iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.55102

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Comparative Evaluation of Risk Adjustment Techniques for Provider Profiling: A Propensity Score Analysis

Wasim Fathima Shah
Wellmark- Blue Cross Blue Shield

Abstract: ***Background: Profiling health care providers requires proper adjustment for case-mix. However, traditional risk adjustment methods may yield suboptimal results, especially in situations with small provider volumes or rare events. Propensity score (PS) methods, widely used in observational studies of binary treatments, have shown promising performance with limited observations and events and can be extended to multiple provider settings. This study aims to evaluate the performance of various risk adjustment methods in profiling multiple health care providers that conduct highly standardized procedures, such as coronary artery bypass grafting. A simulation study compared provider effects estimated through PS adjustment, PS weighting, PS matching, and multivariable logistic regression in terms of bias, coverage, and mean squared error (MSE), while varying event rates, sample sizes, provider volumes, and number of providers. An empirical example from cardiac surgery was used to illustrate the different methods. Overall, PS adjustment, PS weighting, and logistic regression produced provider effects with low bias and good coverage. In contrast, PS matching and PS weighting with trimming yielded biased effects and high MSE in several scenarios. Additionally, PS matching became impractical when the number of providers exceeded three. No PS method significantly outperformed logistic regression, except in scenarios with relatively small sample sizes. Propensity score matching performed inferiorly compared to the other PS methods analyzed.***

Keywords: Provider profiling, risk adjustment, propensity score, logistic regression, healthcare

I. INTRODUCTION

Healthcare providers, including individual physicians and healthcare centers, are increasingly subject to monitoring and evaluation of the quality of care they deliver [1], [2]. Although mortality is a commonly used outcome measure for quality of care, it has been criticized for its simplistic nature [3], [4]. Nevertheless, mortality remains a prominent metric utilized in the widely adopted Hospital Standardized Mortality Ratio model [5]. An essential step in provider profiling is the adjustment for case-mix, often referred to as risk adjustment [1].

The traditional regression-based risk adjustment methods employed in provider profiling have been found to yield inconsistent results, heavily dependent on the specific statistical model chosen [6]–[8]. Additionally, these methods may perform poorly when dealing with low-volume providers or rare outcomes, making it challenging to identify underperforming providers in such situations [9], [10]. As high patient volume has been associated with better patient outcomes, it becomes especially critical to monitor the quality of care delivered by low-volume providers [11], [12]. Presently, providers failing to meet a certain volume threshold are often excluded from analyses and comparisons [10].

An alternative approach to adjusting covariates in observational studies is propensity score (PS) analysis [13]– [15]. PS analysis has demonstrated superior performance over standard multivariable analysis, particularly in cases involving numerous covariates or few events per covariate [16], [17]. Although several PS methods have been extended for multiple treatment comparisons [18], they have received limited consideration in the context of provider profiling [19].

The objective of this study is to evaluate the performance of PS methods, specifically PS adjustment, PS weighting, and PS matching, for risk adjustment in studies involving multiple healthcare providers. To achieve this, a simulation study investigates the impact of sample size, event rate, and provider volume on the risk adjustment performance of PS methods and conventional fixed-effects logistic regression in profiling three providers. Furthermore, the applicability of these PS methods is explored in a more realistic provider profiling setting, encompassing scenarios with up to 20 providers. Finally, the study illustrates different risk adjustment methods through an empirical example from the field of cardiac surgery.

