The Annals of Applied Statistics
2023, Vol. 17, No. 2, 901–928
https://doi.org/10.1214/22-AOAS1629

HOSPITAL QUALITY RISK STANDARDIZATION VIA APPROXIMATE BALANCING WEIGHTS

BY LUKE J. KEELE[1,a], ELI BEN-MICHAEL[2,c], AVI FELLER[3,d], RACHEL KELZ[1,b] AND LUKE MIRATRIX[4,e]

[1]*Hospital of the University of Pennsylvania, University of Pennsylvania,* [a]*luke.keele@gmail.com,* [b]*Rachel.Kelz@pennmedicine.upenn.edu*

[2]*Department of Statistics and Institute for Quantitative Social Science, Harvard University,* [c]*ebenmichael@fas.harvard.edu*

[3]*Goldman School of Public Policy, University of California, Berkeley,* [d]*afeller@berkeley.edu*

[4]*Graduate School of Education, Harvard University,* [e]*lmiratrix@g.harvard.edu*

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 surgery complication rates, can be misleading because hospital case mixes differ—a hospital's overall complication rate may be lower simply because the hospital serves a healthier population overall. In this paper we develop a method of "direct standardization" where we reweight 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. Critically, these balancing weights can also be tuned to preserve sample size for more precise estimates. We also derive principled measures of statistical uncertainty 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.

1. Introduction: Judging hospital quality. How can we assess quality across hospitals? Simple comparisons of hospital-specific outcomes can be misleading: Hospitals that treat patient populations with complex, chronic conditions will generally have worse outcomes than hospitals that treat patients who are healthier. Thus, a hospital's outcomes may be better due to more effective treatments or simply because the hospital serves a healthier clientele.

Risk adjustment, also known as risk standardization, refers to a set of statistical methods that adjust for the hospital patient mix to make hospital outcomes directly comparable (Normand and Shahian (2007)). Risk standardization is widely used to evaluate hospitals and provide the public with information on hospital quality. For example, Medicare's online tool, Hospital Compare, uses risk standardization to help patients identify high-quality hospitals (Medicare.gov (2013)). Moreover, standardized outcomes are used as quality measures that can affect reimbursement rates for medical procedures. For example, the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) uses standardized outcomes to measure quality; clinicians or hospitals that do not meet performance standards can receive lower payments (Centers for Medicare & Medicaid Services (2021)).

Risk standardization comes in two forms, often referred to as direct and indirect standardization. These two methods focus on different questions about how patient mix affects hospital outcomes. Informally, indirect standardization asks, "How should this hospital have

Received February 2021; revised March 2022.
Key words and phrases. Risk adjustment, weighting, direct standardization.

done, given the patients they serve?" Direct standardization asks, "How would this hospital do, if given the same types of patients as everyone else?" That is, each method is focused on a different counterfactual comparison and relies on different statistical methods.

In this paper we develop a suite of new methods for direct standardization using weighting-based methods. In our approach we view each hospital's patient population as a nonrepresentative 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 accounts for systematic differences in patient populations across hospitals while maintaining precision. We also identify a bias-precision tradeoff and find that regularizing the weights can substantially increase precision in our hospital specific estimates while only incurring what appears to be a small increase in bias in estimated hospital quality. We then demonstrate how to use an outcome model both to reduce remaining bias and to improve the precision of the hospital quality estimates even further. Finally, we apply a Bayesian shrinkage estimator as an additional step in order to better account for variability in the size of hospitals.

We compare our weighting-based approach to extant methods for direct standardization. We formalize how direct standardization can be implemented via model based adjustment. We then outline how these model-based forms of direct adjustment are prone to extrapolation in ways that can be limited when weights are used. Next, we compare our method to "template matching" (Silber et al. (2014a)) where an identified set of patients at each hospital are chosen to closely match a template of patients based on a canonical list of patient characteristics. Compared to template matching, we show substantial gains in both bias reduction and precision.

Our paper proceeds as follows. We first discuss our application, and, in Section 2, we review the two primary methods of standardization: direct and indirect. In Section 3 we outline our approach of using balancing weights, a tool taken from the literature on causal inference in observational studies, for direct standardization. We also derive two methods for direct standardization. We then derive methods of variance estimation that are consistent with our estimated weights. This section forms the core of our approach. In Section 4 we apply regression modeling to adjust for remaining imbalance and increase precision in the hospital level estimates. We then outline how to apply a Bayesian shrinkage estimator to account for variation in hospital size and to obtain improved estimates of hospital performance. In Section 5 we evaluate our method using a simulation study, and in Section 6 we apply our methods to claims data on general surgery. We then conclude in Section 7.

1.1. *Hospital quality in PA, FL, and NY on general surgical performance.* General surgery consists of high-volume surgical procedures that are conducted in almost all hospitals, including procedures such as appendectomy (removal of the appendix), cholecystectomy (gall bladder), mastectomy (breast), and hernia repairs. Since deaths are 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 estimating the risk-adjusted rates of such complications. This project is designed to serve as pilot work to develop a quality measure for general surgery. Currently, quality measures have been developed for a number of different surgical specialties but exist for general surgery only within the ambulatory surgical care setting (Centers for Medicare & Medicaid Services (2021a)).

In our analysis we use risk standardization to understand hospital quality for general surgical procedures using claims data from Pennsylvania, New York, and Florida from 2012–2013. The data contain 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

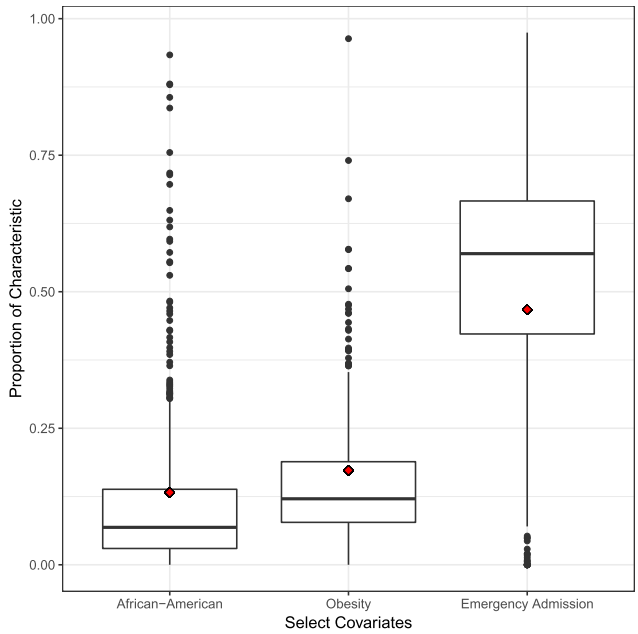


FIG. 1. *Boxplots of Hospital Casemix Distributions for three covariates. Diamond represents population mean for that covariate.*

elective), type of insurance, and age. We analyze 44 general surgery operations.[1] We should note that including socioeconomic and sociodemographic measures in risk adjustment analyses is controversial (National Quality Forum (2014)). In our analysis we include categories for race but no other measures of socioeconomic status.

Across the three states, we have a total of 621,667 patients in 523 hospitals with between 30 and over 8000 general surgery cases for the study period. The median number of patients was 700. 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). Figure 1 displays boxplots of hospital-level proportions of three key patient characteristics: whether a patient is African-American, whether a patient is obese (BMI > 30), and whether the procedure was an emergency admission. All three characteristics are important predictors of complications in the cohort. As the boxplots show, there is substantial variation in all three attributes. For instance, only 17 percent of patients in the population 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.

2. Direct risk standardization for comparing hospital outcomes. There is a large literature in statistics and health services research on statistical methods for risk adjustment via standardization; see Normand and Shahian (2007), Normand et al. (2016) for reviews. Generally, indirect and direct standardization have been viewed as different statistical tools for

[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 use 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).

the same task. However, viewing standardization as a causal inference problem clarifies the key differences between the two methods (Longford (2019)). In the causal inference framework the hospitals are viewed as a vector of treatments, and standardized outcomes serve as the "effect" of each hospital accounting for patient mix differences. The potential outcomes framework can then be used to outline estimands and formally define identification conditions. Notably, direct and indirect standardization target different causal estimands. Given that indirect and direct standardization have different estimands, the two methods are not immediately comparable, and we avoid comparing their results here.

In the clinical literature, however, the choice between the two methods has largely been a function of convenience. Under direct standardization the investigator adjusts the hospital case mix to match a target distribution. This allows the analyst to target the differences between hospital covariate distributions and the population covariate distributions (George et al. (2017)). Direct standardization methods, however, have seen limited use in medical applications. Direct standardization methods have traditionally been viewed as too limited, since it was assumed that they could only adjust for a few patient-level variables (Iezzoni (2012, p. 347)).

Indirect standardization has been much more widely used for risk adjustment, since a widely understood model based estimation method has made it much easier to incorporate larger numbers of patient level covariates (Iezzoni (2012, p. 347)). Under indirect standardization, observed outcomes for patients are compared to expected outcomes derived from a statistical model fit to the larger patient population (Fleiss, Levin and Paik (2003), Iezzoni (2012), Kitagawa (1955), Silber, Rosenbaum and Ross (1995), Silber et al. (2016a)). Typically, an outcome is regressed on patient characteristics via a (generalized) linear model using the entire patient population. This risk adjustment model is used to predict outcomes for patients in a specific hospital. The average of these predictions then serves as the expected outcome for the provider, which 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)).

Contrary to perceptions in the medical literature, direct standardization can also be implemented via regression models to allow for rich specifications. For example, regression-based methods for direct standardization have seen widespread use in education. Goldstein and Spiegelhalter (1996) provide one early, informal description of model-based direct standardization for education data, and McCaffrey et al. (2004) describe direct standardization in the context of value-added modeling. Recently, template matching was developed as a form of direct standardization that can risk adjust for many patient level covariates while also relaxing the stronger functional form assumptions required for regression (Silber et al. (2014a, 2014b, 2016b)). Template matching also makes the goals of hospital standardization clear and focuses the researcher's attention on important considerations, such as whether the covariates are sufficient to justify an attempt at standardization. Under template matching, the investigator seeks to understand how different hospitals would perform with patients similar to a sample ("template") of patients. For each hospital, matching methods are used to find a subset of patients that are highly comparable to the template. Hospital quality is then evaluated based on this matched set.

