Journal of Clinical Epidemiology 152 (2022) 176−184

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study

Jack D. Wilkinson[a,*], Mamas A. Mamas[b], Evangelos Kontopantelis[c]

[a]*Centre for Biostatistics, Manchester Academic Health Science Centre, Faculty of Biology, Medicine, and Health, University of Manchester, Rm 1.307 Jean McFarlane Building, University Place, Oxford Road, Manchester M13 9PL, England*

[b]*Keele Cardiovascular Research Group, Centre for Prognosis Research, Keele University, Keele, England*

[c]*Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, England*

Accepted 13 September 2022; Published online 17 September 2022

Abstract

Objectives: In observational studies, researchers must select a method to control for confounding. Options include propensity score (PS) methods and regression. It remains unclear how dataset characteristics (size, overlap in PSs, and exposure prevalence) influence the relative performance of the methods.

Study Design and Setting: A simulation study to evaluate the role of dataset characteristics on the performance of PS methods, compared to logistic regression, for estimating a marginal odds ratio was conducted. Dataset size, overlap in PSs, and exposure prevalence were varied.

Results: Regression showed poor coverage for small sample sizes, but with large sample sizes was relatively robust to imbalance in PSs and low exposure prevalence. PS methods displayed suboptimal coverage as overlap in PSs decreased, which was exacerbated at larger sample sizes. Power of matching methods was particularly affected by a lack of overlap, low exposure prevalence, and small sample size. The advantage of regression for large data size was reduced in sensitivity analysis with a complementary log−log outcome generation mechanism and unmeasured confounding, with superior bias and error but inferior coverage to matching methods.

Conclusion: Dataset characteristics influence performance of methods for confounder adjustment. In many scenarios, regression may be the preferable option.

Keywords: Confounding; Propensity scores; Odds ratio; Marginal odds ratio; Regression standardization; Logistic regression; Simulation study

1. Introduction

Observational studies employing large, routinely collected datasets are now commonplace in the health sciences, exploiting new opportunities to study the effects of treatments or exposures in representative cohorts. A key concern in observational studies is how to address confounding, to permit the estimation of the effect of the exposure on an outcome [1,2]. Researchers frequently use regression or propensity scores (PSs) for this purpose, with the latter increasing in popularity in the past few decades [3].

The popularity of PS methods for observational health data can be attributed to several attractive features. They offer intuitive checks for balance between groups which are not possible using regression methods [2,4]. In addition, they can be formulated without reference to the outcome. This might reduce bias arising from ‘‘p-hacking’’ (when analyses are selected on the basis of the results they produce) [5] because the impact on the estimated treatment effect is not known during development of the PS model [2]. Regression methods implicitly but heavily rely on extrapolation when exposed and unexposed individuals have very different confounder distributions [6,7]. This is more explicit for PS methods because it manifests in the form of highly variable inverse probability weights or a lack of good matches. It is also appropriately reflected by reduced certainty in the estimated exposure effect [6,7]. A final reason may lie in the fact that PS methods, particularly

Declaration of interest: J.W. was supported by a Wellcome Institutional Strategic Support Fund Award [204796/Z/16/Z]. No other declarations of interest.

* Corresponding author. Centre for Biostatistics, Manchester Academic Health Science Centre, Faculty of Biology, Medicine, and Health, University of Manchester, Rm 1.307 Jean McFarlane Building, University Place, Oxford Road, Manchester, M13 9PL, England. Tel.: +44 0 161 306 8008.

E-mail address: jack.wilkinson@manchester.ac.uk (J.D. Wilkinson).

https://doi.org/10.1016/j.jclinepi.2022.09.009

What is new?

Key findings

- Key features of a dataset (size, exposure prevalence, and imbalance in propensity scores [PSs]) affect the performance of several approaches aiming to address confounding.

What this adds to what was known?

- Multiple logistic regression is relatively robust to low exposure prevalence and imbalance in the PS, except in very small samples ($N = 100$).
- For large sample sizes ($N = 10{,}000$ or more), multiple logistic regression performed better, whereas PS methods performed poorly as imbalance in PS distributions increased.
- Although in some unmeasured confounding scenarios the cumulative coverage and power performance were higher in the nearest neighbor and 1-to-1 PS matching, this was driven by much larger standard errors, and the absolute error and bias in the point estimate were lower with multiple regression.

What is the implication and what should change now?

- In large observation studies of national registries or primary care electronic health records, multiple regression estimation may often be the optimal choice in terms of simplicity and performance.

when PSs are used for matching, are frequently described as ''emulating'' a randomized controlled trial. However, it might not be clear to those using this phrase that the success of this emulation depends on all confounders being included in the estimation of the PS.

The variety of available methods for handling confounding in observational studies creates a challenge for the applied health researcher, who must select the best analytic approach for the particular study at hand. For example, PS matching excludes some data from the analysis, and so its performance might depend on factors such as the study size and the proportion of individuals who are exposed. In addition, PS methods were developed in an era preceding the widespread availability of large health datasets and evaluations of their performance have generally not considered large sample sizes. Evidence is therefore lacking on the relative performance of these methods for the analysis of big data. Previous studies have compared the results obtained by applying different methods [8] or have used simulation without investigating the impact of dataset characteristics on performance [9,10]. We therefore conducted a comprehensive simulation study to evaluate commonly used approaches for confounding control and to investigate the factors affecting their performance, with the aim of providing guidance for health researchers. We considered the roles of data size, imbalance in confounding variables, and the relative number of exposed to unexposed individuals.

