitals.

Keywords: Risk Adjustment, Weighting, Direct Standardization

*Eli Ben-Michael and Avi Feller gratefully acknowledge funding support from the National Science Foundation Grant #1745640. The dataset used for this study was purchased with a grant from the Society of American Gastrointestinal and Endoscopic Surgeons. Although the AMA Physician Masterfile data is the source of the raw physician data, the tables and tabulations were prepared by the authors and do not reflect the work of the AMA. The Pennsylvania Health Cost Containment Council (PHC4) is an independent state agency responsible for addressing the problems of escalating health costs, ensuring the quality of health care, and increasing access to health care for all citizens. While PHC4 has provided data for this study, PHC4 specifically disclaims responsibility for any analyses, interpretations or conclusions. Some of the data used to produce this publication was purchased from or provided by the New York State Department of Health (NYSDOH) Statewide Planning and Research Cooperative System (SPARCS). However, the conclusions derived, and views expressed herein are those of the author(s) and do not reflect the conclusions or views of NYSDOH. NYSDOH, its employees, officers, and agents make no representation, warranty or guarantee as to the accuracy, completeness, currency, or suitability of the information provided here. The authors declare no conflicts.

†University of Pennsylvania, Philadelphia, PA, Email: luke.keele@gmail.com

‡University of California, Berkeley, Berkeley, CA, Email: ebenmichael@berkeley.edu

§University of California, Berkeley, Berkeley, CA, Email: afeller@berkeley.edu

¶University of Pennsylvania, Philadelphia, PA, Email: Rachel.Kelz@pennmedicine.upenn.edu

‖Harvard University, Cambridge, MA, Email: lmiratrix@g.harvard.edu

1 Introduction: Judging Hospital Quality

Measuring hospital quality is a key analytic tool in health services research, both for comparing results across hospitals and for identifying hospitals for improvement. The simplest measure of hospital quality is mortality rate, the proportion of patient fatalities for a fixed period of time. However, comparisons of hospital-level outcomes like mortality rates are complicated by the fact that patient mix varies across hospitals: hospitals with high mortality rates may treat patient populations with complex, chronic conditions, while hospitals with low mortality rates may treat a healthy patient population. In short, good patient outcomes may be due to high quality care or to a healthy patient population requiring less invasive and simpler medical procedures.

Risk adjustment, also known as risk standardization, is a set of methods that adjust the hospital patient mix to make hospital outcomes more comparable (Normand et al., 2007). Typically, risk adjustment uses a statistical model to compare a hospital level outcome such as mortality to what would be expected to occur had the hospital had a patient mix comparable to the overall population. Risk adjustment is widely used to evaluate hospitals and provide the public with information on hospital quality. For example, the online tool, Hospital Compare, provided by Medicare, uses risk standardization to help patients identify high quality hospitals. The most common approach is called indirect standardization via outcome modeling (Iezzoni, 2012). These methods, however, can perform when there is large variation in hospital size and may fail to fully account for differences in patient-mix (George et al., 2017).

In this paper, we develop a new weighting-based approach to risk standardization. In our approach, we view each hospital's patient population as a non-representative sample from the overall patient population. We then generate a set of weights for each hospital so the weighted distribution of its patients matches the overall population. We show that this form of direct standardization both reduces bias due to systematic differences in served populations across hospitals while maintaining reasonable precision due to preserving most of the data.

This method is inspired in part by "template matching" (Silber et al., 2014), where each hospital is rated by the average outcome of an identified set of patients that closely matches a canonical list of patient characteristics that serves as a kind of scorecard. Compared to template matching, we show substantial

gains in both bias control and precision. We also identify a bias-precision tradeoff, and find that by regularizing the weights we can substantially increase precision in our hospital specific estimates while only incurring what appears to be a small cost in bias. Moreover, after standardization via weighting, we then apply a Bayesian shrinkage estimator to better account for variability in the size of hospitals. We use claims data from Pennsylvania, New York, and Florida to judge hospital quality for general surgical procedures and to motivate the methods that we develop. We review the details of the application next.

1.1 Hospital quality in PA, FL, and NY on General Surgical Performance

General surgery consists of high volume surgical procedures that are done in almost all hospitals, including procedures such as appendectomy (removal of the appendix), cholecystectomy (gall bladder), hernia repairs, and mastectomies. Since deaths are (thankfully) rare in general surgery, we use postoperative complications (e.g., infections and bleeds) as an indication of a problematic surgical procedure. We assess hospital quality in general surgery by estimate at risk-adjusted rates of such complications.

In our analysis, we use data based on all-payer hospital discharge claims in New York, Florida, and Pennsylvania from 2012-2013. The data include patient sociodemographic and clinical characteristics including a measure of patient frailty, an indicator for sepsis, and 31 indicators for comorbidities based on Elixhauser indices (Elixhauser et al., 1998), as well as admission type (emergency, urgent, or elective), type of insurance, and age. Overall, we focus on 44 general surgery operations.[1] Across the three states, we have a total of 621,667 patients in 523 hospitals. Our primary outcome of interest is a binary indicator for the development of one or more complications after general surgery (identified using ICD-9-CM diagnosis codes). Our goal is to develop risk-standardized measure of quality for these hospitals based on observed complications. Next, we motivate the need for risk adjustment using our data.

Figure 2 displays boxplots of hospital-level averages of three key patient characteristics: whether a patient is African-American, whether a patient is obese, and whether the procedure was an emergency admission. All three characteristics are important predictors of complications in the cohort. As the

[1]We restrict the patient population to those patients who had a surgical procedure included in the Agency for Healthcare Research and Quality (AHRQ) Clinical Classifications Software (CCS). CCS categories uses International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes to classify whether procedures are surgical or not (Decker et al., 2014). We also removed any hospitals that performed fewer than 30 procedures over the two-year period, which removed 70 hospitals (out of 593) and 605 patients (out of 622,272).

