Investigating Risk Adjustment Methods for Health Care Provider Profiling When Observations are Scarce or Events Rare

Health Services Insights
Volume 11: 1–10

DOI: 10.1177/1178632918785133

Timo B Brakenhoff[1], Karel GM Moons[1], Jolanda Kluin[2] and Rolf HH Groenwold[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [2]Heart Center, Academic Medical Center, Amsterdam, The Netherlands.

ABSTRACT

BACKGROUND: When profiling health care providers, adjustment for case-mix is essential. However, conventional risk adjustment methods may perform poorly, especially when provider volumes are small or events rare. Propensity score (PS) methods, commonly used in observational studies of binary treatments, have been shown to perform well when the amount of observations and/or events are low and can be extended to a multiple provider setting. The objective of this study was to evaluate the performance of different risk adjustment methods when profiling multiple health care providers that perform highly protocolized procedures, such as coronary artery bypass grafting.

METHODS: In a simulation study, provider effects estimated using PS adjustment, PS weighting, PS matching, and multivariable logistic regression were compared in terms of bias, coverage and mean squared error (MSE) when varying the event rate, sample size, provider volumes, and number of providers. An empirical example from the field of cardiac surgery was used to demonstrate the different methods.

RESULTS: Overall, PS adjustment, PS weighting, and logistic regression resulted in provider effects with low amounts of bias and good coverage. The PS matching and PS weighting with trimming led to biased effects and high MSE across several scenarios. Moreover, PS matching is not practical to implement when the number of providers surpasses three.

CONCLUSIONS: None of the PS methods clearly outperformed logistic regression, except when sample sizes were relatively small. Propensity score matching performed worse than the other PS methods considered.

KEYWORDS: Provider profiling, risk adjustment, propensity score, logistic regression, simulation

RECEIVED: January 11, 2018. **ACCEPTED:** May 24, 2018.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: RHH Groenwold was funded by the Netherlands Organization for Scientific Research (NWO-Veni project 916.13.028).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Timo B Brakenhoff, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands. Email: T.B.Brakenhoff-2@umcutrecht.nl; t.brakenhoff@gmail.com

Introduction

Health care providers, whether individual physicians or health care centers, are increasingly being monitored or *profiled* on their quality of care.[1,2] Despite its over-simplistic interpretation of quality of care, mortality is a frequently used outcome measure[3,4] and is used in the widely implemented Hospital Standardized Mortality Ratio model.[5] An essential step when profiling providers is the adjustment for case-mix, often called risk adjustment.[1]

Provider profiling using traditional regression-based risk adjustment has been shown to lead to inconsistent results that are highly dependent on the specific statistical model chosen.[6–8] In addition, these methods may perform poorly when low-volume providers are included or outcomes are rare, leading to an inability to detect poor performing providers in such scenarios.[9,10] As high patient volume has been associated with better patient outcomes,[11,12] it is especially crucial to monitor the quality of care of low-volume providers. Currently, providers that do not reach a certain volume threshold are often omitted from analyses and comparisons.[10]

An alternative to standard regression models when adjusting for covariates in observational studies is propensity score (PS) analysis. When considering dichotomous treatments, PS analysis may outperform standard multivariable analysis,[13–15] especially when dealing with a large number of covariates or few events per covariate.[16,17] Even though several different PS methods have been extended for multiple treatment comparisons (see Linden et al[18] for a comparison), these methods have barely been considered in a provider profiling setting.[19]

The objective of this study was to assess the performance of PS methods (notably PS adjustment, PS weighting, and PS matching) for risk adjustment in studies of multiple providers. A simulation study was performed to investigate the effect of sample size, event rate, and provider volume on the risk adjustment performance of PS methods and conventional fixed-effects logistic regression when profiling 3 providers. Subsequently, the suitability of using these PS methods in a more realistic provider profiling setting was explored by considering scenarios with up to 20 providers. Finally, different risk adjustment methods were illustrated in an empirical example from the field of cardiac surgery.

Risk Adjustment Methods

Fixed-effects logistic regression

Traditionally, multivariable regression models have been used to adjust provider effects for possible case-mix variables. The inclusion of providers as random or fixed effects in the logistic regression model is largely dependent on the goal of provider profiling.[20,21] In this article, a fixed-effects logistic regression model was chosen for both the data generation model and analysis model as the aim was to make direct comparisons between only a few provider effects. In addition, only patient-level case-mix variables were included, reducing the necessity of a hierarchical model. Given the theoretical differences between fixed- and random-effects models, it was deemed inappropriate and unfair to analyze data generated under a fixed-effects model with a random-effects risk adjustment method.

PS models

In 1983, Rosenbaum and Rubin introduced the PS as "the conditional probability of assignment to a particular treatment given a vector of observed covariates"[22] and demonstrated that adjustment using these scores was sufficient to remove bias due to observed covariates if the assumptions of exchangeability and positivity hold. In health care provider profiling, the *treatment* is not a medical intervention but the provider attended by the patient. When comparing 2 providers, the PS can be estimated using a binary logistic regression model where the provider indicator is regressed on the observed case-mix variables. The fitted values of this model, the PSs, can then be used for stratification, covariate adjustment, inverse probability weighting, or matching. PS weighting, PS matching, and to a lesser extent PS adjustment have been shown to lead to a better balance of case-mix between providers and less biased effect estimates when compared with PS stratification.[14,23] PS stratification will therefore not be considered in this article.

Generalized PS models

The PS methods can be extended to a multiple provider setting using the generalized PS (gPS) described by Imbens[24] as the conditional probability of attending a particular provider given case-mix variables.[25] The gPS of each provider can be estimated using multinomial logistic regression including all relevant observed case-mix variables. The application of the aforementioned risk adjustment methods will be described for a setting with 3 providers, yet naturally extends to situations with more than 3 providers. For gPS adjustment, the outcome is regressed on 2 dummy variables of the provider indicator, 2 of the estimated gPSs, and possible interactions. This allows the estimation of the conditional provider effect (for further details, see the following implementations: Spreeuwenberg et al[26]; Feng et al[27]). For gPS weighting, the sample is reweighted by the inverse gPS of the provider actually attended. The outcome is then regressed on 2 dummy variables of the provider indicator to estimate the marginal provider effect.[27] Extreme weights may be trimmed to help reduce the influence of outliers and model misspecification in certain situations but can also decrease the amount of risk adjustment.[28] For gPS matching, the average provider effect of the matched set can be estimated after selecting individuals from each provider based on the overlap of the gPSs, also known as the common support region.[29] Although gPS matching will not necessarily use the same target population as the above-mentioned methods to estimate the provider effect, the estimated average provider effects are comparable if interaction effects are absent (ie, the average provider effect is then equal to the average provider effect on those attending the provider). For the remainder of this article, the standard multivariable logistic regression method will be referred to as *LR*, gPS adjustment as PS_C, gPS weighting as PS_W, gPS weighting with trimming as PS_{WT}, and gPS matching as PS_M. The performance of all these methods was assessed using a simulation study.

