Original Article

Adjusting for confounding in early post-launch settings: going beyond logistic regression models

Amand F Schmidt[a,b,c,d,*], Olaf H Klungel[a,b], Rolf H H Groenwold[a,b], on behalf of the GetReal Consortium.

[a.] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands.

[b.] Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, P.O. Box 80082, 3508 TB Utrecht, The Netherlands.

[c.] Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7, Utrecht 3584 CL, The Netherlands.

[d.] Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, The United Kingdom.

***** Corresponding author: Tel.: +44 20 3549 5625.

E-mail address: amand.schmidt@ucl.ac.uk (A.F.Schmidt).

Running title: Adjusting for confounding

Word count text: 3778

Word count abstract: 253

Number of references: 29

Number of tables: 3

Number of figures: 2

(Web)appendix: 1

Acknowledgement

The research leading to these results was conducted as part of the GetReal consortium. For further information please refer to www.imi-getreal.eu.

Conflict of interest statement

None of the authors of this paper has a financial or personal relationship with other people or organisations that could inappropriately influence or bias the content of the paper.

Author contributions

AFS, RHHG and OHK contributed to the idea and design of the study. AFS performed the analyses and drafted the manuscript. OHK, RHHG provided guidance during initial planning of the paper and during critical revision. AFS had full access to all of the data and takes responsibility for the integrity of the data presented.

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n° [115546], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution

Prior postings and presentations

This study and its results have not been previously published, neither has it been presented at conferences.

Summary

Background To adequately adjust for confounding multivariable models need to be correctly specified and include a sufficient number of subjects per estimated parameter.

Methods A simulation study was conducted which explored the performance of logistic regression (LR), propensity score (PS) analysis and disease risk score (DRS) methods to adjust for confounding. Events/exposed subjects per coefficient (EPC) was set to 10, 5, 2.5, 1.0 and 0.5. Model misspecification was induced by ignoring treatment and/or interaction effects in the DRS training data (i.e., independent dataset to develop the DRS model).

Results At low EPCs of 1.0 and 0.05, the LR estimates had a relative bias of more than 100%. Bias of the DRS estimates was at most 13.40% and 18.84%. For the PS model this was 8.80% and more than 100%, respectfully. Coverage of the LR estimates became less than the nominal level of 0.95 at an EPC of 5 (0.936). For the DRS and PS methods coverage became less than 0.95 at an ECP of 2.5 and 1.0, respectfully. Depending on the direction of the interaction effect relative to the main treatment effect, ignoring the interaction resulted in a bias of 16.94% for the DRS models.

Conclusion In settings with small events/exposed subjects per coefficient, DRS methods can be useful alternative to LR models, especially when PS models cannot be used. However, while in our simulations DRS estimates were the leased unbiased in low EPCs settings, coverage was below acceptable levels after EPC of 2.5 and always less than the more biased PS method.

Keywords: confounding, statistics, simulations study, logistic regression, propensity score, disease risk score.

Background

Nonrandomized studies on (pharmacological) therapeutics are often conducted to complement results from randomized clinical trials (RCTs). For example, nonrandomized studies might be more appropriate to assess the occurrence of rare, but severe, adverse events such as anaphylactic reactions [1-3]. Furthermore, nonrandomized studies are used to estimate the relative effectiveness in real-life clinical practice. Depending on the relationship between the intervention and the outcome, different degrees of confounding can be expected [1-3]. For example, after launch of a new drug it is expected that patients who responded poorly to older drugs cross-over to the new drug (i.e., channelling). In this case the crude association between treatment and outcome is very likely confounded[4].

Frequently, the outcome of interest is dichotomous, such as mortality, in which case multivariable logistic regression (LR) [5] is commonly used to adjust for confounding. One (of many) assumption(s), is that the associations between confounders and the outcome are sufficiently estimated to adjust for confounding bias. In settings (e.g., nonrandomized early post-launch studies) where both the number of events and the number of exposed subjects are small, controlling for confounding can be problematic. Further complicating the matter is that it is not uncommon to consider more than 100 potential confounders [6]. Simulation studies showed, that for prognostic LR models 10 or more events per coefficient (EPC) were needed to get unbiased estimates [7,8]. However, in prognostic studies, the interest lies in correctly estimating all associations between possible predictors and the outcome, whereas in nonrandomized therapeutic studies, the interest is usually in estimating a single association (i.e., the treatment outcome association), while adjusting for numerous potential confounders. Vittinghoff and McCulloch [9] showed that in this case LR models with EPC as small as 6 can adequately adjust for confounding.

In settings where LR models are expected to perform poorly (i.e., EPC smaller than 6), propensity score (PS) [10,11] and disease risk score (DRS) [12-16] methods can be applied to summarize the information of multiple confounders into a single variable. It seems logical that these methods require less events/exposures per coefficient. However, it remains unclear how many events/exposures per variable are needed to sufficiently control for confounding using PS and DRS methods. Furthermore, in training (i.e., developing) DRS models, it is often implicitly assumed that there is no treatment effect or no treatment by confounder interaction. How sensitive DRS models are to violations of these assumptions is unknown, particularly when the DRS model is trained in one dataset and applied in another. We therefore conducted a simulation study to compare LR, PS and four kinds of DRS models with varying amount of EPC and under different levels of model misspecification.

Methods

Simulation set-up

Following the examples given above, we focused on scenarios in which the effects of a new drug (or any other type of medical intervention) were evaluated early post-launch. In addition, pre-launch data on the comparator drug were considered to be available. In each simulation, a training dataset was generated, containing pre-launch information, as well as a test dataset, containing post-launch information. Each training dataset included 5000 subjects of whom approximately 2500 were exposed to the comparator drug C and 2500 to drug B. Approximately 2500 subjects experienced the event of interest. The test dataset included 400 subjects of whom, on average, 200 were exposed to comparator drug C and 200 to the new drug A. Approximately 200 subjects in the test data set experienced an event. The training data were used to train the DRS models. The test data were then used to compare the estimated effect of the intervention (drug A vs. C) obtained through the DRS, LR and PS methods.

Data-generating process

Data of the training and test datasets were generated using the same algorithm. First, j independent confounding variables Z were generated. Z_1 was sampled from a normal distribution with mean 3 and variance 1. The remaining Z_{j-1} variables were sampled from independent Bernoulli distributions, each with a success probability of 0.5. A subject's probability of treatment was given by the model:

$$logit(p_{i,treatment}) = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_j z_{ij} \quad [1]$$

$p_{i,treatment}$ indicates the probability of the ith individual to receive treatment. The value of α_0 was set so that on average 50% of the subjects were exposed. Please see Table 1 and the simulation scenarios section below for an overview parameter values used. For each ith individual the probability of experiencing an event was given by:

$$logit(p_{i,event}) = \delta_0 + \delta_1 x_i + \delta_2 z_{i1} + \cdots + \delta_j z_{ij} + \delta_{int} x_i z_{i1} \quad [2]$$

The intercept (δ_0) was chosen so that on average 50% of the subjects experienced an event. Depending on the value of δ_{int} there was an interaction between treatment and continuous confounder Z_1. The treatment and outcome states were then sampled from Bernoulli distributions:

$$x_i \sim Bernoulli(p_{i,treatment})$$

$$y_i \sim Bernoulli(p_{i,event})$$

Data analyses

The test data contained post launch information on subjects receiving new drug A (indicated by $X = 1$) or drug C (indicated by $X = 0$). To adjust for confounding in the association between treatment and the outcome the subsequently described methods were applied.

Logistic regression confounding adjustment

To adjust for confounding the following LR model was used:

$$logit\big(prob[y_i = 1|x_i, z_{ij}]\big) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_1 z_{i1} + \cdots + \hat{\beta}_j z_{ij} \qquad [3]$$

Where $\hat{\beta}_1$ is an estimate of the ln(odds ratio) of the association between treatment and outcome adjusted for confounders Z.

Propensity score analysis

An alternative to LR models is to first estimate the associations of the confounder with the treatment variable. As a second step, the (logit of the) predicted probability of treatment (i.e., the propensity score) can be used to control for confounding:

$$PS = logit\big(prob[x_i = 1|\, z_{ij}]\big) = \hat{\varphi}_0 + \hat{\varphi}_1 z_{i1} + \cdots + \hat{\varphi}_{j1} z_{ij} \qquad [4]$$

$$logit(prob[y_i = 1|x_i, ps_i]) = \hat{\omega}_0 + \hat{\omega}_1 x_i + \hat{\omega}_2 ps_i \qquad [5]$$

Here $\hat{\omega}_1$, can be interpreted as the ln(odds ratio) of the treatment outcome association adjusted for the confounders included in step 1. Because PS models regresses exposure on the

confounders, instead of the event, PS models is known to be able to adjust for more confounders when the treatment is more common than the outcome.

Disease risk score adjustment

Another approach to control for confounding is adjustment using a disease risk score (DRS). First, the associations between the confounders and outcome are estimated in a training dataset, using equation 3. In the second stage, these associations are used to calculate the logit of the predicted probability of the outcome (the DRS) for the patients included in the test data. Controlling for the DRS variable in a model regressing the outcome on treatment, results in a confounding adjusted treatment outcome association:

$$logit(prob[y_i = 1|x_i, drs_i]) = \hat{\gamma}_0 + \hat{\gamma}_1 x_i + \hat{\gamma}_2 drs_i \quad [6]$$

$\hat{\gamma}_1$, can be interpreted as the ln(odds ratio) of the association between treatment and the outcome, adjusted for the confounders included in the first stage.

Depending on the size of the training dataset a (very) large number of confounder can be including using this DRS method. This makes DRS models particularly interesting for post-launch settings. However, ideally a training dataset is used in which all patients are untreated [16], yet this is often impossible. To explore this, subjects in the simulated training data were exposed to drug C or B. Four DRS model were subsequently applied with the first DRS model ignoring treatment in the training data (DRS 1). In the DRS 2 model, the treatment variable was included in the training model. In the DRS 3 model a treatment by confounder Z_1 interaction was included. Instead of assuming that all interactions are appropriately modelled, DRS 4 prevented interaction by restricting the training dataset to subjects treated with drug C (the reference). Note that these four DRS models differed in how treatment was handled in training data,

however, analysis of the test data did not differ and entailed including the DRS in a logistic regression model regressing outcome on treatment (equation 6).

Simulation scenarios

In all simulations two datasets were generated, a training and test dataset. For both dataset (unless stated otherwise) the association of the continuous confounders Z_1 with treatment and outcome was set to an odds ratio (OR) of 0.60 per unit increase. The associations of the remaining binary confounders with treatment and the outcome were set to an OR of 0.97. The association of treatment with the outcome was set to an OR of 1.00. See Table 1 for an overview.

In **scenario I** different EPCs were generated by increasing the number of coefficients from 20 to 400. EPC was calculated as follows, $EPC = \frac{200}{2+j}$ where $j = \{18,\ 38, 78, 198, 398\}$, the 200 in the numerator representing the expected number of events and the 2 in the denominator representing the intercept and treatment coefficient. For the PS model, the EPC was calculated by taking the number of subjects expected to be exposed to drug A (200) and dividing this by j confounder coefficients plus the intercept coefficient. In **scenarios II** and **III** EPC was set to 10 in the test data, the treatment and interaction OR in the training data were set to 0.30 and 0.30 (for scenario II) or 3.00 and 0.30 (for scenario III). To determine in more detail the susceptibility of the DRS models for misspecification, the interaction effect in the training data was set to 0.30, 0.70, 1.00, 1.50 and 3.00 in **scenario IV**, while the EPC was set to 2.5. In **scenario V** the treatment OR in the training data was set to 0.30, 0.70, 1.00, 1.50 and 3.00 and the interaction effect to 3.00. In **scenario VI** power (i.e., the probability to detect an association if it is present) was explored by setting the treatment OR in the test data to 0.30, 0.70, 1.00, 1.50 and 3.00.

Finally, **scenario VII** was created to explore performance is less extreme settings as those previously explored.

All simulations were repeated 10 000 times and were performed with the statistical package R version 3.0.2 [17]. The number of replication was chosen to ensure sufficient precision to detect small deviations from the typical coverage rate of 0.95 (the 95% lower and upper bounds were 0.946 and 0.954) [18,19]. Furthermore, with 10 000 replications the 95% upper and lower bounds around a mean odds ratio of 1.00 was 0.996 and 1.004 (calculated using the empirical SE of the unadjusted OR which was constant across scenarios).

Performance metrics

The different methods to control for confounding were compared on the mean odds ratio, mean relative bias (see Appendix), the coverage rate, the mean estimated standard error (SE) (see Appendix)[15], the empirical SE (see Appendix), the square root of the mean squared error (RMSE) [18], power, number of models that failed to converge and the number of models with implausible estimates. Mean relative bias was defined as: $E\left[\frac{\widehat{OR}-True\ OR}{True\ OR}\right]*100$, where E indicates the expectation, $\widehat{OR}$ the estimated treatment OR and True OR the simulated treatment OR. The coverage was defined as the number of times the true value was included in the Wald based 95% confidence interval. The mean SE was defined as the mean of the estimated standard errors [15,18]. The empirical SE was estimated by taking the standard deviation of the distribution of $\widehat{OR}$. The RMSE was calculated by taking the square root of the sum of the squared bias and the squared empirical SE[18]. Power equalled the proportion of simulations in which the null-hypothesis of $OR = 1$ was correctly rejected, i.e., when the null-hypothesis was false (scenario VI). Implausible estimates were defined as treatment $\left|\ln(\widehat{OR})\right| > 5$.

Sensitivity analysis.

Instead of using DRS models when the number of EPC is very small, Firth penalized logistic regression (PLR) models have shown promise [20-22] in such settings. To explore this alternative to DRS models, scenario I was repeated with LR, PS and a PLR models. PLR was implement using the package logistf version 1.21 [23]. For comparisons sake Wald based p-values were calculated, however the reader should note that better performance is expected using profile likelihood p-values.

Results

Table 2 shows the results of the simulations evaluating the LR, PS and DRS models under different EPCs (scenario I), in the absence of a treatment effect. Relative bias of the LR and PS models was similar up to and including an EPC of 2.5. After this the LR model showed extreme bias. Relative bias of the PS model increased to 8.80% at an EPC of 1.0. Mean and empirical SE increased for both methods as EPC increased and extreme estimates were seen after EPC of 2.5 (for the LR) and 1.0 (for the PS). The coverage rate of LR model started to deviate from 0.95 at an EPC of 5.0 (0.936), with a more serious deviation at an EPC of 1.0 (0.651). For the PS models the coverage rate started to deviated from 0.05 at an EPC of 1.0 (0.939).

In the same scenario I, the mean odds ratios of the different DRS methods already deviated more than could be explained by random error at an EPC of 10. However, the bias was only small (1.38%) and increased to a maximum of 18.84% at an EPC of 0.5. The relative bias of the DRS model 4 was consistently larger than that of the other DRS models. . After an EPC of 5.0 the coverage rates of the DRS models were smaller than 0.95. Throughout the RMSE increased as the ECP increased.

In scenario II and III model misspecification of DRS 1 and 2 were introduced by adding a treatment by confounder interaction to the training data. In scenario II (interaction OR 0.30) the relative bias was small and the coverage rates were close to 0.95 for all methods (Table 3). In scenario III (interaction OR 3.00) DRS model 1 and 2 showed relative bias of 9.22% and 16.94%. Similarly, the coverage rates of these models were 0.930 and 0.881. On the other hand DRS models 3 and 4 showed coverage rates close to 0.95 and relative bias of 1.87% and 3.25%.

In Figure 1 the relative bias, coverage rates and RMSE of the simulation results of scenarios IV and V are presented. In scenario IV the treatment by confounder interaction effect was iterated from 0.30 to 3.0 at an EPC of 2.5. As expected, the relative bias of the LR and PS was small, and the coverage rate of the LR model was consistently 0.92, while the PS estimates had correct coverage of 0.95 (Figure 1, column 1). The relative bias of DRS model 1 was more or less symmetric and peaked at 14.6% for an interaction effect of 3.0. At an interaction effect of 0.30 DRS model 2 had the least amount of bias (2.52%). This increased to a bias of 19.27% with an interaction effect of 3.00. Relative bias of DRS model 3 and 4 was almost constantly about 5% or 8%. A marked increase was only seen for an interaction effect of 0.30.

In scenario V (Figure 1, column 2) the treatment effect in the training data was iterated from 0.30 to 3.00. All models performed very similar regardless of the treatment effect. The exception being DRS model 1 were the relative bias decreased from 14.42% to 8.66% as treatment increased to 3.00.

Empirical power was explored in scenario VI (depicted in figure 2), EPC 2.5. Power was below 0.40 for treatment effects between 0.70 and 1.50; at treatment ORs of 0.30 and 3.00 power was

almost 1.00. LR models were consistently more powerful than PS models; however previous results showed that in these settings coverage of the LR estimate is less than 0.95.

In scenario VII the DRS models were evaluated with an EPC of 10 with smaller confounder and interaction effects. In these settings the relative bias of the DRS models ranged from 0.86% (DRS 3) to 1.94% (DRS 2), compared to -0.09% for the LR and -0.10% for the PS and coverage rates were close to 0.95 for the DRS and PS methods but not for the LR method (0.941).

In all scenarios every model converged and no estimates were excluded. However, in scenario I extreme values were observed for the LS and PS models. Arbitrarily defining extreme, as an estimate above 5, resulted in excluding 7,243 and 9,421 of the 10,000 estimates for the LR method at an EPC of 1.0 and 0.05. For the PS model this resulted in excluding 4,711 estimates at an EPC of 0.05.

Results of the sensitivity analyses comparing PS and LR models to the PLR methods are presented in Appendix Table 1. Briefly, the PLR model showed a maximum relative bias of -9.13% and coverage rate of 1.000 at an EPC of 0.5. At an EPC of 1.0 relative bias and coverage rate of the PLR model was -0.16% and 0.931, at 0.5 this was -9.13% and 1.000. With regard to coverage the PS model performed similar to the PLR model, however bias was larger (8.75%, 0.942 at an EPC 1.0 and -98.56% and 0.973 at an EPC of 0.05).

Discussion

Our simulations show that, in settings with a relatively small number of events/exposed per coefficient (EPC), disease risk score (DRS) and propensity score (PS) methods provided less biased estimates of the association between treatment and outcome than logistic regression (LR). While DRS methods were more biased than LR and PS methods when EPC was large

(i.e., 10), in smaller settings they outperformed both. However, this was at the cost of a smaller coverage rate than the PS method. Additionally, DRS models were sensitive to misspecification of the treatment by confounder interaction effect. With DRS models, excluding the interaction effect, showing bias of 17% versus 2% to 3% when the interaction was appropriately modelled. However, in settings with less confounding and model misspecification bias was at most 1.94%. Finally, we showed that the PS method needs less exposed subjects per coefficient than LR method needs events per coefficient.

In our simulations the PS estimates were the least biased while keeping a coverage rate closest to the nominal 0.95. Additionally, in a sensitivity analysis PS models had similar coverage rates as penalized logistic regression models; a method which is generally expected to perform best in small EPC settings. Previously, Cepeda et.al.,[24] also explored EPC of PS models, focussing on the number of outcome events. Recognizing that PS model performance is more influence by the number of exposed than the number of event, the present simulations focussed on the number of exposed subjects per coefficient. For comparisons sake the proportion exposed and events was set to 0.50. We recognize that in most empirical studies, the proportion of exposed subjects will be closer to 0.50 than will be the proportion of events. Thus in most empirical studies the benefit of using PS models over LR and DRS models is expected to be greater than shown here. However, in small EPC settings where proportion of exposed subjects is less than the proportion of events, DRS method will likely outperform both LR and PS methods. In setting where EPC was 1.0 or less, DRS estimates were less biased and coverage was closer to 0.95 than estimates from PS and LR methods, however, coverage still deviated from 0.95. Essentially, in these settings, all methods failed and perhaps inclusion of additional subjects would be a more reasonable solution. Furthermore, at the tipping point of an EPC of 2.5, power was less than 40%, unless large treatment effects were present (OR > 1.5). Unless such a large effect is to be expected, inclusion of more subjects might again be the best solution. Finally, we

note that in all simulations the PS model consistently included one coefficient less than the LR model. This resulted in a slightly larger EPC: 10.53, 5.13, 2.53, 1.01, 0.50. This small difference seems unlikely to explain the improved performance of the PS models.

Previous simulation studies on DRS models trained the DRS in the tests data [13,15]. Because the same number of events is available, these DRS models cannot include more confounder than regular LR models and where not considered here. Instead we focussed on DRS models with an independent and larger training dataset. Depending on the size of the training data, these DRS models can potentially adjust for an enormous number of confounders. We expected bias to remain stable over increasing EPCs (due to the size of the training data). However, in our simulations bias did increase, which was probably due to an increase in random difference between the associations in the training and test data (due to an increase in variables as EPC increased). Obviously, this bias could be decreased by increasing the size of the test data. However when the test data increases in size the need for DRS model is less apparent and LR models might be a better choice. Surprisingly, DRS model 4, which limited the derivation dataset to subjects treated with drug C only, consistently showed larger bias than the other DRS models. As Wyss et.al discusses this bias is caused by overfitting the model to the reference group [25].

The simulations presented here are naturally limited. We feel that the following points merit discussion. First, in our simulations we predominantly focused on dichotomous confounders. Because continuous data is less sensitive to small cell counts it seems likely that if the simulations were repeated with only continuous confounders, bias would be smaller. Second, previous studies that explored EPC fixed both the number of event and the number of covariables. In the current paper we only fixed the number of covariables, and the number of events was an average. We feel that this approach more closely follows research practice,

where at the design phase it is possible to specify which and how many confounder would be included, however, only an expected number of event can be specified [26]. Thirdly, while we focussed on the situation where confounders are pre-specified [27], our results are also relevant for researchers wishing to reduce model complexity using e.g., backward selection methods. In the first stage of such an approach a full model is constructed which is equal to the pre-specified model applied here and similar concerns on model misspecification and EPC apply. Note, however, that applying model selection in LR models will increase the type 1 error rate, of the treatment association, beyond the level shown here [28,29]. Finally, all PS and DRS models were implemented using generalized linear models (GLMs). In empirical data, typically, the functional form of the PS or DRS with the outcome is unknown; hence, it seems advisable to use nonparametric methods such as matching or stratification. In our simulations however, the functional form was known and no disadvantage of using GLMs is expected

In conclusion, when the number of events and the number of exposed subjects are equally sparse disease risk models result in the least biased point estimates, however, at the cost of a smaller coverage rate. The propensity score estimates are more biased at an ECP of 1.0 and 0.5, however, coverage levels are close to 0.95. Depending on the settings and aim of the research, estimation or testing, a different method might be preferred. However, at very low EPCs (0.5) all methods had bias and coverage levels below acceptable levels and a better approach would be to include more subjects.

Reference List

(1) Grobbee DE, Hoes AW. Intervention Research: Unintended Effects. *Clinical Epidemiology: Principles, Methods and Applications for Clinical Research*. 2 ed. Burlington: Jones and Bartlett Learning, 2015: 181-214.

(2) Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet.* 2004;363(9422):1728-1731.

(3) Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? *CMAJ.* 2006;174(5):645-646.

(4) Dusetzina SB, Mack CD, Sturmer T. Propensity score estimation to address calendar time-specific channeling in comparative effectiveness research of second generation antipsychotics. *PLoS One.* 2013;8(5):e63973.

(5) Harrell FE, Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 1st ed. New York: Springer; 2001.

(6) Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512-522.

(7) 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(12):1373-1379.

(8) Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol.* 2011;64(9):993-1000.

(9) Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol.* 2007;165(6):710-718.

(10) Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association.* 1984;79(387):516-524.

(11) Sanni AM, Groenwold RH, Pestman WR et al. Time-dependent propensity score and collider-stratification bias: an example of beta2-agonist use and the risk of coronary heart disease. *Eur J Epidemiol.* 2013;28(4):291-299.

(12) Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol.* 1976;104(6):609-620.

(13) Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *J Clin Epidemiol.* 1989;42(4):317-324.

(14) Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf.* 2012;21 Suppl 2:138-147.

(15) Arbogast PG, Kaltenbach L, Ding H, Ray WA. Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology.* 2008;19(1):30-37.

(16) Tadrous M, Gagne JJ, Sturmer T, Cadarette SM. Disease risk score as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiol Drug Saf.* 2013;22(2):122-129.

(17) R Development Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.

(18) Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med.* 2006;25(24):4279-4292.

(19) Schmidt AF, Groenwold RH, Knol MJ et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study. *J Clin Epidemiol.* 2014;67(7):821-829.

(20) Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med.* 2002;21(16):2409-2419.

(21) Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med.* 2006;25(24):4216-4226.

(22) Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. *J Clin Epidemiol.* 2011;64(12):1464-1465.

(23) logistf: Firth's bias reduced logistic regression. Version R package version 1.21. 2013.

(24) 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(3):280-287.

(25) Wyss R, Lunt M, Brookhart MA, Glynn RJ, Sturmer T. Reducing Bias Amplification in the Presence of Unmeasured Confounding Through Out-of-Sample Estimation Strategies for the Disease Risk Score.(2193-3677 (Print)).

(26) Nikolakopoulos S, Roes KC, van der Lee JH, van der Tweel I. Sample size calculations in pediatric clinical trials conducted in an ICU: a systematic review. *Trials.* 2014;15(1):274.

(27) Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol.* 2002;155(2):176-184.

(28) Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society Series A (Statistics in Society).* 1995;158(3):419-466.

(29) Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol.* 1999;52(10):935-942.

Tables

Table 1 Simulation scenarios, assessing performance of different confounding adjustment methods*.

Parameters	Scenario I	Scenario II	Scenario III	Scenario IV	Scenario V	Scenario VI	Scenario VII
Training data							
Sample size [n]	5000	5000	5000	5000	5000	5000	5000
OR of reference treatment A vs. treatment B [δ_1]	1.00	0.30	0.30	0.30	**{0.3, 0.70, 1.00, 1.50, 3.0}**	**0.30**	**0.90**
OR of treatment by Z_1 interaction [δ_{int}]	1.00	**0.30**	**3.0**	**{0.3, 0.70, 1.00, 1.50, 3.0}**	**3.00**	3.00	**1.25**
Confounder Z_1 OR (event [δ_2]/treatment[α_1])	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	**0.80/0.80**
Other confounders (event [δ_{j-1}]/treatment[α_{j-1}])	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97
Test data							
Sample size [n]	400	400	400	400	400	400	400
OR of reference treatment A vs. treatment C [δ_1]	1.00	1.00	1.00	1.00	1.00	**{0.3, 0.70, 1.00, 1.50, 3.0}**	**1.00**
OR of treatment by Z_1 interaction [δ_{int}]	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**
Events per coefficient	{10, 5, 2.5, 1, 0.5}	**10**	10	**2.5**	2.5	2.5	**10**
Number of coefficients	20-400	**20**	20	**80**	80	80	**20**
Confounder Z_1 OR (event [δ_2]/treatment[α_1])	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	**0.80/0.80**
Other confounders (event [δ_{j-1}]/treatment[α_{j-1}])	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97

* Changes from the previous scenario (on the left) are presented in bold.