II. RISK ADJUSTMENT METHODS

Fixed-effects logistic regression has traditionally been utilized to adjust for provider effects concerning potential case-mix variables. Whether providers are included as random or fixed effects in the logistic regression model depends on the objectives of provider profiling [20], [21]. In this study, a fixed-effects logistic regression model was chosen for both the data generation model and analysis model since the aim was to directly compare a limited number of provider effects. Furthermore, the model only considered patient-level case-mix variables, making a hierarchical model less necessary. Given the theoretical distinctions between fixed- and random-effects models, it was considered unsuitable and biased to analyze data generated under a fixed-effects model using a random-effects risk adjustment method. Propensity score (PS) models were introduced by Rosenbaum and Rubin in 1983 as "the conditional probability of assignment to a particular treatment given a vector of observed covariates" [22]. These scores can be used to adjust for bias resulting from observed covariates if the assumptions of exchangeability and positivity are satisfied. In the context of healthcare provider profiling, the "treatment" corresponds to the provider attended by the patient. When comparing two providers, the PS can be estimated using a binary logistic regression model, regressing the provider indicator on the observed case-mix variables. The PSs, i.e., the fitted values of this model, can then be employed for stratification, covariate adjustment, inverse probability weighting, or matching. PS weighting, PS matching, and to a lesser extent PS adjustment have been shown to achieve better case-mix balance among providers and less biased effect estimates compared to PS stratification [14], [23]. Therefore, PS stratification is not considered in this study. The PS methods can be extended to a multiple provider setting using the generalized PS (gPS), defined by Imbens [24] as the conditional probability of attending a particular provider given case-mix variables [25]. For gPS adjustment, multinomial logistic regression is employed to estimate the gPS of each provider by including all relevant observed case-mix variables. The outcome is then regressed on two dummy variables of the provider indicator, two of the estimated gPSs, and possible interactions, allowing the estimation of the conditional provider effect. For gPS weighting, the sample is reweighted using the inverse gPS of the attended provider, and the outcome is regressed on two dummy variables of the provider indicator to estimate the marginal provider effect. Extreme weights may be trimmed to mitigate the impact of outliers and model misspecification in certain situations. For gPS matching, individuals from each provider are selected based on the overlap of the gPSs, known as the common support region, and the average provider effect of the matched set is estimated. The estimated average provider effects are comparable to those of other methods if interaction effects are absent. In this study, the standard multivariable logistic regression method is denoted as LR, gPS adjustment as PSC, gPS weighting as PSW, gPS weighting with trimming as PSWT, and gPS matching as PSM [26], [27], [28], [29]. The performance of all these methods was evaluated through a simulation study.

III.SIMULATION STUDY

A Monte Carlo simulation study was conducted using R (v3.1.2) to evaluate the impact of sample size, event rate, provider volume, and number of providers on the performance of the PS methods and LR. Initially, the first three factors were varied in a limited provider profiling setting involving only three providers. Although this situation is rarely encountered in real-world scenarios, it is analogous to the three treatment settings for which the studied methods have been previously extended, enabling a comprehensive assessment of their performance. Subsequently, the suitability and performance of the studied methods were investigated in settings with up to 20 providers. The simulation study facilitated a comparison of the estimated provider effects of each method with their true (marginal or conditional) effects. The study was designed to ensure comparability of the causal effects estimated by each method by employing properly specified regression models, no interaction terms, and equal coefficients for all included case-mix variables.

A. Data Generation

The data generation process was tailored for a three-provider setting but could be extended to scenarios with more than three providers. Ten case-mix variables (Z1, ..., Z10) were generated from a multivariate standard normal distribution with correlations set either to 0 or 0.1. These variables were then incorporated as covariates in a multinomial logistic regression model to assign each patient to one of the three centers (A, B, or C) within the provider indicator X, with center B acting as the reference category. The coefficients of the case-mix variables for centers A and C, $\beta j1, ..., \beta j10$ ($k = \{A, C\}, j \in k$), were set equal to 1/10. The probabilities for categories A and C of the provider indicator X were generated using the formula presented in equation (1). As the total sample size (N) was fixed, the intercepts of the multinomial model (αA and αC) in equation (1) were manipulated to determine the size of each provider (Nj). A fixed-effects logistic regression model was employed to generate the dichotomous outcome variable (Y), with providers A and C (provider B acting as the reference) included in the model as dummy variables (XA and XC) with relative coefficients (βA and βC) of -0.5 and 0.5, respectively.

In scenarios with more than three providers, provider A served as the reference, and the remaining providers were assigned relative coefficients between -1 and 1 at equidistant intervals based on the number of providers. The case-mix variables Z1, ..., Z10 were included with βZ1, ..., βZ10 = 1/10, following the formulation in equation (2). The case-mix variables acted as confounders of the provider-outcome relation in the data-generating model, and no interaction terms were considered. The provider effects were assumed constant over different levels of the case-mix variables, with unadjusted estimates of βA and βC averaging to approximately -0.40 and 0.60, respectively, across simulations. A total of 16 scenarios were explored, wherein the number of providers, the total sample size over all providers, provider volumes, and the event rate were individually manipulated (refer to Table 1). The variation in the total event rate was achieved by adjusting the intercept (α) of the logistic model (equation (2)), while the intercepts of the multinomial model (αA and αC in equation (1)) were modified to determine the sample size distribution across providers. Each scenario was simulated 2000 times, and scenarios 1 to 12 were repeated with a correlation of 0.1 between all case-mix variables, which is frequently encountered in observational studies [31].

