ELSEVIER

Journal of Clinical Epidemiology ■ (2016) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Propensity score model overfitting led to inflated variance of estimated odds ratios

Tibor Schuster[a,b,c,d,*], Wilfrid Kouokam Lowe[a,b,e], Robert W. Platt[b,f]

[a]*Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, 3755 Chemin de la Côte-Sainte-Catherine, Montréal, Québec H3T 1E2, Canada*

[b]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Avenue West, Montréal, Québec H3A 1A2, Canada*

[c]*Clinical Epidemiology and Biostatistics Unit and the Melbourne Children's Trial Centre, Murdoch Childrens Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia*

[d]*Department of Paediatrics, University of Melbourne, Melbourne, Victoria 3010, Australia*

[e]*UFR de Mathématique et d'Informatique, Université de Strasbourg, 7 Rue René Descartes, 67084 Strasbourg, France*

[f]*Department of Pediatrics, McGill University, Montreal Children's Hospital, 1001 Décarie Boulevard, Montreal, Québec H4A 3J1, Canada*

Accepted 11 May 2016; Published online xxxx

Abstract

Objective: Simulation studies suggest that the ratio of the number of events to the number of estimated parameters in a logistic regression model should be not less than 10 or 20 to 1 to achieve reliable effect estimates. Applications of propensity score approaches for confounding control in practice, however, do often not consider these recommendations.

Study Design and Setting: We conducted extensive Monte Carlo and plasmode simulation studies to investigate the impact of propensity score model overfitting on the performance in estimating conditional and marginal odds ratios using different established propensity score inference approaches. We assessed estimate accuracy and precision as well as associated type I error and type II error rates in testing the null hypothesis of no exposure effect.

Results: For all inference approaches considered, our simulation study revealed considerably inflated standard errors of effect estimates when using overfitted propensity score models. Overfitting did not considerably affect type I error rates for most inference approaches. However, because of residual confounding, estimation performance and type I error probabilities were unsatisfactory when using propensity score quintile adjustment.

Conclusion: Overfitting of propensity score models should be avoided to obtain reliable estimates of treatment or exposure effects in individual studies. © 2016 Elsevier Inc. All rights reserved.

Keywords: Propensity score; Logistic regression; Overfitting; Confounder adjustment; Odds ratio; Inverse probability weighting

1. Introduction

Observational studies are frequently used to estimate treatment or exposure effects in settings where the assignment of subjects into intervention or exposure groups is not under control of the study investigator. A major shortcoming of such studies is that treatment preference or the status of exposure is often linked to individual characteristics that are not independent of the outcome of interest. Therefore, comparison groups may differ in their covariate distributions in ways that will confound the results regarding estimated treatment or exposure effects on the outcome.

Propensity scores can be used to aggregate information about the predictive role of covariates on treatment assignment or exposure status. Formally, the propensity score is

Funding: T.S. and W.K.L. were supported by the Canadian Network for Observational Drug Effect Studies (CNODES). CNODES is funded by a grant from Health Canada, the Drug Safety and Effectiveness Network (DSEN), and the Canadian Institutes of Health Research (CIHR) grant number 111845-1. T.S. was also supported by funding from the Royal Children's Hospital Foundation to the Melbourne Children's Trial Centre. Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program. R.W.P. is supported in part by a National Scholar (Chercheur-national) of the Fonds de Recherche du Quebec—Sante (FQR-S) and is a member of the Research Institute of the McGill University Health Center, which is supported by core funds from FQR-S.

* Corresponding author. Tel.: +61-3-9936-6097; fax: +61-3-9348-1391.

E-mail address: Tibor.Schuster@mcri.edu.au (T. Schuster).

http://dx.doi.org/10.1016/j.jclinepi.2016.05.017

0895-4356/© 2016 Elsevier Inc. All rights reserved.

ARTICLE IN PRESS

What is new?

Key findings

- Overfitting of propensity score models leads to inflation of the variance of effect estimates when applying established conditional and marginal inference methods that use propensity scores for confounder adjustment.

What this adds to what was known?

- Consequently, estimate uncertainty obtained in an individual study can annul alleged unbiasedness due to confounding control if the number of exposed or unexposed individuals per propensity score predictor variable is low.
- Conventional propensity score quintile adjustment is less effective in confounding control than conditioning on propensity score spline functions or using inverse probability of treatment (exposure) weighting.

What is the implication and what should change now?

- We recommend that specification of propensity score models should acknowledge widely accepted guidelines for regression model building to avoid overfitting.
- We discourage the use of propensity score quintile adjustment in favor of modeling propensity score spline functions or using inverse probability of treatment (exposure) weighting.

the probability of receiving treatment (or experiencing a certain exposure status) given individual covariate realizations [1]. There are different ways to use propensity scores to address confounding such as matching based on the propensity score, stratification according to propensity score intervals, ordinary propensity score adjustment in the context of a multivariable binary logistic regression analysis, and performing weighted effect estimation (inverse probability of treatment weighting) in the framework of marginal structural models [2,3].

Because propensity score modeling is undertaken to aggregate multivariate covariate information into a single variable, propensity methods are particularly popular when estimating treatment or exposure effects on rare outcomes using data sets with a large number of potential confounding variables. Binary logistic regression is the most common model used to estimate propensity scores. Previous simulation studies have shown that the number of events relative to the number of parameters in the logistic model should exceed a ratio between 10 or 20 to 1 to avoid inflated standard errors of the parameter estimates [4−6]. Further simulation-based investigations have demonstrated that this rule may be relaxed in sensitivity analyses to demonstrate adequate control of confounding [7].

