

The change in estimate method for selecting confounders: A simulation study

Statistical Methods in Medical Research
2021, Vol. 30(9) 2032–2044
© The Author(s) 2021



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802211034219
journals.sagepub.com/home/smm



Denis Talbot^{1,2} , Awa Diop^{1,2,3},
Mathilde Lavigne-Robichaud^{1,2}  and Chantal Brisson^{1,2,4}

Abstract

Background: The change in estimate is a popular approach for selecting confounders in epidemiology. It is recommended in epidemiologic textbooks and articles over significance test of coefficients, but concerns have been raised concerning its validity. Few simulation studies have been conducted to investigate its performance.

Methods: An extensive simulation study was realized to compare different implementations of the change in estimate method. The implementations were also compared when estimating the association of body mass index with diastolic blood pressure in the PROspective Québec Study on Work and Health.

Results: All methods were susceptible to introduce important bias and to produce confidence intervals that included the true effect much less often than expected in at least some scenarios. Overall mixed results were obtained regarding the accuracy of estimators, as measured by the mean squared error. No implementation adequately differentiated confounders from non-confounders. In the real data analysis, none of the implementation decreased the estimated standard error.

Conclusion: Based on these results, it is questionable whether change in estimate methods are beneficial in general, considering their low ability to improve the precision of estimates without introducing bias and inability to yield valid confidence intervals or to identify true confounders.

Keywords

Confounding, epidemiologic methods, modeling, variable selection

I Background

Adjustment for potential confounders is routinely performed in etiologic studies based on observational data. Subject matter expertise plays a pivotal role in identifying confounders. However, uncertainty often persists regarding whether some covariates are truly confounders or not. In a recent review of studies published in four major epidemiologic journals, only 146/292 (50%) of explicative studies indicated choosing adjustment covariates based on prior knowledge, and 30/146 (20%) of these reported also using data-driven methods.¹ In total, 69/292 (24%) of explicative studies reported using some data driven method to help selecting covariates.¹ This likely underestimates the prevalence of data-driven variable selection since 107/292 (37%) of studies did not

¹Département de médecine sociale et préventive, Université Laval, Québec, Canada

²Unité santé des populations et pratiques optimales en santé, CHU de Québec – Université Laval research center, Québec, Canada

³Département de mathématiques et de statistique, Université Laval, Québec, Canada

⁴Centre de recherche sur les soins et les services de première ligne de l'Université Laval, Québec, Canada

Corresponding author:

Denis Talbot, Département de médecine sociale et préventive, Faculté de médecine, Université Laval, 1050, avenue de la Médecine, Pavillon Ferdinand-Vandry, room 2454, Québec (Québec) G1V 0A6, Canada.

Email: denis.talbot@fmed.ulaval.ca

provide sufficient information to determine how variables were selected. As such, variable selection based on the observed data is frequently attempted in epidemiology.

Also according to this review, the change in estimate (CIE) would be the most popular data-driven method for selecting confounders in epidemiologic studies.¹ Indeed, 34/69 (42%) of studies that used data-driven methods employed the CIE. Studies of varied size and fields of epidemiology were using the CIE. This is unsurprising considering that the CIE is recommended both in modern epidemiologic textbooks and articles over confounder selection methods based on P values in situations where the analyst determines that data-driven selection is warranted (see literature^{2,3} and references therein). For example, in Chapter 15 of the 3rd edition of *Modern Epidemiology*, it is written “Although many have argued against the practice [. . .], one often sees statistical tests used to select confounders (as in stepwise regression), rather than the change-in-estimate criterion just discussed.”³

The most typical implementation of the CIE first entails fitting an outcome model according to the exposure and adjusted for all potential confounders. Potential confounders are then removed from the outcome model one at a time. The procedure stops once it becomes impossible to remove a potential confounder without altering too much the exposure effect estimate as compared to the estimate produced by the initial fully adjusted model. Intuitively, if all confounders are available, the fully adjusted model should yield an estimate that is appropriately adjusted for confounding. Any reduced model that yields an estimate similar to that of the fully adjusted model is thus also expected to provide adequate adjustment for confounding.

While the CIE is appealing because it is intuitive and simple to implement, concerns have been raised concerning its validity. For instance, it has been noted that the change in the effect estimate may partly reflect non-collapsibility instead of confounding when employing effect measures such as the odds ratio or the hazard ratio.^{2,4-6} A further critique of the CIE is that it is susceptible to produce invalid P values and confidence intervals.³ This is because P values and confidence intervals are typically computed by statistical software assuming that the model is known a priori. When the model is selected based on the observed data, this assumption no longer holds. Finally, if the CIE is applied without reflecting on how covariates are causally related to the exposure and the outcome, it may lead to inappropriately controlling for colliders, thus introducing bias.⁴

We are aware of only three simulation studies that have investigated the performance of the CIE.⁷⁻⁹ According to their results, the CIE would yield estimators with small bias and valid confidence intervals when a low change in estimate threshold is used (for example, 10%), but would fail to produce estimators with improved precision as compared to a model adjusting for all potential confounders. However, these simulation studies have a number of limitations. First, they consider scenarios with at most nine potential confounders.⁷⁻⁹ As such, situations with multiple potential confounders have never been investigated. Such situations are those where benefits from performing confounder selection are most expected.² Moreover, the CIE can be implemented in multiple ways. For example, in addition to the backward exclusion described earlier, it is also possible to proceed by forward inclusion of confounders. Forward implementation of the CIE is being used in practice,¹ but its performance has never been investigated as far as we know. Simulation studies have also focused on odds ratio effect measures. To the best of our knowledge, the performance of the CIE with hazard ratios or mean differences has never been examined.

The goal of the current study is thus to provide additional information regarding the performance of the CIE. We do not consider the problems caused by including colliders in the potential confounder set since it has already been shown that is impossible to distinguish between a confounder and a collider from the data alone.^{4,10} Hence, it is expected that the CIE would perform poorly when colliders are included among the potential confounders. Substantive knowledge input is essential for constructing the initial potential confounder set.

We have first conducted an extensive simulation study to investigate and compare the performance of various implementations of the CIE in a wide range of scenarios. Next, we compared the CIE implementations in a real data setting. This illustration, based on data from the PROspective Québec (PROQ) Study on Work and Health led by Brisson,¹¹ concerns the association between body mass index and diastolic blood pressure.