Simulation Study

A Monte Carlo simulation study was performed using R (v3.1.2)[30] to assess the influence of sample size, event rate, provider volume, and number of providers on the performance of the PS methods and *LR*. The first 3 factors were varied in a limited provider profiling setting with only 3 providers. Even though such a situation is rarely encountered in practice, it is analogous to the 3 treatment settings for which the studied methods have been extended previously and allow for a detailed assessment of performance. Subsequently, the suitability and performance of the studied methods in settings with up to 20 providers was explored. Using a simulation study, estimated provider effects of each method could be compared with their true (marginal or conditional) effects. The simulation study was conducted in a controlled setting with properly specified regression models, no interaction terms, and equal coefficients for all included case-mix variables. This ensured comparability of the causal effects estimated by each method.

Data generation

The data generation procedure described in this section was written for a 3-provider setting, yet naturally extends to situations with more than 3 providers.

Ten case-mix variables ($Z_1, \ldots, Z_{10}$) were generated from a multivariate standard normal distribution with correlations either all equal to 0 or 0.1. These variables were then included as covariates in a multinomial logistic regression model to assign each patient to 1 of 3 centers (A, B, or C) within provider indicator X, where center B acted as the reference category. The coefficients of the case-mix variables for center A and C, $\beta_{j1}, \ldots, \beta_{j10}$, were set equal to 1/10, where $k = \{A, C\}, j \in$

Table 1. For each scenario, the number of providers, total sample size over all providers (N), sample size distribution and total event rate was fixed.

SCENARIO	NO. OF PROVIDERS	N	SAMPLE SIZE DISTRIBUTION	TOTAL EVENT RATE, %
1	3	500	$N_A=N_B=N_C$	10
2	3	1000	$N_A=N_B=N_C$	10
3	3	2000	$N_A=N_B=N_C$	10
4	3	5000	$N_A=N_B=N_C$	10
5	3	10 000	$N_A=N_B=N_C$	10
6	3	10 000	$N_A=N_B=N_C$	28
7	3	10 000	$N_A=N_B=N_C$	13
8	3	10 000	$N_A=N_B=N_C$	02
9	3	10 000	$N_A=N_B=N_C$	01
10	3	10 000	$N_A=N_C=800$	10
11	3	10 000	$N_A=N_C=2500$	10
12	3	10 000	$N_A=N_C=4600$	10
13	5	15 000	$N_A=\ldots=N_E$	10
14	10	30 000	$N_A=\ldots=N_J$	10
15	15	45 000	$N_A=\ldots=N_O$	10
16	20	60 000	$N_A=\ldots=N_T$	10

k. The following formula was used to generate probabilities for categories A and C of the provider indicator *X*:

$$\pi_j = \frac{e^{\alpha_j+\beta_{j1}Z_1+\cdots+\beta_{j10}Z_{10}}}{1+\sum_k e^{\alpha_k+\beta_{k1}Z_1+\cdots+\beta_{k10}Z_{10}}} \quad (1)$$

Given that center B was the reference category, $\pi_B = 1-(\pi_A+\pi_C)$. As the total sample size (N) was fixed, the intercepts of the multinomial model (α_A and α_C) shown in equation (1) could be manipulated to determine the size of each provider (N_j). A fixed-effects logistic regression model was used to generate the dichotomous outcome variable (*Y*). Providers *A* and *C* (provider *B* acted as reference) were included in the model as dummy variables (X_A and X_C) with relative coefficients (β_A and β_C) of −0.5 and 0.5, respectively. Thus, irrespective of patient characteristics, the estimated odds of mortality for a patient attending provider *A* or *C* were $e^{-0.5}=0.61$ and $e^{0.5}=1.65$ times the estimated odds for a patient attending provider *B*. In scenarios with more than 3 providers, provider *A* acted as reference with the remaining providers having relative coefficients between −1 and 1 assigned at equidistant intervals based on the number of providers. $Z_1, \ldots, Z_{10}$ were included with $\beta_{Z1}, \ldots, \beta_{Z10}=1/10$:

$$logit\left[P(Y=1)\right] = \alpha + \beta_A X_A + \beta_C X_C + \beta_{Z1}Z_1 + \cdots + \beta_{Z10}Z_{10} \quad (2)$$

Due to this data-generating model, the case-mix variables acted as confounders of the provider-outcome relation. No interaction terms were included in the model. The provider effects were therefore assumed constant over the different levels of the case-mix variables. On average, the unadjusted estimates of β_A and β_C were found to be −0.40 and 0.60, respectively, across simulations.

A total of 16 scenarios were investigated in which the number of providers, the total sample size over all providers, provider volumes, and the event rate were separately manipulated (see Table 1). Varying the total event rate was achieved by manipulating the intercept (α) of the logistic model (equation (2)), whereas the intercepts of the multinomial model (α_A and α_C in equation (1)) were manipulated to determine the distribution of the sample size over the providers. Each scenario was simulated 2000 times. Scenarios 1 through 12 were repeated with a correlation of 0.1 between all case-mix variables. This correlation coefficient is frequently encountered between baseline variables in observational studies.[31]

Methods

This section describes how the methods, which were described above, were applied in the 3-provider setting (scenarios 1-12). For scenarios 13 to 16, PS_M was not applied due to both logistical and computational challenges that arise when trying to find

suitable matches for more than 3 groups. For *LR, Y* was regressed on 2 dummy variables for X (X_A and X_C) and all 10 case-mix variables ($Z_1, \ldots, Z_{10}$) just as in equation (2). The *svyglm* function of the *survey* package (v3.30)[32] was used to estimate the model coefficients and the corresponding standard errors (using Taylor series linearization). For all methods except PS_W, the weight of each individual was set to 1. To alleviate potential problems with separation in the most extreme scenarios of scenarios 1 to 12, results of *LR* were compared with Firth's bias reduced logistic regression[33] (applied using the *logistf* package [v1.21][34]).

