European Journal of Epidemiology (2019) 34:725–730
https://doi.org/10.1007/s10654-019-00529-y

METHODS

A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement

Denis Talbot[1,2] · **Victoria Kubuta Massamba**[1,2]

Received: 9 October 2018 / Accepted: 24 May 2019 / Published online: 3 June 2019

Abstract

A review of epidemiological papers conducted in 2009 concluded that several studies employed variable selection methods susceptible to introduce bias and yield inadequate inferences. Many new confounder selection methods have been developed since then. The goal of the study was to provide an updated descriptive portrait of which variable selection methods are used by epidemiologists for analyzing observational data. Studies published in four major epidemiological journals in 2015 were reviewed. Only articles concerned with a predictive or explicative objective and reporting on the analysis of individual data were included. Method(s) employed for selecting variables were extracted from retained articles. A total of 975 articles were retrieved and 299 met eligibility criteria, 292 of which pursued an explicative objective. Among those, 146 studies (50%) reported using prior knowledge or causal graphs for selecting variables, 34 (12%) used change in effect estimate methods, 26 (9%) used stepwise approaches, 16 (5%) employed univariate analyses, 5 (2%) used various other methods and 107 (37%) did not provide sufficient details to allow classification (more than one method could be employed in a single article). Despite being less frequent than in the previous review, stepwise and univariable analyses, which are susceptible to introduce bias and produce inadequate inferences, were still prevalent. Moreover, 37% studies did not provide sufficient details to assess how variables were selected. We thus believe there is still room for improvement in variable selection methods used by epidemiologists and in their reporting.

Keywords Variable selection · Modeling · Confounding · Epidemiologic methods

Introduction

Variable selection is an inevitable step of epidemiological research. When new data are to be collected, researchers must identify which variables to measure as risk factors or as potential confounders. Similarly, when using available data, researchers are faced with the decision of which variables to consider in their analysis. Variable selection based on the observed data is also sometimes performed. This may be necessary when analysing small data sets or rare events, but can also prove useful in providing more precise estimates than a fully adjusted model (e.g. [1]).

Walter and Tiemeier [2] conducted a review of the variable selection methods used in epidemiological studies. They have reviewed all articles published in 2008 in four major epidemiologic journals (American Journal of Epidemiology, Epidemiology, European Journal of Epidemiology and International journal of Epidemiology), excluding commentaries, genetic association studies and meta-analyses. The authors have assessed which method had been used to select variables among the retained articles and concluded that "methods which have been formally criticized as flawed still prevail in the scientific literature." Using inappropriate variable selection methods can seriously impair the validity of results. Such methods can lead, for example, to seriously biased estimators of the causal exposure effect if the method fails to identify important confounders, imprecise inferences

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-019-00529-y) contains supplementary material, which is available to authorized users.

✉ Denis Talbot
denis.talbot@fmed.ulaval.ca

1 Département de médecine sociale et préventive, Faculté de médecine, Université Laval, 1050, avenue de la Médecine, Pavillon Ferdinand-Vandry, room 2454, Quebec, QC G1V 0A6, Canada

2 Unité santé des populations et pratiques optimales en santé, CHU de Québec – Université Laval Research Center, Quebec, QC, Canada

if too many covariates are selected, inaccurate *p* values and confidence intervals if inference procedures do not account for preliminary testing, or to decreased out-of-sample predictive performance [3–5].

Modelling guidelines in epidemiology have included using prior knowledge to help decide which variables to include in the model for many years (e.g., [6]). In explicative studies, for example, prior knowledge may help to avoid adjusting for variables that lie in the causal pathway between exposure and outcome [6]. Causal graphs are a way to summarize prior knowledge that may further help variable selection for explicative studies. The first introductory articles to causal graphs published in epidemiological journals would date to the late 1990's (e.g., [7]). A disjunctive cause criterion has been introduced in 2011 to perform confounder selection in situations where the causal graph is not fully known, but some knowledge is still available [8, 9].

Since prior knowledge is sometime insufficient to select variables a priori, data-driven approaches may be considered. One such approach is to first screen potential covariates using "univariable analyses." For example, one would measure the crude association between each potential covariate and the outcome and only retain variables that are statistically significant at some threshold in a multivariable model. This approach has the advantage of being simple and applicable even when the sample size is small or if the outcome is rare. However, it also has multiple shortcomings: coefficients of the selected variables tend to be upwardly biased, the predictive ability of the model may be overestimated, confidence intervals tend to be too short (they include the true value less often then they should) and true predictors may fail to be identified because of confounding [10–12].