B. Methods

In the three-provider setting (scenarios 1-12), the methods described earlier were applied. However, for scenarios 13 to 16, PSM was not utilized due to logistical and computational challenges arising when attempting to find suitable matches for more than three groups. For LR, Y was regressed on two dummy variables for X (XA and XC) and all ten case-mix variables (Z1, ..., Z10) as described in equation (2). The svyglm function from the survey package (v3.30) was used to estimate the model coefficients and their corresponding standard errors using Taylor series linearization. For all methods except PSW, the weight of each individual was set to 1. To mitigate potential issues with separation in extreme scenarios of scenarios 1 to 12, results of LR were compared with Firth's bias-reduced logistic regression, applied using the logistf package (v1.21).

For the PS methods, gPSs were first estimated from the data by fitting a multinomial regression model using the multinom function from the nnet package (v7.3). Patients' fitted values for all categories of X were extracted to calculate the gPSs. For PSC, a logistic regression model was fitted with two dummy variables for X (XA and XC) and two gPSs (gPSA and gPSB). For PSW, the gPSs were used to calculate a weight for each patient, with the weight being equal to the inverse of the gPS of the provider actually attended. A weighted logistic regression analysis was performed, similar to LR, but only with the two dummy variables representing X. For PSWT, the highest 2% of weights were trimmed to the 98th percentile to address extreme weights. The optimal trimming threshold was not determined in this study. For PSM, a 1:1:1 matching without replacement strategy was used, and the gPSA and gPSB values of all individuals were divided into equal-sized bins. The bin width was set to 0.2 times the pooled standard deviation of the logit of the gPSA and gPSB values based on the caliper width recommended by Wang et al. A matched set consisted of one random individual from each category of X within the same bin. The number of individuals in the matched set was therefore smaller than the original sample and depended on the overlap of the PS distributions in the three groups. The dataset containing all matched sets was analyzed using marginal logistic regression, including only the two dummy variables of X 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. The consequences of model misspecification were examined elsewhere [37], [38].

C. Reference Values

When determining the reference values for comparing the estimated provider effect (Bj) across methods, it is essential to consider the type of effect each method estimates. LR and PSC estimate a provider effect conditioned on either the observed case-mix variables or the generalized propensity scores (gPSs). The reference provider effects (βj) were set to equal the conditional effects used in the data-generating model. For example, in a 3-provider setting, βA = -0.5 and βC = 0.5, while in a 5-provider setting, βA = -1, βC = -0.33, βD = 0.33, and βE = 1. PSW and PSWT estimate a marginal provider effect, and different reference values were used for scenarios with varying event rates. The reference values were obtained by removing the effect of case-mix variables on X and fitting a marginal logistic regression model using XA and XC over 100 samples of 106 patients generated for each event rate with βj1, ..., βj10 = 0. For PSM, the matched nature of the dataset was not considered in the analysis.

D. Performance Measures

The performance of each method in each scenario was assessed using bias, coverage, and mean squared error (MSE) over 2000 simulations. The bias represents the difference between the average estimated provider effect and the reference value across all simulations. Coverage indicates the proportion of times the reference value falls within the 95% confidence interval (CI) constructed around the estimated provider effect across all simulations.

To assess the reliability of the estimated standard errors for each method in each scenario, the ratio of the average standard error of Bj to the standard deviation of the 2000 estimates of Bj was examined. A ratio of 1 indicates identical values. The MSE is the sum of the average squared standard error of the provider effect and the square of the bias.

E. Results

For scenarios 1 to 12, there were no significant differences in the method performance when the correlation between the case-mix variables was 0 or 0.1. Thus, the results discussed assumed a correlation of 0. Additionally, Firth's bias-reduced logistic regression yielded nearly identical results to LR and was not further discussed. The bias and coverage of BA and BC are shown for all five methods at varying total sample sizes corresponding to scenarios 1 through 5. LR, PSC, and PSW provided unbiased estimates of BA and BC when the total sample size was at least 2000. Conversely, PSM and PSWT slightly overestimated BA and BC consistently. For lower sample sizes, all methods underestimated BA and overestimated BC, consistent with the reference value direction. PSM and LR exhibited the most bias at a total sample size of only 500, with biases of approximately -0.065 (13% of βA) and -0.030 (6% of βA) for BA, and 0.035 (7% of βC) and 0.030 (6% of βC) for BC. Coverage of the 95% CI of BC remained close to 0.95 for all methods and sample sizes, while the coverage of BA was more variable, with LR, PSC, and PSM slightly overcovering when the sample size was 500.