Below, we develop several new additions to the literature on direct standardization. We first develop a weighting-based approach to direct standardization. Then, we demonstrate how direct standardization can be implemented using model based methods, such as modeling the

hospital assignment process or regression modeling of the outcome, and link these methods to the weighting approach. We describe how model-based methods may be particularly sensitive to model misspecification due to extrapolation. We also outline how our weighting-based approach can be combined with outcome modeling. Finally, we can also combine the weighting-based method with a shrinkage estimator to better account for variation in hospital size.

3. Direct standardization via approximate balancing weights. We now develop a weighting method for direct standardization which solves a convex optimization problem to simultaneously optimize for balance and effective sample size. First, we review weighting methods. Next, we outline notation and assumptions and then turn to specific implementation details.

3.1. *Review*: *Balancing weights*. Weighting methods have a long history in survey sampling and causal inference (Horvitz and Thompson (1952), Lohr (2010), Robins and Rotnitzky (1995)). The goal is to weight target groups to have a similar distribution on a given set of covariates. The reweighted groups can then be directly compared on some outcome of interest, since the reweighted groups are comparable on baseline characteristics. 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, Särndal and Sautory (1993)). In this case the resulting weights would achieve exact balance where the reweighted 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 is infeasible, especially for smaller hospitals. One alternative is to use a traditional inverse probability weighting (IPW) estimator which is widely used to estimate treatment effects (Imbens (2004), Robins, Hernan and Brumback (2000)). This approach can be somewhat unstable, however, resulting in reweighted samples that still are not well balanced. To avoid such issues, we build on recent advances in the causal inference literature where analysts use *approximate balancing* weights that are designed to directly target covariate balance in the estimation process. This class of weighting methods solve a convex optimization problem to find a set of weights that target a specific loss function (Hainmueller (2011), Zubizarreta (2015)); see Ben-Michael et al. (2021) for a recent review of these weighting methods.

3.2. *Hospitals as nonrepresentative 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 reflect both differences in hospital quality and differences in the distribution of patient attributes.

Formally, we 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 surface"

[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.

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, $\bar{Y}_j \equiv \frac{1}{n_j} \sum_{Z_i = j} Y_i$. However, these estimates are not directly comparable: even if two hospitals, j and j', have identical quality surfaces, $m_j(x) = m_{j'}(x)$, ρ_j, and $\rho_{j'}$ may differ if they average the quality surfaces over different distributions.

Risk adjustment aims to remove this dependence between the patient and surgery characteristics of a hospital and its overall assessed outcome. We do this by considering a set of hospital estimands that each take the expectation of $m_j(x)$ over a common distribution $X \sim P^*$,

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

These μ_j are more directly comparable, as we have removed systematic differences across distributions. 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 direct standardization estimand as the expected outcome of hospital j if its patient mix were the same as the full population of patients. Importantly, estimands for direct standardization differ from estimands under indirect standardization. Equation (2) examines how hospital performance differs from expected across a canonical distribution of patient characteristics. By contrast, indirect standardization examines how each hospital differs from expected (given the model) for those patients that hospital already serves.

An important question is how to interpret risk-adjusted differences in outcomes across hospitals. One can view risk-adjusted quality measures as being informative of hospital quality without giving them a causal interpretation. With additional assumptions, however, a causal interpretation is possible (Longford (2019)). Specifically, we would need to assume that differences in hospital patient mix are fully captured by X which implies that unobserved differences in patient mix do not contribute to the estimates. This assumption would be violated if the patient mix at some hospitals was significantly at higher risk for complications, but this elevated risk was not captured by the patient level covariates. Understanding the possible role of unobservable differences is critical if risk adjustment is the basis for targeting hospitals for improvement efforts; see Hull (2018) for further discussion.

Even if we wish to perform a noncausal comparison of hospitals with different patient populations, we still have to impose the additional assumption that, at least in principle, any type of patient (as defined by X) in our reference distribution P^* could receive care at any hospital. We formalize this as an overlap assumption.

ASSUMPTION 1 (Overlap). $P(Z = j \mid X = x) > 0$ if $P^*(x) > 0$.

Assumption 1 rules out the possibility that a hospital would never treat a particular type of patient. For example, we assume no hospital treats only women and that all hospitals perform the full range of surgeries we are investigating. As we discuss in Section 3.4, the distribution of estimated weights is a useful diagnostic for assessing overlap in practice. We could further restrict our estimand to a set of patients where there is overlap by restricting P^*. Alternatively,

we could consider a weaker assumption that only requires overlap on the set of covariates that are related to the outcome and are different across hospitals. To find such variables, we could use a double selection procedure (Belloni, Chernozhukov and Hansen (2014)). We leave a thorough investigation of these directions to future work.

3.3. *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 $\hat{\gamma}$ designed to make hospital j representative of the target population,

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

where $\sum_{Z_i=j} \hat{\gamma} = 1$. In our discussion below, we only consider weights that are independent of the outcomes.

Our setup accommodates general function classes for the quality function $m_j(x)$; see Kallus (2020), Hirshberg, Maleki and Zubizarreta (2019), and Hazlett (2020) for further discussion. To motivate our optimization problem, we impose the simplifying restriction that the quality function $m_j(x)$ is a linear function of some transformation of the covariates,

$$m_j(x) = \alpha_j + \beta_j \cdot \phi(x), \tag{4}$$

with $\phi : \mathbb{R}^d \to \mathbb{R}^p$ and $\beta_j \in \mathbb{R}^p$. The vector $\phi(x)$ is our basis and is how we represent the information the covariates provide. We discuss the choice of basis for our application in Section 6.

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

$$\hat{\mu}_j - \mu_j = \underbrace{\beta_j \cdot \left(\sum_{Z_i=j} \hat{\gamma}_i \phi(X_i) - \bar{\phi} \right)}_{\text{bias}} + \underbrace{\sum_{Z_i=j} \hat{\gamma}_i \varepsilon_i}_{\text{variance}}, \tag{5}$$

with $\bar{\phi} \equiv \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$ the overall population mean of the covariate vector $\phi(x)$. There are a range of target populations that we might consider shifting toward. In our application we use the population means across FL, PA, and NY as the target. However, we could instead target the distribution of a different state or the U.S. population. Changing the target population is done by changing $\bar{\phi}$.

The error in (5) has two components: (1) systematic bias due to imbalance in $\phi(X_i)$ between (reweighted) hospital j and the overall sample, and (2) idiosyncratic error due to noise. The goal is to find weights that control both terms. For the first term in equation (5), 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),

$$\left| \mathbb{E}[\hat{\mu}_j - \mu_j \mid X, Z = j] \right| \leq \|\beta_j\|_2 \underbrace{\left\| \sum_{Z_i=j} \hat{\gamma}_i \phi(X_i) - \bar{\phi} \right\|_2}_{\text{imbalance}}. \tag{6}$$

Thus, under equation (6), reducing imbalance controls the bias regardless of the true β_j. If there were only a small number of patient characteristics, we could likely achieve exact balance, forcing equation (6) to zero. This will not generally be feasible in settings with a richer set of covariates and when the range of the weights are restricted. Furthermore,

achieving this goal could come at the cost of extreme weights, which could substantially reduce precision, as we discuss next.

For the second term in Equation (5), the challenge is that the individual ε_i are unknown. However, we can bound the variance (conditional on X and Z) by the sum of the squared weights,

$$\text{Var}(\hat{\mu}_j - \mu_j \mid X, Z = j) = \hat{\gamma}' \Sigma_j \hat{\gamma} \leq \lambda_j \sum_{Z_i = j} \hat{\gamma}_i^2, \tag{7}$$

where Σ_j is the conditional variance-covariance matrix of hospital j's noise terms and λ_j is the maximum eigenvalue of Σ_j. This shows that we can reduce the variance by limiting the spread of the estimated weights. That is, the more homogenous we can make the weights, the more precise the resulting estimates. Here, we choose to penalize the sum of the squared weights, $\sum_{Z_i=j} \hat{\gamma}_i^2$, though other penalties are possible; see Ben-Michael et al. (2021).

3.4. *Weighting via convex optimization*. We can now combine these two objectives into the following optimization problem:

$$\begin{aligned} \min_{\gamma} \ & \sum_{j=1}^{J} \left[\left\| \bar{\phi} - \sum_{Z_i = j} \gamma_i \phi(X_i) \right\|_2^2 + \lambda n_j \sum_{Z_i = j} \gamma_i^2 \right] \\ \text{subject to } & \sum_{Z_i = j} \gamma_i = 1, \\ & \ell \leq \gamma_i \leq u. \end{aligned} \tag{8}$$

The optimization problem (8) trades off two competing terms for each hospital j: better balance (and thus lower bias) and more homogeneous weights (and thus lower variance).[3] 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 6.2.

The constraint set in equation (8) has two components. First, 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. Second, we constrain the weights to have lower bound ℓ and upper bound u. We set the lower bound $\ell = 0$ and the upper bound $u = 1$ so that weights are nonnegative and do not extrapolate outside of the support of the data. Combined with the sum-to-one constraint, this means that each individual weight $\hat{\gamma}_i$ corresponds to the fraction of hospital Z_i's outcome dictated by unit i. These constraints also stabilize the estimate by ensuring sample boundedness; for example, $\hat{\mu}_j$, which estimates a complication rate, is always a valid proportion, $\hat{\mu}_j \in [0, 1]$. In Section 3.5 below, we show that, when the weights are unrestricted ($\ell = -\infty$, $u = \infty$), $\hat{\mu}_j$ is equivalent to a model-based standardization approach using ridge regression.

Under this weighting approach, extreme weights will signal if the covariates for a specific hospital do not overlap with the patient population. We can target such extreme weights using the upper bound u. Setting the upper bound to $u < 1$ would prevent the optimization

[3]For a single hospital the objective in optimization problem (8) reduces to a special case of the minimax linear estimation proposal from Hirshberg, Maleki and Zubizarreta (2019) with a particular choice of function class. The above extends to the case with multiple hospitals.

problem from putting too much weight on any single patient in a hospital. For example, setting $u = 0.2$, would ensure that we do not put more than 20% of the weight on any individual patient. As such, investigators can evaluate the overlap assumption in the estimation process—without reference to outcomes—by inspecting extreme weights and assessing the sensitivity of the estimates to the choice of upper bound u. Importantly, using a *fixed* upper bound may substantively change the estimand for each hospital, depending on the degree of overlap (Li, Morgan and Zaslavsky (2018)). Conversely, allowing the upper bound to change with the sample size (along with assumptions on the propensity score) alleviates this concern (see Athey, Imbens and Wager (2018), who use a varying bound). In our primary results we sidestep this issue and set $u = 1$ (no constraint) and investigate the impact of setting u to be less than 1 in the Supplementary Material (Keele et al. (2023a)).

