Eur J Epidemiol (2009) 24:733–736
DOI 10.1007/s10654-009-9411-2

COMMENTARY

Variable selection: current practice in epidemiological studies

Stefan Walter · Henning Tiemeier

Received: 10 November 2009 / Accepted: 24 November 2009 / Published online: 5 December 2009

Abstract Selection of covariates is among the most controversial and difficult tasks in epidemiologic analysis. Correct variable selection addresses the problem of confounding in etiologic research and allows unbiased estimation of probabilities in prognostic studies. The aim of this commentary is to assess how often different variable selection techniques were applied in contemporary epidemiologic analysis. It was of particular interest to see whether modern methods such as shrinkage or penalized regression were used in recent publications. Stepwise selection methods remained the predominant method for variable selection in publications in epidemiological journals in 2008. Shrinkage methods were not used in any of the reviewed articles. Editors, reviewers and authors have insufficiently promoted the new, less controversial approaches of variable selection in the biomedical literature, whereas statisticians may not have adequately addressed the method's feasibility.

Keywords Variable selection · Covariates · Lasso · Shrinkage · Stepwise · Confounding · Etiology · Prediction

S. Walter · H. Tiemeier (✉)
Department of Epidemiology, Erasmus MC,
Rotterdam, The Netherlands
e-mail: h.tiemeier@erasmusmc.nl

S. Walter
Department of Public Health, Erasmus MC,
Rotterdam, The Netherlands

H. Tiemeier
Department of Child and Adolescent Psychiatry, Erasmus MC,
Rotterdam, The Netherlands

Selection of covariates in epidemiologic analysis is among the most controversial and difficult tasks in epidemiologic analysis. The answer to the question of whether to include or to exclude a covariate from the analysis depends on the research question posed, the design of the study, and ultimately also on the sample size [1]. The goals related to the selection of the best variables are mainly twofold. Firstly, variable selection is used for confounder control to obtain unbiased estimates in etiologic research. Secondly, prediction research depends on variable selection for unbiased estimation of probabilities [1–6].

Prior knowledge from the scientific literature is formally seen as the most important rationale for including or excluding covariates from a statistical analysis but it is not always available for all research questions asked [2, 4, 6]. Statistical science has therefore developed several decision rules and algorithms to achieve selection based on the relations of the data under study: change in the effect estimate, stepwise selection, modern techniques such as shrinkage and penalized regression, and other techniques.

The aim of this commentary is to assess how often different variable selection techniques were applied in contemporary epidemiologic analysis. It was of particular interest to see whether modern methods such as shrinkage or penalized regression were used in recent publications. We screened the methods sections of articles published in four major epidemiologic journals in 2008 (*American Journal of Epidemiology, Epidemiology, European Journal of Epidemiology and the International Journal of Epidemiology*) for a description of the technique used to select variables. We present the frequency of these methods and in addition cited some articles that give a good example of how these selection techniques can be described in the methods section. All articles were categorized by the first author of this commentary into one of the following six

categories: prior knowledge, change-in-estimate, stepwise selection, modern methods, other methods, not described. The second author drew a random sample of 30 articles. Agreement between the two was 87% before the consensus discussion. One study was reclassified afterwards. We excluded commentaries, purely descriptive studies, genetic association studies, and meta-analyses. All other publications were included.

Table 1 shows the frequency of methods used by authors in their publications in these journals in 2008. 300 articles met our inclusion criteria. We could not observe significant differences between the journals (Fisher's exact test, $p = 0.09$).

In 83 (28%) articles, the authors selected the covariates contained in multivariable models based on prior knowledge. Ideally the selection of covariates should be substantiated with references from the literature. This was only the case for 41 of the 83 publications that relied on this method. The remaining 42 studies described the rational for including the covariates without explicit references. Prior knowledge can be documented by referring to a study in the same population that resulted in the identification of risk factors for the outcome under study, as in a study on the impact of smoking on thyroid volume, [7] or by referring to one or more studies that identify each of the potential confounders. An example for this approach is a study examining injury risk in the Swedish population [8].

A total of 59 (20%) of all reviewed publications used stepwise selection procedures with or without univariate pre-screening of potential covariates. These procedures rely on statistical testing of the covariate-disease association to decide which variables to include or to exclude from the model. They have been criticized extensively in the literature because they require arbitrary definitions of thresholds that can lead to bias, overfitting, and exaggerated p values [1–4, 6, 9]. The majority of these studies (66%) explicitly stated the thresholds. Although p values cannot replace prior information to select the best set of covariates, if the exact methods and thresholds used for these procedures are reported as for example in a recent study that derived and validated a mortality index among frail older patients [10], the analysis is at least reproducible and therefore to a certain degree objective [1, 2].