For the PS methods, gPSs were first estimated from the data by fitting a multinomial regression model using the function *multinom* of the *nnet* package (v7.3)[35] and extracting each patients' fitted values for all categories of X. For PS_C, a logistic regression model was fitted with 2 dummy variables for X (X_A and X_C) and 2 gPSs (gPS_A and gPS_B). For PS_W, the gPSs were first used to calculate a weight for each patient. A patient's weight was equal to the inverse of the gPS of the provider actually attended. A weighted logistic regression analysis was performed as in *LR*, except with only the 2 dummy variables representing X. For PS_{WT}, the highest 2% of weights were trimmed to the 98th percentile.[28] The determination of the optimal trimming threshold was beyond the scope of this study. For PS_M, a 1:1:1 matching without replacement strategy was used where the gPS_A and gPS_B values of all individuals were divided into equal-sized bins. The bin width was equal to 0.2 times the pooled standard deviation of the logit of the gPS_A and gPS_B values, based on the caliper width advised by Wang et al.[36] A matched set consisted of one random individual from each category of X that fell within the same bin. The amount of individuals in the matched set was therefore always smaller than the original sample and depended on the overlap of the PS distributions in the 3 groups. The data set containing all matched sets was then analyzed using marginal logistic regression with only the 2 dummy variables of X included in the model. All models used in each method were properly specified as all case-mix variables used to generate the data were also included in the analysis. Investigations into the consequences of model misspecification are discussed elsewhere.[37,38]

Reference values

When determining the reference values to compare the provider effect estimates (B_j) with, it is important to consider the different types of effects each method estimates. *LR* and PS_C both estimate a provider effect conditional on either the observed case-mix variables or a summary in the form of gPSs. The reference provider effects (β_j) were therefore set equal to the conditional effects used in the data-generating model. Thus, for a 3-provider setting, $\beta_A = -0.5$ and $\beta_C = 0.5$, whereas for a 5-provider setting, $\beta_A = -1$, $\beta_C = -0.33$, $\beta_D = 0.33$, and $\beta_E = 1$. PS_W and PS_{WT} both estimate a marginal provider effect. As this effect is influenced by the event rate, different reference values were used for scenarios 5 through 9. The reference values were determined by taking the mean over 100 samples of 10^6 patients generated for each event rate with $\beta_{j1}, \ldots, \beta_{j10} = 0$. The effect of the case-mix variables on X was thus removed and a marginal logistic regression model was fit using X_A and X_C. A similar procedure was used to determine the marginal reference values for scenarios 13 through 16. For PS_M, the analysis did not take into account the matched nature of the data set as suggested by Sjölander and Greenland.[39]

Performance measures

For each scenario, the bias, coverage, and mean squared error (MSE) of the estimated provider effects were assessed over 2000 simulations. The bias is equal to the difference between the average estimated provider effect over all simulations and the reference value. The coverage is equal to the proportion of times that the reference value falls within the 95% confidence interval (CI) constructed around the estimated provider effect over all simulations. To provide a measure of the uncertainty in the results of the simulations, note that in case the true coverage is 95%, the coverage based on 2000 simulation runs is expected to lie between 94% and 96% in 95% of the simulations. To confirm that the standard errors of the provider effects were being properly estimated for each method in each scenario, the ratio of the average standard error of B_j over the standard deviation of the 2000 estimates of B_j was also examined. A ratio of 1 indicates that these values are identical. The MSE is equal to the sum of the average squared standard error of the provider effect over all simulations and the square of the bias.

Results

For scenarios 1 to 12, there were no meaningful differences in the performance of methods when the correlation between the case-mix variables was 0 or 0.1. The figures discussed in this section thus assumed a correlation of 0. Furthermore, Firth's bias reduced logistic regression gave practically identical results to *LR*. As such, the former method is not further discussed.

Total sample size. In Figure 1, the bias and coverage (of the 95% CI) of B_A and B_C are shown for all 5 methods at different total sample sizes, corresponding to scenarios 1 through 5. When the total sample size was at least 2000, *LR*, PS_C, and PS_W gave unbiased estimates of B_A and B_C, whereas PS_M and PS_{WT} consistently slightly overestimated B_A and B_C. For lower sample sizes, all methods underestimated B_A and overestimated B_C, which was in the same direction as the reference value. PS_M and *LR* showed the most bias when the total sample size was only 500, with biases of about −0.065 (13% of β_A) and −0.030 (6% of β_A) for B_A and 0.035 (7% of β_C) and 0.030 (6% of β_C) for B_C. Coverage of the 95% CI of B_C fluctuated closely about 0.95 for all methods and sample sizes. Coverage of B_A was more variable with *LR*, PS_C, and PS_M showing very slight overcoverage when the sample size was 500.