2 Simulation study

2.1 Simulation scenarios

The scenarios we have considered were inspired by the fourth data-generating process in Talbot et al.,¹² but feature more covariates. Let L_1, L_2, \dots, L_{30} represent a set of 30 potential confounders, X the exposure of interest and Y the outcome. For all scenarios, we have first simulated L_6, L_7, \dots, L_{30} as correlated normal variables with mean = 0, variance = 1, and correlations = ρ . Covariates L_1, L_2, \dots, L_5 were then independently generated as

normal variables with mean = $L_{11} + L_{12} + L_{13} + L_{14} + L_{15}$ and variance = 1. X was generated as a normal variable with mean = $L_{11} + L_{12} + \dots + L_{20}$ and variance = 1.

We have considered scenarios where Y was a continuous, binary, or time to event variable. When Y was continuous, it was generated as a normal variable with mean = $0.1L_1 + 0.1L_2 + \dots + 0.1L_{10} + \beta X$ and variance = 1. When Y was binary, $P(Y = 1) = \text{expit}(0.1L_1 + 0.1L_2 + \dots + 0.1L_{10} + \beta X)$, where $\text{expit}(a) = \exp(a)/[1 + \exp(a)]$. When Y was a time to event variable, it was generated as an exponential variable with rate = $\exp(-5 + 0.1L_1 + 0.1L_2 + \dots + 0.1L_{10} + \beta X)$. In these latter scenarios, a random time to censoring variable, C , was also generated according to a log-normal distribution with mean on the log scale = $\log(5)$ and standard deviation on the log scale = $\log(1.5)$. The observed follow-up time was equal to the minimum between Y and C , and observations for which $Y > C$ were treated as right censored. The coefficient β represents the true exposure effect: the mean difference when Y was continuous, the log-odds ratio when Y was binary, and the log-hazard ratio when Y was a time to event.

A causal diagram representing the relationships between the variables is presented in Figure 1. If this diagram was known to the investigator, confounder selection could be performed by ensuring that all backdoor paths from X to Y are blocked (see the appendix of VanderWeele and Shpitser¹³ for an introduction to causal diagrams). For instance, adjusting for either $\{L_1, L_2, \dots, L_5\}$ or $\{L_{11}, L_{12}, \dots, L_{15}\}$ is sufficient to eliminate confounding bias. However, we are interested in a situation where the causal graph is unknown and the investigator is only able to identify $\{L_1, L_2, \dots, L_{30}\}$ as “potential confounders.” That is, the investigator is able to identify the preceding variables as potential risk factors for the exposure or the outcome but is unable to clarify their exact role. Some variables may have been excluded based on substantive knowledge, such as mediators of the effect of X on Y and colliders, or some variables that are assuredly only associated with the exposure. While researchers would generally be able to identify at least some variables as definite confounders (for example, age or sex) and would always adjust for such variables, our previous review revealed that it is not uncommon in practice to include all covariates in the variable selection procedure.¹ The goal was thus to investigate the ability of different CIE implementations to select a subset of these potential confounders that unbiasedly estimate the causal effect. It would typically be expected that the exposure estimate based on the subset is more precise than that based on the model that include all potential confounders, since a more parsimonious model is employed.

A total of 54 different simulation scenarios were constructed by considering all possible combinations of the following factors: (1) sample size of $n = 500$ or $n = 1000$, (2) correlations ρ between L_6, L_7, \dots, L_{30} of either 0, 0.2 or 0.5, (3) effect of X on Y of either $\beta = 0, 0.1$ or 0.5 , and (4) type of outcome as either continuous, binary or time to event. We had initially planned to also consider a sample size of $n = 200$, but this was abandoned due to frequent convergence issues when the outcome was binary or time to event.

2.2 Change in estimate implementations

We have considered six different implementations of the CIE. These six implementations are first described generally, then details specific to the type of outcome are presented.

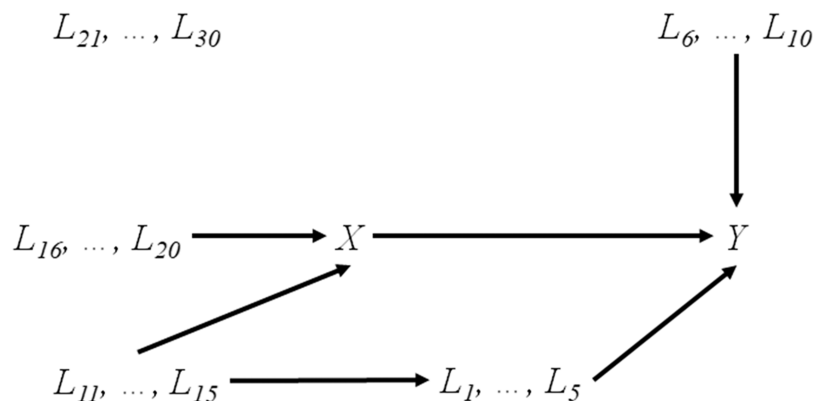


Figure 1. Causal diagram depicting the relationships between the variables in the simulation study. Arrows between groups of variables indicate that each variable of one group is causally affecting each variable in the second group. Variables L_6, \dots, L_{30} are correlated in some scenarios (due to external/unobserved common causes).

Backward – standard (BS). We have first fitted a model for the outcome according to the exposure and all covariates, and computed the effect estimate from this model. All covariates were then considered for exclusion, one at a time, and the relative difference in the effect estimate between the fully adjusted model and the model with one fewer variable was computed. The covariate whose exclusion altered the least the effect estimate was effectively excluded from the model. This process was repeated until it was impossible to exclude a covariate without changing the estimate by more than a pre-specified threshold as compared to the estimate from the fully adjusted model.

Backward – P values (BP). This implementation only differs from BS in that the covariate to be excluded at each step was the one whose associated *P* value was the largest.