The optimization problem of (8) obtains $\hat{\gamma}_i$ without using outcome information. 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 can then simply estimate the adjusted hospital means by taking a weighted average of the patient outcomes, $\hat{\mu}_j = \sum_{Z_i=j} \hat{\gamma}_i Y_i$.

3.4.1. *Dual relation to propensity score estimation.* Finding weights that balance covariates while controlling the variance is well known to be intimately connected with propensity score estimation (for some examples out of many, see Hirshberg and Wager (2021), Wang and Zubizarreta (2020), Yiu and Su (2018), Zhao (2019), Zhao and Percival (2017)). Defining the *generalized propensity score* as $e_j(x) = P(Z = j \mid X = x)$, we can follow Hirshberg and Wager (2021) and Ben-Michael et al. (2021) and inspect the Lagrangian dual of the optimization problem (8). The dual writes the weights as a truncated linear function of the covariates with a dual intercept η_{0j} and dual coefficients η_j,

$$\gamma_i = \left[\eta_{0Z_i} + \eta_{Z_i} \cdot \phi(X_i)\right]_+,$$

where $[x]_+ \equiv \max\{0, x\}$. For each hospital j, the dual optimization problem includes a loss function

$$\begin{aligned}\mathcal{L}_j(\eta) = {} & \frac{1}{2n_j} \sum_{Z_i=j} \left(\frac{1}{e_j(X_i)} - \left[\eta_{0j} + \eta_j \cdot \phi(X_i)\right]_+ \right)^2 \\ & + \sum_{i=1}^{n} \left(\frac{\mathbb{1}\{Z_i = j\}}{n_j e_j(X_i)} - \frac{1}{n} \right) (\eta_{0j} + \eta_j \cdot \phi(X_i)) - \frac{1}{n_j} \sum_{Z_i=j} \frac{1}{e_j(X_i)} \left[\eta_j \cdot \phi(X_i)\right]_-,\end{aligned}$$

where $[x]_- \equiv \min\{0, x\}$. The expected value of the dual objective $\mathcal{L}_j(\eta)$ is the mean square error between the weights for hospital j and the inverse of the propensity score $\frac{1}{e_j(x)}$; the second term in the objective can be viewed as a finite sample adjustment. Overall, the dual variables are found to minimize a ridge-penalized version of the objective, separately for each hospital

$$\min_{\eta} \sum_{j=1}^{J} \left[\mathcal{L}_j(\eta) + \frac{\lambda}{2} \|\eta_j\|_2^2 \right].$$

Therefore, the balancing problem in equation (8) is dual to a propensity score estimation problem where we estimate the inverse propensity score with a (truncated) linear model, separate for each hospital.

3.5. *Direct standardization using parametric models*. As we note above, indirect standardization via regression models has a long history but has seen much less usage for direct standardization in health-care applications (Iezzoni (2012, p. 347)); see Centers for Medicare & Medicaid Services (2021b) for one notable exception. Next, we formally outline methods for model-based direct standardization and show that when the weights are unrestricted (i.e., $\ell = -\infty$ and $u = \infty$), our balancing weights approach is equivalent to using ridge regression to estimate the quality curve. Our formalization of model-based direct standardization differs from that in McCaffrey et al. (2004), since we do not assume longitudinal data and our derivations are with respect to our specific estimand.

One natural model-based approach is to estimate the quality curve $m_j(x)$ directly. Given an estimated $\hat{m}_j(\cdot)$, we can estimate the adjusted hospital mean by averaging the patient-specific predictions across the full population with $\hat{\mu}_j^{\text{out}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_j(X_i)$. A straightforward way to model the quality curve is to use the following *fixed effects* regression model:

$$m_j(x) = \alpha_j + \beta \cdot \phi(x),$$

where the α_j are J hospital fixed effects (implemented via indicator variables) and $\phi(x)$ remains the matrix of patient level covariates. Taken literally, this model assumes that each hospital has a constant shift in patient outcomes, regardless of patient type. If we construct $\phi(x)$ so it is zero-centered with respect to the target population, α_j is then the predicted outcomes for a "typical" patient, with average covariate values, at hospital j. In other words, under this model all hospitals are standardized to a singular common patient type of "the average person." We refer to this as *fixed effect direct standardization* (FEDR).

The FEDR model can lead to biased estimates, however, if the true covariate-outcome relationship differs by hospital. In particular, consider the separate linear quality curve model in equation (4), where for unit i in hospital j, the outcome is $Y_i = \alpha_j + \beta_j \cdot \phi(X_i) + \varepsilon_i$, where the β_j vary. With centered covariates, $\bar{\phi} = 0$ and $\mu_j = \alpha_j$. Then, the estimates $\hat{\alpha}_j$, $\hat{\beta}$ obtained via FEDR will satisfy $\bar{Y}_j = \hat{\alpha}_j + \hat{\beta}\bar{\phi}_j$, with $\bar{\phi}_j \equiv \frac{1}{n_j}\sum_{Z_i=j} \phi(X_i)$. This gives an overall estimation error (including bias) for the standardized hospital quality estimate of

$$\hat{\mu}_j^{\text{out}} - \mu_j = \hat{\alpha}_j - \alpha_j = \bar{Y}_j - \hat{\beta} \cdot \bar{\phi}_j - \alpha_j = (\beta_j - \hat{\beta}) \cdot \bar{\phi}_j + \bar{\varepsilon}_j, \tag{9}$$

where $\bar{\varepsilon}_j \equiv \frac{1}{n_j}\sum_{Z_i=j} \varepsilon_i$. With FEDR, $\hat{\beta}$ is targeting a weighted average of the β_j across hospitals, and the magnitude of $\hat{\beta} - \beta_j$, and thus the bias, depends on both estimation error and the degree to which the outcome-covariate relationship of hospital j (the β_j) differs from this fully pooled average.

An alternative to this more restrictive approach is to allow the quality curves to differ across hospitals and directly fit the varying coefficients model in equation (4). For instance, we can estimate the quality curve via ridge regression, finding intercepts $\hat{\alpha}_j$ and coefficients $\hat{\beta}_j$ for each hospital that solve separate ridge regressions,

$$\min_{\alpha,\beta} \sum_{j=1}^{J} \frac{1}{n_j} \sum_{Z_i=j} \big(Y_i - \alpha_j - \beta_j \cdot \phi(X_i)\big)^2 + \lambda_j \|\beta_j\|_2^2. \tag{10}$$

In this case the estimate of the adjusted mean for hospital j is the prediction for the average covariate value, which for centered covariates is again the estimated fixed effect,

$$\hat{\mu}_j^{\text{out}} = \hat{\alpha}_j + \hat{\beta}_j \cdot \bar{\phi} = \hat{\alpha}_j.$$

However, now the estimation error depends on the error in the hospital-specific coefficients,

$$\hat{\mu}_j^{\text{out}} - \mu_j = (\beta_j - \hat{\beta}_j) \cdot \bar{\phi}_j + \bar{\varepsilon}_j. \tag{11}$$

The difference between β_j and the overall average β has been removed. In other words, if we can consistently estimate the components of the model then the outcome-based standardization estimate will be consistent.

While we have motivated this ridge regression approach through outcome modeling, it is, in fact, a special case of our weighting estimator and a solution to balancing optimization problem (8). Following Ben-Michael, Feller and Rothstein (2021), we can write the outcome based estimate as a weighted average

$$\hat{\mu}_j^{\text{out}} = \sum_{Z_i=j} \hat{\gamma}_i Y_i,$$

with the weight on unit i as

$$\hat{\gamma}_i^{\text{out}} = \frac{1}{n_j} + (\bar{\phi} - \bar{\phi}_j)'(\Sigma_j + \lambda_j I)^{-1}\big(\phi(X_i) - \bar{\phi}_j\big),$$

where $\Sigma_j = \sum_{Z_i=j}(\phi(X_i) - \bar{\phi_j})(\phi(X_i) - \bar{\phi_j})'$. Furthermore, these weights are the solution to the balancing weights problem in equation (8) where we put no bounds on the weights, $\ell = -\infty$ and $u = \infty$. We call this varying-coefficient estimation *ridge-weighted direct standardization* (RWDR). Our initial balancing approach is then a restricted version of RWDR with bounds on the weights. As we discuss in Section 3.4, in our default balancing implementation we set the lower bound $\ell = 0$ to avoid extrapolation. Without this constraint we allow extrapolation from the support of the observed data, meaning we are relying on the linear functional form of the covariate-outcome relationship to use potentially less representative data to predict overall outcomes. We will return to this point in Section 4.1.

Model based methods for direct standardization—or, equivalently, weighting based methods that allow the weights to be negative—may be a particularly good alternative to avoiding extrapolation by setting $\ell = 0$ in applications with smaller samples, where the loss in effective sample size is a concern. When the weights are restricted to be nonnegative, a hospital dissimilar to the overall population will tend to get a small number of units with large weights in an effort to reduce bias between specific hospitals and the target population. Large weights in turn lead to proportionally small effective sample sizes. In such circumstances some extrapolation via a modeling approach may be preferable: instead of relying on a small number of truly representative units, we leverage an assumed covariate-outcome relationship to extrapolate less representative units to help predict the outcome for an overall "typical" patient. That being said, in such circumstances we recommend fitting both approaches, as constraining the weights allows the researcher to examine each hospital to see if the estimate is being driven by a few units with large weight. These diagnostics are less apparent with model-based approaches as negative weights are difficult to interpret.

3.6. *Variance estimation.* We quantify the uncertainty of our estimated hospital 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

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

For each hospital we could then estimate this variance using a plug in,

$$\widehat{\text{se}}(\hat{\mu}_j|\hat{\gamma}_j) = \left[\hat{\sigma}_j^2 \sum_{Z_i=j} \hat{\gamma}_j^2\right]^{1/2} = \frac{\hat{\sigma}_j}{\sqrt{n_j^{\text{eff}}}}, \tag{12}$$

where we estimate the variance of the outcomes as

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

and where

$$n_j^{\text{eff}} \equiv \Big(\sum_{Z_i=j} \hat{\gamma}_i \Big)^2 \Big/ \sum_{Z_i=j} \hat{\gamma}_i^2 = 1 \Big/ \sum_{Z_i=j} \hat{\gamma}_i^2$$

is the *effective sample size* for hospital j (Lohr (2010), Potthoff, Woodbury and Manton (1992)).[4] Conditioned on the weights, that is, with a fixed design, this is the Huber–White heteroskedastic robust standard error for the weighted average and so allows heteroskedasiticty within hospitals. The last equality above assumes the weights sum to 1 within hospital, given under the constraint in the optimization problem in equation (8).