Stepwise selection is another popular data-driven method. Starting with some initial model, stepwise selection entails iteratively including or excluding variables from the model until some stopping criterion is met. Stepwise selection shares both the advantages and the limitations of univariable screening (see [4] Sect. 4.3 and references therein). For predictive models, overestimation of the predictive ability may be corrected to some extent by computing bootstrap estimates of optimism [11].

When the goal of variable selection is to select confounders, the change in estimate approach is often seen as preferable to methods based on statistical significance [3]. The performance of the change in estimate method has been investigated in a few simulation studies [13–15]. Two of these considered scenarios with a single potential confounder and the other with nine potential confounders. Their results suggest that the change in estimate method may provide exposure effect estimates with small bias and appropriate coverage of confidence intervals when a low change in estimate threshold is used (e.g. 10%). However, this approach was also observed to yield little or no gain, or even a decrease, in the precision of estimates as compared to a fully adjusted model. Furthermore, there are concerns regarding the validity of the change in estimate method when considering non-collapsibility effect measures such as the odds ratio [3, 11, 16].

There are also more sophisticated approaches to variable selection that are less common in epidemiology, such as the lasso, the adaptive lasso and Bayesian model averaging [17–19]. Lasso approaches are regression methods that shrink toward zero the coefficients of variables that not associated with the outcome, whereas Bayesian model averaging produces inferences by averaging coefficients or predictions across the multiple possible models. The lasso and Bayesian model averaging have been shown to be superior to methods based on statistical significance for constructing prediction models [17, 19, 20]. Nonetheless, these methods have also been observed to perform poorly for selecting confounders [21–24]. This is because they focus on modeling the outcome as a function of the exposure and potential confounders. Important confounders that are weakly associated with the outcome but strongly associated with the exposure may fail to be adjusted for by these methods. Methods tailored for confounder selection have emerged in the last ten years [1, 21–32]. These methods account for the fact that confounders are associated with both the exposure and the outcome in performing confounding adjustment. For example, the outcome-adaptive lasso is based on exposure (propensity score) modeling to adjust for confounding and shrinks covariates' coefficients toward zero if they are not associated with the outcome. As a results, adjustment tends to be made for true confounders and risk factors of the outcome. This results in estimates that are adequately adjusted for confounding, while being more precise than those of a fully adjusted model or of a model that only includes true confounders [24].

Considering that the previous review of variable selection methods used in epidemiological studies was published in 2009 and that new methods have been introduced since then, we decided to conduct a new review of variable selection methods. More precisely, the main objective of the current study was to provide an updated descriptive portrait of the variable selection methods used in explicative and predictive studies of observational data published in epidemiological journals.

Methods

Our methodology is adapted from the one employed by Walter and Tiemeier in their 2009 review [2]. We have performed a descriptive review of papers published in 2015 in four major epidemiologic journals: American Journal of Epidemiology, Epidemiology, European Journal of

Epidemiology and International Journal of Epidemiology. Articles were initially scanned by one author for eligibility according to predefined inclusion/exclusion criteria. This author indicated in an electronic table whether each article was deemed eligible or not, as well as reasons for exclusion when applicable. The table was revised by the other author and disagreements were resolved by consensus. For each eligible article, the following data were collected by one author (1) name of the journal, (2) primary objective (predictive or explicative) (3) method(s) employed for selecting variables (prior knowledge/causal graphs, change in effect estimate, stepwise regression, univariate analyses, other, not specified), (4) excerpts of the text indicating which method was used. The other author revised the extracted data. Because appropriateness of variable selection methods depends on the objective of the study, explicative and predictive studies were analyzed separately. The data that were extracted are available as online supplementary material.

Inclusion and exclusion criteria

Only articles reporting on the analysis of individual data from observational studies were included. For example, commentaries, replies, study design reports, systematic reviews, and studies performing a meta-analysis of aggregate data were excluded. Studies whose primary objective was neither explicative nor predictive, such as descriptive studies or validation studies, were also excluded. Randomized controlled trials, case-crossover and other studies that control confounding bias through design as well as essentially methodological articles were excluded, because employing variable selection methods isn't as relevant in such studies as it is in substantive studies of observational data. All designs that did not control for confounding were considered eligible, for example cross-sectional, prospective or retrospective cohort and case–control studies. We have further excluded studies that only employed instrumental variable methods and genetic association studies because of the particularities of such applications.