Although there is an ongoing debate and controversy in the literature about correct propensity score model specification, only limited research has been undertaken yet to systematically investigate the role of overfitting logistic propensity score models that are incorporated in different conditional and marginal inference approaches [8−13]. Available simulation studies on the number of variables included in the propensity score did not directly consider the ratio of number of exposed or treated individuals to propensity score predictor variables and were based on real data without knowledge of the true effect of treatment on the outcome [14].

In fact, there is a wide-spread perception that the propensity score is meant to be only descriptive for the data in hand but not to be generalizable to other data sets [15]. We investigate within this article whether inaccurate estimation of the propensity score due to model overfitting leads to considerable bias or inflated variance of estimated effect parameters.

The article is structured as follows: in Sections 2 and 3, we describe the designs of comprehensive Monte Carlo and plasmode simulation studies that investigate to which extent overfitting of propensity score models leads to systematically and randomly erroneous effect estimates. In Section 4, we report the resulting bias, root mean square error, as well as type I and type II error rates in testing the null hypothesis of no treatment effect. Section 5 closes with the discussion of the results and conclusions.

2. Monte Carlo simulation setup

2.1. General data scenario and inference methods to be compared

We consider the scenario of a point-exposure study investigating the effect of a binary treatment E on a dichotomous outcome Y. Within this study, a binary logistic regression model (the propensity score model) is used to estimate every study individual's probability of receiving treatment given the realizations of a prespecified set of covariates $X_1, \ldots, X_k$. The respective propensity score is then used in different ways to account for potential confounding when estimating the conditional or marginal odds ratio as effect parameter. In particular, we consider the following effect estimation approaches within our study: (A) multivariable logistic regression for the binary outcome to estimate the conditional treatment effect (log odds ratio) under adjustment for the entire set of covariates, (B) multivariable logistic regression for the binary outcome to estimate the treatment effect conditioning on binary variables that indicate an individual's membership to one of the quintile-based partitions of the estimated propensity score

distribution, (C) multivariable logistic regression for the binary outcome to estimate the conditional treatment effect under adjustment for the estimated propensity score using a cubic b-spline transformation, (D) inverse probability of treatment weighting (IPW) without weight stabilization, and (E) IPW using weight stabilization by rescaling the individual weights with the marginal prevalence of an individual's treatment realization [2].

The rationale to include propensity score spline functions in our simulation studies was motivated by recent methodological developments in the context of regression modeling using continuous predictor variables such as propensity scores [16–18]. The authors conclude that PS spline functions performed best due to their semiparametric or nonparametric nature which reduces the problem of potential misspecification of the covariate function form.

The use of quintile-based partitions of the propensity score followed the approach for subclassification as implemented in the original propensity score article [1].

2.2. Factorial design, assumed data distributions, and parameter configurations

We used a full-factorial simulation setting that incorporated the following data-generating parameters: the ratio of the number of treated or exposed individuals to the number of covariates in the propensity score model $r \in \{5, 10, 15, 20, 25, 30, 40, 50\}$, the number of variables included in the propensity score model $k \in \{10, 20, 50\}$, the marginal prevalence of treatment or exposure $P(E=1) \in \{0.125, 0.25, 0.5\}$, the marginal prevalence of the binary outcome $P(Y=1) \in \{0.125, 0.25, 0.5\}$, and the assumed underlying treatment effects (odds ratios) $OR \in \{1.0, 1.25, 1.5\}$. We note that the underlying treatment effects are assumed to be full conditional effects, that is, effects that result after adjusting for the entire set of covariates $X_1, \ldots, X_k$. This is important to emphasize because the chosen effect measure (odds ratio) is not collapsible and may take different values for different covariate adjustment sets even in the absence of confounding [19].

Individual covariate values $x_i = (x_{i1}, \ldots, x_{ik})^T$ were sampled from independent standard normal distributions. A binary logistic model was used to generate the individual probabilities for receiving treatment $P(E=1|x_i) = \exp(-\beta_{0E} - x_i^T \beta_E)^{-1}$, that is, the propensity score model. Subsequently, the probabilities for the occurrence of the outcome $P(Y=1|x_i, E) = \exp(-\beta_{0Y} - x_i^T \beta_Y - E \cdot \ln[\mathrm{OR}])^{-1}$ were generated. The treatment status variable E and the binary outcome variable Y were sampled from Bernoulli distributions which used $P(E=1|x_i)$ and $P(Y=1|x_i, E)$ as respective distribution parameters.

The magnitudes of the model coefficients for the covariates in both conditional probability models were set to be uniform and proportional to the numbers of covariates in the model. The proportionality constraint was specified to ensure consistency in terms of the mean predicted values of the different models. Let $\beta_{\mathrm{E,j}} \equiv b_{\mathrm{E}}$ and $\beta_{Y\mathrm{j}} \equiv b_Y$ denote the model coefficients (log odds ratios) of the j^{th} covariate in the treatment model and the outcome model, respectively. The coefficient values b_E and b_Y were set to 0.5 for $k = 10$ covariates, 0.25 for $k = 20$ covariates, and 0.1 for $k = 50$ covariates ($b_E = b_Y = 5/k$).

The coefficients b_E and b_Y ultimately drive the magnitude of confounding induced by a covariate. Irrespective of the signs of the coefficients, larger absolute values of either parameter will inevitably lead to an increased bias factor associated with the respective covariate. However, if coefficient signs are allowed to vary, positive and negative bias induced by two covariates could lead to a dilution or cancellation of the marginal bias. With increasing number of covariates, such dilutions or cancellations become more likely. This can make the simulation of data sets with confounding bias inefficient and cumbersome. Setting both coefficients to equal values reduced the complexity of the simulation design and allowed for a straightforward simulation of desired bias magnitudes for the unadjusted effect estimate.