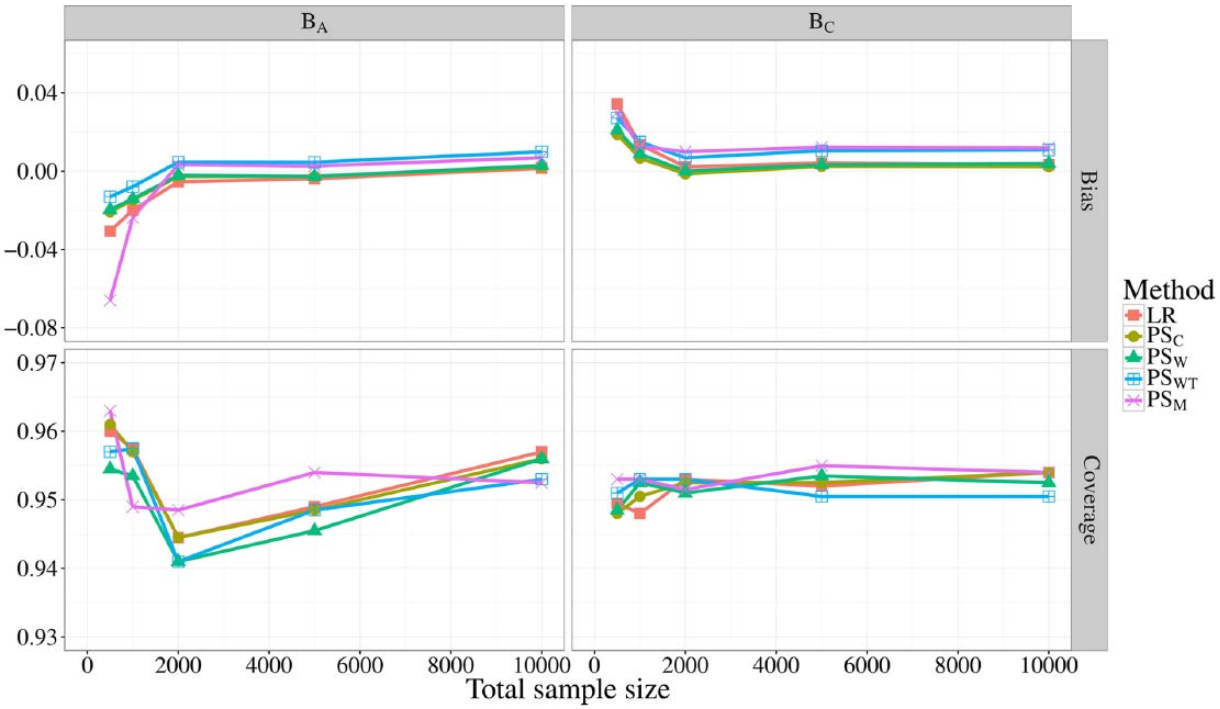

Figure 1. Bias (top) and coverage of the 95% confidence interval (bottom) of B_A and B_C for different total sample sizes. The sample sizes were evenly distributed over providers with a fixed event rate of 10%. Different line colors represent the different risk adjustment methods used.

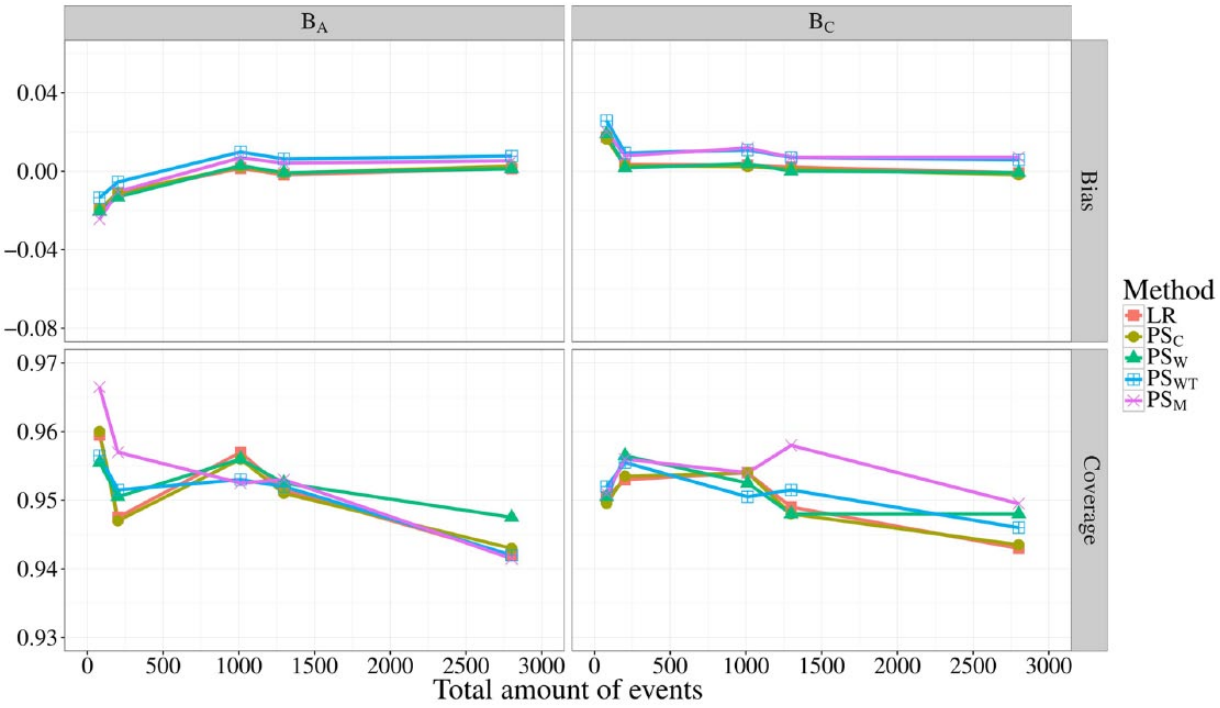

Figure 2. Bias (top) and coverage of the 95% confidence interval (bottom) of B_A and B_C for differing total amounts of events. The total sample size was fixed to 10 000 and distributed evenly over the providers. Different line colors represent the different risk adjustment methods used.

Total event rate. In Figure 2, the bias and coverage of B_A and B_C are shown for all 5 methods at different event rates, corresponding to scenarios 5 through 9. As the total sample size was kept constant at 10 000, the most extreme scenario (9) had an average of 100 total events per simulated data set. Only when the total amount of events decreased below 200, did all methods show slight bias, never exceeding an absolute bias of 0.03. Coverage probabilities of all methods fluctuated between 0.94 and 0.96 and were similar for both B_A and B_C at all event rates, with only PS_M straying beyond 0.96 when the total amount of events was 100.