2. Methods

Methodological details are provided in the supplementary file.

2.1. Propensity score

In a comparison of an exposed with a control group, PS can be estimated using multiple logistic regression where the binary outcome denotes membership of an exposed or a comparator group. Covariates hypothesized to be associated with the outcome should be included in the multiple logistic regression model. The PS is the predicted probability p of exposure. In practice, the regression coefficients must be estimated and hence there is uncertainty in p that is not usually accounted for in the process (although it is possible to do so, e.g., [7,11]). Following estimation, the next concern is to verify that these are balanced across the two groups [2]. Finally, PSs are incorporated into the analysis using one of several approaches (Supplementary File).

2.2. Simulation study

We conducted a simulation study to evaluate PS methods and covariate adjustment for confounding control in observational studies and the dataset characteristics affecting their performance.

2.2.1. Data generating model

To investigate the influence of data size, we considered sample sizes of 100, 1,000, 10,000, and 100,000 to capture scenarios in which large databases are available for analysis. We also investigated various scenarios for the distribution of the exposed and comparator groups: equal group sizes, imbalanced group sizes, and substantially imbalanced group sizes. A third varying parameter was baseline imbalance for the covariates, which took on five different patterns, ranging from well-overlapping PS to almost completely nonoverlapping PS for the two comparison groups. Figure 1 shows the PS distributions when equally exposed and comparator group sizes.

The simulation was implemented in Stata v 15.1 [12]. We used the drawnorm command to draw observations from multivariate normal distributions, which were dichotomized for some variables. The generated variables included binary exposure E, binary covariate X_1, and continuous covariates X_2, X_3, X_4, and X_5. Correlations were set to be low between all variables except for two of the covariates and the exposure. Following that, the outcome Y

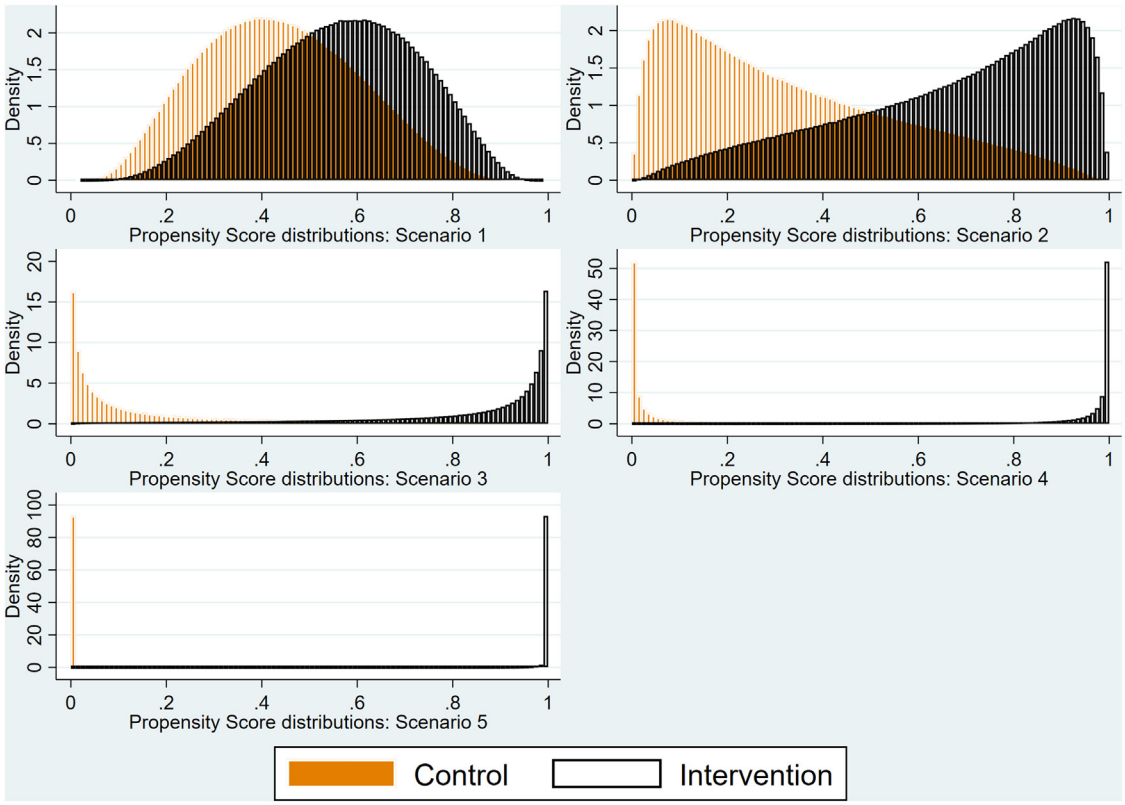

Fig. 1. Simulated propensity score scenarios, when Pr (E = 1) = 0.5.

was generated using a logit model. However, as a sensitivity analysis, we generated Y using a complementary log−log model, to ensure that performance of PS methods and regression was evaluated under more neutral conditions. In additional sensitivity analyses, we included an unmeasured confounder in the outcome generation mechanism (a continuous covariate X_6 which did not feature in the analytical models), for either the logit or complementary log−log model.