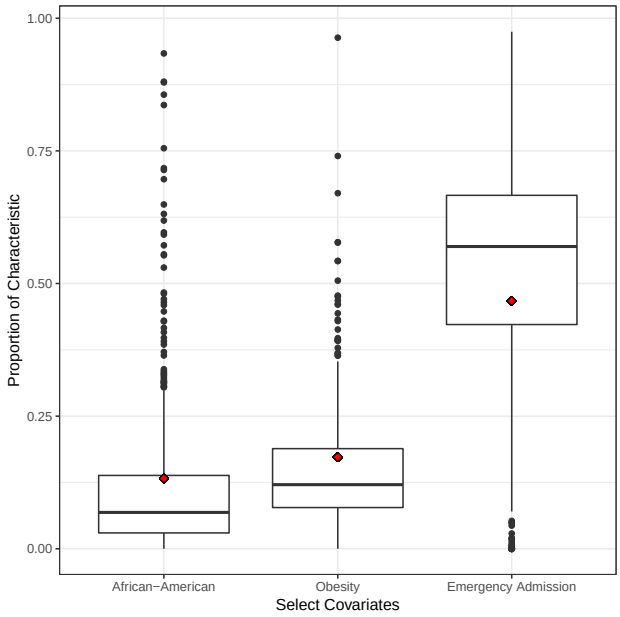


Figure 1: Boxplots of Hospital Casemix Distributions. Diamond represents population mean for that covariate.

boxplots show, there is substantial variation in all three attributes. For instance, only 17 percent of patients in the sample are obese, while several hospitals have patient populations in which more than half are obese. The goal of standardization is to adjust for differences in patient mix like these, allowing us to more directly compare outcomes across hospitals.

Our paper proceeds as follows. First, we review current methods of standardization. Next, we outline our approach to risk standardization using balancing weights. We then derive methods of variance estimation that are consistent with the estimated weights. Next, we show how agnostic regression modeling can be used to increase precision in the hospital level estimates. We then outline how a Bayesian shrinkage estimator can be applied to account for variance in hospital size. We apply our methods to the claims data on general surgery. We then conclude.

2 Risk Standardization for Comparing Hospital Outcomes

There is a large literature in statistics and health services research on risk adjustment; see Normand et al. (2007) for a review. There are two main types of risk adjustment: indirect and direct. Indirect standardization is the current standard in applied research and is used in online tools such as Medicare's Hospital Compare (Ash et al., 2012). Direct standardization is less common, partly because it is thought

to be too limited. Longford (2019) outlines how risk adjustment via both indirect and direct standardization can be viewed as a causal inference problem from within the potential outcomes framework. We next review these two general approaches, highlighting connections to our work and framing.

2.1 Review: Indirect Standardization

Under indirect standardization, observed outcomes for patients are compared to expected outcomes derived from a statistical model fit to the larger patient population (Fleiss et al., 2003; Kitagawa, 1955; Iezzoni, 2012; Silber et al., 1995). More specifically, a linear or generalized linear model is fit where an outcome such as costs or mortality is regressed on patient characteristics using the entire patient population. This risk adjustment model is used to predict outcomes for patients in a specific hospital, and the average of these predictions serve as the expected outcome for the provider. This expected outcome is then compared to the observed outcomes in the same hospital. Commonly, this comparison is computed as the ratio of observed to expected outcomes or O/E ratio. Most research on statistical methods for indirect standardization has focused on the model used for risk adjustment. Early work used classical linear or generalized linear models, but regression models with random effects are now standard (Iezzoni, 2012). The Centers for Medicare Hospital Compare tool, for example, is based on a random effects model (Krumholz et al., 2006). See Ash et al. (2012) for a detailed overview of the tradeoffs associated with the various statistical models used for indirect standardization.

While methods of indirect standardization are widely used, these methods have a number of shortcomings. Indirect standardization relies on generalized linear models that impose strong functional form assumptions. If a hospital's patient population differs from the overall population, the estimated score will rest heavily on model-based extrapolation. The resulting standard errors will also be overly optimistic: uncertainty will be primarily driven by the sample size of patients in a hospital without regard to whether some groups of patients are substantially underrepresented. See George et al. (2017) for a general critique of indirect standardization. In sum, they show that indirect standardization fails to fully account for patient case mix compared to direct standardization.

2.2 Review: Direct Standardization

Rather than fit an outcome model, direct standardization directly adjusts for the hospital case mix typically by weighting outcomes to a target distribution. This allows the analyst to directly target the differences between hospital covariate distributions and the population covariate distributions (George et al., 2017). However, direct standardization methods have traditionally been limited by the inability to incorporate more than a few patient level variables (Iezzoni, 2012). For example, in many direct standardization analyses, outcomes are only risked adjust for age. Direct standardization also allows investigators to easily avoid ratio measures and produce standardized hospital level outcomes on the original scale.

Template matching is a recently developed form of direct standardization that can easily risk adjust for many patient level covariates (Silber et al., 2014). Under template matching, one estimates how different hospitals would perform with patients similar to the overall patient population. For each hospital, matching methods are used to find a subset of patients that are highly comparable to the overall patient population. Hospital quality is judged using this set of matched patients that are representative of the patient population. Template matching serves as an important breakthrough in methods for standardization as it both avoids the strong parametric assumptions of models needed for indirect standardization and provides an estimate of hospital quality that does not rely on a ratio estimate.

However, template matching may be inefficient since it only uses a limited part of the available data. Template matching also requires tuning a large number of hospital specific matches which may be infeasible. For example, in a match, we might need to determine specific propensity score caliper value to reduce imbalances. In a template match, one might have to find over 500 optimal caliper values. Finally, extant work on template matching has not addressed best practice for estimating statistical precision. We address these issues via risk standardization using weighting methods.

2.3 Review: Balancing weights

Weighting methods have a long history in survey sampling and causal inference (Horvitz and Thompson, 1952; Lohr, 2009; Robins and Rotnitzky, 1995). If we were only interested in balancing a small number of patient characteristics, we could directly apply classical calibration approaches from survey sampling

(see e.g. Deming and Stephan, 1940; Deville and Särndal, 1992; Deville et al., 1993). In this case, the resulting weights would achieve *exact balance* where the re-weighted and target covariate averages are equal.