Total event rate. The bias and coverage of BA and BC for all five methods at varying event rates, corresponding to scenarios 5 through 9. The total sample size remained constant at 10,000, and the most extreme scenario (9) had an average of 100 total events per simulated dataset. All methods exhibited only slight bias when the total number of events was above 200, with an absolute bias not exceeding 0.03. The coverage probabilities for all methods fluctuated between 0.94 and 0.96 and were similar for both BA and BC at all event rates, except for PSM, which slightly exceeded 0.96 when the total number of events was 100.

Sample size distribution. The bias and coverage of BA and BC for all five methods at different provider volumes, corresponding to scenarios 5 and 10 through 12. In all scenarios, the total sample size was 10,000, and the volumes of providers A and B were kept equal. LR, PSC, and PSW demonstrated no significant impact of provider volumes on the bias or coverage of BA and BC. However, when using PSM or PSWT, the absolute bias exceeded 0.04 for both provider effects when provider B represented only 7% of the total sample size. While PSWT showed good coverage for both provider effects, PSM exhibited both undercoverage and overcoverage for BC when the volume of provider B was low and high, respectively.

Number of providers. The bias and coverage of 20 provider effects (corresponding to scenario 16) when using LR, PSC, PSW, or PSWT for risk adjustment. Similarly10, or 15 providers (corresponding to scenarios 13 through 15) are available in the Appendix. For LR, PSC, and PSW, the absolute bias of the estimated provider effects never exceeded 0.005 as the number of providers increased from 5 to 20. However, PSWT displayed an increasing bias as the number of providers increased, resulting in biases of approximately 0.04 for many provider effects when profiling 20 providers. The coverage probabilities of provider effects remained close to 0.95 for all methods when considering 5 providers. When profiling 20 providers, PSWT led to under coverage (below 0.94) for most provider effects.

Standard error estimation and MSE. In almost all situations, the ratios of the average estimated standard errors to the standard deviation of provider effects remained close to 1. However, when applying PSM with a sample size of 500, this ratio dropped below 0.7 for BA, indicating an underestimation of the actual variation in provider effects. As expected, the MSE generally decreased as the total sample size or number of events increased. While most methods had almost identical MSE under all conditions, PSM consistently scored higher, particularly when the total sample size decreased to 500. Further details on these outcome measures can be found in the Appendix.

IV. EMPIRICAL EXAMPLE

Open heart surgery has witnessed significant developments in risk-adjusted mortality models for quality control over the past decades. While there have been debates about the suitability of mortality as a proxy for quality, it has been deemed appropriate for profiling procedures like coronary artery bypass grafting (CABG). To demonstrate the evaluated statistical methods for profiling multiple centers, anonymized data from the Adult Cardiac Surgery Database provided by the Netherlands Association of Cardio-Thoracic Surgery (NVT) was utilized. This database is similar to similar databases in other countries, such as the Society of Thoracic Surgeons Adult Cardiac Surgery Database (STS-ACSD) in the United States, which has also been employed in recent provider profiling investigations.

A. Data

The Adult Cardiac Surgery Database of the NVT contains patient and intervention characteristics of all cardiac surgeries performed in 16 centers in the Netherlands from January 1, 2007. In this study, all patients who underwent isolated CABG with an intervention date between January 1, 2007, and December 31, 2009, from all 16 centers, 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 utilized as the dichotomous mortality indicator. Consequently, the final data set comprised 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 encompassed 25,114 patients with an average center mortality rate of 1.4%, ranging from 0.7% to 2.3%.

B. Comparison of risk adjustment methods

While the simulation study allowed for the comparison of different risk adjustment methods, the empirical data set does not provide the true center effects. Nevertheless, the implications of using various risk adjustment methods can be illustrated by ranking centers based on their standardized mortality ratio (SMR). SMR is calculated by dividing the observed mortality by the expected mortality. Expected mortality rates were computed using the aforementioned case-mix variables. After fitting the different risk adjustment models on the complete data set, the predicted probability of mortality was obtained for each center's patients. A similar procedure was applied to a subset of the total data set, only including information from the year 2008, to simulate scenarios with smaller provider volumes or more frequent monitoring. Note that PSM was not included due to logistical issues when dealing with more than three centers.