Results

A total of 975 articles were retrieved, of which 299 were included in our review. Agreement between authors regarding inclusion of articles was excellent (agreement = 96%, $\kappa = 0.91$). The most common reason for exclusion was not reporting on an analysis of individual data (n = 522, 77%), followed by studies whose design controlled for confounding and methodological articles (n = 76, 11%), studies whose objectives were neither explicative nor predictive (n = 48, 7%), and studies that employed instrumental variable methods or genetic association studies (n = 30, 4%). Web Table 1 details articles selection by journal.

Out of the 299 included studies, only 7 were categorized as having a predictive objective. Among those, 3 have selected variables solely based on prior knowledge, 2 used both prior knowledge and stepwise approaches, 1 used only a stepwise approach and 1 didn't provide sufficient information regarding variable selection to allow classification. Of note, the study that selected predictors only with a stepwise approach utilized a bootstrap procedure to obtain an optimism-corrected estimate of the predictive performance of their model [33].

Among the 292 included studies that pursued an explicative objective, 116 (40%) were identified as employing prior knowledge or causal graphs only for selecting variables. An additional 30 studies utilized prior knowledge or causal graphs together with data-driven approaches, for a total of 146 (50%) studies that explicitly based their variable selection procedure on prior knowledge.

A total of 69 explicative studies mentioned utilizing data-driven methods. The change in estimate approach was the most common data-driven variable selection approach (34, 12% of all studies). Various implementations of this approach were reported, but the information provided was generally insufficient to allow replication. For instance, 18 studies used a 10% change in estimate threshold for deciding whether a covariate was a confounder or not, 6 studies utilized thresholds varying between 1% and 20% and 10 studies did not provide a specific threshold. Furthermore, in 19 studies, the comparator was not clearly identified, in 12 studies the comparator was obtained from a partially adjusted model and in 3 studies it was the crude estimate.

Univariable analyses were used in 26 studies (9%), stepwise methods in 16 studies (5%) and one study used both methods together. Other methods were used in five studies: one used a Bayesian approach [34], one used an approach both based on model fit and change in estimate [35], one used an approach that combined prior knowledge, change in estimate and statistical significance [36], one used an approach based on the deviance information criterion and the number of effective parameters [37], and the last based its variable selection on prior knowledge as well as results from previous univariable and multivariable regressions [38].

Although the objective of this study was not to assess the quality of reporting of the variable selection methods, we observed a lack of details in this matter when trying to determine which methods had been used. Notably, out of the 146 studies that mentioned utilizing prior knowledge, only 64 (44%) provided at least one reference in the method section to substantiate their choices of covariates. Moreover, data-driven procedures were often described only in vague terms (e.g., "The 10% change-in-estimate

Table 1 Variable selection methods used in explicative studies published in four major epidemiological journals in 2015

	American Journal of Epidemiology	Epidemiology	European Journal of Epidemiology	International Journal of Epidemiology	Total
Prior knowledge or causal graphs	55 (47%)	33 (59%)	27 (46%)	31 (52%)	146 (50%)
Prior knowledge or causal graphs only	40 (34%)	29 (52%)	19 (32%)	28 (47%)	*116 (40%)*
Change in estimate	20 (17%)	5 (9%)	5 (8%)	4 (7%)	34 (12%)
Stepwise	5 (4%)	3 (5%)	7 (12%)	1 (2%)	16 (5%)
Univariate analyses	16 (14%)	4 (7%)	5 (8%)	1 (2%)	26 (9%)
Other	3 (3%)	1 (2%)	1 (2%)	0 (0%)	5 (2%)
Insufficiently detailed	42 (36%)	16 (29%)	24 (41%)	25 (42%)	107 (37%)
Total	118	56	59	59	292

Results are reported as frequency (%). More than one method could be used in each study; as such, percentages do not add up to 100%

criterion identified confounders") and 37% of studies did not provide sufficient detail for us to determine how variables were selected. Table 1 summarizes the main results.

Discussion

The current review provides insights concerning the methods used in current epidemiology for performing variable selection. Since very few included studies pursued a predictive objective, we focus the discussion around explicative studies. Among the included explicative studies, 50% used causal graphs or prior knowledge for selecting variables, 12% used the change in estimate method, 5% used stepwise methods, 9% used univariable analyses, 2% used other methods and 37% did not provide sufficient information to allow classification of their method.

Although it has long been recognized that prior knowledge should play a role in confounder selection [6, 16, 39], only half of the surveyed articles mentioned doing so. It is however possible that some studies did use prior knowledge but did not report it. While the proportion of studies that used prior knowledge may seem to have substantially increased as compared to the 28% reported in Walter and Tiemeier's 2009 review [2], this might only be due to methodological differences. In fact, we have categorized a study as making use of prior knowledge even when only vague formulations such as "a priori confounders" or "known risk factors" were used, in contrast with the previous review.

We have also observed that 12% of the studies used the change in effect estimate method to select confounders, making it the most popular data-driven approach in the surveyed studies. As mentioned in the introduction, a few simulation results support the validity of the change in estimate method. However, we are unaware of any empirical investigation of the performance of some of the implementations observed in our review, such as when the comparator is the crude estimate.

Univariable and stepwise methods based on statistical significance were employed for selecting covariates in 14% of the surveyed studies, compared to 20% in the 2009 review [2]. These methods are recognized for overestimating exposure effects and underestimating statistical uncertainty (for example, see [3] and references therein). As such, this is an area in which practice have slightly improved in the surveyed journals.

Interestingly, none of the surveyed studies has made use of the newly developed methods for selecting confounders, despite statistical software being readily available in many cases. For instance, Bayesian Adjustment for Confounding [23, 25] is implemented in the R packages *bacr* and *BACprior*, Bayesian Causal Effect Estimation [21] is available in the R package *BCEE* and the model-free algorithms proposed in Ref. [27] are found in the R package *CovSEL*.

Finally, 37% of studies did not provide sufficient details for us to classify how variables were selected. In this area, results are similar to those observed previously [2].

A noteworthy limitation of the current review is that it does not qualify as a systematic review. Performing a systematic review would have increased the validity of our results, but we believe this would have been considerably more time-consuming. Our methods were inspired by those of systematic reviews' guidelines to make them as rigorous as possible. Our review also allowed us to observe that variable selection methods susceptible to introduce bias and to yield inappropriate inferences, such as stepwise and univariable analyses, are still prevalent in the surveyed journals. However, it is unclear how prevalent such methods are in the many journals we did not survey.

A second limitation is that we did not make a critical appraisal of the appropriateness of the methods used in individual studies. Hence, methods that are generally considered as "gold standards," such as using prior knowledge or causal graphs, might not have been applicable in some studies. For instance, this could be the case for exploratory studies whose goals were to generate hypotheses. Also, methods with poor statistical properties may be satisfactory in some contexts. For example, stepwise or univariable analyses may be acceptable for generating hypotheses, despite their tendency to exaggerate statistical significance and to overestimate effects, especially if such limitations are adequately acknowledged. Making a critical appraisal of the appropriateness of the methods used in individual studies would be an important contribution to the scientific literature, but would require carefully considering the goal, context, protocol and design of each individual study.

Despite these limitations, our study sheds light on areas in which variable selection methodology may be improved. First, we believe that additional investigation of the performance of the change in estimate approach is required, since this approach has been the object of only a few simulation studies that do not cover the span of how this approach is used in practice. Next, considering the pitfalls of selection methods based on statistical significance, we call authors to avoid using these methods as much as possible. When variable selection methods based on statistical significance are used, we suggest that the possible impact on the validity of the results be appropriately discussed. We also encourage researchers to consider replacing methods based on statistical significance with novel statistical methods to improve the statistical validity of their results. We further enjoin researchers not to see data-driven variable selection as a mandatory step of their study. For instance, when sample size is large enough that a model adjusting for all available potential confounders provides precise estimates, no variable selection might be required. If data-driven variable selection is performed, we recommend that authors also report the results from the fully adjusted model when possible. Finally, our results showed that variable selection methods were not reported in more than one-third of published studies in the surveyed epidemiological journals during year 2015. We recommend to increase awareness of researchers and publishers on the importance of appropriate reporting of variable selection methods for epidemiological studies.

Funding This work was supported by a start-up Grant from the Fondation du CHU de Québec—Université Laval [#2710 to DT]. DT is a Fonds de Recherche du Québec—Santé Chercheur-Boursier.

References

1. Lefebvre G, Delaney JA, McClelland RL. Extending the Bayesian Adjustment for Confounding algorithm to binary treatment covariates to estimate the effect of smoking on carotid intima-media thickness: the Multi-Ethnic Study of Atherosclerosis. Stat Med. 2014;33(16):2797–813. https://doi.org/10.1002/sim.6123.
2. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6. https://doi.org/10.1007/s10654-009-9411-2.
3. Greenland S, Pearce N. Statistical foundations for model-based adjustments. Annu Rev Public Health. 2015;36:89–108. https://doi.org/10.1146/annurev-publhealth-031914-122559.
4. Harrell FE. Regression modeling strategies, with applications to linear models, survival analysis and logistic regression. 2nd ed. New York: Springer; 2015.
5. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
6. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989;79(3):340–9.
7. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.
8. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011;67(4):1406–13. https://doi.org/10.1111/j.1541-0420.2011.01619.x.
9. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 2019;34(3):211–9. https://doi.org/10.1007/s10654-019-00494-6.
10. Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A Stat Soc. 1995;158(3):419–44.
11. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56(5):441–7.
12. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907–16.
13. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol. 1993;138(11):923–36.
14. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol. 1989;129(1):125–37.
15. Weng H-Y, Hsueh Y-H, Messam LLM, Hertz-Picciotto I. Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. Am J Epidemiol. 2009;169(10):1182–90.
16. 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(2):176–84.
17. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267–88.
18. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
19. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Stat Sci. 1999;14(4):382–401.
20. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059–79.
21. Talbot D, Lefebvre G, Atherton J. The Bayesian causal effect estimation algorithm. J Causal Inference. 2015;3(2):207–36. https://doi.org/10.1515/jci-2014-0035.
22. Crainiceanu CM, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. Biometrika. 2008;95(3):635–51. https://doi.org/10.1093/biomet/asn015.
23. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. Biometrics. 2012;68(3):661–71. https://doi.org/10.1111/j.1541-0420.2011.01731.x.
24. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. Biometrics. 2017;73(4):1111–22. https://doi.org/10.1111/biom.12679.

25. Wang C, Dominici F, Parmigiani G, Zigler CM. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. Biometrics. 2015;71(3):654–65. https://doi.org/10.1111/biom.12315.
26. Cefalu M, Dominici F, Arvold N, Parmigiani G. Model averaged double robust estimation. Biometrics. 2017;73(2):410–21. https://doi.org/10.1111/biom.12622.
27. Persson E, Häggström J, Waernbaum I, de Luna X. Data-driven algorithms for dimension reduction in causal inference. Comput Stat Data Anal. 2017;105:280–92. https://doi.org/10.1016/j.csda.2016.08.012.
28. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Stat Med. 2009;28(1):94–112. https://doi.org/10.1002/sim.3460.
29. Zigler CM, Dominici F. Uncertainty in propensity score estimation: bayesian methods for variable selection and model averaged causal effects. J Am Stat Assoc. 2014;109(505):95–107. https://doi.org/10.1080/01621459.2013.869498.
30. Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Stat. 2011;20(1):217–40. https://doi.org/10.1198/jcgs.2010.08162.
31. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20(4):512–22. https://doi.org/10.1097/EDE.0b013e3181a663cc.
32. Wilson A, Reich BJ. Confounder selection via penalized credible regions. Biometrics. 2014;70(4):852–61.
33. Molinaro AM, Ferrucci LM, Cartmel B, Loftfield E, Leffell DJ, Bale AE, et al. Indoor tanning and the MC1R genotype: risk prediction for basal cell carcinoma risk in young people. Am J Epidemiol. 2015;181(11):908–16. https://doi.org/10.1093/aje/kwu356.
34. Gracia E, Lopez-Quilez A, Marco M, Lladosa S, Lila M. the spatial epidemiology of intimate partner violence: do neighborhoods matter? Am J Epidemiol. 2015;182(1):58–66. https://doi.org/10.1093/aje/kwv016.
35. Zablotska LB, Nadyrov EA, Polyanskaya ON, McConnell RJ, O'Kane P, Lubin J, et al. Risk of thyroid follicular adenoma among children and adolescents in Belarus exposed to iodine-131 after the Chornobyl accident. Am J Epidemiol. 2015;182(9):781–90. https://doi.org/10.1093/aje/kwv127.
36. Zuurbier LA, Luik AI, Hofman A, Franco OH, Van Someren EJ, Tiemeier H. Fragmentation and stability of circadian activity rhythms predict mortality: the Rotterdam study. Am J Epidemiol. 2015;181(1):54–63. https://doi.org/10.1093/aje/kwu245.
37. DiMaggio C. Small-area spatiotemporal analysis of pedestrian and bicyclist injuries in New York City. Epidemiology. 2015;26(2):247–54.
38. Luque-Fernandez MA, Ananth CV, Jaddoe VW, Gaillard R, Albert PS, Schomaker M, et al. Is the fetoplacental ratio a differential marker of fetal growth restriction in small for gestational age infants? Eur J Epidemiol. 2015;30(4):331–41. https://doi.org/10.1007/s10654-015-9993-9.
39. Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol. 1981;114(4):593–603.