Because covariates were sampled independently from standard normal distributions, the variance of the linear predictor of the respective propensity score model (the logit of the propensity score) revealed to be $\sigma^2 = k \cdot b_E^2$, specifically 2.5, 1.25 and 0.5 for $k = 10$, 20, and 50. The intercept parameters β_{0E} and β_{0Y} of the two logistic models were specified to reflect the presumed prevalence of treatment and the outcome. Therefore, the mean of the linear predictor for a propensity score model was given by $\mu = \ln\{P(E=1)/[1 - P(E=1)]\}$, specifically -1.95, -1.10, and 0 for the three desired prevalence conditions. The density curves for the nine resulting propensity score distributions are shown in Fig. 1. The chosen parameterizations led to three different baseline bias scenarios if no covariate adjustment would be considered: relative bias magnitudes (bias factors on the odds ratio scale) of approximately 1.2, 1.5, and 2.

2.3. Simulation repetitions and results processing

Each data simulation scenario was repeatedly generated 1,000 times. For each of the generated data sets, the five previously described estimation approaches (A–E) were applied to retrieve the respective conditional or marginal treatment effect estimates (log odds ratios).

Because the variety of considered treatment prevalence and covariate settings would occasionally lead to convergence problems in effect estimation, we excluded simulation runs where at least one estimation approach would generate an absolute log odds ratio estimate larger than $|\ln(100)|$. This was done to avoid distortion of simulation results (i.e., mean and variance estimates) by unrealistic extreme outliers.

To estimate the empirical bias, we calculated for each single effect estimate the deviation to the true underlying effect parameter value (on the log odds ratio scale) and averaged this deviation over all simulation runs. In the two inference settings which used inverse probability of

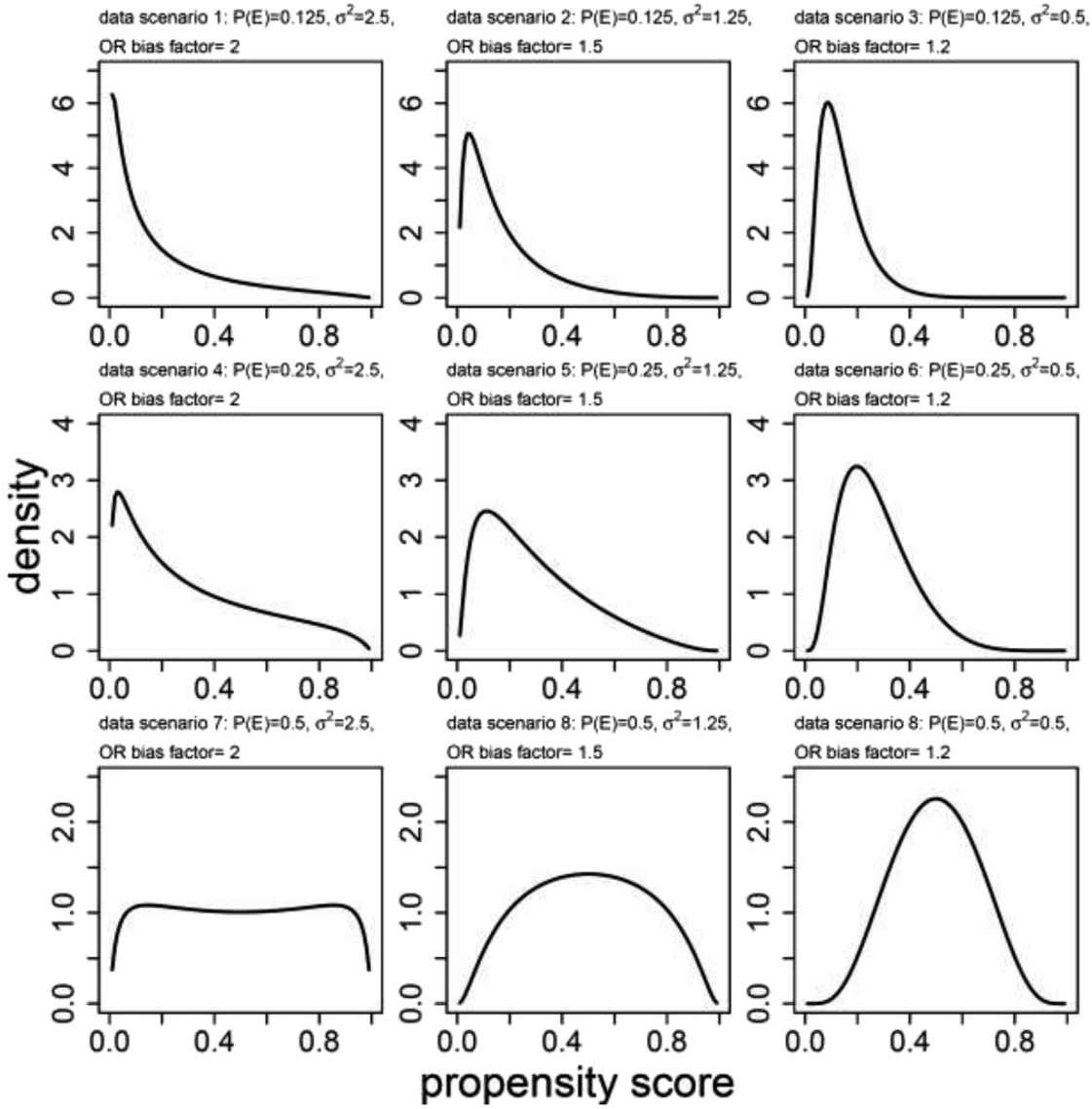

Fig. 1. Density functions of the nine propensity score distributions considered in the Monte Carlo simulation study. OR, odds ratio.

treatment weighting (D and E), the true underlying marginal treatment effect was calculated based on two full conditional (including all covariates $\boldsymbol{X}$) models for the outcome: one model in which the treatment indicator variable was set to one for a large number of individuals ($n = 100{,}000$) and one model in which the treatment variable was set to zero for the same number of individuals. The underlying marginal odds ratio was subsequently calculated as the ratio of the odds of the mean predicted probabilities retrieved from these two models.

