Variable selection should be blinded to the outcome

Tamás Ferenci

Manuscript type: Letter to the Editor

Title: Variable selection should be blinded to the outcome

Author List: Tamás Ferenci * (Physiological Controls Group, John von Neumann Faculty of Informatics, Obuda University; Budapest, Hungary, H-1034 Bécsi út 96/b, Phone number: +36 (1) 666-5553, Fax number: +36 (1) 666-5522, e-mail address: ferenci.tamas@nik.uni-obuda.hu).

Keywords: variable selection, confounder adjustment, multivariate modelling.

Key messages:

- When multivariate models are used for confounder adjustment, all potential confounders should be included in the model, or if not, their selection should be blinded to the outcome.
- Other methods, such as "pre-filtering" variables based on their univariate association with the outcome may give rise to biased regression coefficients, biased standard errors, biased confidence intervals, misspecified test distributions and exaggerated *p*-values.
- If variable selection is needed (for instance due to the high number of potential confounders) data reduction methods that are blinded to the outcome, modern approaches such as Bayesian Model Averaging or the application of shrinkage, i.e. penalization (regularization) of the regression are preferable.

This is a pre-copyedited, author-produced version of an article accepted for publication in International Journal of Epidemiology following peer review. The version of record Tamás Ferenci: Variable selection should be blinded to the outcome, Int J Epidemiol dyx048 is available online at: https://academic.oup.com/ije/article-abstract/doi/10.1093/ije/dyx048/3599366/Variable-selection-should-be-blinded-to-the DOI: 10.1093/ije/dyx048.

Dear Editor,

I've read the paper titled 'Acetaminophen use in pregnancy and neurodevelopment: attention function and autism spectrum symptoms' in your journal [1] with great interest. The article investigates a highly relevant question based on a unique database that allows us to get new insight into this important question of autism research. However, the applied statistical methodology is unfortunately not flawless; pointing out the error is the aim of the present remark.

The 'Statistical analysis' subsection discusses the methodology used, in particular, how the authors used regression models to protect against confounding by including potential confounders in the models. They used two approaches. For some variables they chose to adjust "for a series of predefined variables that were forced to remain in the models" (p. 4). However, "[o]ther covariates were included in the models only when they showed a crude association with both the exposure and the outcome (P-values < 0.20) and caused a change > 5% in the regression coefficient of acetaminophen when they were introduced one by one in the basic model" (p. 4).

It is not explained in the paper why they followed different ways for different variables, nevertheless, their strategy illustrates both the sound and the problematic approach to variable selection. The *most important rule* that should be always kept in mind about variable selection can be summarized in one sentence: *All potential confounders should be included in the model, or if not, then their selection should be blinded to the outcome*. Confounder adjustment should be liberal so as to not have residual confounding. Thus, the first decision of the authors is correct, but the second is unfortunately not.

What they did (checking univariate associations and change in the regression coefficient when adding a covariate) is not blinded to the outcome. The problem with this approach is that such "peeks" at the outcome in reality mean additional tests performed, but this is nowhere accounted for in the final model (which was estimated as if being pre-specified), giving rise to biased regression coefficients, biased standard errors, biased confidence intervals, misspecified test distributions and exaggerated *p*-

values [2,3,4,5]. Apart from theoretical considerations, all of these can be simply checked with simulation nowadays with a personal computer.

There are, of course, legitimate reasons why the authors might have felt that a variable selection is needed; the most important being perhaps the high number of potential confounders (compared to the sample size). Presumably this is case here, i.e. the authors used two approaches because they felt that the variables in the first group are important confounders that must be controlled for, however they were less sure about the variables in the second group, which also happened to include a much higher number of covariates, so they decided that some kind of selection is needed.

This is a legitimate concern. Indeed, including many covariates (relative to the sample size) might increase the risk of overfitting, increase the variance of the estimates or make the clinical applicability/interpretability of the model more complicated. These are actually the very reasons why suggestions of the minimal sample size are often formulated [6,7]. But it is crucially important to emphasize that (1) these considerations make the above rule not a bit less true, (2) variables selection answers none of these problems (except, of course, making the model less complicated), rather "hides" them without giving a solution – actually making the problem even worse.

Many other, valid, solution is available. Apart from *data reduction methods* that are masked to the outcome (variable clustering, principal components regression etc.) [2] and certain modern methods such as Bayesian Model Averaging [8], perhaps the most important is the application of *shrinkage*, i.e. the *penalization* (*regularization*) of the regression. Widely used penalization schemes include L2-penalization (ridge regression), L1-penalization (LASSO) and the combination of both (elastic net) [9], all of which is extensively studied theoretically and implemented in most major statistical software packages.

Finally, no matter how the model was fit, the importance of *validation* and *calibration* [3,2] cannot be overstated. Preferably done with bootstrap [10], the models should have been at least internally validated to get a realistic estimate of their appropriateness.

It is practically impossible to guess the impact of this analytical error on the overall results. Even if it does not change the findings, this represents a bad statistical practice, which should be generally avoided; and in the particular case, it would be important to see a reanalysis of the data taking these considerations into account (i.e. including all covariates, or filtering only blinded to the outcome) so that we can better trust the results of authors' otherwise very important study. It is true for the sake of scientific validity per se, and especially true for such a paper that is already highly quoted, even in the non-medical media.

Funding

The author has no relevant affiliations or financial involvement with any organization or entity with a financial or other relevant interest in or financial or other relevant conflict with the subject matter or materials discussed in the manuscript. No writing assistance was utilized in the production of this manuscript.

References

1. Avella-Garcia CB, Julvez J, Fortuny J, Rebordosa C, García-Esteban R, Galán IR. Acetaminophen use in pregnancy and neurodevelopment: attention function and autism spectrum symptoms. Int J Epidemiol. 2016 Jun 28. pii: dyw115.
2. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Switzerland: Springer, 2015.
3. Steyerberg, E. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer, 2008.
4. Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. The American Statistician 1990; 44:214-217.
5. Copas JB. Regression, prediction and shrinkage. Journal of the Royal Statistical Society. Series B (Methodological) 1983; 45:311-354.

6. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49:1373-1379.
7. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep. 1985; 69:1071-1077.
8. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical science 1999, 14:382-401.
9. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Berlin: Springer, 2001.
10. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54:774-781.