2.2.2. Analyses

A total of five analytical approaches were evaluated. First, PSs were estimated using logistic regression with exposure as the outcome variable [13]. Next, the PSs were used in four logistic regression models: (1) exposure and the PS as independent variables (PS covariate); (2) exposure as the only independent variable, with the number of times each observation appeared in the aforementioned nearest-neighbour−matched dataset as a frequency weight (nearest neighbor matching); (3) exposure as the only independent variable, following one-to-one matching without replacement, when absolute difference on the PS was less than 10^{-2} (Caliper matching); and (4) exposure as the only independent variable and the PS used as an inverse probability treatment weight (IPTW). Note that in this study, we use standard logistic regression following matching rather than a version intended for matched data; we return to this point in the discussion. We also performed logistic regression with the exposure and all five covariates included as independent variables (not using the PS), followed by regression standardization, as a fifth approach [14]. Standardization is necessary so that regression targets the same quantity as PS approaches (Target of inference and Supplementary File). We used the margins command with the postoption following logistic regression to achieve this and used the delta method to compute confidence intervals on the log odds scale.

2.2.3. Target of inference

We evaluated the methods against the marginal odds ratio, which is a measure of the exposure effect at the population level. We calculated the marginal odds ratio for each simulated dataset, using the method described by Austin [9]. When there is no heterogeneity in treatment effect, caliper and nearest neighbor matching, IPTW, and multiple logistic regression with standardization, all estimate the marginal odds ratio [9,14]. PS covariate actually targets a different quantity; the odds ratio conditional on the PS [7,15,16]. We include it here due to its popularity and to compare to other methods.

2.2.4. Performance measures

One thousand datasets were simulated for each scenario. We considered four performance measures: mean absolute error, bias, coverage, and power. Mean absolute error is the mean of the absolute difference between the estimate

and the true parameter. Bias is the mean difference between the estimate and the true parameter. Coverage is the proportion of 95% confidence intervals for the estimate, based on a normal approximation, that contains the true parameter. Finally, we calculated power by the proportion of iterations where the null was rejected when it was actually false. Although power as a metric can be problematic in the presence of bias, it is essential for a complete comparison. However, to obtain a more meaningful metric, power-related statistical significance was calculated one-sided (i.e., statistically more than zero rather than statistically different). We also evaluated model convergence. The other metrics were only computed when convergence for a particular method in a simulation setting was 25% or more, otherwise they were set to missing.

3. Results

Figures 2−7 show the results of the main simulation study. Supplementary Tables S 1−3 give the numerical results, including standard errors for the performance metrics. The performance estimates were sufficiently precise. Results for the sensitivity analyses (neutral comparisons and introduction of an unmeasured confounder) are shown in S Figures 1−19. We describe the results for the main analysis below, noting where sensitivity analyses resulted in departures from the main study results.

3.1. Convergence

As expected, convergence of all methods was adversely affected by reduced exposure prevalence, decreasing overlap in PS, and reduced sample size (Figure 2). All approaches converged infrequently at smaller sample sizes when there was little overlap in PS; IPTW and multiple logistic regression were most robust. With n = 100,000, these two methods generally converged even when the PS distributions were not overlapping (scenario 5) and exposure prevalence was very low (5%), although use of a complementary log−log link for outcome generation adversely affected this behavior (Supplementary File). Convergence for PS covariate was particularly affected by confounding (a lack of PS overlap); convergence was actually reduced when there was little overlap for larger (n = 100,000) compared to smaller (n = 1,000, n = 10,000) sample sizes. This was also observed when comparing datasets of n = 100,000 to n = 10,000 for caliper matching and nearest neighbor matching when exposure prevalence was low (10%) or very low (5%).

3.2. Bias and absolute error

Bias and mean absolute error were consistently low for multiple logistic regression compared to other methods (Figs. 3 and 4), although IPTW was less biased for n = 100, when exposure prevalence was very low (5%) and there was an overlap in PS distributions. Both measures were affected by sample size, with bias and/or error in the presence of nonoverlapping PS distributions actually becoming more pronounced with increasing data size for some methods. IPTW in particular had high bias and error when overlap was low and sample sizes were large. Caliper matching was consistently better than nearest neighbor matching. Despite targeting a different estimate, PS

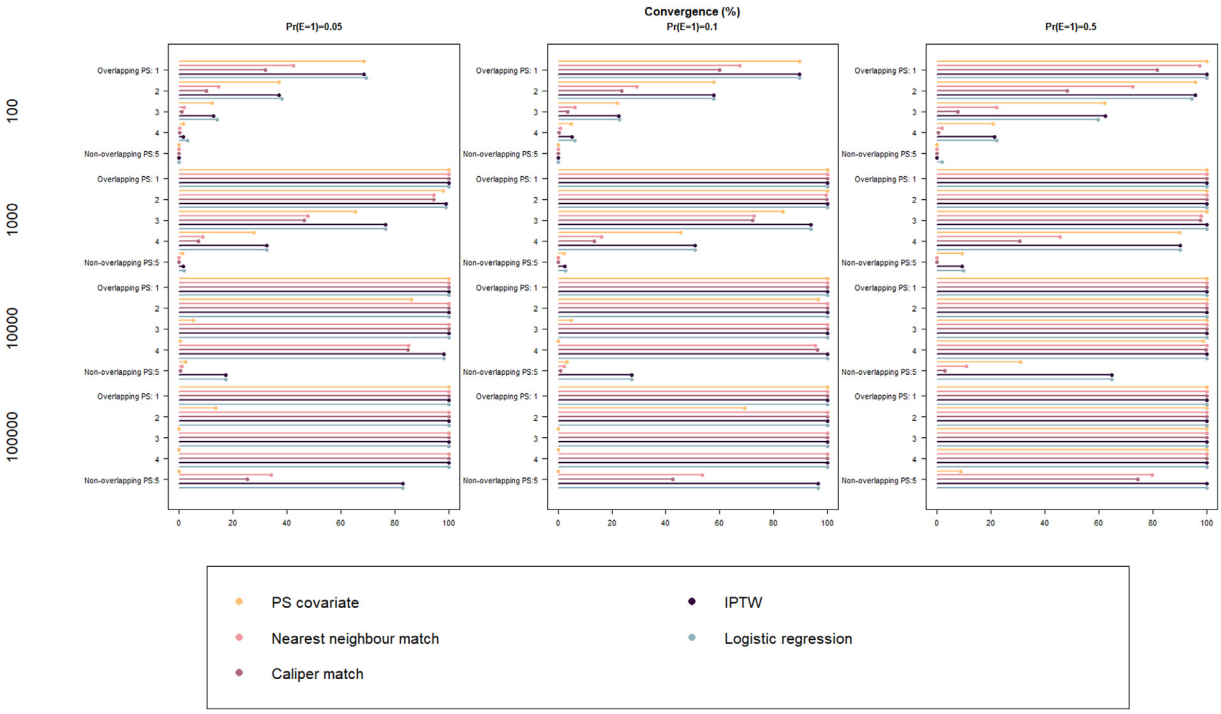

Fig. 2. Convergence (%).

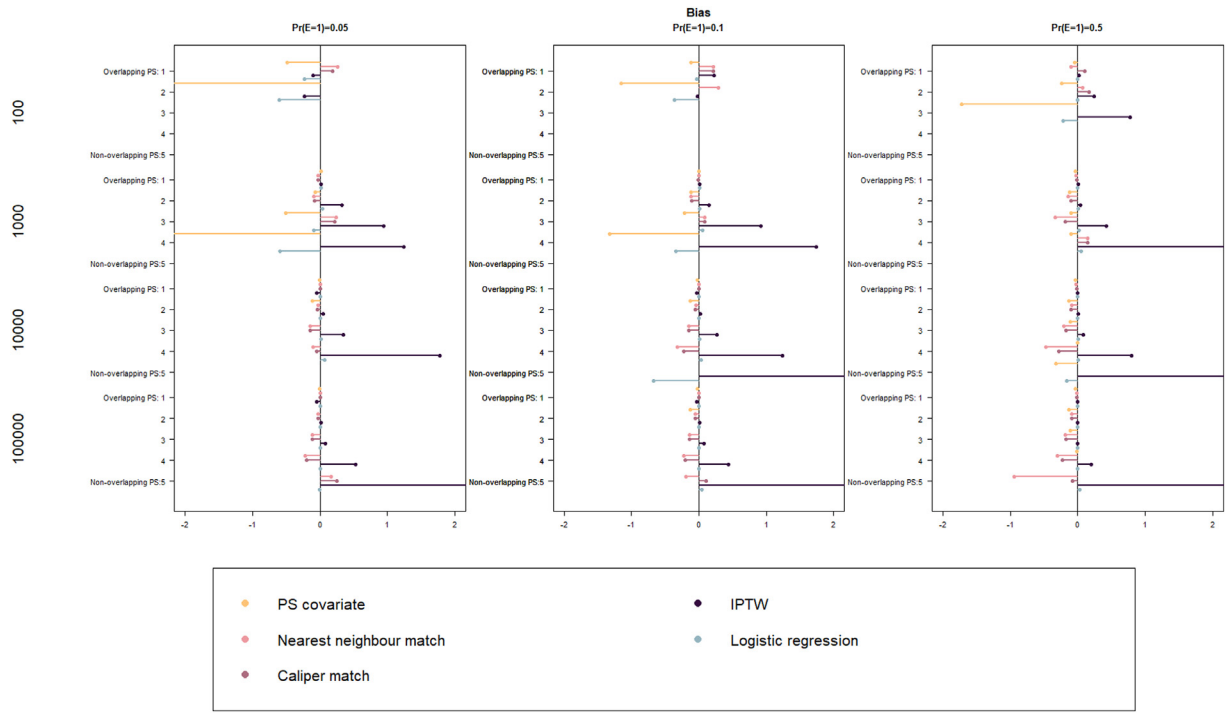

Fig. 3. Bias.

covariate fared relatively well, although performance broke down under challenging circumstances (combinations of low exposure prevalence, small sample size, and little overlap in PS).

3.3. *Power and coverage*

In the main scenario, power was generally as high or higher than other methods for multiple logistic

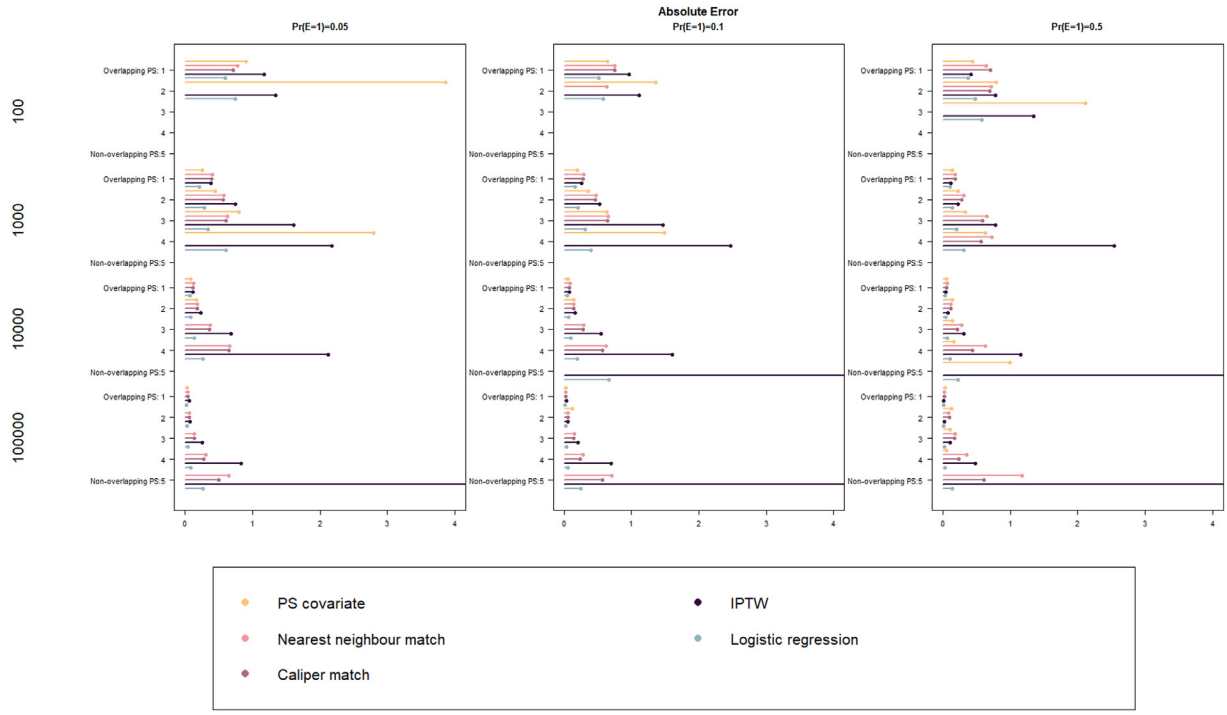

Fig. 4. Mean absolute error.

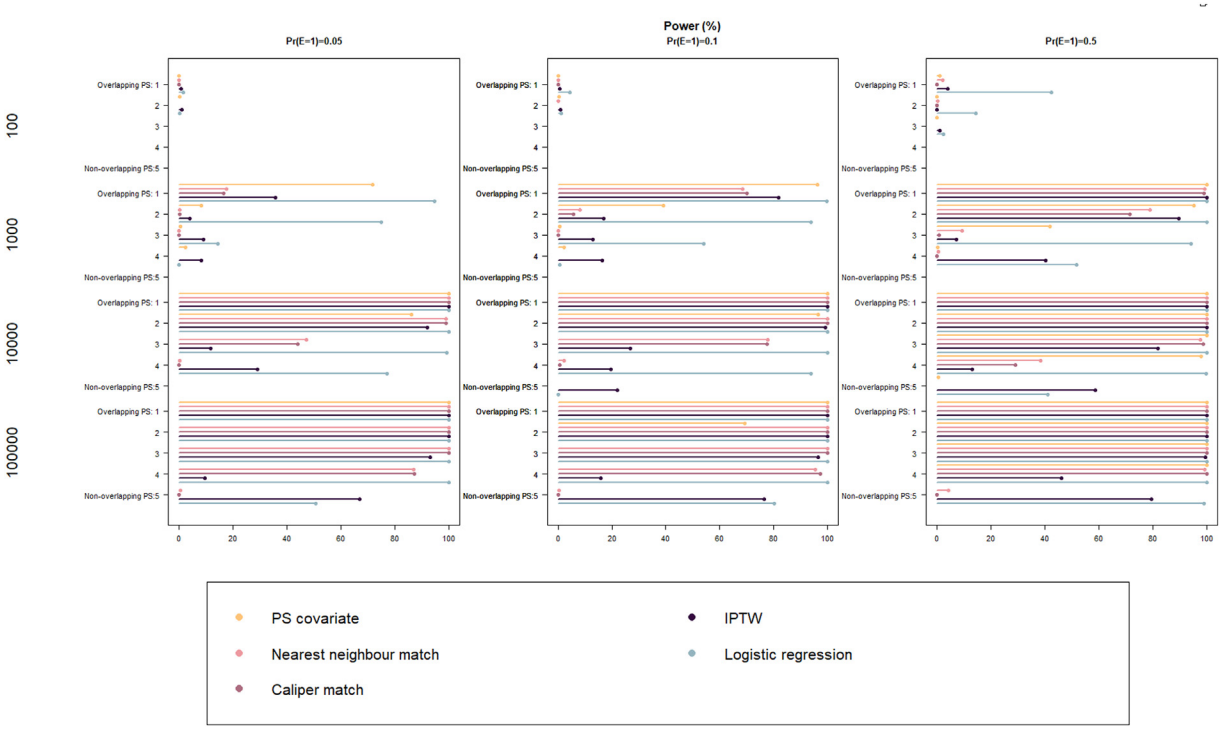

Fig. 5. Power (%).

regression, although IPTW had higher power in several scenarios where PS distributions were nonoverlapping (Figure 5). Coverage of IPTW was generally poor in these scenarios, however (Figure 6), and performance was consistently inferior when both power and coverage were considered (Figure 7). Power of matching methods was greatly affected by a lack of overlap in PS distributions. For caliper matching without replacement, when

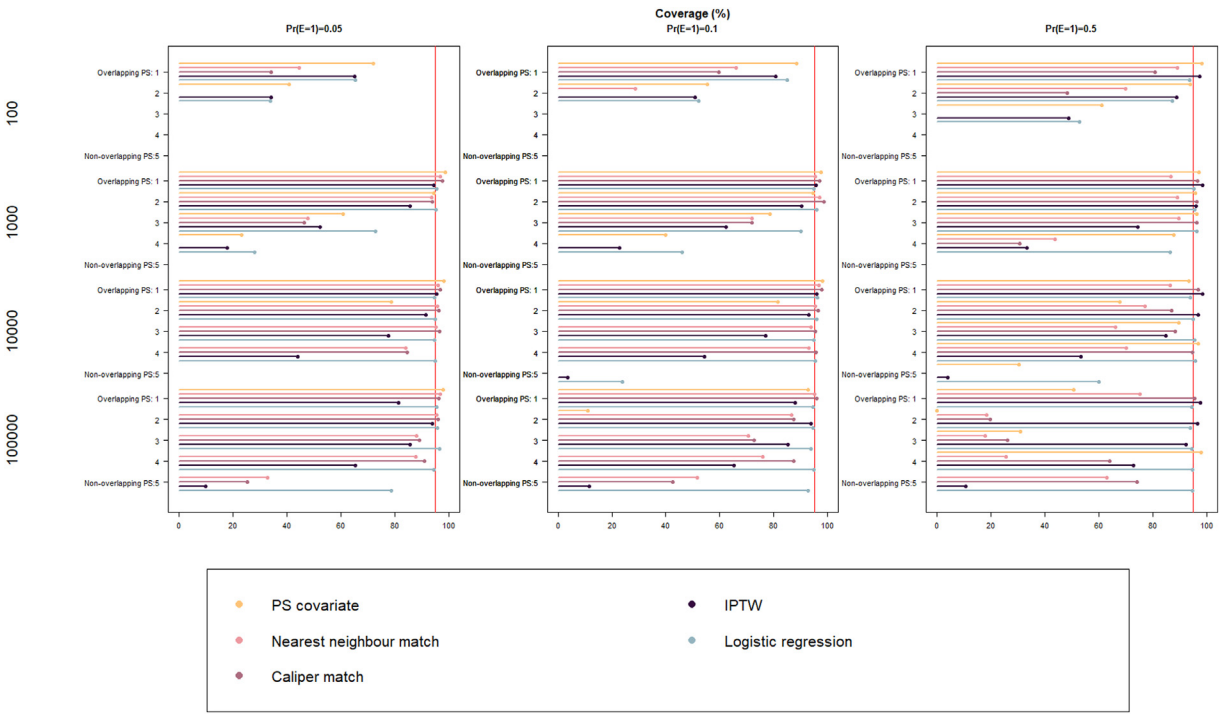

Fig. 6. Coverage (%).

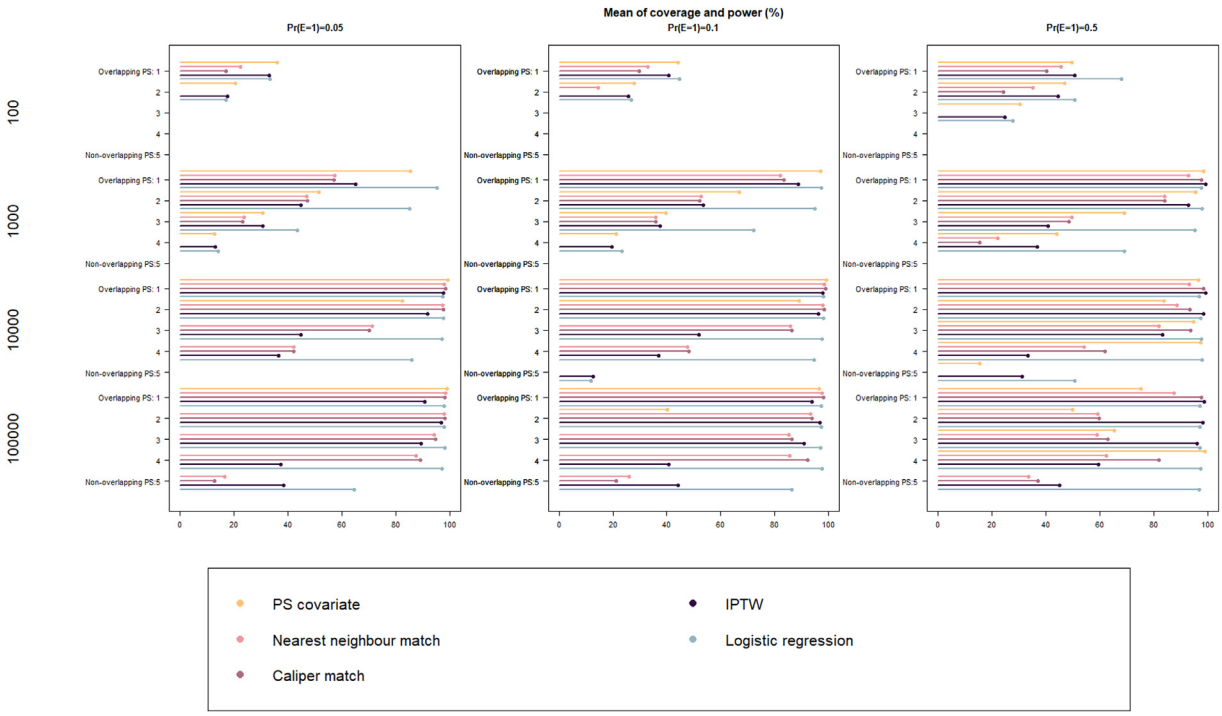

Fig. 7. Mean of coverage and power (%).

there is substantial imbalance, there will tend to be few matches. Consequently, power when there was reduced overlap was sometimes superior for nearest neighbor matching. In addition, sample size following matching is smaller when exposure prevalence is low, and this also affected power for matching compared to other methods, particularly when sample sizes were small to start off with. Coverage was decreased at $n = 100{,}000$ for the matching methods compared to sample sizes of 1,000 and 10,000 when there was imbalance in PS. Power of PS covariate compared to other PS methods was only favorable when there was a considerable overlap in PS distributions or 50% exposure prevalence; coverage was frequently but not consistently superior. However, when a complementary log–log link was used to generate the outcome, coverage was sometimes inferior for logistic regression compared to 1-to-1 and nearest neighbor PS matching, specifically with larger data size and modest or high imbalance in PS. This was exacerbated when an unmeasured confounder was added to the complementary log–log link scenario. Power tended to be poor for matching methods in these scenarios (both with and without unmeasured confounding). Overall, when considering power and coverage as a composite, 1-to-1 and nearest neighbor matching appeared superior to regression for large data size when a complementary log–log model was used to generate the outcome and there was unmeasured confounding. However, logistic regression remained superior in both absolute error and bias across all these scenarios.

4. Discussion

Multiple logistic regression followed by standardization was consistently superior to PS approaches, although coverage still fell short of the advertised level for very small sample sizes ($n = 100$) or for sample sizes of 1,000 when there was a limited overlap in PS distributions. It was observed to be quite robust to imbalance in PS for large sample sizes (10,000 or more), even when exposure prevalence was very low. Findings were broadly consistent in the three sensitivity analyses, with the exception of coverage for some scenarios, although that was driven by inflated standard error for PS methods. We discuss this extensively in the Supplementary File.

Coverage of PS methods was frequently suboptimal, as has been previously observed for a non-null marginal odds ratio [9]. As anticipated, relative performance of PS methods depended on dataset characteristics. Ahead of the study, it was anticipated that matching using PS might be particularly affected by small sample size, imbalance in the PS distributions, and low exposure prevalence. Although power was affected by these factors, overall performance of matching methods withstood these challenges better than IPTW. Matching methods also converged more frequently than did adjusting for the PS as a covariate in the presence of imbalance in PS distributions and large data sizes (10,000 or more). Caliper matching was slightly preferred to nearest neighbor matching overall, although nearest neighbor did achieve superior power and coverage in some scenarios.

IPTW displayed poor performance as an overlap in PS decreased. This suggests that IPTW is only suitable when there is a substantial overlap in PS, highlighting the importance of examining distributions of the estimated PS [2,17]. Vandersteedt and Daniels similarly found poor performance of IPTW when there was a limited overlap in PS, up to a sample size of 1,000; our results show that sample sizes considerably larger than this do not alleviate the problems [7]. However, we did not consider the use of stabilized or truncated weights [18], and it is unclear whether these would have led to improved performance. We have also not considered methods accounting for uncertainty in the PS estimation in this study.

PS covariate does not estimate the marginal odds ratio but instead targets a different and arguably unusual quantity—the odds ratio conditional on the PS. Performance against the marginal odds ratio was generally unacceptable, frequently falling short of the advertised coverage level when there was no high overlap in PS distributions and when there were small numbers of exposed participants, and for large data sizes (10,000 or more), balance did not guarantee appropriate coverage. PS covariate did not usually converge for larger data sizes when there was less overlap in PS distribution.

Although we have considered the role of dataset characteristics in selecting a method for controlling confounding, there are several outstanding questions. One question is whether a paired or unpaired version of regression should be used after PS matching. There is uncertainty around this point because people matched on their PS may nonetheless differ in terms of their covariate values. This question has been addressed in relation to continuous outcomes [19], but it remains to settle the issue in relation to binary outcomes. Suboptimal coverage against a non-null marginal odds ratio has been previously observed for several methods for analysing paired data following PS matching [9], and in the present study we also found this to hold when using an unpaired regression method. A direct comparison would be useful for future work.

By design, we did not include covariates in the regression models incorporating the PS (as a covariate, with IPTW, or following matching). Including covariates in the PS covariate approach results in a doubly robust estimator, which offers valid inference in relation to some summary measures of the exposure effect (other than odds ratios) provided that one of the PS model and outcome model are correctly specified [20,21]. Aside from this protection against misspecification, Vansteelandt and Davies found some power advantages when additionally adjusting for covariates in the outcome regression model compared to adjusting for the PS alone [7]. Consistent benefits of adjusting for covariates when using IPTW were not observed.

We have not considered the case where there is heterogeneity in the exposure effect across strata defined by the PS. When there is heterogeneity, different PS methods target different quantities and might produce substantially different answers as a result [22,23]. For example, matching estimates the exposure effect in the population corresponding to those who were exposed in the study because the matching process produces a sample with similar PS distributions to the exposed group [23] (or rather, for caliper matching, similar to the exposed participants for whom unexposed matches could be found [22]). These considerations motivate the suggestion that the possibility of an interaction between the PS and exposure should be routinely examined [24].

The study has some other limitations. For example, we have not considered data sizes larger than 100,000, and so we are extrapolating by supposing our results would apply to even larger datasets. Although we have considered several data-generating mechanisms, they are all fairly simple, which could have plausibly favored regression. We have also not considered scenarios with large numbers of confounders. PS methods are likely to be more useful here, particularly when the number of outcome events is relatively low. We have also not considered possible variations or alternatives to PS matching, such as use of different calipers, or coarsened exact matching [25]. Finally, we note that the implications of the current results for estimating alternative effect measures, such as the risk difference, warrant consideration.

5. Conclusion

Researchers analysing observational data often face difficult analytical choices, whereas PS approaches are not easy to implement in large databases of electronic health records. Our results show how key features of a dataset (size, exposure prevalence, and imbalance in PS) affect the performance of several approaches aiming to address confounding. This study suggests that multiple logistic regression is relatively robust to low exposure prevalence and imbalance in PS, outside of very small sample sizes. For large sample sizes, multiple logistic regression was clearly the preferred method, especially in the main scenario, whereas PS methods performed poorly as imbalance in PS distributions increased, and this was not mitigated by large sample size or balanced group sizes. This highlights the importance of examining overlap in PS if these methods are to be used but also suggests that their performance is worst when the problem they are intended to solve is most severe. Coverage of logistic regression was inferior to 1-to-1 and nearest neighbor PS matching methods in some large-data scenarios, but what was driven by much larger standard errors in these two matching approaches, whereas logistic regression remained the best performer in mean absolute error and bias. In large observational studies, multiple regression estimation appears to be the optimal choice, both in terms of simplicity and performance.

CRediT authorship contribution statement

All authors conceived the idea for the manuscript. E.K. and J.W. wrote code to undertake the simulation study. All authors critically interpreted the results, coauthored and revised the manuscript, and approved the submitted version.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.09.009.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41−55.
[2] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399−424.
[3] Sturmer 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;59:437−47.
[4] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757−63.
[5] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359−66.
[6] Tan Z. Comment: understanding OR, PS and DR. Stat Sci 2007;22:560−8.
[7] Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. Stat Med 2014;33:4053−72.
[8] Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, et al. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. J Am Coll Cardiol 2017;69:345−57.
[9] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007;26:3078−94.
[10] Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009;29:661−77.
[11] Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat Med 2014;33:721−37.
[12] StataCorp. In: Stata Statistical Software: Release 15. 1st ed.15 College Station, TX: StataCorp LLC; 2015.
[13] Leuven E, Sianesi B. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing 2017.
[14] Sjolander A. Regression standardization with the R package stdReg. Eur J Epidemiol 2016;31:563−74.
[15] Wan F, Mitra N. An evaluation of bias in propensity score-adjusted non-linear regression models. Stat Methods Med Res 2018;27:846−62.
[16] Wan F. An interpretation of the properties of the propensity score in the regression framework. Commun Stat - Theor Methods 2019;50:2096−105.
[17] Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015;34:3661−79.
[18] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656−64.
[19] Wan F. Matched or unmatched analyses with propensity-score-matched data? Stat Med 2019;38:289−300.
[20] Van der Laan M, Robins JM. Unified methods for censored longitudinal data and causality. New York, NY: Springer; 2003.
[21] Tan Z. A distributional approach for causal inference using propensity scores. J Am Stat Assoc 2006;101:1619−37.
[22] Lunt M, Solomon D, Rothman K, Glynn R, Hyrich K, Symmons DP, et al. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. Am J Epidemiol 2009;169:909−17.
[23] Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006;163:262−70.
[24] Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98:253−9.
[25] Iacus SM, King G, Porro G. Causal inference without balance checking: coarsened exact matching. Polit Anal 2012;20:1−24.