C. Results

The ranked SMRs for all 16 centers for the year 2008 and the years 2007 through 2009 combined. In the total data set, slight differences were observed in the rankings of the centers across all methods. This disparity increased in the reduced data set due to larger uncertainty around the SMRs caused by smaller center volumes. The similarity in rankings of LR and PSC in both panes aligns with the similar performance observed in the simulation study. However, the marginal methods (PSW and PSWT) resulted in different conclusions, particularly for lower-ranked centers.

V. DISCUSSION

A. Key Findings

The simulation study comparing risk adjustment methods for provider profiling indicated that, among the four PS methods considered, PS adjustment and PS weighting performed the best. Both demonstrated similar or slightly less absolute bias compared to conventional logistic regression across all scenarios. However, PS matching performed notably worse in terms of bias and coverage than the other methods, especially when the number of observations decreased. Additionally, PS matching and PS weighting with trimming were the only methods significantly influenced by the distribution of volume across providers. As the number of providers to be profiled increased beyond three, PS weighting led to increasingly biased provider effect estimates.

B. Relation to Previous Work

In line with Spreeuwenberg et al. [26], PS adjustment consistently demonstrated similar performance to logistic regression. Additionally, both PS adjustment and PS weighting experienced slight performance decline when dealing with low sample sizes, which was previously suggested by Feng et al. [27]. However, PS weighting with trimming did not yield better results than untrimmed weighting. This might be attributed to the investigation of only one trimming threshold. Nevertheless, determining the optimal variance-bias trade-off trimming threshold was beyond the scope of this study. Other potential enhancements for PS weighting include using stabilized weights [48].

These findings contrasted with simulation studies that compared PS methods with conventional regression analysis in settings with only two exposure levels (e.g., providers) and where observations or events were rare. In our study, none of the PS methods clearly outperformed logistic regression [13, 15]. Furthermore, PS matching exhibited slightly inferior performance compared to all other methods, especially with very low total sample sizes [14]. This difference could be attributed to the increased complexity of applying PS methods in settings with multiple providers, as studies have yet to find significant performance improvements of various PS methods over conventional regression analyses when comparing multiple treatment options [25–27, 29].

C. PS Matching

The performance of PS matching was likely influenced by the specific matching procedure employed, as previous simulation studies have shown that risk adjustment performance can heavily depend on the matching algorithm used [10]. The chosen matching procedure was designed for quick application in a simulation study, emphasizing ease of use, and might not have effectively minimized distances between potential matches, leading to occasional biased estimates and consistently higher MSE, which was not found by Rassen et al. [29] when using a computationally more intensive matching algorithm that implemented local minimization.

D. Strengths and Limitations

A strength of our simulation study is that the scenarios selected reflected realistic situations that may arise in practice. Extreme scenarios with smaller total sample sizes or lower event rates were considered unnecessary. However, an obvious limitation of this simulation study was that most scenarios were still simplifications of reality, where often more than 20 providers are profiled. Nevertheless, these 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, but further investigation is required to assess the practical consequences, such as outlier detection rates, of using these methods with many providers or when there are unobserved case-mix differences.

E. Unobserved case-mix

In all performed simulations, all relevant case-mix variables were observed and appropriately included in the model (either the PS model or the outcome regression model). While the performance of the methods differed for relatively small samples, it is not surprising that for relatively large samples, the different methods yield similar results. However, in the presence of nonignorable, yet unobserved, case-mix differences across providers, the different methods may yield biased results, even in relatively large samples. Investigating whether the methods are differentially affected by unobserved case-mix was beyond the scope of this study.

F. Directions for Future Research

The PS methods investigated in this article are the most commonly encountered and easiest to apply in the literature. However, there are alternative and more complex methods that may be used for risk adjustment. For instance, the gPSs can also be estimated using machine learning procedures, such as generalized boosted models [49]. These methods can estimate larger numbers of gPSs with higher accuracy than conventional multinomial regression models, though they are computationally more intensive and were not included in the simulation study. Additionally, there are alternative ways to use the estimated gPSs, like marginal mean weighting through stratification, which computes weights based on stratified PSs and has been suggested as a suitable risk adjustment method [50]. Further research is needed into these alternative gPS estimation and risk adjustment methods to determine whether they offer improvements over the risk adjustment methods presented in this study.