In our setting, however, we need to find weights that balance a large number of patient characteristics and comorbidities, so achieving exact balance in infeasible, especially for smaller hospitals. We therefore adapt recent advances in the causal inference literature to allow for *approximate balance*, which yields a feasible optimization problem at the cost of accepting some small bias (Zubizarreta, 2015; Hirshberg et al., 2019). See Ben-Michael et al. (2020) for a recent review of these weighting methods.

3 Direct Standardization via Approximate Balancing Weights

We now develop a weighting method for direct standardization, which solves a convex optimization problem to optimize for balance and effective sample size. We briefly describe the basic setup and then turn to specific implementation details. The notation we outline below is consistent with the framework outlined in Longford (2019) using potential outcomes. Longford (2019) conceives of risk adjustment as a counterfactual question about how a hospital would perform with a different set of patients. In this framework, hospitals are themselves the treatment of interest.

3.1 Hospitals as Non-Representative Samples

In our data, we observe $i = 1, \ldots, n$ patients nested in hospitals $j = 1, \ldots, J$, with patient hospital indicator $Z_i \in \{1, \ldots, J\}$ and n_j patients in each hospital.[2] For each patient, we observe a vector of background covariates $X_i \in \mathbb{R}^d$. We also observe an outcome Y_i, which in our data is a binary indicator for a postoperative complication.

The primary statistical problem is that the distribution of patient- and surgery-level characteristics vary across hospitals — $p(X \mid Z = j) \neq p(x \mid Z = j')$ for $j \neq j'$ — so the difference between the average outcomes between two hospitals reflects both differences in hospital quality and differences in the distribution of patient attributes. We develop a notion of standardized hospital quality by considering the hypothetical situation in which a hospital's patient mix is instead the population patient mix; this

[2] In principle, each observation is a patient-surgery pair. Since we focus only on one surgery per patient, we ignore this complication in our exposition.

is the motivating spirit behind template matching (Silber et al., 2014).

Formally, denote the expected value of our outcome given observed covariates x and hospital j as $m_j(x) = \mathbb{E}[Y \mid X = x, Z = j]$. We can think of $m_j(x)$ as a "quality curve" of the hospital: it describes our expected outcome for hospital j when serving a patient with characteristics x. The expected overall average outcome in hospital j is then

$$\rho_j = \mathbb{E}[Y \mid Z = j] = \int m_j(x) dP(x \mid Z = j).$$

This quantity is easily estimated by the raw mean of hospital j, $\bar{Y}_j \equiv \frac{1}{n_j} \sum_{Z_i = j} Y_i$. These ρ_j are not directly comparable across hospitals: they could be systematically different as they are averaging quality curves over different distributions.

The goal of risk adjustment is to remove the dependence between X and Z. We do this by considering a set of hospital estimands that each take the expectation of $m_j(x)$ over a common distribution $X \sim P^*$:

$$\mu_j^{P^*} = \int m_j(x) dP^*(x). \tag{1}$$

These μ_j are more directly comparable as we have removed systematic differences in distribution. There are many distributions that we may consider shifting towards, e.g. a region of high overlap between hospital patient distributions, or the marginal distribution of patients. Here, we focus on one simple estimand: the empirical distribution of the covariates across all hospitals. This gives

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} m_j(X_i). \tag{2}$$

We can view this estimand as the expected outcome of hospital j if it were given a patient mix the same as the full population of patients in our sample. To avoid complications, we assume that any type of patient, as defined by X, could receive care at any hospital, at least in principle. We formalize this as an overlap assumption:[3]

Assumption 1 (Overlap)**.** $0 < P(Z = j \mid X = x)$

Assumption 1 rules out the possibility that a hospital would never treat a particular type of patient; e.g.,

[3] In principle, we could restrict our estimand to a set of patients where there is overlap. We leave this to future work.

a hospital that only treats women or that doesn't perform a certain type of surgery.

3.2 Estimating hospital means

With direct standardization, we estimate the average population outcome for hospital j, μ_j, with a weighted average of observed outcomes for hospital j, using normalized weights $\hat{\gamma}$:

$$\hat{\mu}_j = \sum_{Z_i=j} \hat{\gamma}_i Y_i, \tag{3}$$

with $\sum_{Z_i=j} \hat{\gamma} = 1$. This equation demonstrates how our method method of direct standardization differs from indirect standardization: we examine how hospital performance differs from expected across patient characteristics, giving the same weight for each patient type to each hospital. By contrast, indirect standardization examines how hospitals differ from expected, given the model, weighting by those patients the hospitals serve.

To make this more concrete, we further assume our quality function $m_j(x)$ is a linear function of some transformation of the covariates:

$$m_j(x) = \beta_j \cdot \phi(x),$$

with $\phi : \mathbb{R}^d \to \mathbb{R}^p$ and $\beta_j \in \mathbb{R}^p$. See Kallus (2016); Hirshberg et al. (2019) and Hazlett (2019) for generalizations to the fully non-parametric case, although we note the $\phi(x)$ function allows for easy inclusion of interactions and higher order terms.

Given $m_j(x)$, $\varepsilon_i \equiv Y_i - \beta_j \cdot \phi(X_i)$ is the *residual* of outcome Y_i given covariates X_i and hospital $Z_i = j$. We can then express the difference between the observed raw average in hospital j, $\bar{Y}_j$ and the target estimand μ_j:

$$\hat{\mu}_j - \mu_j = \underbrace{\beta_j \cdot \left(\bar{\phi} - \frac{1}{n_j} \sum_{Z_i=j} \hat{\gamma}_i \phi(X_i) \right)}_{\text{bias}} + \underbrace{\sum_{Z_i=j} \hat{\gamma}_i \varepsilon_i}_{\text{variance}}, \tag{4}$$

with $\bar{\phi} \equiv \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$ the overall population mean of the covariate vector $\phi(x)$.

Inspecting Equation (4), we see that the error decomposes into two components: (1) systematic bias due to imbalance in $\phi(X_i)$ between hospital j and the overall sample; and (2) idiosyncratic variance due to noise. The goal is to find weights that control both terms.