Backward – confidence intervals (BC). We have started by fitting a model for the outcome according to the exposure and all potential confounders and determined the lower and upper bounds of the 95% confidence interval for the effect estimate. Next, covariates were considered for exclusion, one at a time. For each candidate, we calculated the relative change in each bound of the effect estimate's confidence interval following the exclusion of the covariate and computed the maximum of these two changes. The covariate whose exclusion altered the least both bounds – the one with the smallest maximum change in bounds – was effectively excluded. This exclusion procedure was repeated until the maximum change in bounds of all candidates for exclusion was larger than a pre-specified threshold, comparing to the bounds of the fully adjusted model. Some authors have proposed that such an implementation may be superior to those focusing on estimates, because the confidence interval is usually the final product of an analysis.³

Backward – mean squared error (BM). Again, a model for the outcome according to the exposure and all potential confounders was first fitted. The exposure coefficient and its estimated standard error are computed. The mean squared error (MSE) of this initial model is estimated as the square of the exposure coefficient's standard error. Then, covariates are considered for exclusion, one at a time. For each candidate, we computed the estimated MSE as the square of the difference between the exposure coefficient from the reduced model and that of the initial model, plus the square of the exposure coefficient's standard error in the reduced model. The covariate whose exclusion yielded the lowest estimated MSE was effectively excluded. The exclusion procedure stopped once excluding any of the candidates for exclusion increased the estimated MSE as compared to the one in the previous step. This is a slightly adapted version of the procedure proposed by Greenland et al.,¹⁴ which is designed to focus on accurate exposure effect estimation, as measured by the MSE.

Forward – crude (FC). We have first fitted a model for the outcome according to the exposure only and calculated the exposure effect estimate. All covariates were then considered for inclusion, one at a time. The covariate whose inclusion altered the most the estimate was effectively included. The estimate from the model adjusting for one additional covariate became the new comparator. This inclusion process was repeated until all candidates for inclusion altered the effect estimate by less than a pre-specified threshold.

Forward – partial (FP). This implementation only differs from FC in that the initial model was adjusted for L_1 , L_2 , L_3 , L_6 , L_7 . This implementation seeks to imitate a situation where the investigator is able to identify some confounders and risk factors of the outcome based on prior knowledge but is unsure about the status of the other potential confounders.

For all implementations, the model was a linear regression when Y was continuous, a logistic regression when Y was binary, and a Cox regression when Y was a time to event. For all implementations, except BM, the exposure effect estimate for determining the change in estimate was a mean difference, an odds ratio or a hazard ratio, when Y was continuous, binary or a time to event, respectively. For BM, the change in MSE was based on the regression coefficient, regardless of the type of outcome. Three different changes in estimate thresholds were used for each implementation, except BM: 1%, 5% and 10%. For BM, there was no threshold. Percentages following implementation abbreviations are henceforth used to indicate the threshold.

2.3 Analysis

For each scenario, 1000 datasets were generated. The exposure effect was estimated with each of the 16 combinations of CIE implementation and threshold value, as well as with an unadjusted model and a fully adjusted model. For each of these 18 analysis methods (16 CIE implementations, unadjusted model and fully adjusted model), we first computed bias as the difference between the estimated exposure coefficient and the true exposure effect coefficient. For scenarios with a continuous outcome, this true exposure coefficient is the value of β in the fully adjusted model. However, in scenarios with a binary outcome or a time to event outcome, the true exposure effect coefficient depends on the covariates that are selected, because of the non-collapsibility of the odds ratio

and hazard ratio. To estimate the true effect, we thus simulated a large sample of 1,000,000 observations where none of the covariates affected the exposure, but all other data-generating equations were the same as in the corresponding scenario. This allowed simulating a very large randomized experiment. The true effect, for a given set of covariates, was then estimated as the exposure coefficient in the regression of the outcome on the exposure and selected covariates in this randomized experiment. For each scenario and analysis method, we also estimated the standard error (SE) as the standard deviation of the exposure coefficient estimates, and the proportion of the time the 95% confidence intervals included the true exposure coefficient (cover) across the 1000 simulated datasets. The root-mean-squared-error (RMSE) was computed as the square root of the sum of the squared bias and the squared SE. For each method, we compared the RMSE to the one of the fully adjusted model to determine if the confounder selection was able to improve the accuracy of estimates (RMSE ratio). We also determined the proportion of the time that each covariate was included by each CIE implementation in each scenario. We used this to calculate the proportion of the time a set sufficient to control confounding bias was selected (sufficient sets must at least include either $\{L_1, L_2, \dots, L_5\}$ or $\{L_{11}, L_{12}, \dots, L_{15}\}$). Also, adjustment for $\{L_{16}, L_{17}, \dots, L_{20}\}$ may be particularly harmful to estimation, since these variables are so-called instruments (variables only associated with exposure). Adjusting for instruments is susceptible to inflate the variance and bias.¹⁵ We thus computed the proportion of the time that each instrument was included and report the average across all instruments. Similarly, we computed the average proportion of inclusion of other variables, that is, those that are neither part of a sufficient set, nor an instrument.

3 Results

The Monte Carlo standard error was less than 0.008 for estimating bias, 0.006 for estimating standard error, and 0.016 for estimating coverage.¹⁶ The results did not vary much according to sample size and amount of correlation between covariates. Therefore, only tables presenting the results for $n = 500$ and $\rho = 0.2$ are included in the manuscript. Tables reporting the results of the other scenarios are devolved to online supplemental material.

3.1 Continuous outcome

Table 1 and Web Tables 1–5 summarize the results of the scenarios with a continuous outcome. When $\beta = 0$, all methods except the unadjusted model produced unbiased estimates. While all CIE have much lower bias than the unadjusted model, FC10% and FP10% introduced substantial bias as compared to the fully adjusted model in many scenarios when $\beta = 0.1$. The other CIE methods yielded estimates with little or no bias. BS10%, BP10%, FC5%, FC10%, FP5% and FP10% had substantial bias in many scenarios when $\beta = 0.5$, but the other methods remained essentially unbiased. Bias tended to increase in scenarios with greater correlations between covariates (ρ).

All CIE implementations produced confidence intervals that included the true effect in less than 90% of replications in at least some scenarios, except for BC1% and BC5%. Overall, the coverage of confidence intervals tended to be lower when the true effect increased, when a larger threshold was used or for the larger sample size. Forward implementations generally had confidence intervals with the lowest coverage, sometimes as low as 0%.