Another approach that is often combined with stepwise selection procedures is using a pre-specified change-in-estimate criterion ($n = 44$, 15%). This approach has been judged more favorable than stepwise procedures particularly when using the change of the interval estimate instead of the point estimate of the effect under study [4, 11]. It takes into account the covariate—disease association but also the change in the estimate, i.e. the exposure—covariate association, upon removal of the covariate [4]. The decision on the adequacy of the threshold depends on the context of the study and requires prior knowledge. Reporting of the criterion used is essential for this procedure and it is up to the researcher and the audience to decide whether, e.g., a 10% change in the risk measure of the association between income and recurrent coronary events [12] is reasonable, whereas for the association between socioeconomic position and pre-term birth a 5% change was seen as more adequate [13]. Together, stepwise procedures and change-in-estimate, represent 34% of all the methods used and virtually all of the data-driven statistical methods. Several variants of these procedures are implemented in standard software packages used to analyze epidemiological data. The ease of using these methods and the dominance in the existing literature, albeit years of criticism by leading epidemiologists, have probably hindered the breakthrough of other less controversial methods such as shrinkage.

In 9 articles other, very diverse methods for variable selection were applied (4 studies used principal components, [14–17] 1 study used propensity scores, [18] 1 study explicitly included all variables in the regression [19], 2 studies used causal diagrams, [20, 21] and 1 study used Deletion/Substitution/Addition algorithm [22]).

Not a single study used shrinkage procedures. Selection due to shrinkage in particular the Least Absolute Shrinkage

Table 1 Variable selection methods used in major epidemiologic journals in 2008

Selection technique	American Journal of Epidemiology		Epidemiology		European Journal of Epidemiology		International Journal of Epidemiology	
	n	%	*n*	%	*n*	%	*n*	%
Prior knowledge	50	29	11	28	13	30	9	20
Effect estimate change	31	18	6	15	3	7	4	9
Stepwise selection	27	16	9	23	10	23	13	29
Modern methods (shrinkage, penalized regression)	0	0	0	0	0	0	0	0
Other (e.g., principal components, propensity scores)	2	1	4	10	1	2	2	4
Not described	61	36	10	25	17	39	17	38
Total	171		40		44		45	

and Selection Operator (LASSO) [23] was welcomed in the literature [3, 4]. In the case of the Cox proportional hazard model, this algorithm maximizes the partial likelihood of the regression coefficients subject to a constraint imposed on the sum of the absolute value of all regression coefficients in the model. The constraint itself can be estimated via cross validation [23]. The LASSO technique has been labelled "shrinkage with selection" [24]. It corrects the extremes in the distribution of all variables and thus shrinks very unstable estimates towards zero. This effectively excludes some variables without the need for formal statistical testing [24–26]. Although LASSO and similar methods have been lauded in the epidemiologic literature because of these positive attributes, they were not applied in the selected articles in 2008. Admittedly it is a tedious procedure to implement LASSO for variable selection in the R program. Also, there is no consensus on the interpretation of estimates nor on how confidence intervals can be reliably estimated for penalized regression results such as those obtained by LASSO [3]. But as is the case with multiple imputation, which is now implemented as a routine in SAS and SPSS, the implementation of shrinkage and LASSO in commonly applied analysis software may help the dissemination of these modern methods for variable selection.

A total of 105 publications did not describe the method in sufficient detail. While it is remarkable to see that 35% of all selected articles in these epidemiologic journals scored in this category, this does not mean that the research is flawed. It is merely an indication of the quality of information in the methods section. One of the common reasons why we categorized the selection technique as "not described" were the use of vague formulations such as "based on prior knowledge" or "a priori". When the selection of variables is based on prior knowledge, this knowledge needs to be made explicit with references or explanations; otherwise the selection cannot be judged or discussed.