For the bias term, the challenge is that the coefficients β_j are unknown. Using the Cauchy-Schwarz inequality, we can see that controlling the imbalance in $\phi(X_i)$ also controls the systematic bias (conditional on X and Z):

$$\left|\mathbb{E}\left[\mu_j - \hat{\mu}_j \mid X, Z = j\right]\right| \leq \|\beta_j\|_2 \underbrace{\left\|\bar{\phi} - \frac{1}{n_j}\sum_{Z_i=j}\hat{\gamma}_i\phi(X_i)\right\|_2}_{\text{Imbalance}}. \tag{5}$$

Thus, we can directly reduce bias regardless of the true β_j. If there were only a small number of patient characteristics, we could likely achieve exact balance, where the imbalance term in Equation (5) is exactly zero. However, this is not feasible in our setting; the goal is to make the imbalance term as small as possible, all else equal.

For the variance term, the individual ε_i are unknown. To make this tractable, we therefore assume that, within each hospital, the residuals ε_i are homoskedastic and independent with variance σ_j^2. Under this assumption, the variance of $\hat{\mu}_j$ (conditional on the weights) is $\text{Var}\left\{\hat{\mu}_j | X, Z\right\} = \sum_{Z_i=j}\hat{\gamma}^2\sigma_j^2$; in turn, the sum of squared weights controls the variance:[4]

$$\text{Var}\left(\mu_j - \hat{\mu}_j \mid X, Z\right) = \sigma_j^2 \sum_{Z_i=j}\hat{\gamma}_i^2. \tag{6}$$

All else equal, the more we can reduce this sum of squares for a given hospital — or, equivalently, make the weights more homogenous — the more precise the estimate.

3.3 Weighting via convex optimization

We can now combine these two objectives into the following optimization problem:

$$\begin{aligned} \min_{\gamma} \quad & \sum_{j=1}^{J}\left[\left\|\bar{\phi} - \sum_{Z_i=j}\gamma_i\phi(X_i)\right\|_2^2 + \lambda n_j \sum_{Z_i=j}\gamma_i^2\right] \\ \text{subject to} \quad & \sum_{i=1}^{n} n_{Z_i}\gamma_i\phi(X_i) = n\bar{\phi} \\ & \sum_{Z_i=j}\gamma_i = 1 \\ & \ell \leq \gamma_i \leq u \end{aligned} \quad . \tag{7}$$

[4] Allowing for heteroskadasticity within hospitals, we can instead *bound* the variance by replacing σ_j^2 with the maximum variance in the hospital $\max_{Z_i=j}\sigma_i^2$.

The optimization problem (7) trades off two competing terms for each hospital j: better balance (and thus lower bias) and more homogeneous weights (and thus lower variance).[5] A global hyperparameter λ negotiates the tradeoff: when λ is large the optimization problem will prioritize variance reduction and search for more uniform weights, when λ is small it will instead prioritize bias reduction. We explore the role of λ in the bias-variance tradeoff empirically in Section 5.2.

The constraint set in Equation (7) has three components. First, we constrain the weights so that the overall weighted average across all hospitals is equal to the unweighted average; this ensures that weighting procedure does not change the overall population. Second, we constrain the weights to sum to one within each hospital, ensuring that each hospital estimate is in fact a weighted average of its outcomes. This constraint also stabilizes the estimates and ensures that they are sample-bounded, that is, that $\hat{\mu}_j \in [\min_{Z_i=j} Y_i, \max_{Z_i=j} Y_i]$. Third, we constrain the weights to have lower bound ℓ and upper bound u. We set the lower bound $\ell = 0$ so that weights are non-negative and do not extrapolate outside of the support of the data. Setting the upper bound to $u < 1$ would prevent the optimization problem from putting too much weight on any single patient in a hospital. For example, setting $u = .2$, would ensure that we do not put more than 20% of the weight on any individual patient. In our primary results, we set $u = 1$ (no constraint) and investigate the impact of setting u to be less than 1 in our supplementary material. With both the nonnegative constraint and the unit sum constraint, each individual weight $\hat{\gamma}_i$ corresponds to the fraction of hospital Z_i's outcome dictated by unit i.

The optimization problem above obtains $\hat{\gamma}_i$ without regard to outcome. Similar to matching and propensity score methods in observational studies, this is a design step where we set up our final evaluation using covariate information alone. We then simply estimate the adjusted hospital means by taking a weighted average of the patient outcomes, $\hat{\mu}_j = \sum_{Z_i=j} \hat{\gamma}_i Y_i$. Finally, one concern under direct standardization is a lack of overlap between the covariate distribution of a specific hospital and the patient population. Under our weighting approach, extreme weights will signal if a covariates for a specific hospital do not overlap with the patient population.

[5] For a single hospital, the objective in optimization problem (7) reduces to a special case of the minimax linear estimation proposal from Hirshberg et al. (2019), with a particular choice of function class. The above extends to the case with multiple hospitals.

4 Additional Extensions

Thus far our methodology allows analysts to estimate risk standardized hospital outcomes. However, accounting for variability in our estimates is also critical. In this section, we consider variance estimation and two additional methods that can be used to improve precision and account for variation in hospital size.

4.1 Variance estimation

We can obtain uncertainty for the estimated means using standard results from survey sampling. Under the assumption that individual outcomes are independent within a hospital, the sampling variance (conditional on the weights) is:

$$\mathsf{Var}\left\{\hat{\mu}_j|\hat{\gamma}_j\right\} = \mathsf{Var}\left\{\sum_{Z_i=j}\hat{\gamma}_j Y_i\right\} = \sum_{Z_i=j}\hat{\gamma}_j^2 \mathsf{Var}\left\{Y_i\right\}.$$

Under the further assumption that individual observations are either homoskedastic or are independent of the weights we can then estimate this variance using a plug in:

$$\widehat{\mathsf{se}}(\hat{\mu}_j|\hat{\gamma}_j) = \left[\hat{\sigma}_j^2 \sum_{Z_i=j}\hat{\gamma}_j^2\right]^{1/2} = \frac{\hat{\sigma}_j}{\sqrt{n_j^{\mathsf{eff}}}}, \tag{8}$$

where

$$n_j^{\mathsf{eff}} \;\equiv\; \left(\sum_{Z_i=j}\gamma_i\right)^2 \Big/ \sum_{Z_i=j}\gamma_i^2 \;=\; 1 \Big/ \sum_{Z_i=j}\gamma_i^2$$

is the *effective sample size* for hospital j (Lohr, 2009), and where the last equality above assumes the weights sum to 1 within hospital. The effective sample size is the inverse of the dispersion penalty in the balancing weights optimization problem in Equation (7).