If all of the hospitals in our sample were large, the individual σ_j, estimated separately for each hospital, would be stable. In practice, however, the $\hat{\sigma}_j$ estimates from smaller hospitals may be noisy, which will complicate subsequent adjustments, especially partial pooling across hospitals. In particular, if a small hospital has an unusually small estimated standard error, its point estimate will receive excessive weight when trying to estimate cross-hospital variation. In our empirical example, for instance, some hospitals had naïve estimates of 0 for the standard error since they had no complications observed.

Therefore, we instead pool the individual standard deviation estimates into a global estimate. Specifically, we estimate the pooled standard deviation as

$$\hat{\sigma}_{\text{pool}}^2 = \frac{1}{N^{\text{eff}}} \sum_{j=1}^{J} n_j^{\text{eff}} \hat{\sigma}_j^2, \tag{13}$$

where $N^{\text{eff}} = \sum_j n_j^{\text{eff}}$ is the pooled effective sample size. The pooled variance is a weighted average of the noisy hospital-specific variance estimates. The hospital specific standard errors are then $\hat{\sigma}_{\text{pool}}^2 / \sqrt{n_j^{\text{eff}}}$.

See Weiss et al. (2017) and Bloom et al. (2016) for extended discussion of this approach for stabilizing estimates of impact variation in multisite trials; they find that the potential bias from ignoring heteroskedasticity is small. If heteroskedasticity were a concern, we could also merge this variance estimation step with a Bayesian model, as discussed below. For binary outcomes we can also conduct a two-step process, where we use preliminary shrunken estimates of the hospital means to recalculate hospital-specific standard errors and then refit using those improved uncertainty estimates; see the online suppement for the results based on this correction (Keele et al. (2023a)). In general, we found that the bias from heteroskedasticity was minimal.

4. Extensions. We now consider two extensions to the basic weighting approach: bias correction and partial pooling. These can be used together or separately to improve risk adjustment based on weighting alone.

4.1. *Incorporating additional bias correction.* As with other methods of direct standardization, the weighted estimator in Section 3.4 has the benefit of being *design-based*, that is, the weights $\hat{\gamma}$ solving the optimization problem in equation (8) are independent of the outcomes (Rubin (2008)). There are, however, reasons to utilize outcome information in the risk

[4]The effective sample size is the inverse of the dispersion penalty in the balancing weights optimization problem in equation (8).

adjusted estimates. We now describe how to include outcome information into the hospital level estimates.

Especially in smaller hospitals, the weighted mix of patients in hospital j may still not quite match the target distribution. We can use an outcome model to estimate how far off our weighted average outcome might be, and, given this remaining imbalance, we can adjust outcomes for this remaining imbalance. Specifically, given an estimate of the quality surface, $\hat{m}_j(x)$, for hospital j we adjust our estimated outcome as follows:

$$\hat{\mu}_j = \sum_{Z_i=j} \hat{\gamma}_i Y_i + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{m}_j(X_i) - \sum_{Z_i=j} \hat{\gamma}_i \hat{m}_j(X_i)}_{\text{imbalance in } \hat{m}_j(\cdot)}. \tag{14}$$

Analogous to bias correction for matching (Abadie and Imbens (2011), Rubin (1973)), the bias is estimated as the imbalance in the estimated quality surface, $\hat{m}_j(\cdot)$. This bias is then removed from the risk adjusted hospital outcome. If $\hat{m}_j(\cdot)$ is a good estimator for the true quality surface $m_j(\cdot)$, then the adjustment term in equation (14) will reduce any bias due to remaining imbalance (see Athey, Imbens and Wager (2018), Hirshberg and Wager (2021), for more discussion on bias-corrected balancing weights).

One obvious way to obtain $\hat{m}_j(\cdot)$ is via least squares, with hospital-specific intercepts but common coefficients across hospitals, $\hat{m}_j(x) = \hat{\alpha}_j + \hat{\beta} \cdot \phi(x)$ (Normand et al. (2016)). This allows the model to share information on the relationship between the covariates and the outcome, while still allowing for systematic differences in hospital quality. However, like the linear model outlined in Section 3.5, this model is fairly restrictive. Following our assumption on the form of the quality surface (4) as $\hat{m}_j = \hat{\alpha}_j + \hat{\beta}_j \cdot \phi(x)$, we instead estimate $\hat{m}_j$ via ridge regression separately within each hospital, as in equation (10), which allows the quality curves to differ across hospitals. More elaborate pooling procedures are possible, for example, directly using hierarchical Bayesian modeling (George et al. (2017)).

Plugging the linear model into equation (14), the bias-corrected estimator is

$$\hat{\mu}_j = \sum_{Z_i=j} \hat{\gamma}_i Y_i + \underbrace{\hat{\beta}_j \cdot \left(\bar{\phi} - \sum_{Z_i=j} \hat{\gamma}_i \phi(X_i)\right)}_{\text{adjustment for remaining imbalance}}. \tag{15}$$

The hospital fixed effects $\hat{\alpha}_j$ drop out due to the sum-to-one constraint, leaving us with only the coefficients $\hat{\beta}_j$ determining how to aggregate the imbalance in each of the components of $\phi(X_i)$.

This estimator is a special case of the general augmented minimax linear estimator proposed by Hirshberg and Wager (2021), specialized to the separate linear model and this setting with multiple hospitals. We can similarly view this estimator, as related to the approximate residual balancing proposal from Athey, Imbens and Wager (2018), controlling the L^2 norm imbalance and using ridge regression rather than the Lasso. Following Athey, Imbens and Wager (2018), we can inspect the estimation error of this bias corrected estimator and see that the bias is controlled by *both* the imbalance and the error in estimating β_j,

$$\hat{\mu}_j - \mu_j = (\hat{\beta}_j - \beta_j) \cdot \left(\frac{1}{n_j}\sum_{Z_i=j} \hat{\gamma}_i \phi(X_i) - \bar{\phi}\right) + \sum_{Z_i=j} \hat{\gamma}_i \varepsilon_i. \tag{16}$$

Therefore, if the error in the coefficients $\|\hat{\beta}_j - \beta_j\|$ is smaller than the magnitude $\|\beta_j\|$, by including estimates from the outcome model we can reduce the bias arising from any remaining imbalance. Furthermore, under additional assumptions on the propensity score, this bias-corrected estimator can be as efficient as possible (see, e.g., Hirshberg and Wager

(2021)). Compared to the estimation error for outcome modeling alone in equation (11), the bias of the bias-corrected estimator depends on the *product* of the error in the coefficients and the imbalance which will lead to lower bias than outcome modeling alone.

In addition, Ben-Michael, Feller and Rothstein (2021) show that this bias-corrected estimator is itself a weighting estimator with weights

$$\hat{\gamma}_i^{\text{bc}} = \hat{\gamma}_i + \left(\bar{\phi} - \sum_{Z_i=j} \hat{\gamma}_i \phi(X_i)\right)' (\Sigma_j + \lambda_j I)^{-1} (\phi(X_i) - \bar{\phi}_j).$$

So the bias correction adjusts the weights to achieve better balance by allowing for some extrapolation away from the observed patient mix. Compared to only using an outcome model for direct standardization, as in Section 3.5, bias-correcting in this manner will only extrapolate when necessary to achieve better balance. If it is possible to achieve good balance without extrapolating, then the bias-corrected estimator will not extrapolate; otherwise, the level of regularization in the outcome model will control the degree of extrapolation. This differs from using the outcome model alone, which can extrapolate away from the support of the data, even when an adequate solution exists that does not extrapolate; see Ben-Michael, Feller and Rothstein (2021) for further discussion on extrapolation and bias correction.

In general, bias adjustment will be more aggressive for hospitals with larger imbalances. This will be especially true for hospitals with smaller effective sample sizes, as there are fewer patients available for trying to match the population distribution. The adjustments will be smaller for hospitals with excellent balance, for example, hospitals with large patient populations or a high degree of overlap. We also use the model to aid in estimating the variance. Rather than using the residual $Y_i - \hat{\mu}_j$ when estimating the hospital-specific variance $\hat{\sigma}_j^2$, we use the empirical residual $Y_i - \hat{m}_j(X_i)$. Specifically, for the separate linear model the hospital-specific variance is

$$\hat{\sigma}_j^2 = \frac{1}{\sum_{Z_i=j} \hat{\gamma}_i^2 - 1} \sum_{Z_i=j} \hat{\gamma}_i^2 (Y_{ij} - \hat{\alpha}_j - \hat{\beta}_j \cdot \phi(X_i))^2.$$

We then use these hospital variance estimates in the overall pooled estimate $\hat{\sigma}^2_{\text{pool}}$ in equation (13).

In sum, we can use an outcome model to incorporate additional risk adjustment into our estimates of hospital quality, albeit at the price of allowing for extrapolation away from the support of the data. We refer to risk adjustment using weights alone as *weighted risk adjustment* and risk adjustment using both weights and additional outcome modeling as *bias-corrected risk adjustment*.

4.2. *Partially pooling hospital-specific estimates.* Thus far, our approach estimates hospital-specific means $\hat{\mu}_j$ in relative isolation. These estimates can be unstable, especially for the smaller hospitals and those hospitals with low effective sample size. Following standard practice in hospital quality research, we, therefore, partially pool the estimates via a hierarchical Bayesian model (George et al. (2017), Iezzoni (2012), Normand and Shahian (2007)). We can use this approach either with or without the bias correction step in Section 4.1.

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

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

with estimated estimated standard error, $\widehat{\text{se}}_j = \hat{\sigma}_{\text{pool}}/\sqrt{n_j^{\text{eff}}}$. This is a "modular" Bayesian procedure that treats $\widehat{\text{se}}_j$ as known which avoids some complications that arise from estimating hospital-specific variances in a fully Bayesian setup (Jacob et al. (2017)). Nonetheless, it is straightforward to extend this approach to a fully Bayesian or empirical Bayes setup, such as triple goal estimation (Paddock et al. (2006), Shen and Louis (1998)).

Partially pooling the final estimates in this manner differs from other generalized linear mixed models used for assessing hospital quality (Normand and Shahian (2007), Normand et al. (2016)) that partially pool parameters in the outcome model (e.g., they fit varying intercept models) rather than partially pooling the estimated means themselves. We view our approach as more transparent, since the analyst controls the level of partial pooling directly.