We conclude that variable selection methods which have been formally criticized as flawed still prevail in the scientific literature. This may be due to the ease of implementation, slow knowledge transfer, or because of the fear that editors or reviewers do not appreciate new approaches. We call for more cooperation between the academic research into methodology and ask statisticians to cooperate with research groups to demonstrate the usability of new algorithms in real data instead of simulation studies. Journals may wish not only to publish criticism of these methods but also to actively encourage the use of less controversial selection routines. At least, more referencing could be required when prior knowledge is used as a selection criterion. In addition, we encourage researchers to not simply use stepwise regression because of its availability in standardized software packages but rather to explore the new methods for variable selection in their research.

References

1. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1990.
2. Steyerberg EW. Clinical prediction models. New York: Springer; 2009.
3. Hesterberg TC, Choi NH, Meier L, Fraley C. Least angle and L1 penalized regression: a review. Stat Surv. 2008;2:61–93.
4. Greenland S. Invited Commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008;167(5):623–9.
5. Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
6. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol. 2002;155(2):176–84.
7. Vejbjerg P, Knudsen N, Perrild H, Carle A, Laurberg P, Pedersen IB, et al. The impact of smoking on thyroid volume and function in relation to a shift towards iodine sufficiency. Eur J Epidemiol. 2008;23(6):423–9.
8. Li X, Sundquist S, Johansson SE. Effects of neighbourhood and individual factors on injury risk in the entire Swedish population: a 12-month multilevel follow-up study. Eur J Epidemiol. 2008;23(3):191–203.
9. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol. 1989;129(1):125–37.
10. Drame M, Novella JL, Lang PO, Somme D, Jovenin N, Laniece I, et al. Derivation and validation of a mortality-risk index from a cohort of frail elderly patients hospitalised in medical wards via emergencies: the SAFES study. Eur J Epidemiol. 2008;23(12):783–91.
11. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989;79(3):340–9.
12. Laszlo KD, Janszky I, Ahnve S. Income and recurrent events after a coronary event in women. Eur J Epidemiol. 2008;23(10):669–80.
13. Morgen CS, Bjork C, Andersen PK, Mortensen LH, Nybo Andersen A-M. Socioeconomic position and the risk of preterm birth—a study within the Danish National Birth Cohort. Int J Epidemiol. 2008;37(5):1109–20.
14. Kolaczinski JH, Reithinger R, Worku DT, Ocheng A, Kasimiro J, Kabatereine N, et al. Risk factors of visceral leishmaniasis in East Africa: a case-control study in Pokot territory of Kenya and Uganda. Int J Epidemiol. 2008;37(2):344–52.
15. Bogin B, Varela-Silva MI. Fatness biases the use of estimated leg length as an epidemiological marker for adults in the NHANES III sample. Int J Epidemiol. 2008;37(1):201–9.
16. Kubo A, Levin TR, Block G, Rumore GJ, Quesenberry CP Jr, Buffler P, et al. Dietary patterns and the risk of Barrett's esophagus. Am J Epidemiol. 2008;167(7):839–46.
17. Wade TJ, Calderon RL, Brenner KP, Sams E, Beach M, Haugland R, et al. High sensitivity of children to swimming-

associated gastrointestinal illness: results using a rapid assay of recreational water quality. Epidemiology. 2008;19(3):375–83.
18. Harder VS, Stuart EA, Anthony JC. Adolescent cannabis problems and young adult depression: male–female stratified propensity score analyses. Am J Epidemiol. 2008;168(6):592–601.
19. Winkelmayer WC, Bucsics AE, Schautzer A, Wieninger P, Pogantsch M. Pharmacoeconomics Advisory Council of the Austrian Sickness Funds, Use of recommended medications after myocardial infarction in Austria. Eur J Epidemiol. 2008;23(2):153–62.
20. Wernli KJ, Ray RM, Gao DL, Fitzgibbons ED, Camp JE, Astrakianakis G, et al. Occupational exposures and ovarian cancer in textile workers. Epidemiology. 2008;19(2):244–50.
21. Hoffman CS, Mendola P, Savitz DA, Herring AH, Loomis D, Hartmann KE, et al. Drinking water disinfection by-product exposure and fetal growth. Epidemiology. 2008;19(5):729–37.
22. Mortimer K, Neugebauer R, Lurmann F, Alcorn S, Balmes J, Tager I. Air pollution and pulmonary function in asthmatic children: effects of prenatal and lifetime exposures. Epidemiology. 2008;19(4):550–7.
23. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.
24. Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. Stat Neerlandica. 2001;55(1):76–88.
25. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.
26. Houwelingen JCv. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Stat Neerlandica. 2001;55(1):17–34.