Additional to the empirical bias, we calculated empirical standard errors as well as root mean square errors for the different effect estimates. The estimated type I and type II error probabilities were calculated based on the proportion of rejections of the null hypothesis of no treatment effect using a two-sided level of significance of 0.05. Test statistics for the marginal effect estimates (approaches D and E) were computed using robust (''sandwich'') variance estimates to account for the individual clustering induced by the weighting.

All analyses were performed using R software (R version 3.1.3 (2015-03-09)—''Smooth Sidewalk'') [20].

Simulation runs with nonconvergence warnings reported by the software routine (glm function in R, positive convergence tolerance set to a value of 1e−06, maximum number of iterations set to 25) were substituted by convergent runs which were sampled from the same data scenario. Nonconvergence of effect estimates is a result of either (quasi-)completely data separation or multicollinearity and occurs under specific data configurations [21,22].

3. Plasmode simulation study

Because simulation studies have limited capacity to emulate real-world data sets with complex correlation and

effect structures, we conducted a so-called plasmode simulation study [23].

The plasmode approach resamples from an original data set without affecting the correlation structure among covariates. Based on the covariate matrix and originally estimated regression coefficients, different prediction models for exposure and outcome variables can be specified. These prediction models are then used to generate new exposure and outcome variables with desired characteristics such as exposure or outcome prevalence and conditional exposure effects on the outcome.

We used a publicly available data set on critically ill patients who whether or not received right heart catheterization (rch) during the first 24 hours of care in an intensive care unit [24]. The data set has been described before and propensity score models were used in the original data analysis [25]. A total of 2,184 individuals (38%) received rch (the exposure of interest) and 3,722 individuals (65%) died within 30 days (binary outcome of interest). We estimated the coefficients $\beta_{E,j}$ and $\beta_{Y,j}$ from the data set using 49 prespecified variables (66 estimated parameters in the propensity score model for rch). Modified propensity score models were then used to generate new exposure variables with different desired exposure prevalences. Alteration of exposure prevalence was achieved by changing the intercept of the propensity score model while keeping all other model coefficients at their original value.

The observed covariates in conjunction with the originally estimated coefficients and the generated exposure variable were then used as linear predictor in a logistic model to simulate a new outcome variable. This approach allowed for an investigation how propensity score overfitting can affect effect estimates, and related standard errors in settings were (1) covariates are likely not to be independent and (2) magnitudes and signs of the coefficients $\beta_{E,j}$ and $\beta_{Y,j}$ vary.

We investigated the effect of propensity score overfitting on effect estimates and standard errors by varying the ratio of rch-treated to covariates from 1:1 to 50:1 with 100 repeated samples per setting.

We have provided detailed R code for the plasmode simulation study as Supplementary Material/Appendix at www.jclinepi.com.

4. Results

4.1. Simulation study results

4.1.1. Prevalence of extreme estimation inaccuracy despite estimate convergence

The proportion of extreme log odds ratio estimates as defined by $|\ln \mathrm{OR}| > \ln(100)$ was less than 0.0004 over all simulation runs. Such estimates occurred although simulation runs with nonconvergence warnings reported by the software routine were substituted by convergent runs. For the purpose of sensitivity analysis, we considered two even more conservative estimate thresholds $\ln(20)$ and $\ln(10)$ for the exclusion of simulation runs. The results did not measurably differ between the three choices of thresholds.

4.1.2. Estimation accuracy

The distributions of deviations of effect estimates from the true parameter values, marginal over all simulation settings, are displayed as boxplots in Fig. 2. The boxplot whiskers depict the respective 0.025 and 0.975 distribution quantiles. The x-axes represent the ratio (r) of the number of treated or exposed individuals to the number of covariates in the propensity score model. The panels show the results for the five different underlying inference approaches (A–E).

The results of the inverse probability weighting approaches D and E were essentially the same: scaling of the weights with a constant factor, the prevalence of treatment or exposure received, had not impact on the effect estimation. We therefore collapsed the results of the inference approaches D and E in this and the following representations to one graph.

The relative bias over all unconditional simulation settings averaged to a factor of 1.5 (bias factor on the odds ratio scale). Marginally over the nine propensity score distribution scenarios, all approaches except propensity score quintile adjustment (approach B) were able to diminish this bias to a negligible level even for low r values. Propensity score quintile adjustment yielded an average residual bias of about 5%. The relative bias showed to be only slightly affected by the true underlying effect size.

Across all estimation approaches, the accuracy in effect estimation demonstrated to be strongly associated with the number of treated or exposed individuals per variable included in the propensity score model. For r values less or equal to 10, the lower and upper 2.5% percentiles of relative deviations to the true effect parameter (odds ratio scale) ranged from 1.75 to 3.5.

Adjustment for a cubic b-spline transformation of the propensity score (approach C) demonstrated the best estimation performance across all scenarios. For r values of 30 and above, all inference approaches except PS quintile adjustment, performed similarly well.

A more detailed investigation of the relative accuracy in effect estimation was performed by separating the results according to the nine different underlying propensity score distributions (Appendix Figs. 2A–C at www.jclinepi.com). The relative performance of the inference approaches A–E was comparable with the marginal results depicted in Fig. 2. The prevalence of exposure and the variance of the linear predictor of the propensity score showed a strong impact on estimation performance. The inaccuracy in effect estimation was maximal in settings with prevalence of exposure closer to 0.5, high linear predictor variance, and using a conventional multivariable regression model for the outcome (approach A).

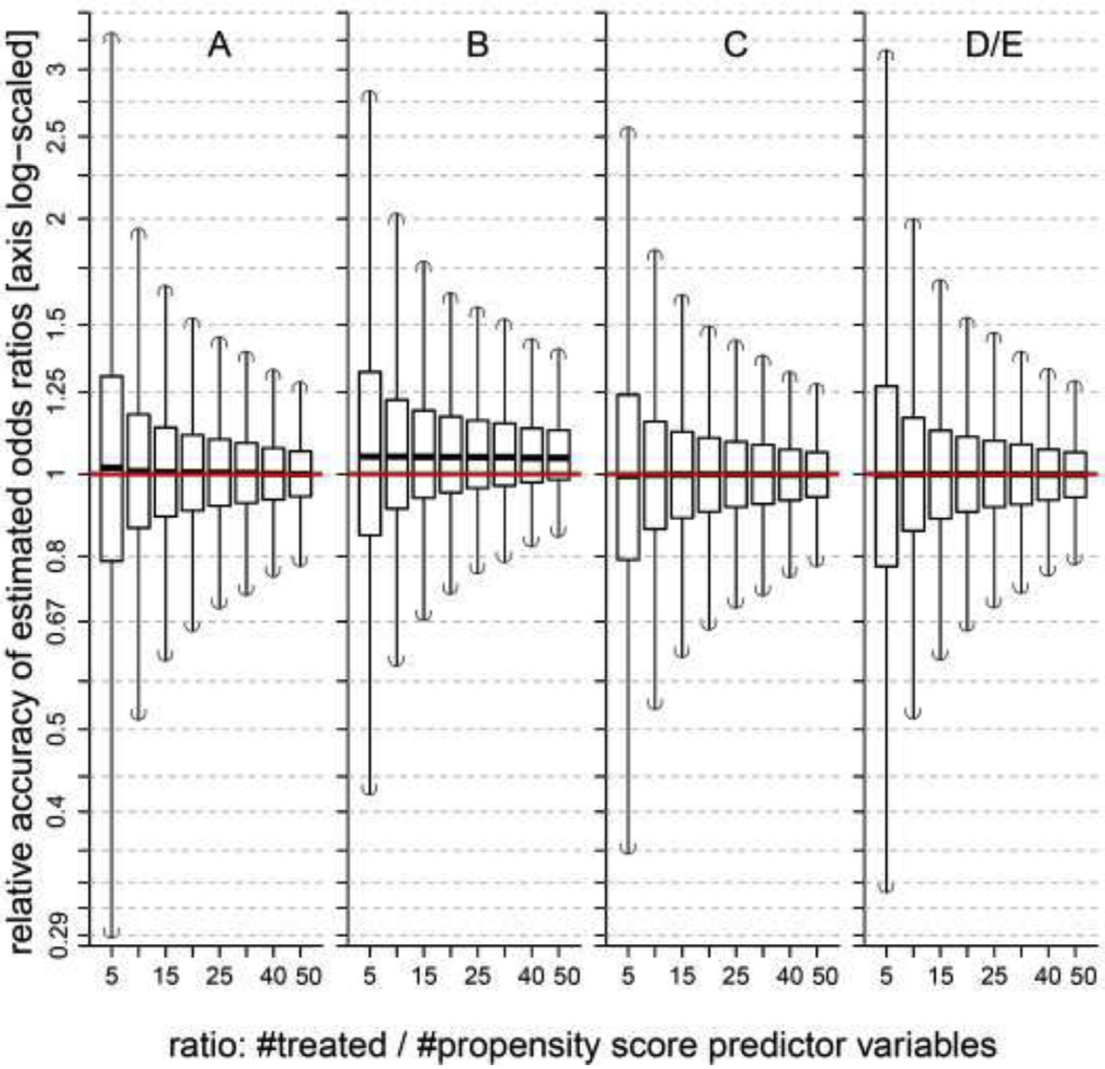

Fig. 2. Relative accuracy of estimated odds ratios, marginal over all data simulation configurations (n = 1,000 repeated samples). The lower and upper boxplot whiskers represent the 0.025 and 0.975 distribution quantiles. (A) Multivariable logistic regression for the binary outcome including the binary treatment indicator variable and all confounding variables (correctly specified outcome model). (B) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on propensity score quintiles (correctly specified propensity score model). (C) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on cubic b-spline transformation of the propensity score (correctly specified propensity score model). (D and E) Logistic regression for the binary outcome including the binary treatment indicator variable and using (stabilized) inverse probability of treatment weights (correctly specified propensity score model).

4.1.3. Empirical root mean square error

The empirically estimated marginal root mean square errors (RMSEs) for all five effect estimation approaches are displayed in the first row of Fig. 3. For large values of r, all methods performed similarly with slightly worse RMSE values yielded by approach B (propensity score quintile adjustment). For small values of r, conventional multivariable covariate adjustment showed the highest average RMSE values followed by the two inverse probability weighting approaches (D and E). All marginal RMSE curves showed a steep slope for the lower range of r values ($r \leq 20$), notably attenuating to more acceptable levels of estimate imprecision for r values of 30 and above.

Similar RMSE curve patterns were observed with each of the nine propensity score distribution scenarios (Appendix Figs. 3A–C at www.jclinepi.com). Naturally, the average magnitude of the RMSE strongly depended on the variance of the linear predictor of the propensity score as well as on the prevalence of treatment. In particular, data scenarios with higher densities for propensity score values close to 0.5 showed larger RMSE values and also steeper slopes of the RMSE curves in dependence on the r values.

We note that the RMSE is measured on the log odds ratio scale. The exponential transformation of the RMSE, however, can be directly converted to the relative inaccuracy (due to bias and variance) of an estimate to be

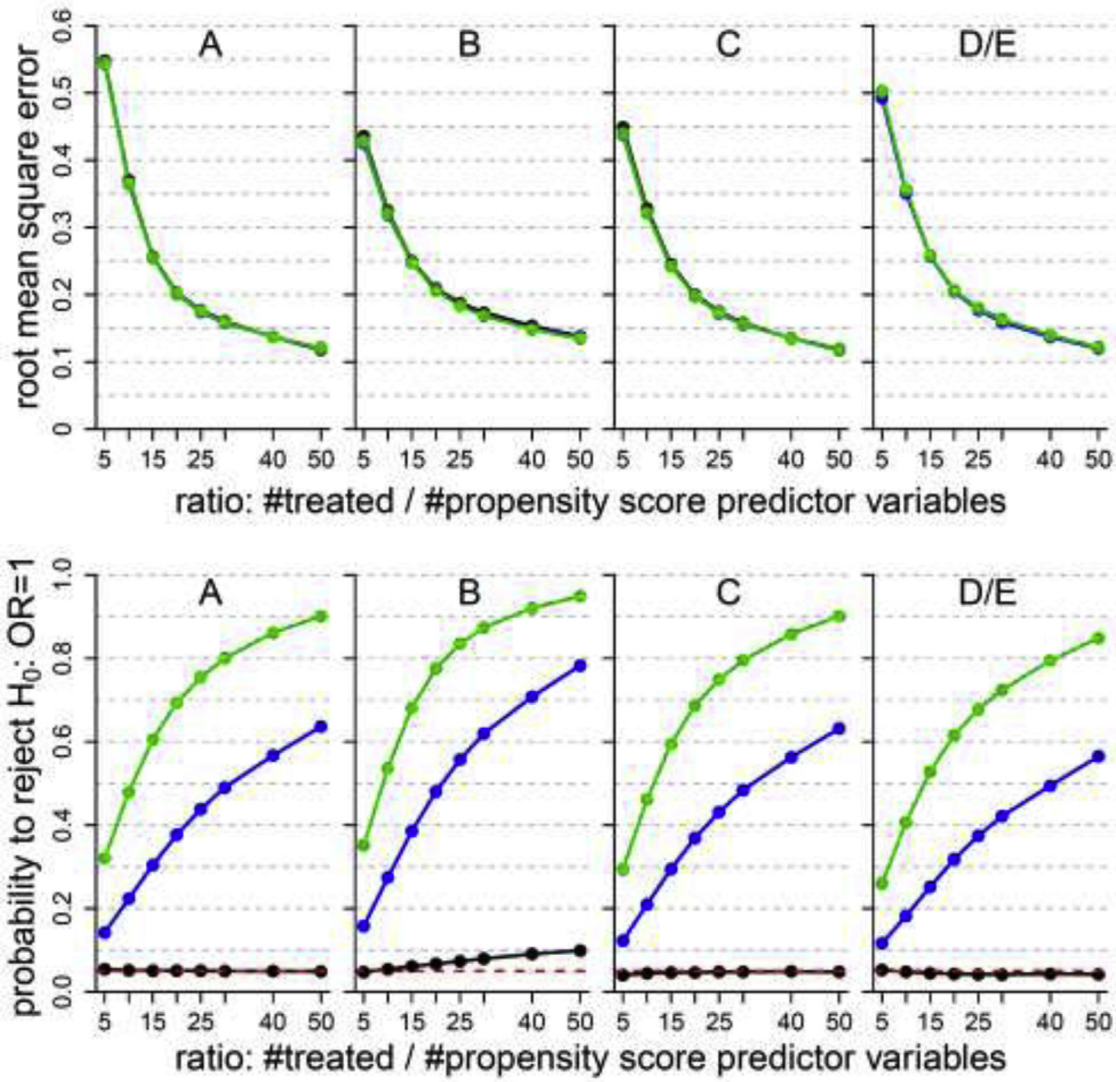

Fig. 3. First row: empirical root mean square error in estimating log odds ratios, marginal over all data simulation configurations (n = 1,000 repeated samples). Second row: empirical power functions for testing the null hypothesis of no treatment effect, marginal over all data simulation configurations (n = 1,000 repeated samples). Underlying treatment effects (odds ratios): 1 (black lines), 1.25 (blue lines), 1.5 (green lines). (A) Multivariable logistic regression for the binary outcome including the binary treatment indicator variable and all confounding variables (correctly specified outcome model). (B) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on propensity score quintiles (correctly specified propensity score model). (C) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on cubic b-spline transformation of the propensity score (correctly specified propensity score model). (D and E) Logistic regression for the binary outcome including the binary treatment indicator variable and using (stabilized) inverse probability of treatment weights (correctly specified propensity score model). OR, odds ratio. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

expected in a single study of a given sample size. RMSE values of 0.2 and 0.4, for example, translate to relative estimate uncertainty of ±22% and ±49% on the original odds ratio scale.

4.1.4. Empirical type I and type II error probabilities

The second row in Fig. 3 displays the marginal empirical type I error rates (black lines) and type II error rates (colored lines) for the statistical tests of no treatment effect retrieved from the five evaluated inference approaches. The dashed red line indicates the desired 5% level of falsewise rejections of the two-sided null hypothesis.