Under no pooling across hospitals, we estimate σ_j for the above using the (weighted) residual variance:

$$\hat{\sigma}_j^2 = \frac{1}{\sum_{Z_i=j}\hat{\gamma}_i^2 - 1}\sum_{Z_i=j}\hat{\gamma}_i^2(Y_{ij} - \hat{\mu}_j)^2.$$

In practice, however, individually estimated $\hat{\sigma}_j$ can behave poorly, particularly for smaller hospitals. This

complicates subsequent adjustments, especially partial pooling across hospitals, as we discuss in Section 4.3. We instead propose a fully pooled estimate for the residual variances:

$$\hat{\sigma}^2_{\mathsf{pool}} = \frac{1}{N^{\mathsf{eff}}} \sum_{j=1}^{J} n_j^{\mathsf{eff}} \hat{\sigma}_j^2,$$

where $N^{\mathsf{eff}} = \sum_j n_j^{\mathsf{eff}}$ is the pooled effective sample size. This method is adapted from Weiss et al. (2017), who pool to stabilize Empirical Bayes estimates of cross-site heterogeneity in the context of multi-site trials. Finally, we then use $\hat{\sigma}_{pool}$ to compute the standard error in Equation (8) rather than $\hat{\sigma}_j$.

4.2 Covariate adjustment

The optimization problem in Section 3 reweights each hospital based on the covariates but does not exploit any relationship between covariate and outcome to improve precision. By contrast, indirect standardization removes variation in the outcome explained by the covariates in order to increase overall precision, albeit at the price of additional model dependence.Borrowing from model-assisted survey sampling (Sarndal et al., 2003), we can also improve the precision of the risk adjusted hospital means via agnostic regression adjustment — and do so without requiring that the adjustment model be correctly specified. To do this we fit an initial working least squares regression model to the full data,

$$Y_i = \phi(X_i)'\eta,$$

where $\phi(X_i)$ are the same (possibly transformed) patient-level background characteristics used for standardization. After we fit this model, we generate empirical residuals for Y_i,

$$\hat{\epsilon}_i = Y_i - \phi(X_i)'\hat{\eta}.$$

This regression uses only patient level information and *does not* incorporate hospital fixed effects. Thus, we generally expect these residuals to be positively correlated with hospital quality. As in model-assisted survey sampling, we can then adjust the overall prediction from our model with the weighted

residuals:

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)'\hat{\eta} + \sum_{Z_i=j}\hat{\gamma}_i(Y_i - \phi(X_i)'\hat{\eta}).$$

By apply the weights to these residuals, this should reduce variation in the risk adjusted measures using outcome information and lead to more precise hospital level estimates after risk standardization. We also use $\hat{\epsilon}_i$ to estimate $\hat{\sigma}^2_{\text{pool}}$.

4.3 Partially Pooling Hospital-Specific Estimates

Finally, our estimation strategy gives individual estimates of $\hat{\mu}_j$ in relative isolation. These estimates can be unstable, especially for the smaller hospitals and those hospitals with low n_j^{eff}. Following standard practice in hospital quality research, we therefore partially pool the estimates via a hierarchical Bayesian model (Normand et al., 2007; Iezzoni, 2012; George et al., 2017). For our application, we focus on a simple, modular Bayesian procedure (Jacob et al., 2017), which allows for partial pooling while maintaining transparency.[6]

The important change from the no pooled estimate is that we now assume that the hospital-specific scores are drawn from an underlying random effects distribution, G:

$$\hat{\mu}_j \sim N(\mu_j, \widehat{\text{se}}_j^2)$$
$$\mu_j \sim G,$$

where we plug in the estimated standard error, $\widehat{\text{se}}_j$, which we treat as known. Taken together, our approach to risk adjustment therefore achieves several distinct goals: (1) adjust for possible bias via weights; (2) account for sampling variability due to differences in the number of cases per hospital; and (3) partially pool across hospitals via hierarchical models.

[6] It is straightforward to extend this setup to a fully Bayesian approach or to an empirical Bayes setup, such as via triple goal estimation (Paddock et al., 2006).

5 Hospital Performance on Postoperative Complication in General Surgical Performance

5.1 Setup

We now apply our apply our approach to estimate risk-adjusted complication rates for general surgery patients in Pennsylvania, New York, and Florida. In addition to the sample restrictions in Section 1.1, we also preprocess the patient characteristics by standardizing. This is important in practice because covariates with high variances implicitly receive greater weight in the optimization problem. For continuous covariates and binary covariates with estimated proportion $\hat{p} \geq 0.05$, we standardize by subtracting the mean and dividing by the standard deviation. For binary variables with rare outcomes, $\hat{p} < 0.05$, we standardize by $\sqrt{0.05 \cdot 0.95}$ instead of $\sqrt{\hat{p}(1-\hat{p})}$, which prevents extremely rare covariates from receiving too much weight in the optimization process.

To assess the success of our weighting procedure, we focus on the increase in precision and reduction in bias. For precision, we calculate the implied effective sample size, n_j^{eff}. For bias, we calculate the improvement in (weighted) covariate imbalance by hospital. For each hospital, we calculate $\overline{\phi(X)}_h$, the unadjusted covariate means, and $\overline{\phi(X)}_{h,w}$, the weighted hospital means, using the weights from our procedure. Next, we regress Y_i on $\phi(X_i)$ to obtain a vector of regression coefficients $\hat{\eta}$, which give us variable importance weights. Using these estimates, we can then then estimate initial bias, $\Delta_h = (\overline{\phi(X)}_h - \overline{\phi(X)})'\hat{\beta}$, and final bias, $\Delta_{h,w} = (\overline{\phi(X)}_{h,w} - \overline{\phi(X)})'\hat{\beta}$, for each hospital h. The first quantity is the estimated bias due to baseline differences in case mix; the second is the remaining bias due to case mix after weighting. Using these two quantities, we then calculate the Percent Bias Reduction (PBR):

$$PBR = 100\% \times \left[\frac{1}{H} \sum_h |\Delta_{h,w}| \Big/ \frac{1}{H} \sum_h |\Delta_h| \right]$$

This measure describes the change in bias due to direct standardixation while also accounting for the strength of the association between the different covariates and the outcomes.