Sample size distribution. In Figure 3, the bias and coverage of B_A and B_C are shown for all 5 methods at different provider

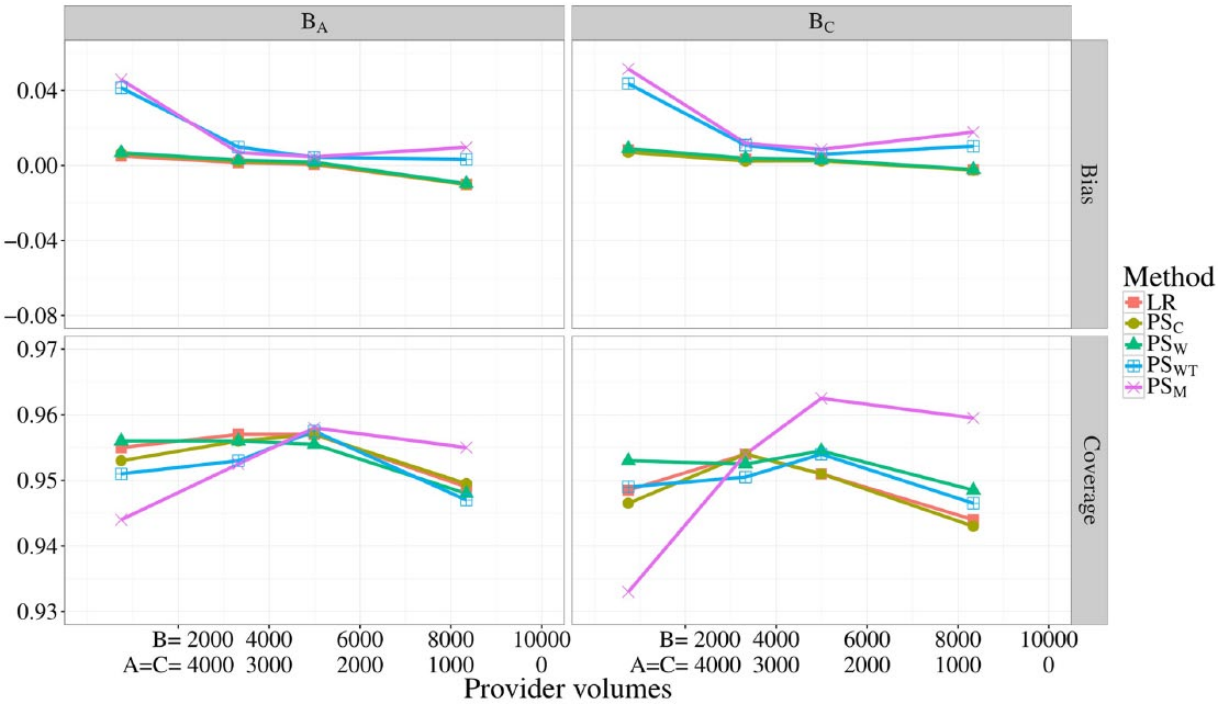

Figure 3. Bias (top) and coverage of the 95% confidence interval (bottom) of B_A and B_C for differing provider volumes. The total sample size was fixed to 10 000. Different line colors represent the different risk adjustment methods used.

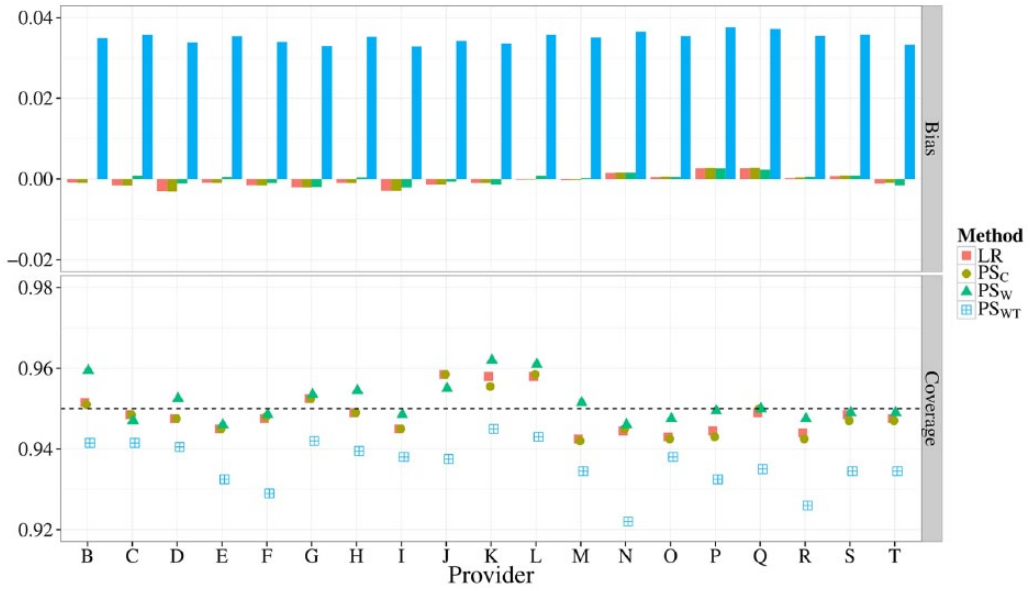

Figure 4. Bias (top) and coverage of the 95% confidence interval (bottom) of 19 estimated provider effects when using different risk adjustment methods. All provider volumes were fixed to 3000 with a total event rate of 10%.

volumes, corresponding to scenarios 5, and 10 through 12. Note that for all scenarios, the total sample size was 10 000 and the volumes of providers *A* and *B* were kept equal. Provider volumes seemed to have no meaningful effect on the bias or coverage of B_A and B_C when using *LR*, PS_C, or PS_W. When using, PS_M or PS_{WT}, however, the absolute bias exceeded 0.04 for both provider effects when provider *B* had only 7% of the total sample size. Although PS_{WT} demonstrated good coverage for both provider effects, PS_M showed both undercoverage and overcoverage for B_C when the volume of provider *B* was low and high, respectively.

Number of providers. In Figure 4, the bias and coverage of 20 provider effects (corresponding to 16) is shown when using *LR*, PS_C, PS_W, or PS_{WT} for risk adjustment. Similar figures comparing 5, 10, or 15 providers (corresponding to scenarios 13 through 15) are shown in the Appendix (Supplemental Figures 1 to 3). The absolute bias of *LR*, PS_C, and PS_W never exceeded 0.005 for all

estimated provider effects as the number of providers increased from 5 to 20. For PS_{WT}, however, the overall bias of the provider effects increased as the number of providers increased, culminating in biases of about 0.04 for many provider effects when profiling 20 providers. Coverage probabilities of provider effects fluctuated closely about 0.95 for all methods when considering 5 providers. When profiling 20 providers, however, PS_{WT} lead to undercoverage (below 0.94) for most provider effects.

Standard error estimation and MSE. Ratios of the average estimated standard errors over the standard deviation of provider effects fluctuated closely about 1 in almost all situations. Only when applying PS_M with a sample size of 500, did this ratio drop below 0.7 for B_A indicating that the average standard error was an underestimation of the actual variation in provider effects simulated. As expected, the MSE generally declined as the total sample size or amount of events increased. Although most methods had an almost identical MSE under all conditions, PS_M consistently scored higher, especially when the total sample size decreased to 500. Figures for these outcome measures can be found in the Appendix (Supplemental Figures 4 to 10).