Only the BM implementation achieved a modest RMSE reduction as compared to the fully adjusted model, around 10%, in all scenarios. When $\beta = 0$, all other methods had a negligible impact on the RMSE. When $\beta = 0.1$, BS, BP and BC implementations as well as FC1%, and FP1% also produced RMSE similar to those of the fully adjusted models. FC5%, FC10%, FP5% and FP10% yielded estimates with a lower RMSE than the fully adjusted model in some scenarios and with a larger RMSE in others. When $\beta = 0.5$, BS1%, BP1%, and BC1% produced a RMSE comparable to the one of the fully adjusted model. BP5%, BC5%, BC10%, FC1%, and FP1% allowed some reduction of the RMSE in all scenarios, especially BC10%. The other CIE implementations had a variable effect on the RMSE, sometimes producing reduced RMSE and sometimes increased RMSE. FC5% and FC10% were particularly prone to yield greatly increased RMSE, whereas FP5% and FP10% had an RMSE much smaller than the fully adjusted model in most scenarios.

3.2 Binary outcome

The results for the scenarios with a binary outcome are presented in Table 2 and Web Tables 6–10. When $\beta = 0$, only BS1%, BP1%, BC1%, BC5%, BC10% and BM produced estimates with low bias in all scenarios. When $\beta = 0.1$, only BC5% and BC10% had low bias overall. For $\beta = 0.5$, all methods had substantial bias in most

Table 1. Results of scenarios with continuous outcome, $n = 500$ and $\rho = 0.2$.

Method	$\beta = 0$				$\beta = 0.1$				$\beta = 0.5$			
	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio
Crude	.276	.011	0.0	4.41	.276	.011	0.0	4.41	.276	.011	0.0	4.41
Full	.001	.046	94.9	1.00	.001	.046	94.9	1.00	.001	.046	94.9	1.00
BS1	.001	.046	93.2	1.00	.001	.046	92.0	1.00	.002	.046	85.2	1.00
BS5	.001	.046	89.1	1.00	.002	.046	82.3	1.01	.013	.043	66.7	0.97
BS10	.002	.046	85.2	0.99	.003	.046	74.0	0.99	.031	.044	54.0	1.16
BPI	.001	.046	93.7	1.00	.001	.046	93.0	1.00	.001	.046	89.7	1.00
BP5	.001	.046	91.0	1.00	.001	.046	90.0	0.99	.007	.040	73.2	0.88
BP10	.001	.045	88.8	0.99	.000	.044	86.7	0.95	.032	.040	43.5	1.11
BC1	.001	.046	94.9	1.00	.001	.046	94.8	1.00	.002	.046	95.3	1.00
BC5	.001	.046	94.9	1.01	.001	.046	95.2	1.00	.001	.045	93.0	0.98
BC10	.001	.047	94.9	1.01	.001	.046	95.2	1.00	-.002	.042	86.3	0.90
BM	.002	.040	79.9	0.87	.002	.040	79.9	0.87	.002	.040	79.9	0.87
FC1	.001	.046	94.6	1.00	.001	.045	94.6	0.98	.004	.042	89.6	0.91
FC5	.001	.044	93.3	0.95	.009	.046	79.2	1.01	.056	.023	12.5	1.31
FC10	.002	.044	88.7	0.96	.048	.035	25.0	1.28	.082	.015	0.1	1.80
FPI	.001	.046	94.3	1.00	.001	.045	94.6	0.98	.004	.041	89.7	0.90
FP5	.001	.044	93.0	0.95	.006	.043	84.4	0.95	.029	.018	58.1	0.74
FP10	.002	.043	89.0	0.94	.028	.036	54.2	0.98	.029	.017	58.9	0.73

Crude: Unadjusted model; Full: fully adjusted model; BS: backward – standard; BP: backward – P values; BC: backward – confidence intervals; BM: backward – MSE; FC: forward – crude; FP: forward – partial; %: change in estimate threshold; SE: Monte Carlo standard error; Cover: proportion of 95% confidence intervals that included the true effect; RMSE ratio: root mean squared error of the analysis method/root mean squared error of the fully adjusted model.

Table 2. Results of scenarios with binary outcome, $n = 500$ and $\rho = 0.2$.

Method	$\beta = 0$				$\beta = 0.1$				$\beta = 0.5$			
	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio
Crude	.243	.026	0.0	1.48	.269	.033	0.0	1.46	.380	.068	0.0	1.09
Full	-.003	.126	94.2	1.00	.005	.142	93.7	1.00	.095	.211	91.1	1.00
BS1	-.002	.124	85.0	0.99	.005	.141	83.0	0.99	.094	.210	79.0	0.99
BS5	.015	.112	65.3	0.90	.020	.127	57.7	0.90	.096	.194	51.5	0.93
BS10	.046	.110	57.4	0.95	.049	.122	49.8	0.92	.117	.173	34.0	0.90
BPI	-.003	.125	89.4	1.00	.005	.141	89.4	1.00	.095	.210	84.7	1.00
BP5	.007	.111	77.3	0.88	.012	.127	74.1	0.90	.093	.200	71.9	0.95
BP10	.031	.095	67.0	0.79	.034	.105	59.8	0.78	.096	.173	55.8	0.86
BC1	-.003	.126	93.8	1.00	.004	.142	93.1	1.00	.095	.211	89.2	1.00
BC5	-.003	.119	94.3	0.95	.001	.134	92.9	0.95	.087	.207	88.3	0.97
BC10	-.002	.117	89.9	0.93	.002	.132	88.9	0.93	.072	.204	86.7	0.94
BM	-.001	.111	75.2	0.89	.007	.124	73.4	0.87	.088	.186	65.1	0.89
FC1	-.001	.113	90.6	0.90	.004	.128	88.4	0.90	.091	.199	87.4	0.95
FC5	.076	.042	53.8	0.69	.077	.054	57.5	0.66	.111	.107	63.9	0.67
FC10	.071	.035	58.4	0.63	.068	.042	62.5	0.56	.079	.084	72.5	0.50
FPI	-.001	.113	91.4	0.90	.004	.128	89.9	0.90	.094	.198	86.8	0.95
FP5	.033	.049	87.8	0.47	.036	.062	82.9	0.51	.078	.113	75.8	0.59
FP10	.029	.041	91.7	0.40	.030	.048	88.6	0.40	.053	.087	83.7	0.44

Crude: Unadjusted model; Full: Fully adjusted model; BS: backward – standard; BP: backward – P values; BC: backward – confidence intervals; BM: backward – MSE; FC: forward – crude; FP: forward – partial; %: change in estimate threshold; SE: Monte Carlo standard error; Cover: proportion of 95% confidence intervals that included the true effect; RMSE ratio: root mean squared error of the analysis method/root mean squared error of the fully adjusted model.