5.2 The Role of λ

An important tuning parameter in our approach is λ, the global hyperparameter that controls the bias-variance tradeoff: when λ is large the optimization problem prioritizes variance reduction and searches

for more uniform weights; when λ is small it instead prioritizes bias reduction, allowing extreme weights that can reduce n_j^{eff} and thus increase the variance in the performance estimates for each hospital. To investigate the role of λ, we estimated weights for each of a series of λ values ranging between 0 and 3.5. For each λ value, we computed the average PBR and average effective sample size across all hospitals. In this analysis, we do not apply additional adjustment via regression modeling. As we noted above, we set $u = 1$ (no constraint) which does not limit the amount of weight assigned to any one patient. In the supplemental materials, we present results for an analysis with $u = .2$. We found the results were unchanged.

Figure 2 contains a summary of the results from this analysis and demonstrates the trade-off between bias reduction and effective sample size. When $\lambda = 0$, bias is reduced by nearly 80%, but the average effective sample size is less than 200. We cannot achieve perfect balance due to the constraints on the weights within hospitals — some hospitals do not have a patient mix that allows for perfect covariate balance after reweighting. Conversely, when $\lambda = 3.5$ bias reduction is approximately 47% but the average effective sample size is nearly 1000, an increase of more than a factor of 3. The results in Figure 2 suggest that we might reasonably select a value for λ of around 0.05, which decreases bias reduction from our maximum possible of 80% by approximately three percentage points, but essentially doubles the average effective sample size.

As a comparator, we also implemented a template match. Following Silber et al. (2014), we first created a template by taking 500 random samples from the patient population each with a sample size of 300. Of these 500 random samples, we selected the sample with the smallest discrepancy between the random sample and the overall population means; this set of 300 patients serve as the template. Next, we matched patients from from each hospital to the template, using optimal match with refined covariate balance, which is an extension of fine or near-fine balance designed to balance the joint distribution of many nominal covariates (Pimentel et al., 2015). In the match, we employed both a propensity score caliper and optimal subsetting. Thus for each hospital, we obtain the largest set of patients that are similar to the patients in the template. This implies that for each match, we can obtain up to 300 matched pairs, since that is the size of the template. However, the number of matched pairs may be smaller if only a subset of patients are comparable to the template. The patients selected from each hospital serve as the risk adjusted population for that hospital. The subsequent risk adjusted measure of

complications is then the proportion of complications in the matched sample. For the template match, we also calculate the PBR and average sample size across hospitals.

Figure 2 shows that template matching, which does not directly minimize imbalance, does not have comparable performance: bias reduction is under 50%, less than some of the most regularized λ considered, and the average effective sample size is less than 300, only slightly above the fully unregularized $\lambda = 0$. See the supplemental materials for more detailed results from this analysis.

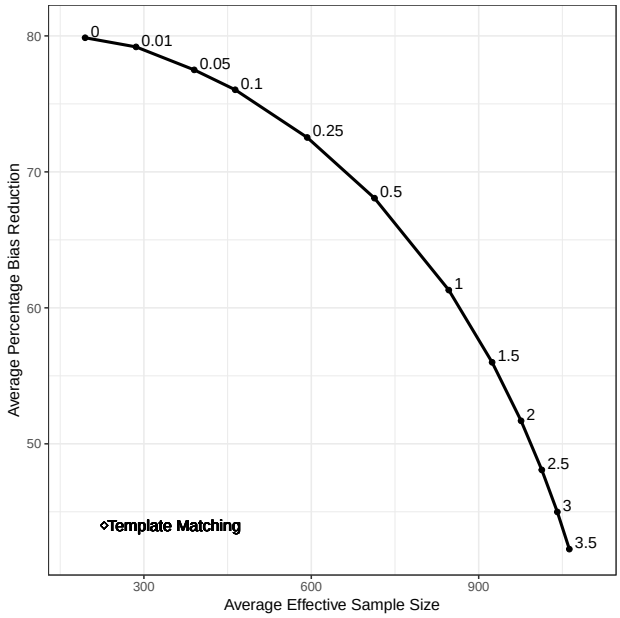
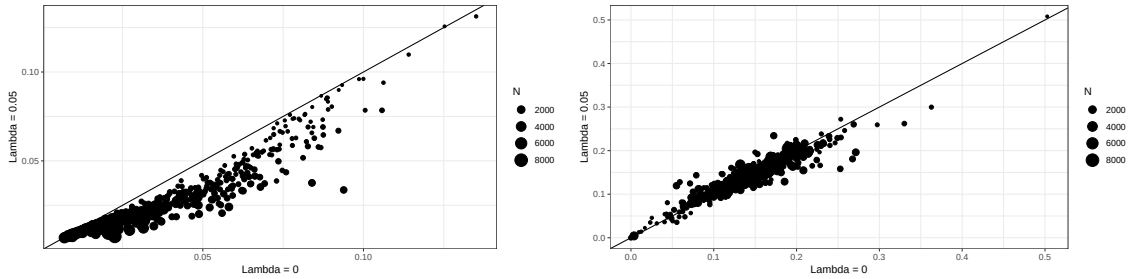


Figure 2: Estimated Bias-Variance Tradeoff as a Function of λ values. Each dot represents the PBR and average effective size for the approximate balancing weights with different λ and template matching.

5.3 Adjustment and precision

We next explore the benefits of model adjustment, and further explore the gains of small amounts of regularization. See the supplemental materials for a review of overlap diagnostics. We first calculated final standard errors for $\lambda = 0$, both with and without model adjustment via least squares. We then standardize by applying the estimated weights to the residuals instead of the outcome. We found that hospital level standard errors were on average 15% smaller with model adjustment.

We then compared the results for when $\lambda = 0$ vs. $\lambda = 0.05$. Figure 3 contains two scatterplots. In the first one, we plot the pairs of hospital standard errors under each evaluation. The standard error estimates allows us to clearly identify the role of increasing the effective sample size by increasing λ.



(a) Standard Errors (b) Risk Adjusted Estimates

Figure 3: Scatterplots of hospital level standard errors and point estimates for $\lambda = 0.05$ and $\lambda = 0$.

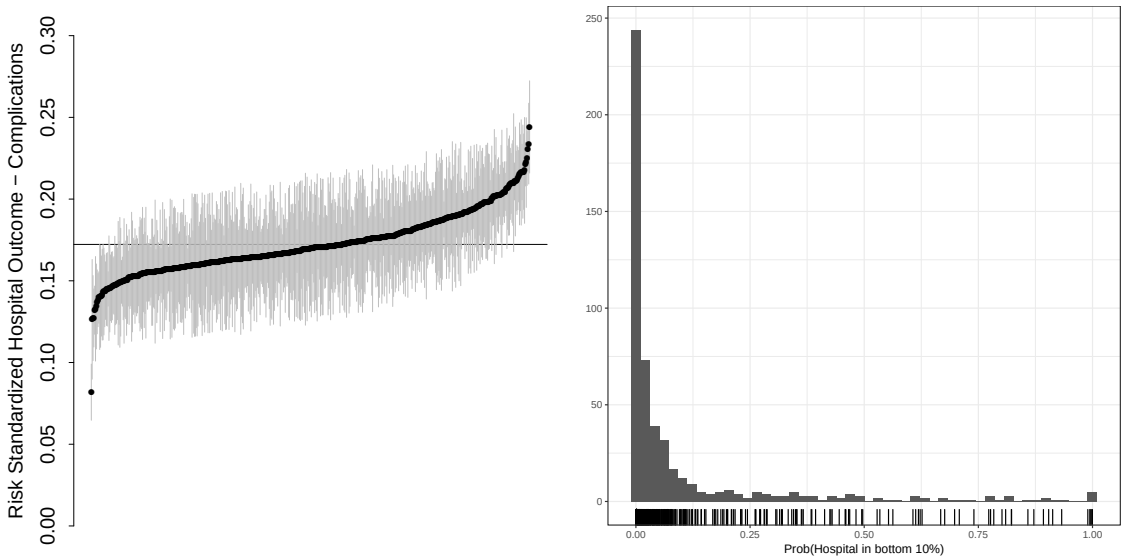
That is, in the scatterplot the standard error estimates are uniformly smaller when $\lambda = 0.05$. In the second scatterplot, we plot the risk adjusted complication rates for each hospital, showing no clear pattern. Taken together, these plots suggest that there are gains from selecting a value for λ larger than 0: Small increases in λ produce smaller hospital level standard errors, but have little effect on the adjusted estimates for each hospital.

5.4 Estimates After Bayesian Shrinkage

As described in Section 4.3, we now use a Bayesian hierarchical model to partially pool the hospital-specific estimates; we estimate this model using Stan, a Bayesian software package (Carpenter et al., 2017). We initially set the random effect G as a simple Normal, $G = N(\alpha_\mu, \tau_\mu^2)$, consistent with prior work on hospital quality (Normand et al., 2007). Possible alternative parameterizations include a t_7 and a mixture of Normal distributions; see Miratrix and Feller (2020). For Normal G we impose a uniform prior over the random effect standard deviation, $\tau_\mu \in [0, \infty)$, and a uniform prior over the random effect mean, which we constrain to be in the unit interval, $\alpha_\mu \sim \text{Unif}[0, 1]$, since we focus on binary outcomes. Results are largely unchanged with other prior choices; see McCulloch and Neuhaus (2011) for seminal discussion.

Figure 4a shows the posterior means and corresponding 95% uncertainty intervals for the set of μ_j, the risk-standardized hospital complication rate.[7] We see variation in both the point estimates as well as the width of the hospital-specific uncertainty intervals. While there is a large mass of hospitals in the center of the distribution, there are clearly some hospitals with consistently above- or below-average quality as defined by the risk adjusted proportion of within-hospital surgical complications.

[7] See Paddock et al. (2006) for alternative approaches to summarizing the posterior, especially triple goal estimation.



(a) Catepillar Plot of Risk Standardized Complications for $\lambda = 0.05$

(b) Histogram of the Probability of Being in the 10% of Ranks $\lambda = 0.05$

Figure 4: Hospital quality results after applying Bayesian shrinkage.

As a further verification, we also estimated the amount of cross-hospital variation in complication rate using the "Q statistic" approach from meta-analysis (see, e.g., **?**) as implemented in the `blkvar` package.[8] To assess cross-hospital variation, we treat each hospital as a "study" and invert a series of tests, testing whether cross-hospital variation equals a sequence of values from 0 (all hospitals are the same) and up. We use the raw estimates of the weighted (unshrunk) hospital outcomes and associated standard errors. The core idea is the hospital deviations from the common mean , standardized with a combination of estimation error and assumed cross-hospital variation, should tend to average at around 1; if they are larger we conclude variation must be more substantial than the tested null. We find, with 95% confidence, that the degree of cross-hospital variation is between 1.7pp and 2.2pp, with a Hodges-Lehman point estimate (corresponding to the value with the largest p-value) of 2.0pp. Overall we find the overall estimated average complication rate is 17%, with hospitals 1 standard deviation more inferior having complication rates of 19%, and hospitals 1 standard deviation superior having rates of 15%. A 95% prediction interval (assuming roughly normal distribution of quality) is a range of average complication rates from 13% to 21%.

While the primary aim of our analysis is to produce risk standardized measure of hospital performance,

[8] See `https://github.com/lmiratrix/blkvar/` for package source.

one additional aim is to identify institutions that are outliers. For example, hospitals that are identified as underperforming may be targeted for quality improvement efforts. Next, we outline how the result from the Bayesian hierarchical model can also be used for identifying outliers. Figure 4b shows the posterior probability that each hospital is in the highest decile — that is, the *worst* performing 10 percent — of (standardized) surgical complication rates. For the vast majority of hospitals, the probability of being in this "danger zone" is quite low. More concretely, 98.5% of hospitals have a less than 10% chance of being in this low performing group of institutions. However, some hospitals are very likely to be low-performing. Specifically, there are 8 hospitals that have at least a 90% chance of being in this low performing group, 5 of which with at least a 99% chance.

6 Conclusion

Methods of risk adjustment are widely used to compare the performance of hospitals and physicians. The current standard for risk adjustment is model based and may suffer from model misspecification. Risk adjustment via direct standardization avoids modeling, but has previously been limited by data dimensionality. Here, we develop a new method of direct standardization based on weighting. We treat each hospital as a sample from the overall patient population and find weights such that the re-weighted hospital patient mix matches the overall population. We obtain these weights via a convex optimization problem that trades off covariate balance and effective sample size. Finally, we applied our approach to data on general surgery in Pennsylvania, Florida, and New York.

This approach to risk adjustment offers several critical advantages. The risk adjusted outputs are readily interpretable. Principled methods of variance estimation are easily adapted from the literature on survey sampling and weighted regression. Compared to other direct standardization approaches, risk adjustment via weighting substantially reduces bias. We also found large increases in effective sample size for a slight increase in possible bias. This method of direct standardization can also be easily combined with shrinkage methods to account for the variation in hospital size when comparing hospitals to each other and identifying high and low performing hospitals. Overall, the estimation process is not computationally intensive, and requires little user input outside of selecting the penalty. Estimating a set of weights for over 600,000 patients required less than five minutes on a desktop computer. Template matching, however, required fine tuning of over five hundred different matches, and was a much more

time consuming process.

A limitation is that, as with all design-based methods, the weights do not use outcome information. While we propose a model-assisted approach to incorporate outcome modeling, the user must still select which covariates to include and the corresponding balance measure. We can easily extend the proposed approach to allow for a richer covariate basis, including interactions and higher-order terms, as well as to prioritize balance in some covariates (see Section 5.1). Another strategy is to use external data to fit an outcome model and then use our approach to balance the predicted value from that model (D'Amour and Franks, 2019).

A number of other extensions are possible. Currently the weights are only target population level means. More complex population targets are possible that depend on overlap in the population. Penalty parameter selection could also be optimized to find an optimal tradeoff between bias reduction and effective sample size. We have also explored best practice in terms of the application of shrinkage methods.

Bibliography & References Cited

Ash, A. S., M. Shwartz, E. A. Pekoz, and A. D. Hanchate (2012). Comparing outcomes across providers. In L. I. Iezzoni (Ed.), *Risk adjustment for measuring health care outcomes* (4th ed.). Chicago, IL: Health Administration Press.

Ben-Michael, E., D. Hirschberg, A. Feller, and J. Zubizarreta (2020). The balancing act for causal inference.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*(1).

D'Amour, A. and A. Franks (2019). Covariate reduction for weighted causal effect estimation with deconfounding scores.

Decker, M. R., C. M. Dodgion, A. C. Kwok, Y.-Y. Hu, J. A. Havlena, W. Jiang, S. R. Lipsitz, K. C. Kent, and C. C. Greenberg (2014). Specialization and the current practices of general surgeons. *Journal of the American College of Surgeons 218*(1), 8–15.

Deming, W. E. and F. F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics 11*(4), 427–444.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J. C., C. E. Särndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association 88*(423), 1013–1020.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey (1998). Comorbidity measures for use with administrative data. *Medical care 36*(1), 8–27.

Fleiss, J., B. Levin, and M. Paik (2003). The standardization of rates. In *Statistical Methods for Rates and Proportions*, Chapter 19, pp. 627–647. New York, NY: John Wiley and Sons.

George, E. I., V. Ročková, P. R. Rosenbaum, V. A. Satopää, and J. H. Silber (2017). Mortality rate

estimation and standardization for public reporting: MedicareâĂŹs hospital compare. *Journal of the American Statistical Association 112*(519), 933–947.

Hazlett, C. (2019). Kernel Balancing : A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sincia*.

Hirshberg, D. A., A. Maleki, and J. Zubizarreta (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

Iezzoni, L. I. (2012). *Risk adjustment for measuring health care outcomes* (4th ed.). Chicago, IL: Health Administration Press.

Jacob, P. E., L. M. Murray, C. C. Holmes, and C. P. Robert (2017). Better together? statistical learning in models made of modules. *arXiv:1708.08719*.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the american statistical association 50*(272), 1168–1194.

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 an acute myocardial infarction. *Circulation 113*(13), 1683–1692.

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education.

Longford, N. T. (2019). Performance assessment as an application of causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

McCulloch, C. E. and J. M. Neuhaus (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical science*, 388–402.

Miratrix, L. W. and A. Feller (2020). reatment effect distributions in multi-site trials.

Normand, S.-L. T., D. M. Shahian, et al. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science 22*(2), 206–226.

Paddock, S. M., G. Ridgeway, R. Lin, and T. A. Louis (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational statistics & data analysis 50*(11), 3243–3262.

Pimentel, S. D., R. R. Kelz, J. H. Silber, and P. R. Rosenbaum (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association 110*(510), 515–527.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Sarndal, C.-E., B. Swensson, and J. Wretman (2003). *Model assisted survey sampling*. Springer.

Silber, J. H., P. R. Rosenbaum, and R. N. Ross (1995). Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics? *Journal of the American Statistical Association 90*(429), 7–18.

Silber, J. H., P. R. Rosenbaum, R. N. Ross, J. M. Ludwig, W. Wang, B. A. Niknam, N. Mukherjee, P. A. Saynisch, O. Even-Shoshan, R. R. Kelz, et al. (2014). Template matching for auditing hospital cost and quality. *Health services research 49*(5), 1446–1474.

Weiss, M. J., H. S. Bloom, N. Verbitsky-Savitz, H. Gupta, A. E. Vigil, and D. N. Cullinan (2017, November). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials. *Journal of Research on Educational Effectiveness 10*(4), 843–876.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.