Empirical Example—Profiling Cardiac Surgery Centers

Open heart surgery is a field that has been subject to many developments in risk-adjusted mortality models for quality control in the past decades.[40,41] Although many have disputed the legitimacy of mortality as a proxy for quality,[42–44] mortality is considered appropriate when profiling procedures such as coronary artery bypass grafting (CABG).[40,41,45] A selection of anonymized data from the Adult Cardiac Surgery Database provided by the Netherlands Association of Cardio-Thoracic Surgery (NVT; www.nvtnet.nl) was used to illustrate the statistical methods evaluated above when profiling multiple centers. This database is similar to databases in other countries, such as the Society of Thoracic Surgeons Adult Cardiac Surgery Database (STS-ACSD) maintained in the United States which has also been used for recent provider profiling investigations.[41]

Data

The Adult Cardiac Surgery Database of the NVT contains patient and intervention characteristics of all cardiac surgery performed in 16 centers in the Netherlands as of January 1, 2007. This data set has previously been described and used by Siregar et al[46,47] for benchmarking. In the current study, all patients, from all 16 centers, who underwent isolated CABG with an intervention date between January 1, 2007, and December 31, 2009, were included in the cohort. Case-mix variables were selected based on the EuroSCORE prediction model. Dichotomous case-mix variables with an overall prevalence below 5% were excluded from the analysis. In-hospital mortality was used as the dichotomous mortality indicator. As a result, the final data set included 8 case-mix variables (age [centered], sex, chronic pulmonary disease, extracardiac arteriopathy, unstable angina, LV dysfunction moderate, recent myocardial infarction, and emergency intervention), 1 mortality indicator, and 1 anonymized center indicator (with centers labeled A through P). This data set contained 25 114 patients with an average center mortality rate of 1.4%, ranging from 0.7% to 2.3%.

Comparison of risk adjustment methods

Although the performance of the different risk adjustment methods could be compared in the simulation study described earlier, this is not possible in an empirical data set as the *true* center effects are unknown. Nevertheless, the consequences of using different risk adjustment methods can be illustrated by ranking centers based on their standardized mortality ratio (SMR). This ratio is calculated by dividing the observed by the expected mortality. Expected mortality rates were calculated using the case-mix variables mentioned above. After fitting the different risk adjustment models on the full data set, the predicted probability of mortality was extracted for the patients attending each center. To mimic a situation with smaller provider volumes or more frequent monitoring, the same procedure was applied to a selection of the total data set only including information from the year 2008. Note that PS_M was not included as a risk adjustment method due to logistical issues that arise when dealing with more than 3 centers.

Results

In Figure 5, the SMRs for all 16 centers are ranked for the year 2008 as well as the years 2007 through 2009 combined. In the total data set, all methods showed slight differences in the rankings of the centers. This disagreement increased in the reduced data set as the uncertainty around the SMRs became much larger due to the smaller center volumes. The similarity in rankings of LR and PS_C in both panes echoes the similar performance observed in the simulation study. The marginal methods (PS_W and PS_{WT}) led to quite different conclusions, especially in the lower ranked centers.

Discussion

Key findings

Our simulation study, in which risk adjustment methods for provider profiling were compared, showed that of the 4 PS methods considered, PS adjustment and PS weighting performed best. Both showed similar or slightly less absolute bias as compared with conventional logistic regression across all scenarios of the simulation study. The PS matching clearly performed worse in terms of bias and coverage than the other methods when the number of observations decreased. Furthermore, PS matching and PS weighting with trimming were the only methods strongly affected by the distribution of volume across providers. When the number of providers to be profiled increased beyond 3, PS weighting was the only method that led to increasingly biased provider effect estimates.

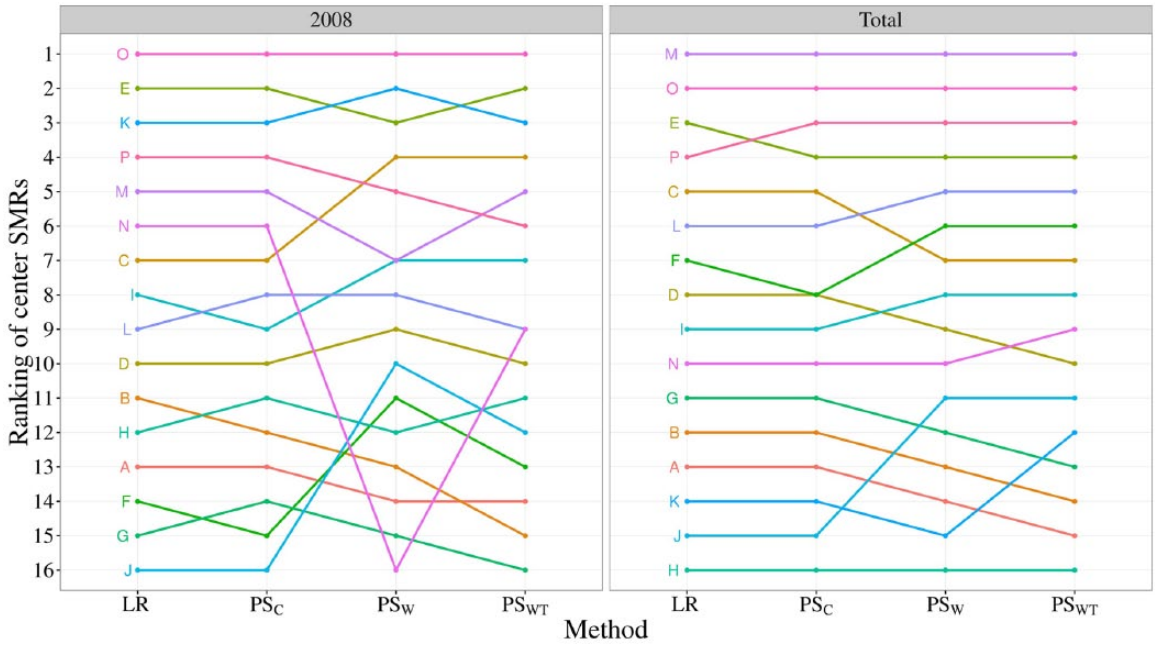

Figure 5. Ranking of SMRs of all 16 centers for the total data set and for 2008 separately. A rank of 1 is given to the center with the lowest SMR. SMR indicates standardized mortality ratio.