5. Simulation study. We conduct a simulation study to explore the benefits of additional regression adjustment compared to weighting alone. For each scenario considered, we first generate a fixed set of J hospitals, where each hospital has a specific distribution of patients, relationship of patient characteristics and outcomes, and overall quality score. We then repeatedly sample patients from each hospital, generate a binary adverse effect for those patients, and fit the resulting data using three strategies: simple averaging (no adjustment), our weighting adjustment, and our weighting adjustment with an additional covariate adjustment. We then compare the sets of resulting hospital specific quality estimates to the true quality scores and measure variability, bias, and root-mean squared prediction error (RMSPE). Overall, we verify, as expected, that in contexts where hospitals serve different patient demographics, weighting does reduce bias, but this comes at a cost of increased variance. Regression adjustment provides additional benefit via further bias reduction. The additional regression adjustment does not increase estimator variance as the covariates—all at the individual level—are estimated with high precision, given the large sample size. We now provide specific details of our model and simulation and then present the results.

5.1. *The data generation process.* We characterize each hospital with three independent variables, $u_0, u_1, u_2 \overset{\text{iid}}{\sim} \text{Unif}[-0.5, 0.5]$, representing latent hospital characteristics. We denote this set of variables as u_k. These variables jointly drive four main aspects of the data generation process: hospital size, hospital quality, hospital patient population served, and the hospital-specific outcome-covariate relationship of the hospital's patients. We have four primary simulation factors: $\bar{\alpha}$, σ_α^2, $\bar{\beta}$, and σ_β^2. Conceptually, $\bar{\alpha}$ is the average hospital quality (i.e., is the mean hospital intercept in a linear model), controlling what proportion of the patients at a median hospital would have a complication, and σ_α^2 is the variance of the individual hospital-level intercepts. The relationship between the covariates and the outcome is controlled by $\bar{\beta}$ and σ_β^2, with $\bar{\beta}$ controlling how strongly the patient level covariates predict patient level risk on average and σ_β^2 controlling the variation across hospitals in the strength of this relationship. A nonzero σ_β^2 allows different hospitals to have different relationships between their patient covariates and outcomes. Critically, when $\sigma_\beta^2 > 0$, increasing $\bar{\beta}$ will increase model misspecification, which should increase the bias for model-based estimates of hospital quality.

Within each hospital j, a patient i, has a vector of seven covariates and the following patient-level risk π_i:

$$\text{logit}(\pi_i) = \bar{\alpha} + \alpha_j + \bar{\beta} v'(X_i - \bar{X}) + (\beta_j - \bar{\beta}) v' X_i, \tag{17}$$

where $v = (0.4, 0.3, 0.4, 0.2, 0.2, 0.2, 0.2)$ is a vector of coefficients for our seven covariates, X_{ij} is the demographic vector of covariates for patient i in hospital j, and $\bar{X}$ is expected average of the X_{ij} across the full population. The α_j is the deviation of how much more (or

less) effective hospital j is than average, and β_j controls how much stronger (or weaker) the covariate-outcome relationship is. We link the u_k to α_j with

$$\alpha_j = \bar{\alpha} + \sigma_\alpha 4(u_0 + u_1 + u_2 - 1.5)$$

and

$$\beta_j = \bar{\beta} + \sigma_\beta 6(u_0 + u_1 - 1),$$

where the centering of -1.5 and -1 and scaling of 4 and 6 to make the α_j and β_j centered at zero with variances of σ_α^2 and σ_β^2.

Given a set of J hospitals, we generate a sample of $N = 80J$ patients across the hospitals, with each hospital's size being proportional to $u_0 + 0.3$, allowing some hospitals to be more than four times larger than others, with an average size of 80. For each patient, we generated covariates with varying relationships to the latent u_k which induces different patient populations across the different hospitals. Our first covariate is a count, with $X_1 \sim \text{Pois}(\lambda)$, with $\lambda = 1 + 1.5\exp^{-1}(u_0)$, where $\exp^{-1}(\cdot)$ is the quantile function of the exponential distribution. Our second and third characteristics are binary indicators with $X_2 \sim \text{Bern}(\frac{1}{3}(u_1 + u_2 + 0.5))$ and $X_2 \sim \text{Bern}(\frac{1}{3}(u_0 + u_1 + u_2))$. The four additional binary covariates are not related to the hospital characteristics and are all drawn as independent Bern(1/2). We then calculate patient level risks using equation (17) and, finally, patient outcomes Y_{ij}, as Bernoulli draws with the given π_{ij}. In sum, patient level covariates predict the risk of a complication; however, there is a hospital level component that varies from hospital to hospital in a stochastic fashion.

5.2. *Simulation implementation*. In each simulation we apply the range of possible methods of direct standardization. First, we include hospital-level estimates without risk adjustment. Next, we apply the two model-based methods of estimation: fixed-effect direct standardization (FEDR) and ridge-weighted direct standardization (RWDR). For these two methods we expect FEDR to display higher levels of bias but have lower variance. Recall that RWDR is equivalent to using a separate ridge regression model to adjust each hospital case-mix. Next, we include two weighting-based methods. First, we applied weighted risk adjustment. Next, we applied bias-corrected risk adjustment—weighting-based estimates combined with modeling of the outcome. As we outlined in Section 4.1, we performed outcome modeling with hospital specific ridge regression fits. We selected the ridge regression tuning parameter via cross-validation. In the simulation we did not apply the Bayesian shrinkage methods.

We focus on $\bar{\beta}$ as the primary factor of our simulation. As $\bar{\beta}$ gets larger, the bias in the naïve estimates of hospital quality should increase, because patient level risk will depend to a greater extent on the covariates, causing hospitals serving different types of patients to diverge. In our simulations we used 10 different values for $\bar{\beta}$ that varied from 0 to 3 in increments of 2/3. We set $\bar{\alpha} = -1$ and vary σ_α^2 and σ_β^2 across 0 and 1. For each set of simulation parameters, we generate a fixed set of hospitals and repeat the simulation 1000 times, resampling the patients with each trial. We fix the hospitals in order to be able to directly estimate the RMSPEs for the individual hospital-specific quality estimates. For each simulation we calculated the standard error, bias, the coverage rate for the 95% confidence interval, and RMSPE for each hospital, and then average across hospitals.

5.3. *Results*. Figure 2 reports results for when $\sigma_\alpha^2 = \sigma_\beta^2 = 1$; see the online Supplementary Material for results with the full set of simulation parameters (Keele et al. (2023b)). As expected, the unadjusted results always have higher bias than any of the direct standardization methods. (we still have bias when $\bar{\beta} = 0$ due to the individual β_j varying and being correlated with hospital size). Both FEDR and RWDR reduce bias, compared to unadjusted results, but

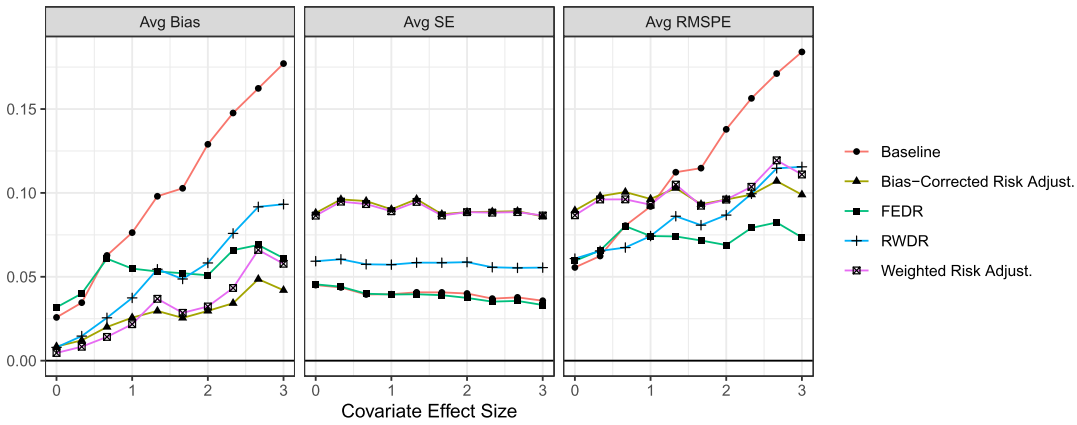


FIG. 2. *Average bias, standard error, and RMSPE across all hospitals and simulation runs. X-axis is* $\bar{\beta}$ *which controls the strength of the relationship between the outcome and covariates. Baseline represents unadjusted results.*

as model misspecfication increases both of these model-based methods underperform relative to weighting. In terms of variance, however, both model-based methods produce the lowest variances across all scenarios. In fact, FEDR produces standard error estimates that are nearly identical to unadjusted results.

Despite the bias reduction advantages, weighting can have a higher RMSPE than no adjustment when the covariates are not generally predictive (i.e., $\bar{\beta}$ is low); in these cases the bias-variance trade-off of weighting is too expensive in terms of variance. We see the variance cost is essentially independent of the covariate-outcome relationship, as illustrated by the flat SE curves. The model approaches also generally outperform weighting with respect to RMSPE given this variance cost. The model adjustment on top of weighting appears to be essentially free: the standard error plot shows no increase in variance for bias-corrected risk adjustment relative to weighting alone, and bias is further reduced, making model adjusted bias correction the lowest bias approach overall. This reduction in bias—with no corresponding increase in variance—provides an overall reduction in RMSPE from model adjustment. For the higher $\bar{\beta}$, RMSPE reduction for bias-corrected risk adjustment is a bit more than 10% than for weighting alone. Depending on the purpose of our estimates, we may choose to reduce bias over reducing overall RMSPE. For example, remaining bias can undermine efforts to estimate cross-hospital variation, as the bias, if it is not systematic across hospitals, would be counted as hospital variation.

Bias also seriously undermines coverage, as shown by Figure 3. We see both weighting methods generally maintain nominal coverage across all the simulation scenarios. By contrast, the smaller variances produced by the model-based methods come at a significant cost: they produce confidence intervals that generally fail to cover the truth.