Table 3. Results of scenarios with time to event outcome, $n = 500$ and $\rho = 0.2$.

Method	$\beta = 0$				$\beta = 0.1$				$\beta = 0.5$			
	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio	Bias	SE	Cover	RMSE Ratio
Crude	.255	.033	0.0	1.02	.290	.032	0.0	1.31	.414	.043	0.0	1.16
Full	-.001	.201	91.9	1.00	.014	.172	90.9	1.00	.058	.127	89.6	1.00
BS1	-.001	.200	82.2	1.00	.014	.171	79.5	0.99	.058	.126	55.1	0.99
BS5	.010	.185	63.4	0.92	.025	.154	55.1	0.90	.065	.110	25.5	0.92
BS10	.035	.167	60.3	0.85	.052	.140	44.6	0.86	.090	.092	14.1	0.92
BP1	-.001	.201	91.9	1.00	.014	.172	88.3	1.00	.058	.127	29.9	1.00
BP5	-.002	.198	92.2	0.99	.011	.169	90.4	0.98	.052	.126	67.6	0.98
BP10	.009	.197	91.1	0.98	.029	.168	87.9	0.99	.078	.121	57.6	1.03
BC1	-.001	.201	91.7	1.00	.014	.172	90.0	1.00	.058	.127	75.4	1.00
BC5	-.003	.197	89.8	0.98	.011	.168	88.6	0.97	.047	.124	74.3	0.95
BC10	-.004	.187	88.8	0.93	.010	.161	87.0	0.93	.049	.121	59.1	0.94
BM	.002	.178	67.3	0.89	.016	.152	65.6	0.89	.057	.115	41.8	0.92
FC1	.000	.192	89.7	0.96	.012	.162	88.6	0.94	.057	.117	66.7	0.93
FC5	.074	.081	67.6	0.55	.076	.065	48.6	0.58	.059	.062	21.1	0.61
FC10	.066	.050	75.0	0.41	.065	.045	55.0	0.46	.097	.069	13.1	0.85
FPI	.001	.193	89.1	0.96	.014	.165	88.3	0.96	.056	.116	64.3	0.92
FP5	.037	.097	83.9	0.52	.040	.072	77.5	0.48	.037	.057	36.5	0.49
FPI0	.033	.060	89.0	0.34	.032	.051	83.1	0.35	.029	.049	39.1	0.41

Crude: Unadjusted model, Full: Fully adjusted model; BS: backward – standard; BP: backward – P values; BC: backward – confidence intervals; BM: backward – MSE; FC: forward – crude; FP: forward – partial; %: change in estimate threshold; SE: Monte Carlo standard error; Cover: proportion of 95% confidence intervals that included the true effect; RMSE ratio: root mean squared error of the analysis method/root mean squared error of the fully adjusted model.

Table 4. Proportion of simulation replicates in which a set sufficient to control confounding (S) was selected and average proportion of inclusion of each instrument (I) and other variable (O).

	Continuous			Binary			Time to event		
	S	I	O	S	I	O	S	I	O
BS1	0.53	0.93	0.73	0.27	0.94	0.22	0.22	0.92	0.37
BS5	0.41	0.71	0.41	0.17	0.32	0.00	0.14	0.34	0.01
BS10	0.36	0.56	0.27	0.11	0.12	0.00	0.09	0.13	0.00
BP1	0.64	0.79	0.81	0.44	0.98	0.99	0.26	0.37	0.37
BP5	0.53	0.68	0.82	0.11	0.52	0.69	0.44	0.88	0.88
BP10	0.43	0.56	0.69	0.03	0.17	0.28	0.22	0.91	0.91
BC1	0.66	0.98	0.77	0.50	1.00	0.58	0.50	1.00	0.87
BC5	0.53	1.00	0.52	0.49	1.00	0.04	0.46	1.00	0.05
BC10	0.50	0.85	0.28	0.46	0.89	0.00	0.42	0.79	0.00
BM	0.63	0.57	1.00	0.41	0.54	0.97	0.37	0.54	0.99
FC1	0.68	0.94	0.86	0.07	0.91	0.35	0.07	0.94	0.60
FC5	0.33	0.58	0.51	0.00	0.02	0.00	0.00	0.01	0.00
FC10	0.20	0.36	0.29	0.00	0.00	0.00	0.00	0.00	0.00
FPI	0.70	0.93	1.00	0.18	0.91	1.00	0.18	0.92	1.00
FP5	0.35	0.58	1.00	0.00	0.02	1.00	0.00	0.01	1.00
FPI0	0.21	0.37	1.00	0.00	0.00	1.00	0.00	0.00	1.00

S: Sufficient; I: instrumentals; O: others (not in a sufficient set and not an instrument); BS: backward – standard; BP: backward – P values; BC: backward – confidence intervals; BM: backward – MSE; FC: forward – crude; FP: forward – partial.

scenarios. FC5% and FC10% generally had the largest bias, although this bias was much lower than that of the unadjusted model. Of note, even the fully adjusted model produced biased estimates when $\beta = 0.1$ or $\beta = 0.5$. This is likely due to the finite sample bias associated with fitting a large model with a relatively small sample size, since the bias was lower for scenarios with $n = 1000$ than in scenarios with $n = 500$.

Table 5. Characteristics of the extracted sample from the PROspective Québec (PROQ) Study on Work and Health according to body mass index.

	BMI < median	BMI > median
Sex (women)	2033 (62.3)	1120 (34.3)
Age, mean (SD)	45.3 (7.5)	49.0 (8.5)
Income, mean (SD)	53.1 (18.4)	55.8 (17.4)
Hours of work per week		
≤20 h/week	16 (0.5)	10 (0.3)
21–34 h/week	193 (5.9)	153 (4.7)
35–40 h/week	2877 (88.2)	2800 (85.8)
≥41 h/week	177 (5.4)	300 (9.2)
Education		
High school or less	982 (30.1)	827 (25.3)
College	983 (30.1)	895 (27.4)
University	1298 (39.8)	1541 (47.2)
Occupation		
White collars	1167 (35.8)	809 (24.8)
Technicians	725 (22.2)	656 (20.1)
Professionals	1067 (32.7)	1258 (38.6)
Managers / Directors	225 (6.9)	454 (13.9)
Others	79 (2.4)	86 (2.6)
Job strain	686 (21.0)	680 (20.8)
Hypertension or hypertensive medication	328 (10.1)	836 (25.6)
Diabetes	40 (1.2)	71 (2.2)
Family history of cardiovascular disease	1036 (31.7)	1177 (36.1)
Physical activity, mean (SD)	5.3 (4.3)	5.2 (4.4)
Alcohol, mean (SD)	2.8 (4.01)	3.6 (5.20)
Smoking	809 (24.8)	621 (19.0)

Note: All results are *n* (%), unless otherwise indicated, SD: Standard deviation, BMI: body mass index.