Relation to previous work

In line with Spreeuwenberg et al,[26] PS adjustment consistently showed similar performance to logistic regression. Also, PS adjustment as well as PS weighting suffered, albeit slightly, in performance when sample sizes were low, which was suggested before by Feng et al.[27] The PS weighting with trimming did not improve on the performance of untrimmed weighting. This may be due to the fact that only one trimming threshold was investigated. However, the arduous task of determining the trimming threshold that has the optimal variance-bias trade-off was beyond the scope of this study. Other alternatives to enhance PS weighting include using stabilized weights.[48]

These findings contrasted those of simulation studies comparing PS methods with conventional regression analysis in settings with 2 exposure levels (eg, providers) in which observations or events were rare. In our study, none of the PS methods clearly outperformed logistic regression.[13,15] Furthermore, PS matching performed slightly worse than all other methods, especially when the total sample size was very low.[14] This is most likely due to the added complexity of applying PS methods in settings with multiple providers, as studies are yet to find noteworthy performance improvements of different PS methods over conventional regression analyses when comparing multiple treatment options.[25–27,29]

PS matching

The performance of PS matching was likely influenced by the specific matching procedure applied, as earlier simulation studies have shown that risk adjustment performance can depend highly on the specific matching algorithm used.[10] As the matching procedure was developed for quick application in a simulation study and ease of use, it failed to locally minimize distances between potential matches. This could have led to the occasionally biased estimates and the consistently higher MSE, which were not found by Rassen et al[29] when using a computationally more intensive matching algorithm that did implement local minimization.

Strengths and limitations

A strength of our simulation study is that the scenarios investigated were chosen to reflect realistic situations that may be encountered in practice. As such, more extreme scenarios with smaller total sample sizes or lower event rates were deemed unnecessary. However, an obvious limitation of this simulation study was that most scenarios are still a simplification of reality, in which often more than 20 providers are profiled. Nevertheless, the studied scenarios allowed for a fair technical comparison of PS methods and multivariable regression in a provider profiling context. Scenarios 13 to 16 suggest that most PS methods can also be applied in settings with more providers, yet further investigation into the practical consequences (eg, in terms of outlier detection rates) of using these methods in settings with many providers or when there are unobserved case-mix differences is required.

Unobserved case-mix

In all simulations that were performed, all relevant case-mix variables were observed and appropriately included in the model (either the PS model or the outcome regression model). All methods used the same amount of case-mix information and all models were correctly specified. Although the

performance of the methods differs for relatively small samples, it may not come as a surprise that for relatively large samples the different methods yield similar results. However, in the presence of nonignorable, yet unobserved, differences in case-mix across providers, the different methods may yield biased results, also in relatively large samples. The question whether the methods are differentially affected by unobserved case-mix was beyond the scope of this study.

Directions for future research

The PS methods investigated in this article are the ones most commonly encountered in the literature and easiest to apply. There are, however, alternative and more complex methods that may be used for risk adjustment. First of all, the gPSs can also be estimated using machine learning procedures such as generalized boosted models.[49] These methods are able to estimate larger numbers of gPSs with higher accuracy than conventional multinomial regression models but are computationally more intensive and therefore not included in the simulation study. There are also alternative ways to use the estimated gPSs. One such example is marginal mean weighting through stratification, which computes weights based on stratified PSs and has recently been suggested as a suitable risk adjustment method.[50] To limit the scope of this study, these methods were not considered. Further research is required into these alternative gPS estimation and risk adjustment methods to determine whether they are better than the risk adjustment methods presented in this study.

Recommendations

Inherent advantages of PS methods compared with covariate adjustment to correct for differences in case-mix have been described before.[14] The PS methods separate the design from the analysis of a study, allowing the assessment of balance and overlap of case-mix variables across different providers; an assessment that can be performed independent of the outcome variable. Furthermore, once balance is achieved, eg, through PS matching, it becomes relatively easy to study multiple outcomes. However, due to unfamiliarity with the methods, application of PS methods may be more prone to error compared with more traditional covariate adjustment through regression analysis. Considering the similarity in performance between PS methods and covariate adjustment through logistic regression that we observed in our simulations, neither of the methods can be clearly recommended instead of the others.

Conclusions

None of the PS methods clearly outperformed logistic regression, except for relatively small sample sizes. The PS matching performed slightly worse than all other methods, especially when the total sample size was very low.

Availability of Data and Material

Anonymized data was used for the empirical example with permission of the NVT. The original (nonanonymized) data are not available to the authors or the public and cannot be published in full form due to privacy concerns of the surgery centers included in the data set. The NVT has very strict terms to which need to be agreed for use of their data. This means that even the anonymized or de-identified data cannot be provided to third parties as the aggregated numbers can already allow direct identification of surgery centers. R code used for the simulation results can be found at https://github.com/timobrakenhoff/ProviderProfiling1 and an R simulation results file can be found at https://figshare.com/s/35b9b9b4e450e95b9b07.

Author Contributions

TBB performed the simulation study, executed the analysis of the empirical example, and wrote the majority of the manuscript. KGMM supervised the research, contributed to the design of the simulation study, and assisted in the assessment of societal impact of the research. JK was closely involved with the acquisition of the data for the empirical example and proofread the paper. RHHG was the initiator of the research project, closely supervised all technical analyses included in the paper and was a significant contributor to the decisions made for the analysis as well as the writing of the manuscript. All authors read and approved the final manuscript.

Supplemental Material

Supplementary material for this article is available online.