6. Hospital performance on postoperative complication in general surgical performance.

6.1. *Setup.* We now apply our approach to estimate risk-adjusted complication rates for general surgery patients in Pennsylvania, New York, and Florida. Our first step is to build $\phi(x)$, the basis for assessing covariate imbalance; we set this to be a standardized version of the full set of covariates. Standardizing is important because dimensions of $\phi(x)$ with high variances will implicitly receive greater weight in the optimization problem. For binary covariates with estimated proportion $\hat{p} \geq 0.05$, we standardize by subtracting the mean and

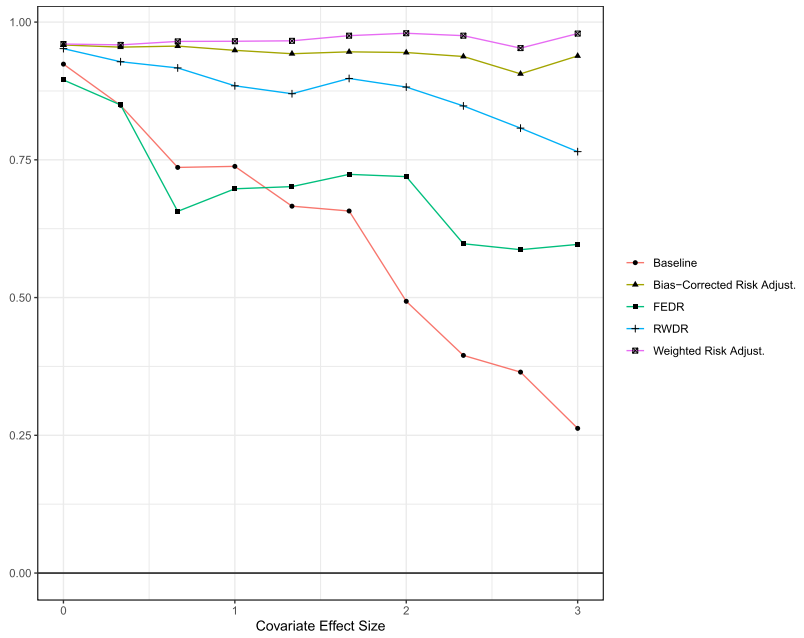


FIG. 3. *Coverage of* 95% *confidence intervals across* $\tilde{\beta}$ *which controls the strength of the relationship between the outcome and covariates. Baseline represents unadjusted hospital level results.*

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. For continuous covariates, we standardize by subtracting the mean and dividing by the standard deviation. We also augment our $\phi(x)$ with one aggregate measure: in our context, many of our covariates are indicators for relatively rare comorbidities, and matching on all of them separately may be difficult. We, therefore, add a key summary measure of this risk, the number of total comorbidities for each patient, and include this as an additional covariate in $\phi(x)$. Our setup is general and, in principle, we could generate a richer basis.

To assess the performance 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 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, the initial approximated bias is $\Delta_h = (\overline{\phi(X)}_h - \overline{\phi(X)})'\hat{\eta}$, and the final bias is $\Delta_{h,w} = (\overline{\phi(X)}_{h,w} - \overline{\phi(X)})'\hat{\eta}$ for each hospital h. The first quantity is the estimated bias due to baseline differences in case mix; the second is the estimated remaining bias due to case mix after weighting. Using these two quantities, we then calculate the percent bias reduction (PBR),

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

This measure describes the estimated change in bias, due to risk standardization, while also accounting for the strength of the association between the different covariates and the outcomes.

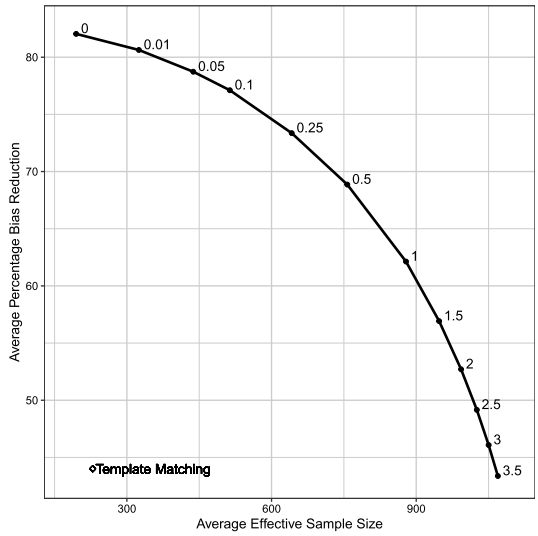


FIG. 4. *Estimated bias-variance tradeoff as a function of λ values. Each dot represents the estimated percent bias reduction and average effective size for balancing weights with different values of λ. The comparable values for template matching are shown in the bottom left, suggesting large gains in both bias reduction and effective sample size from using the proposed weighting approach relative to template matching.*

6.2. *The bias-variance tradeoff and 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} 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 focus on risk adjustment via weighting alone. We also set $u = 1$ (no constraint) which does not limit the amount of weight assigned to any one patient. In the Supplementary Material, we present results for an analysis with $u = 0.2$ and found the results were unchanged (Keele et al. (2023a)).

Figure 4 summarizes the results and demonstrates the trade-off between bias reduction and effective sample size. When $\lambda = 0$ (no attempt to control variance), the estimated percent bias reduction is 80% relative to the unadjusted estimate with an average effective sample size less than 200. While this bias reduction is large, we cannot achieve perfect balance, especially for smaller hospitals. 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 4 suggest a value for λ 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 comparison, we also implemented a template match. Following Silber et al. (2014a), 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 serves as the template. Next, we matched patients from each hospital to the template, using optimal match with refined covariate balance. This approach 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. For each hospital we optimally match individual patients to patients in

the template, dropping template patients that have no match within a given caliper. For each match, therefore, we can obtain up to 300 matched pairs, the size of the template; the number of matched pairs may be smaller, however, if only a subset of patients are comparable. The patients selected from each hospital serve as the risk adjusted population for that hospital, and the subsequent risk adjusted measure of complications is the proportion of complications in that matched sample. For the template match we also calculate the PBR and average sample size across hospitals. We did not discard hospitals where the matches were poor, as was implemented in Silber et al. (2014b). We did this to measure the performance of both methods across all hospitals in the sample.

Figure 4 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 Supplementary Material for more detailed results from this analysis (Keele et al. (2023a)).

We next compared the estimated outcomes and estimated standard errors for $\lambda = 0$ and $\lambda = 0.05$, estimating average bias reduction and average effective sample size for these two scenarios. When $\lambda = 0$, the bias reduction was 86.8% with an average effective sample size of 195. When $\lambda = 0.05$, the bias reduction was 79% with an average effective sample size of 438. Overall, these results suggest that there are gains from selecting a value for λ larger than 0: with a small increase in λ, we lost little in terms of bias reduction while more than doubling the effective sample size.

The results from this analysis raise the question of how users should select a value for λ in applied clinical research. While we do not yet have a data-driven approach for selecting λ, the approach we use here seems reasonable for assessing the bias-variance trade-off. More specifically, users can start with $\lambda = 0$ as a reference point, since this will maximally reduce bias. Users can then estimate additional fits with larger λ values to find the point where bias increases are modest but effective sample size is maximized. Importantly, such choices can be done without respect to outcome information and, therefore, in a principled way.

Finally, we demonstrate how the weights affect the distribution of covariates at the hospital level. Figure 5 shows box plots of hospital means before and after weighting (when $\lambda = 0.05$) for the three key covariates shown in Figure 1. After weighting, the distributions of hospital means are clearly much closer to the population means, though some variation remains. After weighting, the distributions of hospital means are clearly much closer to the population

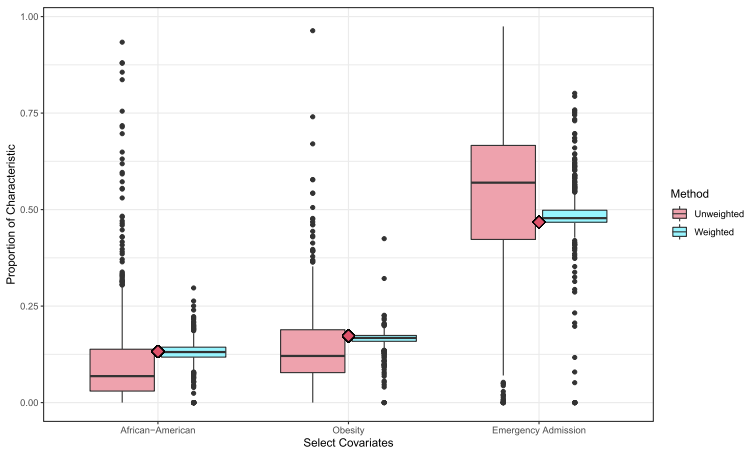


FIG. 5. *Boxplots of hospital casemix distributions for three covariates before and after weighting when* $\lambda = 0.05$. *Diamond represents population mean for that covariate.*

means. For some hospitals, however, there is still some amount of bias that is not eliminated by weighting. These residual imbalances motivate the use of further outcome modeling for additional bias reduction. Moreover, if used in a quality comparison setting, we might elect to remove such hospitals from further cross-hospital outcome comparisons, given that these hospitals do not appear to be directly comparable, even after adjustment via weighting. One area for future investigation is to develop a metric for identifying (and possibly removing) hospitals that are outliers in terms of imbalance after weighting.

6.3. *The role of risk adjustment on hospital quality.* We now focus more directly on how risk adjustment alters hospital level outcomes. To that end, we examine the effects of weighted risk adjustment and then explore the utility of the additional bias-correction step. Figure 6 shows the proportion of complications for each hospital before risk adjustment against the proportion of complications after weighted risk adjustment.[5] Hospitals close to the 45-degree line saw little change; hospitals off the line were changed more. Generally, weighted-risk adjustment induces relatively small changes, although several smaller hospitals changed quite a bit. The nonparametric trend line shows that, in the middle of the distribution, risk adjustment tends to move complication rates up.

Adjustment under the outcome model should also improve the precision of our estimates, depending on the model's predictive power. We calculated an R^2 value for our model by comparing the pooled variances of weighted risk adjustment to that of bias-corrected risk adjustment. Specifically, we compare the overall pooled estimate $\hat{\sigma}^2_{\text{pool}}$ in equation (13) from

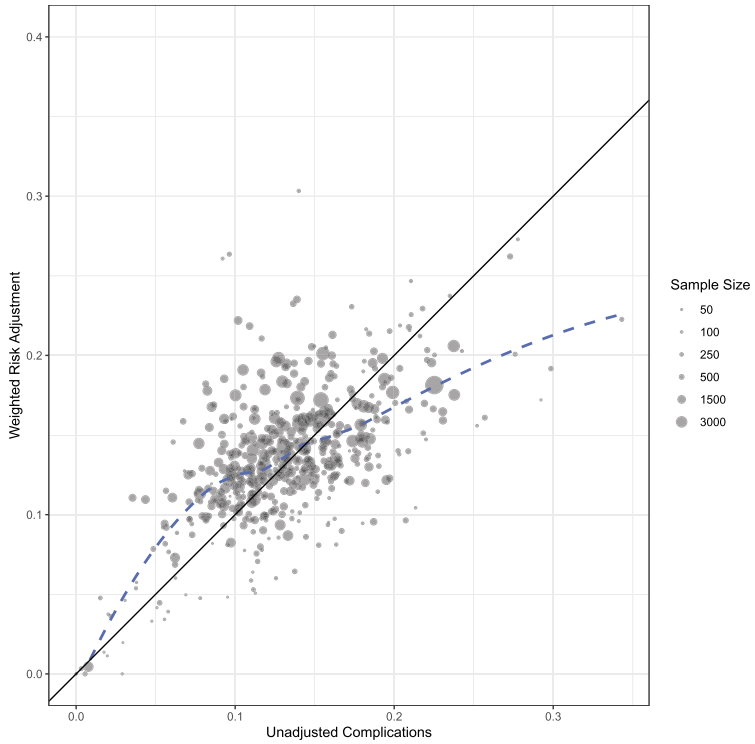


FIG. 6. *Scatterplot of the proportion of complications before risk adjustment against estimates after weighted risk adjustment. Dashed line is a loess fit.*

[5]Figure 6 reflects both actual variation as well as measurement error; we focus on accounting for measurement error with the shrinkage estimator below.

weighting alone $\hat{\sigma}^2_{\text{weighted pool}}$ to the pooled estimate with bias-correction $\hat{\sigma}^2_{\text{bias corrected pool}}$: $1 - \hat{\sigma}^2_{\text{weighted pool}}/\hat{\sigma}^2_{\text{bias corrected pool}} \approx 0.32$. We find that model adjustment, on average, removed 32% of the variation within hospital. This variance reduction will lead to more precise standard errors due to removing variation that we can predict by specific case characteristics.

Next, we investigate two hospitals where risk standardization resulted in large changes in the estimated complication rates. Hospital A had more than 500 general surgery patients,[6] with an unadjusted complication percentage of 6.7%, well below the sample average of 13.5%. After we apply the weights, the risk-adjusted complication percentage increases to 16%, above the sample average. Looking at this hospital's case mix makes this shift clear. The average age of the general surgery patients in this hospital is 51, somewhat low compared to the overall population of 55, and these patients typically have only a single, preexisting medical condition. Moreover, only 15% of the patients at this hospital are admitted for urgent procedures. Hospital A confirms the value of risk adjustment: due to a relatively healthy patient population, hospital A appears to have a low complication percentage; once we risk adjust to match the population as a whole, the estimated complication percentage is much higher.

Compare that to hospital B which had more than 500 general surgery patients. The unadjusted complication percentage is 20.7%, and the risk adjusted complication percentage is 9.6%—approximately a third lower than the percentage at hospital A. However, the patient case mix at hospital B is much different than for hospital A. The average age of the patient at this hospital is 63, higher than the overall average, and patients typically have around three preexisting medical conditions. Moreover, 60% of the procedures at hospital B were urgent admissions. Thus, once we risk adjust to match the characteristics of the population as a whole, we find the estimated performance for hospital B actually improves.

6.4. *Assessing the range of hospital quality.* We next assess the variation in estimated hospital quality and how that variation changes when we adjust hospital complication rates to account for different patient mixes. To do this, we use the "Q statistic" approach from meta-analysis (see, e.g., Hedges and Pigott (2001)). The Q statistic is calculated as

$$Q = \sum_{j=1}^{J} \frac{(\hat{\mu}_j - \bar{\mu})^2}{\widehat{\text{se}}_j^2 + \tau^2},$$

where $\bar{\mu}$ is an estimate of the overall average outcome across hospitals (we use the simple mean) and τ^2 is a hypothesized degree of cross-hospital variation in the true quality measures μ_j.[7]

Under the null hypothesis, $H_0 : \tau = \tau_0$, the Q statistic has an approximate χ^2_{n-1} distribution. We can then estimate τ, using a Hodges–Lehman point estimate (corresponding to the τ with the largest p-value; here, the value where $Q = n - 1$) and generate a confidence interval via test inversion. We used this approach on three sets of hospital estimates: the raw mean outcomes of the hospitals without any adjustment, the mean outcomes of the hospitals after weighted risk adjustment, and the mean outcomes of the hospitals after both weighting and bias correction. Results of these three analyses are summarized in Table 1. When we do not adjust for patient characteristics, the estimated standard deviation is over five percentage

[6]We cannot disclose the specific hospital size in print per our data use agreement.

[7]To build intuition for this estimator, notice that we can decompose the difference of $\hat{\mu}_j$ and $\bar{\mu}$, as $\hat{\mu}_j - \bar{\mu} = (\hat{\mu}_j - \mu_j) + (\mu_j - \bar{\mu})$. The two terms in the denominator correspond to uncertainty in $\hat{\mu}_j - \mu_j$, captured by the estimation uncertainty $\widehat{\text{se}}_j$, and to uncertainty in $\mu_j - \bar{\mu}$, captured by the structural variation in hospital quality, τ. For implementation, see the `blkvar` package: https://github.com/lmiratrix/blkvar/.

TABLE 1

Estimated average and standard deviation of hospital complication rates. First column is overall average across hospitals. The Std. Dev. is the estimated amount of variation in true complication rates across hospitals. The CI is the confidence interval for this amount of variation. The prediction interval estimates the range of the inner 80% *of hospitals, assuming normality in the complication rates. For the rows, Raw includes variation induced by different patient mix*

Estimates	Grand Average	Std. Dev.	CI	80% Prediction Interval
Raw	13.2%	5.1%	(4.9%–5.4%)	7%–20%
Weighted Risk Adjustment	13.5%	2.7%	(2.4%–2.9%)	10%–17%
Bias-corrected Risk Adjustment	13.7%	3.2%	(2.9%–3.5%)	10%–17%

points. Under a Normality assumption, this suggests the complication rate for the middle 80% of hospitals ranges from approximately 7% to 20%.

When we standardize using weighting alone, the standard deviation falls sharply to 2.7 percentage points. Relative to the raw estimates, this suggests that hospitals would have more similar outcomes if treating similar patients. We also see that 70% of the variation in hospital quality is explained by the mix of patients, as measured by an R^2 type statistic of $R^2 = 1 - \sigma^2_{\text{adj}}/\sigma^2_{\text{raw}}$.[8] Finally, the estimates are largely unchanged when we also incorporate bias correction; see the online Supplementary Material for additional results (Keele et al. (2023a)).

6.5. *Results after partial pooling.* As described in Section 4.2, 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 set the random effect G as a simple Normal, $G = N(\alpha_\mu, \tau^2_\mu)$, consistent with prior work on hospital quality (Normand and Shahian (2007), Normand et al. (2016)). 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.

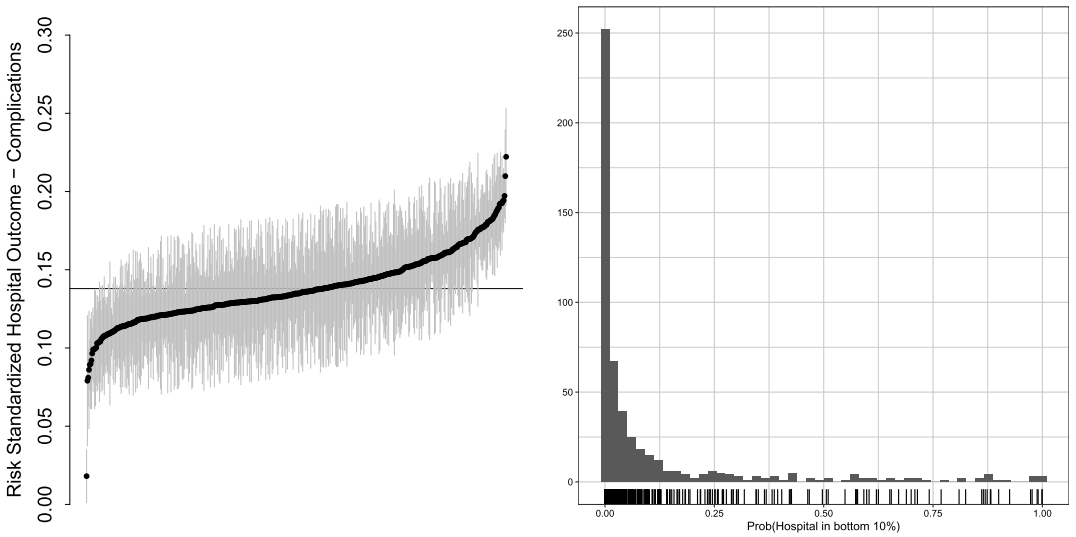
Figure 7a shows the posterior means and corresponding 95% uncertainty intervals for the set of μ_j, the risk-standardized hospital complication rates.[9] We see variation in both the point estimates as well as the width of the hospital-specific uncertainty intervals. While there is a large mass of hospitals in the center of the distribution, there are clearly some hospitals with consistently above- or below-average estimated complication rates.

While the primary aim of our analysis is to produce risk standardized measure of hospital performance, risk adjustment is also used to identify institutions that are outliers (Ridgeway and MacDonald (2009)). For example, hospitals that are identified as underperforming may be targeted for quality improvement efforts. The Bayesian hierarchical model we fit can be used for this purpose. Figure 7b 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: 98.5% of hospitals have a less than 10% chance of being in this low-performing group.

Some hospitals, however, are very likely to be low performing: there are nine hospitals that have at least a 90% chance of being in this low-performing group, four of which with at least

[8]This is a distinct quantity from how predictive individual covariates are for the outcome, which is the R^2 reported in the model adjustment section above.

[9]See Paddock et al. (2006) for a discussion of alternative approaches to summarizing the posterior in terms of the "triple goals" of estimating hospital-specific means, hospital-specific ranks, and the overall distributions.



(a) Catepillar Plot of Risk Standardized Complications

(b) Histogram of the probability of being in the bottom 10% of ranks

FIG. 7. *Hospital quality results after applying Bayesian shrinkage* ($\lambda = 0.05$).

a 99% chance. Among the hospitals with at least a 90% chance of being a low performer, the average adjusted complication rate was 23%—relative to 13.5% overall. Moreover, the average patient volume in this group of hospitals was over 2700 patients, suggesting that, at least in our data, the low performing hospitals are not the lowest volume hospitals.

7. Conclusion. Methods of risk adjustment are widely used to compare the performance of hospitals and physicians. 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 reweighted 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. In the simulations we found that risk adjustment via weighting substantially reduces bias compared to model based methods. In the application we further found weighting outperformed template matching in terms of bias reduction. We also obtained large increases in effective sample size by allowing a slight increase in possible bias. We proposed a bias-correction approach to incorporate outcome modeling as well. Finally, our method of direct standardization can be 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, by contrast, required fine tuning of over five hundred different matches and was a much more time consuming process.

We can extend the proposed approach to allow for a richer covariate basis, including interactions and higher-order terms, and to prioritize balance in some covariates (see Section 6.1). 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)). We expect that this will be a fruitful direction for future work. Currently, we only use the weights to target population level means. We could easily apply this procedure to target a specific subset of patients, such as the average African-American patient in the population. Alternatively, the target for balance need not be the overall patient population. We could, instead, target the patient mix of a specific hospital, a different state, or the U.S. national population. The choice of target population is primarily a substantive question. That is, for what set of patients should hospital be rendered comparable? For example, given that much medical administration is overseen by state governments, it may make sense to restrict the target population to the state in which a hospital is located. Parameter selection could also be optimized to find an optimal tradeoff between bias reduction and effective sample size. We could also explore best practice in terms of the application of the shrinkage methods.

Acknowledgements. We thank Peng Ding, Skip Hirshberg, and Sam Pimentel for helpful feedback as well as seminar participants in the Penn Causal Group. 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 authors 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. This publication was derived, in part, from a limited data set supplied by the Florida Agency for Health Care Administration (AHCA) which specifically disclaims responsibility for any analysis, interpretations, or conclusions that may be created as a result of the limited data set. The authors declare no conflicts.

Funding. Eli Ben-Michael and Avi Feller gratefully are funded by National Science Foundation Grant #1745640.

Rachel Kelz is funded by a grant from the National Institute on Aging, R01AG049757-01A1.

SUPPLEMENTARY MATERIAL

Supplement A: Complete simulation results (DOI: 10.1214/22-AOAS1629SUPPA; .pdf). Complete Simulation Results.

Supplement B: Additional empirical results (DOI: 10.1214/22-AOAS1629SUPPB; .pdf). Full set of results from the empirical application.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386 https://doi.org/10.1198/jbes.2009.07333

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2021). The augmented synthetic control method. *J. Amer. Statist. Assoc.* **116** 1789–1803. MR4353714 https://doi.org/10.1080/01621459.2021.1929245

BEN-MICHAEL, E., HIRSCHBERG, D., FELLER, A. and ZUBIZARRETA, J. (2021). The balancing act for causal inference.

BLOOM, H. S., RAUDENBUSH, S. W., WEISS, M. and PORTER, K. E. (2016). Using multi-site experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *J. Res. Educ. Eff.* 1–66.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76**.

CENTERS FOR MEDICARE AND MEDICAID SERVICES (2021). Quality payment program website. Available at https://qpp.cms.gov/about/qpp-overview.

CENTERS FOR MEDICARE & MEDICAID SERVICES (2021a). Facility-level 7-day hospital visits after general surgery procedures performed at ambulatory surgical centers. Technical report, Washington, D.C.

CENTERS FOR MEDICARE & MEDICAID SERVICES (2021b). Patient-mix coefficients for January 2022 (3Q20 through 1Q21 Discharges) publicly reported HCAHPS results. Available at http://https://www.hcahpsonline.org/ (accessed February 11, 2022).

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

DECKER, M. R., DODGION, C. M., KWOK, A. C., HU, Y.-Y., HAVLENA, J. A., JIANG, W., LIPSITZ, S. R., KENT, K. C. and GREENBERG, C. C. (2014). Specialization and the current practices of general surgeons. *J. Am. Coll. Surg.* **218** 8–15.

DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11** 427–444. MR0003527 https://doi.org/10.1214/aoms/1177731829

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

DEVILLE, J. C., SÄRNDAL, C. E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

ELIXHAUSER, A., STEINER, C., HARRIS, D. R. and COFFEY, R. M. (1998). Comorbidity measures for use with administrative data. *Med. Care* **36** 8–27. https://doi.org/10.1097/00005650-199801000-00004

FLEISS, J. L., LEVIN, B. and PAIK, M. C. (2003). *Statistical Methods for Rates and Proportions*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2001202 https://doi.org/10.1002/0471445428

GEORGE, E. I., ROČKOVÁ, V., ROSENBAUM, P. R., SATOPÄÄ, V. A. and SILBER, J. H. (2017). Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. *J. Amer. Statist. Assoc.* **112** 933–947. MR3735351 https://doi.org/10.1080/01621459.2016.1276021

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. Roy. Statist. Soc. Ser. A* **159** 385–409.

HAINMUELLER, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46. https://doi.org/10.1093/pan/mpr025

HAZLETT, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statist. Sinica* **30** 1155–1189. MR4257528 https://doi.org/10.5705/ss.20

HEDGES, L. V. and PIGOTT, T. D. (2001). The power of statistical tests in meta-analysis. *Psychol. Methods* **6** 203–217.

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. arXiv preprint arXiv:1901.10296.

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HULL, P. (2018). Estimating hospital quality with quasi-experimental data. SSRN 3118358.

IEZZONI, L. I. (2012). *Risk Adjustment for Measuring Health Care Outcomes*, 4th ed. Health Administration Press, Chicago, IL.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules. arXiv:1708.08719.

KALLUS, N. (2020). Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.* **21** Paper No. 62, 54. MR4095341

KEELE, L. J., BEN-MICHAEL, E., FELLER, A., KELZ, R. and MIRATRIX, L. (2023a). Supplement to "Hospital quality risk standardization via approximate balancing weights." https://doi.org/10.1214/22-AOAS1629SUPPA

KEELE, L. J., BEN-MICHAEL, E., FELLER, A., KELZ, R. and MIRATRIX, L. (2023b). Supplement to "Hospital quality risk standardization via approximate balancing weights." https://doi.org/10.1214/22-AOAS1629SUPPB

KITAGAWA, E. M. (1955). Components of a difference between two rates. *J. Amer. Statist. Assoc.* **50** 1168–1194.

KRUMHOLZ, H. M., WANG, Y., MATTERA, J. A., WANG, Y., HAN, L. F., INGBER, M. J., ROMAN, S. and NORMAND, S.-L. T. (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** 1683–1692. https://doi.org/10.1161/CIRCULATIONAHA.105.611186

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LOHR, S. L. (2010). *Sampling*: *Design and Analysis*, 2nd ed. Brooks/Cole, Cengage Learning, Boston, MA. MR3057878

LONGFORD, N. T. (2019). Performance assessment as an application of causal inference. *J. Roy. Statist. Soc. Ser. A*.

MCCAFFREY, D. F., LOCKWOOD, J., KORETZ, D., LOUIS, T. A. and HAMILTON, L. (2004). Models for value-added modeling of teacher effects. *J. Educ. Behav. Stat.* **29** 67–101.

MEDICARE.GOV (2013). Hospital compare. Available at https://www.medicare.gov/hospitalcompare/search.html (accessed February 14, 2022).

MIRATRIX, L. W. and FELLER, A. (2020). Reatment effect distributions in multi-site trials.

NATIONAL QUALITY FORUM (2014). Risk adjustment for socioeconomic status or other sociodemographic factors. Technical report, Washington, D.C.

NORMAND, S.-L. T. and SHAHIAN, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statist. Sci.* **22** 206–226. MR2408959 https://doi.org/10.1214/088342307000000096

NORMAND, S.-L. T., ASH, A. S., FIENBERG, S. E., STUKEL, T. A., UTTS, J. and LOUIS, T. A. (2016). League tables for hospital comparisons. *Annu. Rev. Stat. Appl.* **3** 21–50.

PADDOCK, S. M., RIDGEWAY, G., LIN, R. and LOUIS, T. A. (2006). Flexible distributions of triple-goal estimates in two-stage hierarchical models. *Comput. Statist. Data Anal.* **50** 3243–3262. MR2239666 https://doi.org/10.1016/j.csda.2005.05.008

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244 https://doi.org/10.1080/01621459.2014.997879

POTTHOFF, R. F., WOODBURY, M. A. and MANTON, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *J. Amer. Statist. Assoc.* **87** 383–396. MR1173805

RIDGEWAY, G. and MACDONALD, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *J. Amer. Statist. Assoc.* **104** 661–668. MR2751446 https://doi.org/10.1198/jasa.2009.0034

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

RUBIN, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 185–203.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061 https://doi.org/10.1111/1467-9868.00135

SILBER, J. H., ROSENBAUM, P. R. and ROSS, R. N. (1995). Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics? *J. Amer. Statist. Assoc.* **90** 7–18.

SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., MUKHERJEE, N., SAYNISCH, P. A., EVEN-SHOSHAN, O. et al. (2014a). Template matching for auditing hospital cost and quality. *Health Serv. Res.* **49** 1446–1474.

SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., SAYNISCH, P. A., EVEN-SHOSHAN, O., KELZ, R. R. et al. (2014b). A hospital-specific template for benchmarking its cost and quality. *Health Serv. Res*. **49** 1475–1497.

SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., HILL, A. S., EVEN-SHOSHAN, O., KELZ, R. R. et al. (2016a). Indirect standardization matching: Assessing specific advantage and risk synergy. *Health Serv. Res*. **51** 2330–2357.

SILBER, J. H., ROSENBAUM, P. R., WANG, W., LUDWIG, J. M., CALHOUN, S., GUEVARA, J. P., ZORC, J. J., ZEIGLER, A. and EVEN-SHOSHAN, O. (2016b). Auditing practice style variation in pediatric inpatient asthma care. *JAMA*, *J. Am. Med. Assoc. Southeast Asia*, *Suppl., Pediatr.* **170** 878–886.

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WEISS, M. J., BLOOM, H. S., VERBITSKY-SAVITZ, N., GUPTA, H., VIGIL, A. E. and CULLINAN, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *J. Res. Educ. Eff.* **10** 843–876.

YIU, S. and SU, L. (2018). Covariate association eliminating weights: A unified weighting framework for causal effect estimation. *Biometrika* **105** 709–722. MR3842894 https://doi.org/10.1093/biomet/asy015

ZHAO, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist*. **47** 965–993. MR3909957 https://doi.org/10.1214/18-AOS1698

ZHAO, Q. and PERCIVAL, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5** Art. No. 20160010. MR4323812 https://doi.org/10.1515/jci-2016-0010

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc*. **110** 910–922. MR3420672 https://doi.org/10.1080/01621459.2015.1023805