VI. RECOMMENDATIONS

Inherent advantages of PS methods compared to covariate adjustment for correcting case-mix differences have been previously described [14]. The PS methods separate the study's design from its analysis, enabling the assessment of balance and overlap of case-mix variables across different providers, independently of the outcome variable. Moreover, once balance is achieved, for example, through PS matching, it becomes relatively easy to study multiple outcomes. However, due to the unfamiliarity with the methods, applying PS methods may be more susceptible to errors compared to more traditional covariate adjustment through regression analysis. Considering the similar performance between PS methods and covariate adjustment through logistic regression observed in our simulations, neither of the methods can be clearly recommended over the others.

VII. CONCLUSIONS

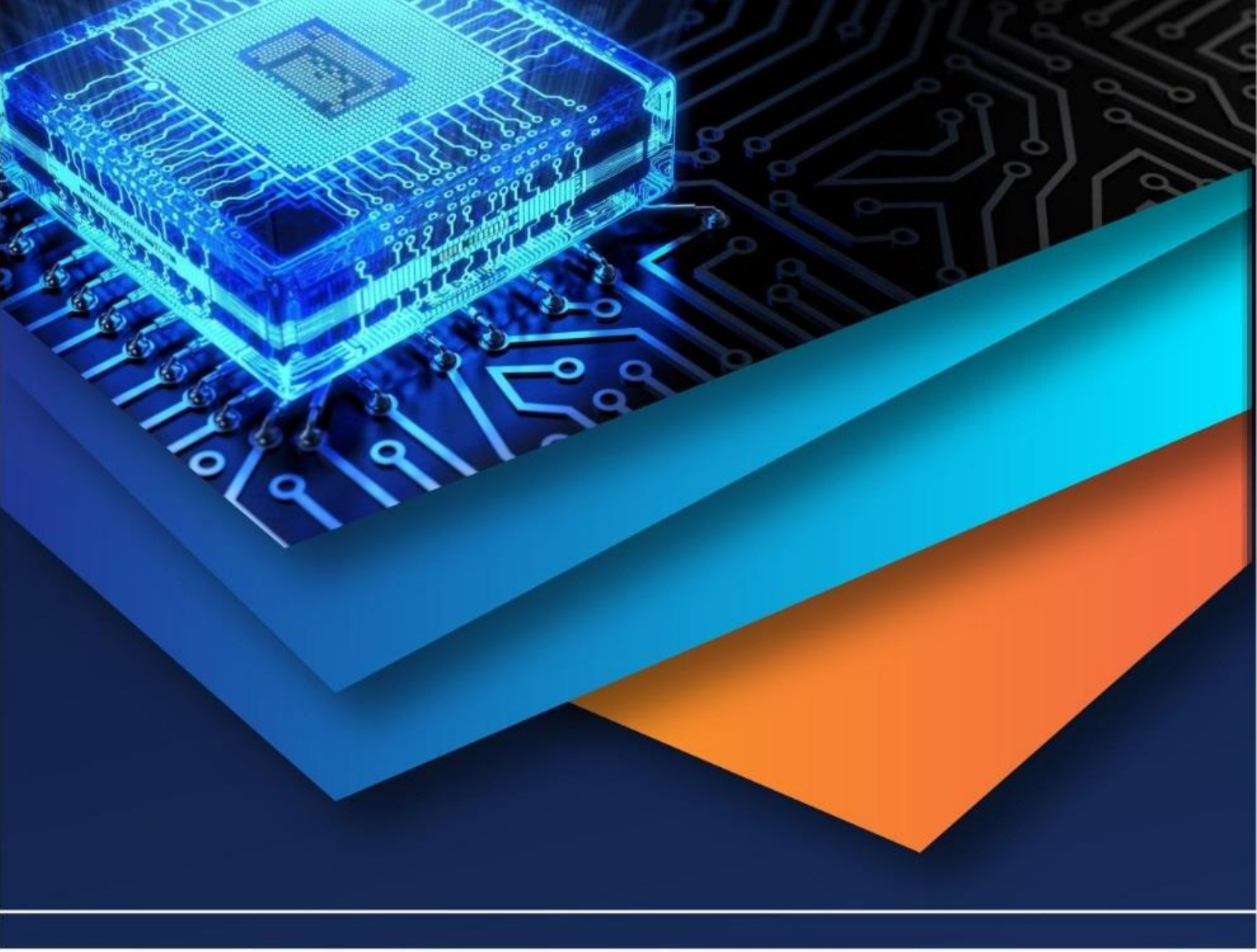
None of the PS methods clearly outperformed logistic regression, except for relatively small sample sizes. PS matching exhibited slightly inferior performance compared to all other methods, particularly with very low total sample sizes.

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:10.1 177/0272989X13498825.
[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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY