

# Adjusting for Confounding in Early Postlaunch Settings Going Beyond Logistic Regression Models

Amand F. Schmidt,<sup>a,b,c,d</sup> Olaf H. Klungel,<sup>a,b</sup> and Rolf H. H. Groenwold<sup>a,b</sup>; on behalf of the GetReal Consortium

**Background:** Postlaunch data on medical treatments can be analyzed to explore adverse events or relative effectiveness in real-life settings. These analyses are often complicated by the number of potential confounders and the possibility of model misspecification.

**Methods:** We conducted a simulation study to compare the performance of logistic regression, propensity score, disease risk score, and stabilized inverse probability weighting methods to adjust for confounding. Model misspecification was induced in the independent derivation dataset. We evaluated performance using relative bias confidence interval coverage of the true effect, among other metrics.

**Results:** At low events per coefficient (1.0 and 0.5), the logistic regression estimates had a large relative bias (greater than -100%). Bias of the disease risk score estimates was at most 13.48% and 18.83%. For the propensity score model, this was 8.74% and >100%, respectively. At events per coefficient of 1.0 and 0.5, inverse probability weighting frequently failed or reduced to a crude regression, resulting in biases of -8.49% and 24.55%. Coverage of logistic regression estimates became less than the nominal level at events per coefficient  $\leq 5$ . For the disease risk score, inverse probability weighting, and propensity score, coverage became less than nominal

at events per coefficient  $\leq 2.5$ ,  $\leq 1.0$ , and  $\leq 1.0$ , respectively. Bias of misspecified disease risk score models was 16.55%.

**Conclusion:** In settings with low events/exposed subjects per coefficient, disease risk score methods can be useful alternatives to logistic regression models, especially when propensity score models cannot be used. Despite better performance of disease risk score methods than logistic regression and propensity score models in small events per coefficient settings, bias, and coverage still deviated from nominal.

(*Epidemiology* 2016;27: 133–142)

Nonrandomized studies on (pharmacologic) 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.<sup>1–3</sup> Furthermore, nonrandomized studies can be 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.<sup>1–3</sup> For example, it might be expected that patients who responded poorly to older drugs will cross over to the new drug.<sup>4</sup> Alternatively, as shown by Mack et al.,<sup>5</sup> physicians might be hesitant to prescribe a novel drug to patients with comorbidities. Furthermore, depending (among other factors) on the speed of uptake, differences in patient populations pre- and post-launch or difference between early and late adopters may increase the potential for effect modification, further obstructing comparison of a new drug to older compounds.<sup>6,7</sup>

Frequently, the outcome of interest is dichotomous, such as mortality, in which case multivariable logistic regression<sup>8</sup> is commonly used to adjust for confounding. One (of many) assumption(s) is that associations between confounders and the outcome are sufficiently estimated to adjust for confounding bias. In settings (e.g., nonrandomized early postlaunch 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.<sup>9</sup> Simulation studies showed that for prognostic logistic regression models, 10 or more events per coefficient were needed to

Submitted 28 November 2014; accepted 25 August 2015.

From the <sup>a</sup>Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; <sup>b</sup>Division of Pharmacoeconomics and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht, The Netherlands; <sup>c</sup>Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands; and <sup>d</sup>Institute of Cardiovascular Science, Faculty of Population Health, University College London, London, United Kingdom.

Supported by the Innovative Medicines Initiative Joint Undertaking under Grant Agreement No. 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.

A.F.S., R.H.H.G., and O.H.K. contributed to the idea and design of the study. A.F.S. performed the analyses and drafted the manuscript. O.H.K. and R.H.H.G. provided guidance during initial planning of the paper and during critical revision. A.F.S. had full access to all of the data and takes responsibility for the integrity of the data presented.

The research leading to these results was conducted as part of the GetReal consortium. For further information, please refer to [www.imi-getreal.eu](http://www.imi-getreal.eu).

The authors report no conflicts of interest.

**SDC** Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article ([www.epidem.com](http://www.epidem.com)).

Correspondence: Amand F. Schmidt, Institute of Cardiovascular Science, Faculty of Population Health, University College London, London, United Kingdom. E-mail: [amand.schmidt@ucl.ac.uk](mailto:amand.schmidt@ucl.ac.uk).

Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 1044-3983/16/2701-0133

DOI: 10.1097/EDE.0000000000000388

get unbiased estimates.<sup>10,11</sup> 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 adjusted for potential confounders. Vittinghoff and McCulloch<sup>12</sup> showed that in this case, logistic regression models with events per coefficient as small as six can adequately adjust for confounding.

In settings where logistic regression models are expected to perform poorly (i.e., events per coefficient smaller than six), propensity score,<sup>13,14</sup> disease risk score,<sup>15–19</sup> or inverse probability weighting,<sup>20–22</sup> methods can be applied to summarize information of multiple confounders into a single variable. It remains unclear how many events/exposures per variable are needed to sufficiently control for confounding using propensity score, disease risk score, and inverse probability-weighting methods. Furthermore, in training (i.e., developing) disease risk score models, it is often implicitly assumed that there is no treatment effect or no treatment by confounder interaction. However, there might be considerable differences between post- and pre-launch patients, increasing the potential for previously unrecognized interaction. The sensitivity of disease risk score models to model misspecification due to omitting a main or interaction effect is unknown, particularly when the disease risk score model is trained in a prelaunch dataset and applied in a postlaunch dataset (two-sample disease risk score). To explore these issues, we conducted a simulation study comparing logistic regression, propensity score, and inverse probability weighting and four kinds of disease risk score models with varying amounts of events or exposed subjects per coefficient, under different levels of model misspecification.

## METHODS

### Simulation Set-up

We focused on a scenario in which the effect of a new drug (or any other type of medical intervention) is evaluated postlaunch in an observational study. In each simulation, a training dataset was generated containing prelaunch information ( $n = 5,000$ ) and a test dataset ( $n = 400$ ) containing postlaunch information. In both datasets, half of the subjects experienced an event and, independent of outcome status, half were exposed to the comparator drug C. In the training data, the other half were exposed to drug B, and in the test data to drug A. The training and test datasets were generated using the same algorithm, containing a single continuous confounder  $Z_1$  and  $Z_{j-1}$  binary confounders (see eAppendix 1 for the algorithm applied; <http://links.lww.com/EDE/A972>).

Disease risk score models were derived in the training data. We then used to the test data to compare the estimated effect of the intervention (drug A vs. C) obtained through the disease risk score, logistic regression, inverse probability weighting, and propensity score methods (see eAppendix 1

for a description of the methods; <http://links.lww.com/EDE/A972>). Depending on the size of the training dataset, a (very) large number of confounders can be included. Ideally, the training dataset consists exclusively of untreated patients.<sup>19</sup> Otherwise disease risk score models might be biased by possible treatment by confounder interaction. To explore how sensitive this method is to unobserved interaction, we compared four disease risk score models. The first disease risk score model, DRS 1, reflecting current practice in most prediction models,<sup>23</sup> ignored treatment in the training data. In DRS 2, the treatment variable was included in the training model. In DRS 3, a treatment by confounder  $Z_1$  interaction was included. Instead of assuming that all interactions are appropriately modeled, DRS 4 prevents interaction by restricting the training dataset to subjects treated with drug C (the reference). We implemented propensity score adjustment by including the estimated propensity score as a continuous covariate in a regression model.

### Simulation Scenarios

In both the training and test datasets (unless stated otherwise), the association of confounder  $Z_1$  with treatment and outcome was set to an odds ratio (OR) of 0.60. The associations of the remaining confounders with treatment and the outcome were set to an OR of 0.97, to minimize the difference in the amount of confounding bias between the different events per coefficient scenarios. 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 events per coefficient were generated by increasing the number of coefficients from 20 to 400. In *scenarios II and III*, events per coefficient was set to 10, the treatment and interaction OR in the training data were set to 0.30 and 0.30 (for scenario II) or 0.30 and 3.00 (for scenario III). To further determine the susceptibility of the disease risk score 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 events per coefficient 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. Note that we focus exclusively on power because the conditional OR will differ from the marginal OR due to noncollapsibility<sup>24</sup>; other performance metrics (see below) are presented in the eTables 1–6 (<http://links.lww.com/EDE/A972>). *Scenario VII* explores performance in less extreme settings (Table 1). Finally, to gain insight into the influence of the size of the disease risk score training data, scenario 1 was repeated with training sample sizes of 5,000, 2,500, 1,000, and 400.

All simulations were repeated 10,000 times and were performed with the statistical package R version 3.1.1<sup>25</sup> for Linux. We chose the number of replications to ensure

**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	5,000	5,000	5,000	5,000	5,000	5,000	5,000
OR of reference treatment C vs. treatment	1.00	0.30	0.30	0.30	<b>{0.3, 0.70, 1.00, 1.50, 3.0}</b>	<b>0.30</b>	<b>0.90</b>
OR of treatment by $Z_1$ interaction	1.00	<b>0.30</b>	<b>3.0</b>	<b>{0.3, 0.70, 1.00, 1.50, 3.0}</b>	<b>3.00</b>	3.00	<b>1.25</b>
Confounder $Z_1$ OR (event/treatment)	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	<b>0.80/0.80</b>
Other confounders OR (event/treatment)	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	<b>0.80/0.80</b>
Test data							
Sample size	400	400	400	400	400	400	400
OR of reference treatment C vs. treatment A	1.00	1.00	1.00	1.00	1.00	<b>{0.3, 0.70, 1.00, 1.50, 3.0}</b>	<b>1.00</b>
OR of treatment by $Z_1$ interaction	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Events per coefficient	{10, 5, 2.5, 1, 0.5}	<b>10</b>	10	<b>2.5</b>	2.5	2.5	<b>10</b>
Number of coefficients	20–400	<b>20</b>	20	<b>80</b>	80	80	<b>20</b>
Confounder $Z_1$ OR (event/treatment)	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	0.60/0.60	<b>0.80/0.80</b>
Other confounders OR (event/treatment)	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	0.97/0.97	<b>0.80/0.80</b>

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

sufficient precision to detect small deviations from the typical confidence interval coverage rate of 0.95 (the 95% lower and upper bounds are 0.946 and 0.954),<sup>26,27</sup> and a mean OR 1.00 (the 95% lower and upper bounds are 0.996 and 1.004).

## Performance Metrics

The different methods were compared on the mean OR, mean relative bias (presented in eTables 1–6; <http://links.lww.com/EDE/A972>), the coverage rate, the mean estimated standard error (SE; presented in eTables 1–6; <http://links.lww.com/EDE/A972>),<sup>18</sup> the empirical SE (presented in eTables 1–6; <http://links.lww.com/EDE/A972>), the square root of the mean squared error,<sup>26</sup> power, number of models that failed to converge and the number of models with implausible estimates.

Mean relative bias was defined as  $\frac{\overline{\text{OR}} - \text{trueOR}}{\text{trueOR}} \times 100$ ,

where  $\overline{\text{OR}}$  indicates the average estimated treatment OR, and true OR the simulated treatment OR. 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 SEs.<sup>18,26</sup> The empirical SE was estimated by taking the standard deviation of the  $\overline{\text{OR}}$  distribution. The root mean squared error was calculated by taking the square root of the sum of the squared bias and the squared empirical SE.<sup>26</sup> Power equaled the proportion of simulations in which the null hypothesis of OR = 1 was correctly rejected (scenario VI). Implausible estimates were defined as treatment  $|\ln(\overline{\text{OR}})| > 5$ . Finally, we note that model convergence depends on the convergence criterion. For the current analysis, the R default criterion,  $<10^{-8}$ , was used.

## Sensitivity Analysis

Instead of using disease risk score models when the number of events per coefficient is very small, Firth penalized logistic regression models have shown promise.<sup>28–30</sup> Firth suggested to use a Jeffrey's prior to penalize the size of the regression coefficient toward zero. To evaluate this penalization in low events per coefficient settings, scenario I was repeated implementing logistic regression, propensity score, inverse probability weighting, and disease risk score methods using Firth's penalized logistic regression. For purposes of comparison, Wald-based *P* values and confidence intervals were calculated; however, better performance is expected using profile likelihood *P* values. All estimates were derived using the R package logistf version 1.21.<sup>31</sup>

In all the above-mentioned scenarios, the covariates included in the various models are true confounders (i.e., a common cause of both the exposure and the outcome). Obviously, in empirical settings, it is unknown whether variables are true confounders or not. Erroneous inclusion of instrumental variables (i.e., variables independent of true confounders and only related to the outcome via the exposure), in the presence of residual confounding, increases bias. However, this bias has been shown, typically, to be minimal compared with the amount of residual confounding.<sup>32</sup> To evaluate the susceptibility of the different methods to erroneous inclusion of instruments, scenario 1 was repeated including two binary instruments; probability 0.10 and 0.35 and an OR with treatment of 5.0.

Finally, while it is tempting to distill a rule of thumb for the number of events per coefficient required, performance is probably highly dependent on the simulation scenario used; therefore, it is more useful to focus on relative

(i.e., between methods) performance. To underline this, we repeated scenario 1 with the OR for  $Z_1$  set to 1.5 for both the outcome and exposure, and the remainder of the confounder ORs set to 0.90.

## RESULTS

Table 2 shows the results of the simulations evaluating the logistic regression, propensity score, inverse probability weighting, and disease risk score models under different

events per coefficient (scenario I), in the absence of a treatment effect. Relative bias of the logistic regression, inverse probability weighting, and propensity score models were similar up to and including an events per coefficient value of 2.5. After this, the logistic regression model showed extreme bias. Relative bias of the propensity score and inverse probability weighting models increased to 8.74% and -8.49%, respectively, at an events per coefficient of 1.0. Mean and empirical SE (eTable 1; <http://links.lww.com/EDE/A972>) increased

**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 <sup>b</sup>	0.5 EPC
<b>Mean odds ratio</b>					
Crude	1.18	1.19	1.19	1.21	1.25
LR	1.00	1.00	1.00	N/A <sup>a</sup>	N/A <sup>a</sup>
PS	1.00	1.00	1.00	1.09	N/A <sup>a</sup>
DRS 1	1.01	1.03	1.05	1.10	1.15
DRS 2	1.01	1.03	1.05	1.10	1.15
DRS 3	1.01	1.03	1.05	1.10	1.15
DRS 4	1.03	1.05	1.08	1.13	1.19
IPW	1.00	1.00	0.99	0.92	1.25
<b>Relative bias (%)</b>					
Crude	18.44	18.56	19.26	21.46	25.55
LR	-0.03	-0.28	-0.32	N/A <sup>a</sup>	-100
PS	-0.08	-0.28	-0.19	8.74	-100
DRS 1	1.36	2.52	4.72	9.58	14.75
DRS 2	1.36	2.52	4.72	9.59	14.75
DRS 3	1.38	2.54	4.76	9.62	14.76
DRS 4	2.55	4.53	7.77	13.48	18.83
IPW	0.09	-0.15	-0.51	-8.49	24.55
<b>Coverage</b>					
Crude	0.865	0.858	0.862	0.834	0.810
LR	0.950	0.941	0.917	0.656	1.000
PS	0.957	0.956	0.954	0.945	0.974
DRS 1	0.952	0.945	0.946	0.924	0.897
DRS 2	0.952	0.945	0.946	0.925	0.898
DRS 3	0.951	0.945	0.944	0.925	0.897
DRS 4	0.948	0.941	0.933	0.903	0.867
IPW	0.958	0.954	0.952	0.656	0.810
<b>RMSE</b>					
Crude	0.26	0.26	0.27	0.28	0.30
LR	0.22	0.25	0.30	$2.7 \times 10^{14}$	$4.5 \times 10^4$
PS	0.21	0.21	0.23	0.27	$4.8 \times 10^4$
DRS 1	0.21	0.21	0.21	0.23	0.25
DRS 2	0.21	0.21	0.21	0.23	0.25
DRS 3	0.21	0.21	0.21	0.23	0.25
DRS 4	0.21	0.21	0.22	0.24	0.27
IPW	0.21	0.22	0.26	4.19	0.30

Scenario 1 consisted of a training dataset of 5,000 subjects (for the DRS models only), a 400 subject test dataset, both with 50% exposed and 50% of the subjects experiencing an event.