On average, over all considered data simulation scenarios, only propensity score quintile adjustment failed to maintain the predefined type I error level. This finding is not surprising considering the previously explored bias magnitudes in effect estimation associated with this approach. With increasing sample size for a given covariate set in the propensity score, the standard error decreases so that the residual bias is more likely to lead to a rejection of the null hypothesis of no treatment effect. For this reason, a

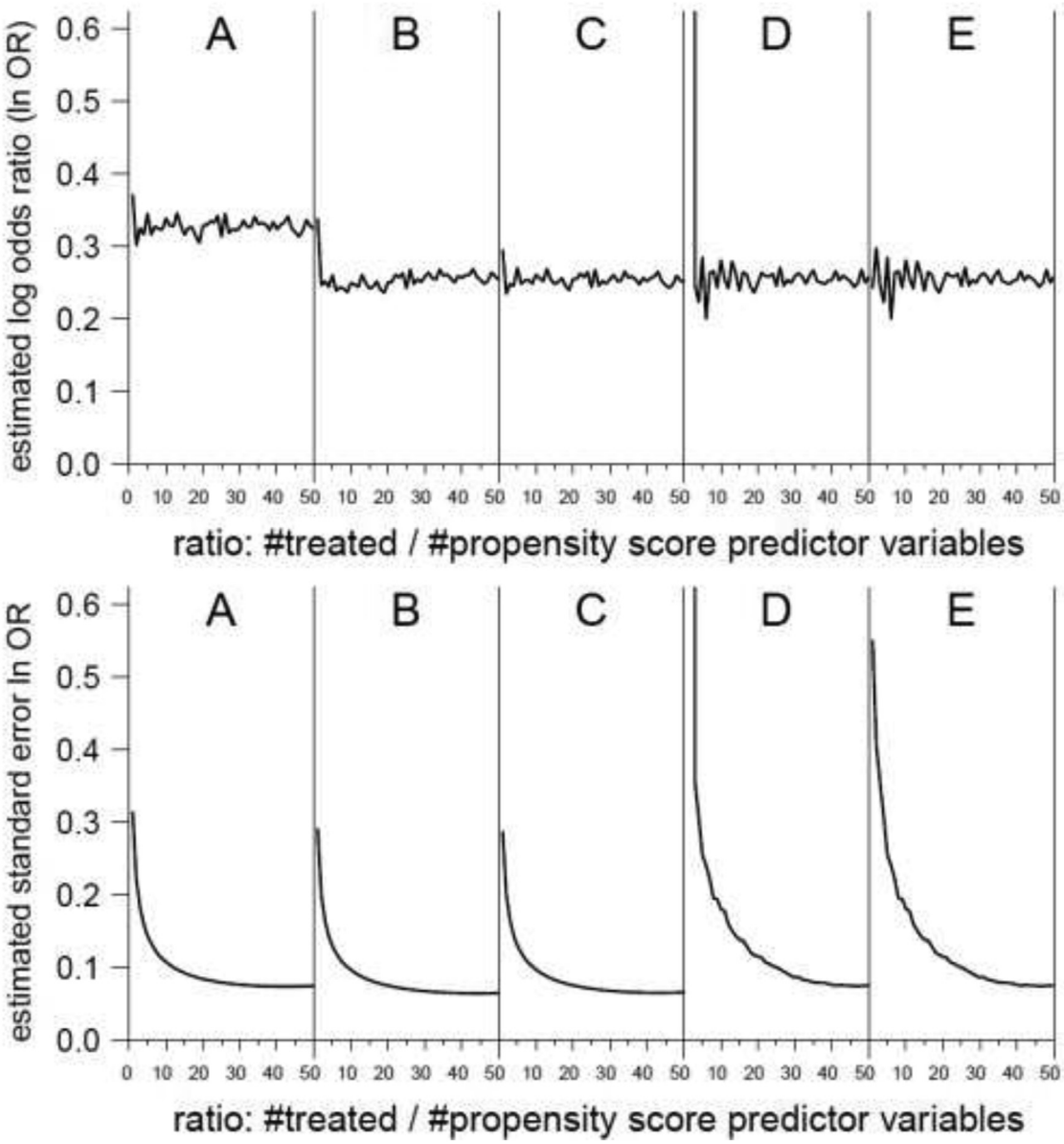

Fig. 4. First row: mean log odds ratio based on plasmode simulation using real data on critically ill patients and the effect of right heart catheterization on 30-day hospital mortality (n = 100 repeated samples for each r value). Second row: mean standard errors for the log odds ratios. (A) Multivariable logistic regression for the binary outcome including the binary treatment indicator variable and 49 prespecified covariates. (B) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on propensity score quintiles. (C) Logistic regression for the binary outcome including the binary treatment indicator variable and conditioning on cubic b-spline transformation of the propensity score. (D and E) Logistic regression for the binary outcome including the binary treatment indicator variable and using unstabilized/stabilized inverse probability of treatment weights.

slight increment of the empirical type I error probability can be observed for larger r values.

In terms of statistical power, the inverse probability of treatment weighting approaches (D and E) performed slightly worse compared to the remaining inference procedures. Conventional multivariable covariate adjustment (approach A) and adjustment for a propensity score b-spline approximation (approach C) performed equally well. The apparently higher average power functions for the propensity score quintile adjustment method were expected because of inherent bias of this estimation procedure.

For the nine propensity score distribution scenarios, similar discrepancies between the power functions of the different estimation procedures resulted as compared to the overall evaluation (Appendix Figs. 3D–F at www.jclinepi.com). Notably, the relative frequency of false rejections of the null hypothesis of no treatment effect under approach B was large (up to 20%) in settings with low

prevalent treatment and high variance of the linear predictor of the propensity score.

4.2. Plasmode simulation study results

The estimated log odds ratios and standard errors from the plasmode simulation study are displayed in Fig. 4. The results are consistent with the results of the conventional simulation study. The effect estimates were not measurably affected by the ratio of exposed individuals to covariates in the propensity score model. However, the standard errors of the effect estimates were monotonously decreasing with larger numbers of treated individuals per covariate in the propensity score model. Effect estimates based on propensity score quintile adjustment (panel B, upper graph) were comparable to the other propensity score inference approaches. For very low r values, unstabilized inverse probability weighting (panels D) led to extreme average effect estimates. In general, both inverse probability weighting approaches showed stronger relative inefficiency compared to the other methods than in the conventional simulation settings.

5. Discussion

In this article, we presented extensive Monte Carlo and plasmode simulation studies to assess the impact of overfitting propensity score models in common inference approaches that use propensity scores for estimating conditional or marginal odds ratios as treatment or exposure effects on binary outcomes. The results of our simulation study suggest that overfitting of propensity score models can lead to inflated variance of effect estimates and therefore to estimation inaccuracy in situations where relatively many covariates are included in the propensity score model. Our simulation results show that inverse probability of treatment approaches can be particularly sensitive to propensity score overfitting. One likely explanation for this finding are practical (near) violations of the positivity assumption [26]: overfitting of a logistic regression model for treatment preference is likely to yield predicted probabilities close to zero or one, which then results in large weights and unstable variance of the IPW estimator. The standard sandwich variance estimator does not allow for the incorporation of the uncertainty in estimating the propensity score and is typically conservative [27]. This reason as well as the inefficiency of IPW estimators are explanations for the observation that the IPW approaches in our simulation study achieved lower power levels than the other methods.

We showed that the classical method of propensity score quintile adjustment may lead to considerable residual bias and therefore to severely inflated type I error rates. In the plasmode simulation study, however, effect estimates under propensity score quintile adjustment were comparable to the other inference approaches.

There are certain limitations to our study. In general, simulation studies can never capture the whole space of possible data scenarios as they occur in reality. We explored the statistical properties of different estimators for treatment or exposure effects on a binary outcome using both simulated and real-data scenarios. We implemented a variety of different propensity score distribution scenarios with different bias magnitudes and considered a wide range of simulation configurations including diverse treatment effects, various prevalence values for exposure and outcome, as well as varying absolute samples sizes.

The observed dependence of effect estimate variance on relative propensity score model complexity was consistent for all simulation configurations and all inference approaches considered. It is very unlikely that this steady observation is explained by a randomly favorable (or unfavorable) choice of parameter settings being considered in the simulation study.

In a recent commentary, concern has been raised that selection bias due to exclusion of nonconvergent result can play a dominant role in bias assessment [28]. Because our simulation study results did not indicate changing bias with increasing number of events per covariate, we conclude that possible bias due to selection is unlikely.

For several reasons, we did not consider matching on the propensity score as an additional method within our simulation study: effect estimates based on matched samples are subject to various restrictions introduced by the matching procedure. Design-related constraints such as caliper definitions, matching with or without replacement, as well as the ratio of matched cases to controls would make comparisons of the results with the other procedures complex and difficult. Similarly, we neither considered doubly robust estimators in our comparisons nor data scenarios that incorporated missing data issues.

The findings of our simulation study suggest that generally accepted recommendations on the maximum number of covariates to be included in a binary regression model should be applied with the same rigor to propensity score models that are used for control of confounding. Failure to do so may lead to undesirable inaccuracy of single study effect estimates due to inflation of estimate variance.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.05.017.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

ARTICLE IN PRESS

[2] 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.
[3] D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265−81.
[4] Cepeda S, Boston R, Farrar JT, Strom B. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:3.
[5] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.
[6] Harrell FE. Regression modeling strategies. New York: Springer Science & Business Media; 2001:61.
[7] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165:710−8.
[8] Rubin D, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249−64.
[9] Bryson A, Dorsett R, Purdon S. The use of propensity score matching in the evaluation of labour market policies. Working Paper 4. London, England: Dept. for Work and Pensions; 2002.
[10] Zhao Z. Sensitivity of propensity score methods to the specifications. Econ Lett 2008;98:309−19.
[11] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149−56.
[12] Clarke KA, Kenkel B, Rueda MR. Misspecification and the propensity score: the possibility of overadjustment, 2011. Available at https://www.rochester.edu/college/psc/clarke/MissProp.pdf. Accessed January 06, 2016.
[13] Millimet DL, Tchernis R. On the specification of propensity scores, with applications to the analysis of trade policies. J Bus Econ Stat 2009;27:397−415.
[14] Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol 2011;173:1404−13.
[15] Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B, Mukhopadhyay P. Variable selection and raking in propensity scoring. Stat Med 2007;26:1022−33.
[16] Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ Jr. Splines for trend analysis and continuous confounder control. Epidemiology 2011;22:874−5.
[17] Hade EM, Lu Bo. Bias associated with using the estimated propensity score as a regression covariate. Stat Med 2014;33:74−87.
[18] Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8:551−61.
[19] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat Sci 1999;14:29−46.
[20] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available at http://www.R-project.org/. Accessed September 05, 2016.
[21] Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984;71:1−10.
[22] Lesaffre E, Albert A. Partial separation in logistic discrimination. Journal of the Royal Statistical Society. Series B (Methodological) 1989;51:109−16.
[23] Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal 2014;72:219−26.
[24] Available at http://biostat.mc.vanderbilt.edu/DataSets. Accessed April 29, 2016.
[25] Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically III patients. JAMA 1996;276:889−97.
[26] Westreich D, Stephen RC. Invited commentary: positivity in practice. Am J Epidemiol 2010;171:674−7.
[27] 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.
[28] Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. J Clin Epidemiol 2011;64:1464−5.