Most methods had coverage rate below 90% in most scenarios. BC1% and BC5% had close to appropriate coverage in all scenarios with either $\beta = 0$ or 0.1 and in all but one scenario with $\beta = 0.5$. While the fully adjusted model had important bias in some circumstances, as previously noted, its coverage remained adequate in all scenarios.

BS1%, BP1%, BC1% and BC5% had an RMSE similar to the one of the fully adjusted model in all scenarios. BS10%, FC5% and FC10% had mixed results, sometimes increasing, and sometimes decreasing the RMSE. BP10%, BM, FP5% and FP10% were the implementations that most consistently decreased the RMSE across scenarios. FP5% and FP10% were the most susceptible to yield an important decrease in the RMSE.

3.3 Time to event outcome

The results for the scenarios with a time to event outcome are displayed in Table 3 and Web Tables 11–15. In scenarios with $\beta = 0$, most CIE methods had low bias, except BS10%, FC5%, FC10%, FP5% and FP10% which had notable bias in at least some cases. In contrast, when $\beta = 0.1$ or 0.5, all methods had important bias in at least some scenarios, including the fully adjusted model. The bias was smaller for the larger sample size, again suggesting a finite sample bias.

The coverage of 95% confidence intervals was below 90% in at least some scenarios for all methods. Only the fully adjusted model had close to appropriate coverage in all scenarios.

As in the binary outcome scenarios, BS1%, BP1%, BC1% and BC5% had an RMSE similar to the one of the fully adjusted model. BS10%, BP5%, BP10%, FC5% and FC10% increased the RMSE by more than 10% in one or more of the scenarios we considered. FP5% and FP10% were the methods that most consistently decreased the RMSE and that induced the greatest reduction in the RMSE.

3.4 Covariates inclusion

Table 4 presents the proportion of inclusion of covariates for each CIE implementation across simulation scenarios. The proportion of replicates in which a set sufficient to control confounding was effectively selected was less than 70% for all methods. This proportion was generally greater for smaller thresholds than for larger ones. On the other hand, most methods included instruments relatively often, sometimes as much as including all instruments in 100% of the replicates. This was particularly the case for smaller thresholds. Similarly, most methods also often included variables that are neither confounders nor instruments (“other” variables), especially with the 1% threshold. These results indicate that none of the implementation adequately differentiated confounders and non-confounders in our simulation study.

3.5 PROspective Québec (PROQ) study on work and health

To illustrate the change in estimate method in a real data setting, we explored the association between body mass index (BMI) and diastolic blood pressure (BP) in the PROspective Quebec (PROQ) Study on Work and Health. We note it would be possible (in fact, preferable) to instead draw a causal graph based on our subject-matter knowledge. However, since our goal is to illustrate and compare the different CIE implementations, we applied a naïve approach where all variables that were identified as potential confounders based on substantive knowledge are included in the variable selection procedures.

In Quebec City (QC, Canada) 9188 white-collar workers (48.5% women) enrolled in the PROQ cohort, were followed over a period of 25 years.¹¹ At recruitment, in 1991–1993, participants worked in one of 19 public and parapublic organizations in the Quebec City region. The eligibility criteria were to work at least 21 hours a week and not to hold another paid job of more than 10 h a week. The PROQ cohort had a participation rate of 75%. A first follow-up occurred in 1999–2001 with a participation rate of 89%. A second follow-up happened in 2015–2018, but is not considered in this analysis. This study was approved by the CHU de Québec – Université Laval’s ethical review board (#2012-1674).

BMI, a measure of body fat based on height and weight, has been positively associated with systolic and diastolic BP as well as with hypertension.^{17–20} The reduction of BMI combined with lifestyle modification are preferred strategies to reduce BP.^{21–23} Weight management strategies are supported by prospective evidence reporting that BMI reduction is associated with BP reduction, implying a causal relationship.^{18,24,25} Since BMI and BP are important cardiovascular risk factors,²⁶ this relationship has important implication, especially in population with high obesity rates.

We estimated the association between BMI at baseline and diastolic BP at the first follow-up using an unadjusted model, a fully adjusted linear regression, as well as the 16 CIE implementations considered in the previous section. Based on subject-matter knowledge, a set of 13 potential confounders measured at baseline was considered. This set included sex, age, income (in 1000 Canadian dollars), hours of work per week, education, occupation, exposure to job strain (a psychosocial work stressor), hypertension diagnosis or use of hypertension medication, self-reported diabetes, family history of cardiovascular disease, frequency of 20–30 minutes physical activity per month, number of alcohol drinks per week and current smoking (yes or no). Weight (kg) and height (cm) were measured by a trained research assistant. BMI was calculated by dividing a participant’s weight by their height in metres squared. BP was measured according to recognized protocol.²⁷ In brief, participant’s BP was measured at rest after they had been sitting for 5 min. The average of three BP measurements taken 1–2 minutes apart was recorded. More information is available elsewhere.¹¹ For the FP implementation, age and sex were forced to be included. To simplify the illustration, we considered a subsample of 6526 participants without missing data on any of the considered variables.

Table 5 provides descriptive statistics on these data. Among others, the proportion of female participants was lower among those with a BMI over the median (24.1 kg/m²) than those with a BMI below the median. Those with a greater BMI were also older, more likely to have a university diploma, suffered from hypertension or diabetes in a greater proportion, drank more alcohol and were less likely to be smokers. The linear association between BMI (in kg/m²) and diastolic BP (in mm Hg) estimated using the different methods is reported in Table 6. In this illustration, all methods produced similar point estimates, wherein each increase of 1 kg/m² of BMI was associated with an increase of approximately 0.78 mm Hg of diastolic BP. Interestingly, all methods yielded virtually identical standard errors (0.042). Hence, in this example, employing the CIE to select potential confounders did not improve the apparent precision of the estimation. The covariates selected by the CIE implementations were also

Table 6. Estimate and selected covariates for the association between body mass index and diastolic blood pressure in the PROspective Québec (PROQ) Study on Work and Health according to change-in-estimate implementation.