<sup>a</sup>While all LR samples converged, the odd ratio estimate was  $\exp(5.42 \times 10^{12})$  resulting in an error when calculating the mean odds ratio and relative bias.

<sup>b</sup>At an EPC of 1.0, the IPW model failed to converge 2,763 time out of 10,000 replications the other methods did converge.

DRS indicates disease risk score; IPW, inverse probability weights; LR, logistic regression; PS, propensity score; RMSE, square root of the mean squared error; EPC, events per coefficient.



**TABLE 3.** Simulation Results from Scenarios II and III Comparing Different DRS Models in the Presence of an Interaction Effect in the Training Data

	Crude	LR	PS	DRS 1	DRS 2	DRS 3	DRS 4	IPW
<b>Scenario II<sup>a</sup></b>								
Mean odds ratio	1.18	1.00	1.00	1.05	1.00	1.02	1.04	1.00
Relative bias (%)	18.16	−0.29	−0.32	4.95	0.36	2.04	3.98	−0.15
Coverage	0.861	0.942	0.951	0.940	0.946	0.945	0.941	0.953
RMSE	0.26	0.23	0.21	0.22	0.21	0.21	0.21	0.21
<b>Scenario III<sup>b</sup></b>								
Mean odds ratio	1.18	1.00	1.00	1.09	1.17	1.01	1.03	1.00
Relative bias (%)	18.42	−0.04	−0.08	8.75	16.55	1.49	2.83	0.11
Coverage	0.863	0.946	0.953	0.931	0.879	0.948	0.948	0.955
RMSE	0.26	0.23	0.21	0.22	0.26	0.21	0.21	0.21

Scenarios II and III consisted of a training dataset of 5,000 subjects (for the DRS models only), a 400 subject test dataset, both with 50% exposed and 50% of the subjects experiencing an event, 10 events per coefficients, and different amounts of treatment by confounder 1 interaction in the training dataset.

<sup>a</sup>Treatment by confounder 1 interaction OR of 0.30.

<sup>b</sup>Treatment by confounder 1 interaction OR of 3.0.

DRS indicates disease risk score; IPW, inverse probability weights; LR, logistic regression; PS, propensity score; RMSE, square root of the mean squared error.

for these methods as events per coefficient increased and extreme estimates were seen after events per coefficient of 2.5 (for the logistic regression model) and 1.0 (for the propensity score model). At an events per coefficient of 0.5, the propensity score and logistic regression methods both showed extreme bias of −100%, with 24.55% the bias of the inverse probability weighting approximated that of the crude analysis 25.55%, indicating that the inverse probability weighting failed to adjust for confounding in this setting. The coverage rate of logistic regression models started to deviate from 0.95 at events per coefficient of 5.0 (0.941), with more serious deviation at an events per coefficient of 1.0 (0.656). For the propensity score models, the coverage rate started to deviate from 0.05 at an events per coefficient of 1.0 (0.945).

In the same scenario I, the mean ORs of the different disease risk score methods deviated more than could be explained by random error at an events per coefficient of 10. However, the bias was small (1.36%), with a maximum of 18.83% at an events per coefficient of 0.5. The relative bias of disease risk score model 4 was consistently larger than that of the other disease risk score models. After an events per coefficient of 5.0, the coverage rates of the disease risk score models were less than 0.95.

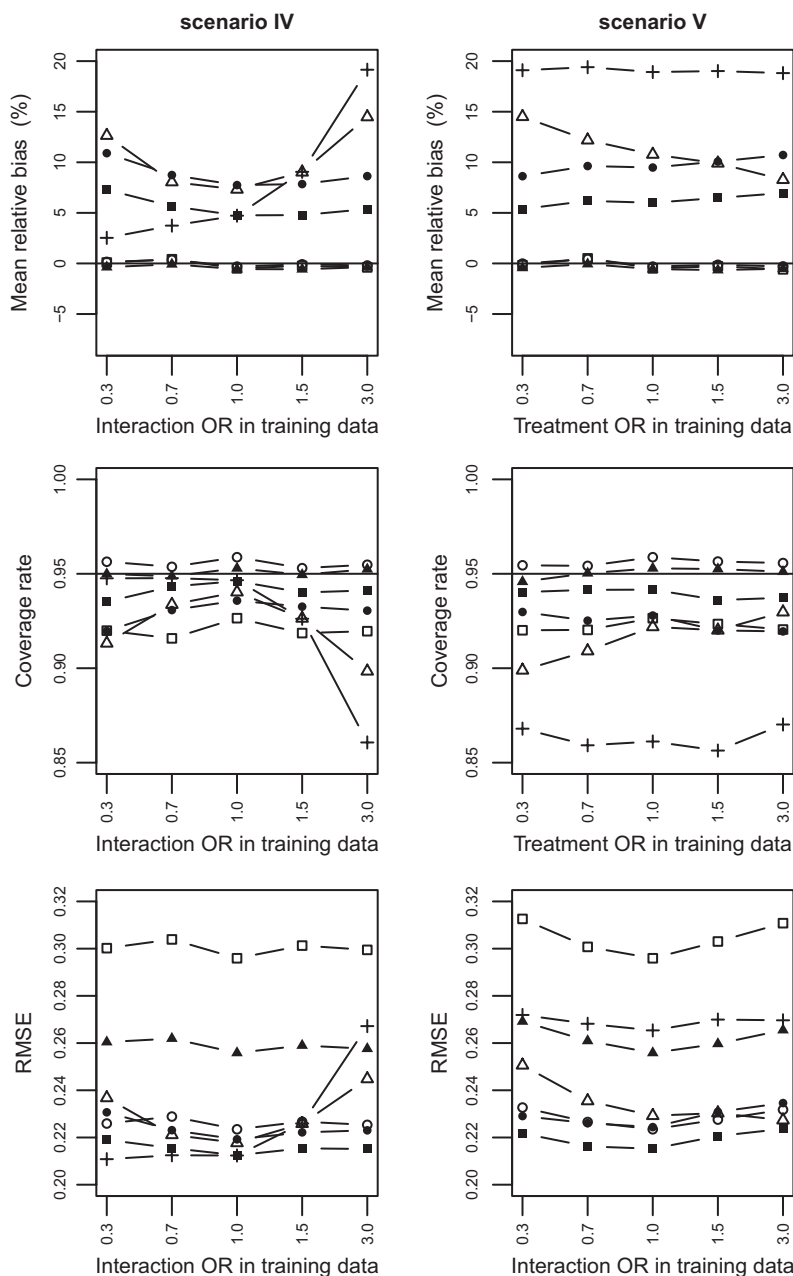
In scenarios II and III, model misspecification of DRS 1 and 2 were introduced by adding a treatment by confounder interaction to the training data. In scenario II (interaction OR 0.30), the relative bias was small and the coverage rates were close to 0.95 for all methods (Table 3 and eTable 2; <http://links.lww.com/EDE/A972>). In scenario III (interaction OR 3.00), DRS 1 and 2 showed relative bias of 8.75% and 16.55%, respectively. Similarly, the coverage rates of these models were 0.931 and 0.879. Disease risk score models 3 and 4 showed coverage rates close to 0.95 and relative bias of 1.49% and 2.83%.

In Figure 1 the relative bias, coverage rates and root mean squared error 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 logistic regression, inverse probability weighting, and propensity score models were small, and the coverage rate of the logistic regression model was consistently 0.92, while the inverse probability weighting and propensity score estimates had coverages of 0.95 (Figure 1, left-hand side). The relative bias of disease risk score model 1 was more or less symmetric and peaked at 14.48% for an interaction effect of 3.0. At an interaction effect of 0.30, disease risk score model 2 had the least amount of bias (2.53%). This increased to a bias of 19.15% with an interaction effect of 3.00. The relative bias of disease risk score models 3 and 4 was constantly between 7% and 5% or 10% and 8%, respectively.

In scenario V (Figure 1, right-hand side), the treatment effect in the training data ranged from 0.30 to 3.00, while the interaction effect was kept constant at OR 3.00. All models performed similarly regardless of the treatment effect, except, disease risk score model 1 where the relative bias decreased from 14.50% to 8.28% as treatment increased to 3.00.

We explored empirical power in scenario VI (Figure 2 and eTable 3; <http://links.lww.com/EDE/A972>), with events per coefficient 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. Disease risk score and logistic regression models were consistently more powerful than propensity score and inverse probability weighting models.

In scenario VII, disease risk score models were evaluated with an events per coefficient of 10 with smaller confounder and interaction effects. In these settings, the relative bias of the disease risk score models ranged from 1.37% (DRS 3)



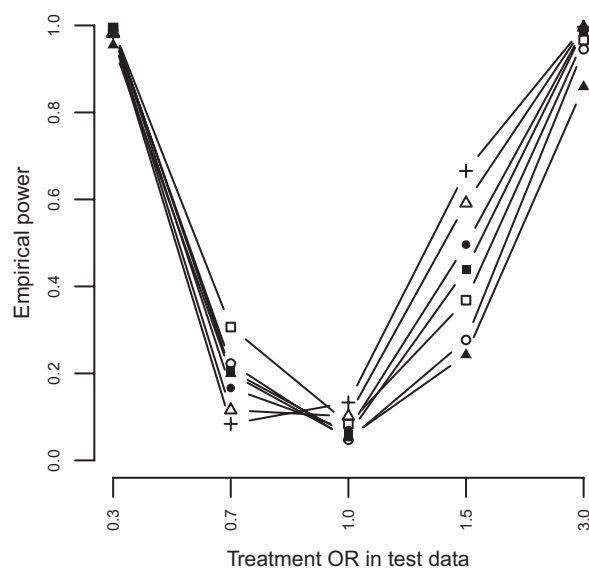
**FIGURE 1.** Simulation results from scenarios IV and V comparing different disease risk score models to inverse probability weights, propensity score, and logistic regression models on relative bias, coverage rate, and RMSE; the square root of the mean squared error. Solid line with a square logistic regression; solid line with a circle propensity score; solid line with a triangle DRS 1 model; solid line with a plus DRS 2; solid line with a filled out square DRS 3; solid line with a filled out circle DRS 4; solid line with a filled out triangle inverse probability weights. DRS indicates disease risk score.

to 2.50% (DRS 4), compared with 0.01% for the logistic regression, 0.06 for the inverse probability weighting, and  $-0.02\%$  for the propensity score models, with coverage rates close to 0.95 for all methods.

To assess the impact of the disease risk score training sample size, scenario 1 was repeated with different sample sizes (Figure 3). Given the similar performance of the disease risk score methods in this scenario, we focused on DRS 2. Obviously, relative bias and root mean squared error increased as sample size decreased. However, even with a training dataset of 400 subjects (equal to the test sample size), the disease risk score method outperformed the logistic regression method at events per coefficient of 1.0 and 0.5. The root mean

square error of disease risk score trained with 5,000 and 2,500 subjects was relatively similar up to an events per coefficient of 2.5 indicating sufficiency of both sample sizes.

In all scenarios, all disease risk score, propensity score, and logistic regression models converged and no estimates were excluded. However, in scenario 1, 2,763 inverse probability weighting models failed to converge at events per coefficient of 1.0, while at events per coefficient of 0.5 all models converged again. Arbitrarily defining extreme estimates, as an absolute estimate above five on the natural logarithmic scale, resulted in 7,219 and 9,415 extreme estimates for the logistic regression method (events per coefficient of 1.0 and 0.5). For the propensity score model, 4,706 extreme estimates occurred



**FIGURE 2.** Simulation results from scenario VI comparing different disease risk score models to inverse probability weights, propensity score, and logistic regression models on power. Solid line with a square symbol logistic regression; solid line with a circle propensity score; solid with triangle DRS 1 model; solid line with a plus DRS 2; solid line with a filled out square DRS 3; solid line with a filled out circle DRS 4; solid line with a filled out triangle inverse probability weights. DRS indicates disease risk score.

at events per coefficient of 0.53 and 672 extreme inverse probability weighted estimates were observed at an events per coefficient of 1.0 and zero at an events per coefficient of 0.5.

Results of the sensitivity analysis replacing logistic regression by Firth's penalized logistic regression showed improved performance of logistic regression and propensity score models (eTable 4; <http://links.lww.com/EDE/A972>). The other models performed similar as in scenario 1 using logistic regression, while disease risk score model 3 (modeling an interaction term) performed worse. Inclusion of two instrument variables did not impact results (eTable 5; <http://links.lww.com/EDE/A972>), except for the inverse probability weighting approximating the crude estimate at an EPC of 1.0 instead of 0.5 (in the original simulation). Finally, repeating scenario 1 with extreme levels of confounding (eTable 6; <http://links.lww.com/EDE/A972>) showed the same relative performance with the logistic regression method being the first to show suboptimal coverage and bias, followed by the disease risk score and the propensity score models. However, due to the extreme bias, this occurred at higher events per coefficient than previous. Similarly, due to extreme estimates, the inverse probability weighting method already approximated the crude analysis at events per coefficient of 1.0.

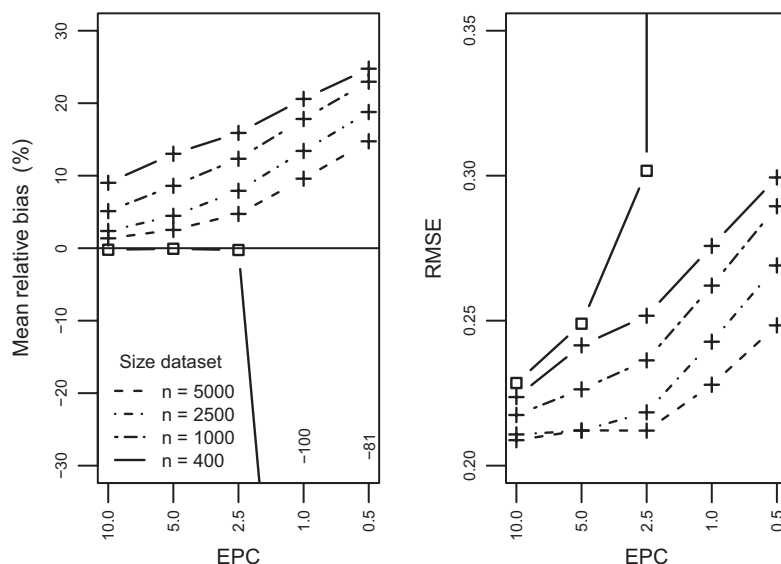
## DISCUSSION

This simulation study showed that in observational settings with a relatively small number of events/exposed per

coefficient, disease risk score, and propensity score methods provided less biased estimates than did logistic regression and inverse probability weighted regression. In larger events per coefficient settings (e.g., 10), disease risk score methods remained biased while the other methods were not. Penalization of the likelihood markedly improved performance of the logistic regression, and propensity score, but not the performance of inverse probability weighted and disease risk score methods. In addition, depending on the magnitude and direction of treatment by covariate interaction in the disease risk score training data, disease risk score models ignoring this interaction (i.e., misspecified disease risk score models) were more biased.

Previous simulation studies on disease risk score models trained the disease risk score in the same test data.<sup>16,18</sup> In contrast, we evaluated two-sample disease risk score models using an independent training dataset, which we showed increased performance of these models as long as the training dataset was sufficiently large and the model was correctly specified. Simulations with differently sized training data revealed that the disease risk score methods outperformed logistic regression even when the test and training dataset had the same sample size, and that the increased bias of DRS 4 was due to a reduction in sample size from 5,000 to 2,500. We note, however, that for most events per coefficient (except 0.5), the propensity score method was the least biased method. Increasing the size of the training data will likely increase disease risk score performance further, however, due to the two-sample nature, some residual bias will remain even in infinitely large training data. This residual confounding is due to the training data estimates being suboptimally tailored to the specific confounding structure of test dataset. Because the random difference between training and test data remained, penalized disease risk score models performed similar as the regular disease risk score models. The worse performance of the penalized DRS 3 (including an interaction term) is likely due to the fact that the Jeffrey's prior does not perform well in multidimensional settings. Performance of other penalization methods obviously deserves further attention (e.g., Lasso or Ridge penalizations or the use of informative Bayesian priors<sup>33</sup>).

Previously, Cepeda et al.<sup>34</sup> also explored events per coefficient of propensity score models, focusing on the number of outcome events. Instead, the present simulations focused on the number of exposed subjects per coefficient, which is more influential in propensity score estimation. For comparisons sake, the expected number of exposed and events was kept equal. In most empirical studies, the proportion of exposed subjects will be closer to 0.50 than the proportion of events, thus the benefit of using propensity score models over logistic regression and disease risk score models is expected to be greater than shown here. We also note that in all simulations, the propensity score and inverse probability weighting models consistently included one coefficient less than the logistic



**FIGURE 3.** Simulation results from scenario 1 with different sized training datasets, comparing disease risk score model 2 to a logistic regression model. Line with a square logistic regression; line with a plus disease risk score model 2. Vertical values above the x axis represent the y values of the logistic regression model. RMSE, square root of the mean squared error; EPC, events per coefficient.

regression and disease risk score models, resulting in a slightly larger events per coefficient: 10.53, 5.13, 2.53, 1.01, 0.50. This small difference seems unlikely to explain the improved performance of the propensity score model. At events per coefficient of 1.0, extreme inverse probability weights resulted in a great number of model failures. At an events per coefficient of 0.5, all models converged; however, the estimated propensity of treatment was approximately 0.5, resulting in inverse probability weighted estimates equal to the crude (unadjusted for confounding) logistic regression estimates. In low events per coefficient settings, essentially all methods failed and perhaps inclusion of additional subjects would be a more reasonable solution. Alternatively, the penalized regression model greatly improved performance and should more often be considered. One should take into account, however, that unless large effects ( $OR > 1.5$ ) can be expected, power is likely limited in such settings. This underperformance of all methods at low events per coefficient settings is due to a combination of separation (i.e., no variation in outcome at certain levels of the predictors)<sup>28,29</sup> and nonpositivity<sup>35</sup> (i.e., absence of variation in exposure at certain levels of the covariates). In our simulation, nonpositivity was random by design, so interpolation of the (confounder) estimates is appropriate. In empirical data, one should assess, case by case, whether nonpositivity might be deterministic, which may invalidate any interpolation or extrapolation. Similarly, while separation can be dealt with analytically, for example, using penalized Firth regression, analytic strategies should only be applied after careful consideration of potential biological reasons for the lack of observed outcomes, which can be a valid result by itself.

In empirical analyses, besides deciding on the confounding adjustment method, it is important to take account of the inherent time-dependent nature of most confounding biases, especially in pharmacoepidemiology.<sup>7</sup> An important initial assessment might be to explore how patient characteristics

change over time and monitor the proportion of new initiators.<sup>4,36</sup> When there is a large fluctuation between time points, options are to model time explicitly,<sup>5</sup> focus on a subset of patients that show similar characteristics,<sup>6</sup> or refrain from performing a comparison until a more stable pattern emerges.<sup>4</sup> Finally, data over multiple years of follow-up might be used to emulate an RCT as described by Hernán et al.<sup>37</sup> While erroneous inclusion of instrumental variables did not impact our results (eTable 5; <http://links.lww.com/EDE/A972>), likely due to the already large bias caused by nonpositivity and separation, if instruments are available researchers could consider performing an instrumental variable analysis.<sup>38,39</sup> Similar to an RCT, results from an instrumental variable analysis are unaffected by observed and unobserved confounders. Conditional on the speed of uptake and the observed difference between early and late initiating patients (or prescribers), the impact of treatment effect modification should also be considered.<sup>6</sup> We showed that the disease risk score method is fairly robust for unobserved interaction (or exclusion of the main effect of treatment) in the training data, unless of course this interaction effect is very large, which is unlikely to go unrecognized for drugs that have been marketed for a longer period. Depending on the likelihood of (unobserved) interaction in the test data, one might apply logistic regression, disease risk score, inverse probability weighting, or propensity score (included as a covariate) to estimate a population average effect,<sup>40,41</sup> or a matched propensity score adjustment to estimate the treatment effect in the exposed.<sup>42</sup>

The simulations presented here are naturally limited and the following points merit discussion. First, as we showed in a sensitivity analysis, changing the simulation parameters will result in worse or better performance at a particular events per coefficient value. Therefore, instead of distilling a rule of thumb for the events per coefficient required, one should focus on the relative (i.e., between methods) difference in performance,



which we showed to be more robust. Second, previous studies that explored events per coefficient fixed both the number of events and the number of covariates. In this article, the number of covariates was fixed, 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 confounders would be considered, but only an expected number of events can be specified.<sup>43</sup> Third, while we focused on the situation where confounders are prespecified,<sup>44</sup> results are also relevant for researchers wishing to reduce model complexity using, for example, backward selection methods. In the first stage of such an approach, a full model is constructed which is equal to the prespecified model applied here, hence similar concerns about model misspecification and events per coefficient apply. Note, however, that applying model selection in logistic regression models will increase the type 1 error rate of the treatment associations beyond the level shown.<sup>45,46</sup>

In conclusion, when the number of events and the number of exposed subjects are equally sparse relative to the number of coefficients (e.g., events per coefficient of 0.5), disease risk models result in the least biased point estimates, albeit at the cost of a smaller coverage rate. While propensity score estimates were more biased at low events per coefficient, coverage remained closest to 0.95. At higher events per coefficient, propensity score models typically performed better than disease risk and logistic regression models. Depending on the settings and aim of the research, estimation or testing, a different method might be preferred. However, at very low events per coefficient (0.5), all methods had unacceptable levels of bias and coverage and a better approach might be to include more subjects or to use penalized likelihood methods.

## REFERENCES

- Grobee DE, Hoes AW. Intervention research: unintended effects. In: Falivene C, ed. *Clinical Epidemiology: Principles, Methods and Applications for Clinical Research*. 2nd ed. Burlington, VT: Jones and Bartlett Learning; 2015:181–214.
- Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363:1728–1731.
- Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? *CMAJ*. 2006;174:645–646.
- Reams BD, O'Malley CD, Critchlow CW, Lauffenburger JC, Brookhart MA. Changing patterns of use of osteoporosis medications in the years after launch: implications for comparative effectiveness research. *Pharmacoepidemiol Drug Saf*. 2014;23:251–260.
- Mack CD, Glynn RJ, Brookhart MA, et al. Calendar time-specific propensity scores and comparative effectiveness research for stage III colon cancer chemotherapy. *Pharmacoepidemiol Drug Saf*. 2013;22:810–818.
- Franklin JM, Rassen JA, Bartels DB, Schneeweiss S. Prospective cohort studies of newly marketed medications: using covariate data to inform the design of large-scale studies. *Epidemiology*. 2014;25:126–133.
- Rassen JA, Schneeweiss S. Newly marketed medications present unique challenges for nonrandomized comparative effectiveness analyses. *J Comp Eff Res*. 2012;1:109–111.
- Harrell FE, Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 1st ed. New York, NY: Springer; 2001.
- 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;20:512–522.
- 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;49:1373–1379.
- 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;64:993–1000.
- Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007;165:710–718.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516–524.
- Sanni Ali M, 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;28:291–299.
- Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol*. 1976;104:609–620.
- 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;42:317–324.
- 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.
- Arbogast PG, Kaltenbach L, Ding H, Ray WA. Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology*. 2008;19:30–37.
- Tadrous M, Gagne JJ, Stürmer T, Cadarette SM. Disease risk score as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiol Drug Saf*. 2013;22:122–129.
- Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
- Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664.
- van der Wal WM, Noordzij M, Dekker FW, et al. Comparing mortality in renal patients on hemodialysis versus peritoneal dialysis using a marginal structural model. *Int J Biostat*. 2010;6:Article 2.
- Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart*. 2011;97:689–697.
- Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14:29–46.
- R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.
- Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25:4279–4292.
- 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;67:821–829.
- Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21:2409–2419.
- Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med*. 2006;25:4216–4226.
- Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. *J Clin Epidemiol*. 2011;64:1464–1465; author reply 1463.
- logistf: Firth's bias reduced logistic regression. Version R package version 1.21. 2013.
- Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174:1213–1222.
- Schmidt AF, Klugkist I, Klungel OH, et al. Bayesian methods including nonrandomized study data increased the efficiency of postlaunch RCTs. *J Clin Epidemiol*. 2015;68:387–396.
- Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287.
- Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171:674–677.
- Gagne JJ, Bykov K, Willke RJ, Kahler KH, Subedi P, Schneeweiss S. Treatment dynamics of newly marketed drugs and implications for comparative effectiveness research. *Value Health*. 2013;16:1054–1062.

37. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19:766–779.
38. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology*. 2006;17:260–267.
39. Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *Am J Epidemiol*. 2009;169:273–284.
40. Schmidt AF, Hoes AW, Groenwold RH. Comments on “the use of propensity scores and observational data to estimate randomized controlled trial generalizability bias” by Taylor R. Pressler and Eloise E. Kaizar, *Statistics in Medicine* 2013. *Stat Med*. 2014;33:536–537.
41. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Stat Med*. 2013;32:3552–3568.
42. Stürmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*. 2006;15:698–709.
43. 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;15:274.
44. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
45. Chatfield C. Model uncertainty, data mining and statistical inference. *J R Stat Soc Series A*. 1995;158:419–466.
46. 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;52:935–942.