REFERENCES

1. Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes*. 4th ed. Chicago, IL: Health Administration Press; 2013.
2. Shahian DM, He X, Jacobs JP, et al. Issues in quality measurement: target population, risk adjustment, and ratings. *Ann Thorac Surg*. 2013;96:718–726. doi:10.1016/j.athoracsur.2013.03.029.
3. Lilford RJ, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet*. 2004;363:1147–1155.
4. Lilford RJ, Brown CA, Nicholl J. Use of process measures to monitor the quality of clinical practice. *BMJ*. 2007;335:648–650.
5. Jarman B, Gault S, Alves B, et al. Explaining differences in English hospital death rates using routinely collected data. *BMJ*. 1999;318:1515–1520.
6. Shahian DM, Wolf RE, Iezzoni LI, Leslie Kirle MPH, Normand ST. Variability in the measurement of hospital-wide mortality rates. *N Engl J Med*. 2010;363:2530–2539.
7. Eijkenaar F, van Vliet RCJA. Performance profiling in primary care: does the choice of statistical model matter? *Med Decis Making*. 2014;34:192–205. doi:10.1177/0272989X13498825.
8. Glance LG, Dick AW, Osler TM, Li Y, Mukamel DB. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State cardiac surgery report card. *Med Care*. 2006;44:311–319.
9. Krell RW, Hozain A, Kao LS, Dimick JB. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg*. 2014;149:467–474. doi:10.1001/jamasurg.2013.4249.
10. Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: a Monte Carlo study. *Circulation*. 2014;7:299–305. doi:10.1161/CIRCOUTCOMES.113.000685.
11. Birkmeyer JD, Siewers AE. Hospital volume and surgical mortality in the United States. *N Engl J Med*. 2002;346:1128–1137.

12. Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Ann Intern Med*. 2002;137:511–520. doi:10.7326/0003-4819-137-6-200209170-00012.
13. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59:437–447. doi:10.1016/j.jclinepi.2005.07.004.
14. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Mult Behav Res*. 2006;46:399–424. doi:10.1080/00273171.2011.568786.
15. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *Int J Epidemiol*. 2008;37:1142–1147. doi:10.1093/ije/dyn079.
16. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287. doi:10.1093/aje/kwg115.
17. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006;163:262–270. doi:10.1093/aje/kwj047.
18. Linden A, Uysal SD, Ryan A, Adams JL. Estimating causal effects for multivalued treatments: a comparison of approaches. *Stat Med*. 2015;35:534–552. doi:10.1002/sim.6768.
19. Huang IC, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv Res*. 2005;40:253–278.
20. MacKenzie TA, Grunkemeier GL, Grunwald GK, et al. A primer on using shrinkage to compare in-hospital mortality between centers. *Ann Thorac Surg*. 2015;99:757–761. doi:10.1016/j.athoracsur.2014.11.039.
21. Austin PC, Alter DA, Tu JV. The use of fixed-and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Med Decis Making*. 2003;23:526–539. doi:10.1177/0272989X03258443.
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
23. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23:2937–2960. doi:10.1002/sim.1903.
24. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87:706–710.
25. Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *J Am Stat Assoc*. 2004;99:854–866. doi:10.1198/016214504000001187.
26. Spreeuwenberg MD, Bartak A, Croon MA, et al. The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Med Care*. 2010;48:166–174. doi:10.1097/MLR.0b013e3181c1328f.
27. Feng P, Zhou X-H, Zou QM, Fan MY, Li XS. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med*. 2012;31:681–697. doi:10.1002/sim.4168.
28. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS ONE*. 2011;6:e18174. doi:10.1371/journal.pone.0018174.
29. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*. 2013;24:401–409. doi:10.1097/EDE.0b013e318289dedf.
30. R Core Team. R: a language and environment for statistical computing. 2015. Organization: R Foundation for Statistical Computing. CIty: Vienna, Austria.
31. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med*. 2017;4:1985–1992. doi:10.1371/journal.pmed.0040352.
32. Lumley T. Analysis of complex survey samples. *J Stat Softw*. 2004;9:1–19.
33. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80:27–38.
34. Georg Heinze and Meinhard Ploner (2016). logistf: Firth's Bias-Reduced Logistic Regression. R package version 1.22. https://CRAN.R-project.org/package=logistf
35. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York: Springer; 2002.
36. Wang Y, Cai H, Li C, et al. Optimal caliper width for propensity score matching of three treatment groups: a Monte Carlo study. *PLoS ONE*. 2013;8:e81045. doi:10.1371/journal.pone.0081045.
37. Landsman V, Pfeiffer RM. On estimating average effects for multiple treatment groups. *Stat Med*. 2013;32:1829–1841. doi:10.1002/sim.5690.
38. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Meth Med Res*. 2017;26:1654–1670.
39. Sjölander A, Greenland S. Ignoring the matching variables in cohort studies—when is it valid and why? *Stat Med*. 2013;32:4696–4708. doi:10.1002/sim.5879.
40. Shahian DM, Normand S-LT, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155–2168.
41. Englum BR, Saha-Chaudhuri P, Shahian DM, et al. The impact of high-risk cases on hospitals' risk-adjusted coronary artery bypass grafting mortality rankings. *Ann Thorac Surg*. 2015;99:856–862.
42. Shackford SR, Hyman N, Ben-Jacob T, Ratliff J. Is risk-adjusted mortality an indicator of quality of care in general surgery? A comparison of risk adjustment to peer review. *Ann Surg*. 2010;252:452–459. doi:10.1097/SLA.0b013e3181f10a66.
43. Lilford RJ, Pronovost PJ. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ*. 2010;340:c2016.
44. van Gestel YR, Lemmens VE, Lingsma HF, de Hingh IH, Rutten HJ, Coebergh JW. The hospital standardized mortality ratio fallacy: a narrative review. *Med Care*. 2012;50:662–667. doi:10.1097/MLR.0b013e31824ebd9f.
45. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci*. 2007;22:206–226. doi:10.1214/088342307000000096.
46. Siregar S, Groenwold RHH, Jansen EK, Bots ML, van der Graaf Y, van Herwerden LA. Limitations of ranking lists based on cardiac surgery mortality rates. *Circ Cardiovasc Qual Outcomes* 2012;5(3):403–409. doi:10.1161/CIRCOUTCOMES.111.964460.
47. Siregar S, Groenwold RRH, Versteegh MIM, et al. Data resource profile: adult cardiac surgery database of the Netherlands Association for Cardio-Thoracic Surgery. *Int J Epidemiol*. 2013;42:142–149. doi:10.1093/ije/dys241.
48. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664. doi:10.1093/aje/kwn164.
49. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32:3388–3414. doi:10.1002/sim.5753.
50. Hong G. Marginal mean weighting through stratification: a generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychol Methods*. 2012;17:44–60. doi:10.1037/a0024918.