Method	Estimate (SE)	Selected covariates												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Crude	1.411 (0.044)													
Full	0.774 (0.042)	X	X	X	X	X	X	X	X	X	X	X	X	X
BS1%	0.776 (0.042)	X	X			X			X					
BS5%	0.787 (0.042)	X	X						X					
BS10%	0.787 (0.042)	X	X						X					
BP1%	0.773 (0.042)	X	X	X	X	X		X	X	X	X	X	X	X
BP5%	0.788 (0.042)	X	X						X			X	X	X
BP10%	0.788 (0.042)	X	X						X			X	X	X
BC1%	0.776 (0.042)	X	X			X			X					
BC5%	0.787 (0.042)	X	X						X					
BC10%	0.787 (0.042)	X	X						X					
BM	0.776 (0.042)	X	X	X	X	X		X	X		X		X	
FC1%	0.776 (0.042)	X	X			X			X					
FC5%	0.787 (0.042)	X	X						X					
FC10%	0.787 (0.042)	X	X						X					
FPI%	0.776 (0.042)	X	X			X			X					
FP5%	0.787 (0.042)	X	X						X					
FP10%	0.787 (0.042)	X	X						X					

Crude: Unadjusted model, Full: fully adjusted model; BS: backward – standard; BP: backward – P values; BC: backward – confidence intervals; BM: Backward – MSE; FC: forward – crude; FP: forward – partial; %: change in estimate threshold; SE: Standard error; 1 = sex; 2 = age; 3 = income; 4 = hours of work per week; 5 = education; 6 = occupation; 7: job strain; 8 = hypertension diagnosis or use of hypertensive medication; 9 =diabetes; 10 = family history of cardiovascular disease; 11 = physical activity; 12 = alcohol; 13 = smoking.

very similar. Age, sex and diagnosis of hypertension were always selected, and education was additionally included when a 1% threshold was used. Only the BP and BM implementations included further covariates.

4 Discussion

We have conducted an extensive simulation study that aimed at addressing a gap in knowledge regarding the performance of the CIE for selecting potential confounders. To the best of our knowledge, this is the first simulation study to consider scenarios with multiple potential confounders, with a continuous or a time to event outcome and to investigate the performance of forward inclusion and backward exclusion based on changes in confidence intervals implementations of the CIE.

In summary, we have observed that all CIE implementations are at risk of introducing substantial bias and to produce confidence intervals that include the true effect much less often than expected. Forward inclusion methods produced particularly poor results in terms of bias and coverage. Our results also indicate mixed performance of many CIE implementations in terms of reducing the RMSE. While most methods were able to achieve at least some reduction of the RMSE as compared to the fully adjusted models in some scenarios, the reduction was often modest and an increase in the RMSE could also occur. In such cases, the data-driven selection of confounders was harmful. In this regard, our results are similar to those of previous simulation studies.^{7–9} Forward CIE implementations were the most susceptible to substantially reduce the RMSE but were also the most likely to introduce important bias and yield invalid inferences. We have also observed that all CIE implementations failed to adequately differentiate confounders from non-confounders. The methods that were the most likely to include confounders were also the most likely to include instruments, which are particularly harmful for effect estimation. In the real data illustration, no reduction of the estimated standard error was observed, regardless of the implementation or threshold.

A limitation to consider when interpreting these results is that they arose from a synthetic data simulation study. Although a total of 54 different scenarios were considered, alternative scenarios may yield different results. Moreover, it is difficult to adequately replicate the complexity of real data using synthetic data simulations. Future studies may consider plasmode simulations,²⁸ which combine real and synthetic data, to address this

issue. Another limitation concerns the non-collapsibility of the odds ratio and the hazard ratio. When non-collapsible effect measures are used, the change in estimate between different covariate adjustment sets may reflect non-collapsibility in addition to confounding. We attempted to mitigate this issue by estimating a true effect specific to the variables that were included in the final model. It would have also been possible to employ adjustment methods that circumvent the non-collapsibility issue by estimating marginal effects instead of conditional ones. When estimating a marginal effect, the effect measure is no longer conditional on the covariates that are included in the model, thus bypassing the non-collapsibility issue. Such methods notably include inverse probability weighting, standardization/g-formula, augmented inverse probability weighting and targeted maximum likelihood estimation.^{29,30} We did not consider these methods in the current study in an attempt to evaluate the CIE methods as they are currently being implemented in practice.

Despite these limitations, our results provide important insights concerning the CIE for selecting confounders. Considering the low ability of all CIE methods we have explored to yield unbiased estimates with improved precision and their inability to identify true confounders or valid inferences, there seems to be little or no benefit in employing any of the CIE implementation. An important point to consider is that CIE may give a false impression of improved precision when the estimated standard error is reduced after variable selection. However, this estimated standard error is invalid and underestimates the true standard error, since it does not account for the variability associated with the variable selection. Unfortunately, adequately accounting for the variable selection when estimating the standard error and producing confidence intervals is theoretically challenging.^{31,32} Although the usual bootstrap has been proposed,¹⁴ this solution is not theoretically supported when the post-selection estimator is insufficiently smooth.³³

More sophisticated data-driven methods for confounder selection have been developed in recent years. Some of these methods address multiple of the shortcomings of the CIE that were observed in our study. For instance, they target an unbiased estimation of the casual effect with improved precision as compared to a fully adjusted model, and offer theoretical guarantees, under some assumptions, concerning the identification of true confounders in large sample sizes.^{12,34-40} However, causal thinking is always essential to adequately control confounding, notably to avoid adjusting for colliders or for variables lying on the causal pathway of interest (mediators) and to ensure that all known confounders are adjusted for.

Multiple tools and methods have been proposed to facilitate knowledge-based selection of covariates.^{5,13,41,42} Sensitivity analyses can also be used to explore the robustness of results to different choices of covariates when the role of some of them is unclear. As such, we believe that data-driven confounder selection should be considered as a potential complement to substantive knowledge only when it is unclear if some variables are true confounders or not; it should not be considered as a mandatory step of exposure effect estimation. Data-driven confounder selection methods may prove particularly helpful in new areas of research where expert knowledge is scarce, such as an emerging disease like COVID-19. When data-driven confounder selection seems warranted, we believe the CIE should be avoided since it offers little or no benefits and it can yield invalid estimates and inferences. Instead, the novel methods with a stronger theoretical background we mentioned earlier should be considered.

Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the Fonds de recherche du Québec – Santé [#265385 to DT] and the Natural Sciences and Engineering Research Council of Canada [# 2016-06295 to DT]. DT is a Fonds de recherche du Québec – Santé Chercheur-Boursier (FRQS). MLR was supported by a FRQS training award for health professionals. The illustration's data come from the PROspective Québec (PROQ) Study on Work and Health funded in large part by the Canadian Institutes of Health Research. Funding sources had no role in the design of the study, analysis and interpretation of the data, or in writing the manuscript.

ORCID iDs

Denis Talbot  <https://orcid.org/0000-0003-0431-3314>

Mathilde Lavigne-Robichaud  <https://orcid.org/0000-0002-5130-9118>

Supplemental material

Supplemental material for this article is available online.

References

1. Talbot D and Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *Eur J Epidemiol* 2019; **34**: 725–730.
2. Greenland S and Pearce N. Statistical foundations for model-based adjustments. *Annu Rev Public Health* 2015; **36**: 89–108.
3. Rothman KJ, Greenland S and Lash TL. *Modern epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins, 2008, p.x, 758 p.
4. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002; **155**: 176–184.
5. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol* 2019; **34**: 211–219.
6. Martinussen T and Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Anal* 2013; **19**: 279–296.
7. Maldonado G and Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993; **138**: 923–936.
8. Mickey RM and Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989; **129**: 125–137.
9. Weng H-Y, Hsueh Y-H, Messam LLM, et al. Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *Am J Epidemiol* 2009; **169**: 1182–1190.
10. Pearl J. *Causality: models, reasoning and inference*. New York, NY: Cambridge University Press, 2009, p.464.
11. Trudel X, Gilbert-Ouimet M, Milot A, et al. Cohort profile: the PROspective Quebec (PROQ) study on work and health. *Int J Epidemiol* 2018; **47**: 693–693i.
12. Talbot D, Lefebvre G and Atherton J. The Bayesian causal effect estimation algorithm. *J Causal Inference* 2015; **3**: 207–236.
13. VanderWeele TJ and Shpitser I. A new criterion for confounder selection. *Biometrics* 2011; **67**: 1406–1413.
14. Greenland S, Daniel R and Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *Int J Epidemiol* 2016; **45**: 565–575.
15. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol* 2011; **174**: 1223–1227.
16. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; **38**: 2074–2102.
17. Linderman GC, Lu J, Lu Y, et al. Association of body mass index with blood pressure among 1.7 million Chinese adults. *JAMA Netw Open* 2018; **1**: e181271-e181271.
18. Dua S, Bhuker M, Sharma P, et al. Body mass index relates to blood pressure among adults. *N Am J Med Sci* 2014; **6**: 89–95.
19. Gelber RP, Gaziano JM, Manson JE, et al. A prospective study of body mass index and the risk of developing hypertension in men. *Am J Hypertens* 2007; **20**: 370–377.
20. Shuger SL, Sui X, Church TS, et al. Body mass index as a predictor of hypertension incidence among initially healthy normotensive women. *Am J Hypertens* 2008; **21**: 613–619.
21. Nerenberg KA, Zarnke KB, Leung AA, et al. Hypertension Canada’s 2018 Guidelines for Diagnosis, Risk Assessment, Prevention, and Treatment of Hypertension in Adults and Children. *Can J Cardiol* 2018; **34**: 506–525.
22. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *Eur Heart J* 2018; **39**: 3021–3104.
23. Flack JM, Calhoun D and Schiffrin EL. The new ACC/AHA hypertension guidelines for the prevention, detection, evaluation, and management of high blood pressure in adults. *Am J Hypertens* 2017; **31**: 133–135.
24. Drøyvold W, Midthjell K, Nilsen T, et al. Change in body mass index and its impact on blood pressure: a prospective population study. *Int J Obes* 2005; **29**: 650–655.
25. Blumenthal JA, Sherwood A, Gullette ECD, et al. Exercise and weight loss reduce blood pressure in men and women with mild hypertension: effects on cardiovascular, metabolic, and hemodynamic functioning. *Arch Intern Med* 2000; **160**: 1947–1958.
26. O’Donnell CJ and Elosua R. Cardiovascular risk factors. Insights from Framingham Heart Study. *Rev Esp Cardiol (Engl Ed)* 2008; **61**: 299–310.
27. Frohlich ED. Recommendations for blood pressure determination by sphygmomanometry. *Ann Intern Med* 1988; **109**: 612–612.
28. Gadbury GL, Xiang Q, Yang L, et al. Evaluating statistical methods using plasmode data sets in the age of massive public databases: an illustration using false discovery rates. *PLoS Genet* 2008; **4**: e1000098.
29. Hernán MA and Robins JM. *Causal inference: what if*. Boca Raton, FL: Chapman & Hill/CRC, 2020.

30. Luque-Fernandez MA, Schomaker M, Rachet B, et al. Targeted maximum likelihood estimation for a binary treatment: a tutorial. *Stat Med* 2018; **37**: 2530–2546.
31. Leeb H and Pötscher BM. Can one estimate the unconditional distribution of post-model-selection estimators? *Econ Theory* 2008; **24**: 338–376.
32. Leeb H and Pötscher BM. Model selection and inference: facts and fiction. *Econ Theory* 2005: 21–59.
33. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc* 2014; **109**: 991–1007.
34. Shortreed SM and Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* 2017; **73**: 1111–1122.
35. Koch B, Vock DM and Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* 2018; **74**: 8–17.
36. Wang C, Dominici F, Parmigiani G, et al. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* 2015; **71**: 654–665.
37. Wang C, Parmigiani G and Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 2012; **68**: 661–671.
38. Ertefaie A, Asgharian M and Stephens DA. Variable selection in causal inference using a simultaneous penalization method. *J Causal Inference* 2018; **6**.
39. Cefalu M, Dominici F, Arvold N, et al. Model averaged double robust estimation. *Biometrics* 2017; **73**: 410–421.
40. van der Laan MJ and Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat* 2010; **6**.
41. Shrier I and Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008; **8**: 70.
42. Textor J, van der Zander B, Gilthorpe MS, et al. Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *Int J Epidemiol* 2016; **45**: 1887–1894.