Table 2 Simulation results from scenario I assessing performance of different confounding adjustment methods with different events per coefficient*.

	10 EPC	5 EPC	2.5 EPC	1 EPC	0.5 EPC
Mean odds ratio					
LR	1.00	1.00	1.00	N/A#	$9.2*10^{118}$
PS	1.00	1.00	1.00	1.09	$1.2*10^{9}$
DRS1	1.01	1.03	1.05	1.10	1.15
DRS2	1.01	1.03	1.05	1.10	1.15
DRS3	1.01	1.03	1.05	1.10	1.15
DRS4	1.03	1.05	1.08	1.13	1.19
Relative bias					
LR	0.05	-0.06	-0.49	N/A#	$9.2*10^{120}$
PS	0.00	-0.12	-0.30	8.80	$1.2*10^{11}$
DRS1	1.38	2.57	4.57	9.53	14.70
DRS2	1.38	2.57	4.57	9.53	14.71
DRS3	1.39	2.59	4.61	9.56	14.72
DRS4	2.55	4.62	7.52	13.40	18.84
Coverage					
LR	0.946	0.936	0.920	0.651	1.000
PS	0.954	0.950	0.954	0.939	0.975
DRS1	0.951	0.949	0.945	0.926	0.898
DRS2	0.951	0.949	0.945	0.926	0.898
DRS3	0.950	0.949	0.945	0.927	0.898
DRS4	0.948	0.945	0.936	0.904	0.867
SMSE					
LR	0.22	0.25	0.30	$2.9*10^{14}$	$1.7*10^{4}$
PS	0.21	0.22	0.23	0.27	1317.99
DRS1	0.21	0.21	0.21	0.23	0.25
DRS2	0.21	0.21	0.21	0.23	0.25
DRS3	0.21	0.21	0.21	0.23	0.25
DRS4	0.21	0.21	0.22	0.24	0.25

* SMSE = square root of the mean squared error. # While all LR samples converged, the OR estimate was $\exp(5.42*10^{12})$ resulting in an error when calculating the mean OR and relative bias.

Table 3 Simulation results from scenario II and III comparing different DRS models in the presence of an interaction effect in the training data*.

	LR	PS	DRS1	DRS2	DRS3	DRS4
Scenario II#						
Mean odds ratio	1.00	1.00	1.05	1.01	1.02	1.04
Relative bias	-0.16	-0.19	5.10	0.51	2.19	4.14
Coverage	0.950	0.956	0.947	0.954	0.954	0.949
RMSE	0.22	0.21	0.21	0.21	0.21	0.21
Scenario III^						
Mean odds ratio	1.00	1.00	1.09	1.17	1.02	1.03
Relative bias	0.27	0.23	9.22	16.94	1.87	3.25
Coverage	0.948	0.954	0.930	0.881	0.952	0.950
RMSE	0.22	0.21	0.22	0.26	0.21	0.21

* SMSE = square root of the mean squared error. # Treatment by confounder 1 interaction OR of 0.30.
^ Treatment by confounder 1 interaction OR of 3.0

Figure captions

Figure 1 Simulation results from scenarios IV and V comparing different DRS models to PS and LR models on relative bias, coverage rate and square root of the mean squared error (RMSE). *

[Figure 1 here]

* line number 1 logistic regression; line number 2 propensity score; line number 3 disease risk score (DRS) 1 model; line number 4 DRS 2; line number 5 DRS 3 and line number 6 DRS 4.

Figure 2 Simulation results from scenario VI comparing different DRS models to PS and LR models on power.*

[Figure 2 here]

* line number 1 logistic regression; line number 2 propensity score; line number 3 disease risk score (DRS) 1 model; line number 4 DRS 2; line number 5 DRS 3 and line number 6 DRS 4.