REVIEW

Five myths about variable selection

Georg Heinze & Daniela Dunkler

Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Correspondence
Georg Heinze PhD, Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Tel.: +4314040066880;
fax: +4314040066870;
e-mail: georg.heinze@meduniwien.ac.at

SUMMARY

Multivariable regression models are often used in transplantation research to identify or to confirm baseline variables which have an independent association, causally or only evidenced by statistical correlation, with transplantation outcome. Although sound theory is lacking, variable selection is a popular statistical method which seemingly reduces the complexity of such models. However, in fact, variable selection often complicates analysis as it invalidates common tools of statistical inference such as *P*-values and confidence intervals. This is a particular problem in transplantation research where sample sizes are often only small to moderate. Furthermore, variable selection requires computer-intensive stability investigations and a particularly cautious interpretation of results. We discuss how five common misconceptions often lead to inappropriate application of variable selection. We emphasize that variable selection and all problems related with it can often be avoided by the use of expert knowledge.

Transplant International 2017; 30: 6–10

Key words
association, explanatory models, multivariable modeling, prediction, statistical analysis

Received: 12 September 2016; Revision requested: 14 October 2016; Accepted: 25 November 2016

Observational studies are often conducted in a prognostic or etiologic research context [1]. To this end, many papers dealing with observational studies use multivariable regression approaches to identify important predictors of an outcome [2–4] or to assess effects of new markers adjusted for known clinical predictors [5–7], respectively. As the number of candidate predictor variables or known confounders can be large, a "full" model including all candidate predictors as explanatory variables is often considered impractical for clinical use or, in the extreme case, is even impossible to estimate because of multicollinearity. Therefore, variable selection approaches are often employed, mostly based on evaluating *P*-values for testing regression coefficients against zero. For example, Martinez-Selles *et al.* [8] used univariate screening of effects to build a multivariable model for survival after heart transplantation. After univariate selection, Rodriguez-Perálvarez *et al.* [9] employed "backward elimination" which means that first a multivariable model was built with all predictors selected by univariate screening in a first step. Then, nonsignificant predictor variables were sequentially eliminated and models re-estimated until all variables remaining in the model show significant association with the outcome. The technique of backward selection is sometimes also applied directly to a set of predictor variables [10–13], or the process of variable selection is reversed by "forward selection" [5,14], meaning that candidate predictors are sequentially included in a model if their association with the outcome variable, on top of the set of variables already in the model, is significant. Box 1 provides an overview of the most common approaches of variable selection and explains some statistical notions commonly used in this context.

Whatever technique applied, the approach of letting statistics decide which variables should be included in a model is popular among scientists. Among all 89 clinical research articles published in *Transplant*

doi:10.1111/tri.12895

Box 1. Glossary

AIC	Akaike information criterion. A number based on information theory expressing the model fit (log likelihood) discounted by the number of unknown parameters
Augmented backward elimination	An extension of backward elimination recently proposed by Dunkler *et al.* [23], which can consider the change-in-estimate as an additional selection criterion. Usually leads to selection of more variables than standard backward elimination. Preferable method for etiologic models
Backward elimination	Remove insignificant predictors from a model one-by-one until all variables are significant. The preferable method for prognostic models if enough data are available
Bayes factor	Ratio of likelihood of two competing models
Change-in-estimate	The magnitude by which the regression coefficient of a variable *X* changes if a variable *Z* is removed from a multivariable model
EPV	Events per variable. A simple measure to define the amount of information in a data set (the number of events or the sample size) relative to the number of regression coefficients to be estimated (the number of variables). Note that nonselection of a variable corresponds to an estimated regression coefficient of zero, and thus, this formula should always consider all candidate variables
Etiologic models	Statistical models used to explain the role of a risk factor or treatment in its (possibly causal) effect on patient outcome
Forward selection	Add significant candidate predictors to a model one-by-one until no further predictors can be added. Usually leads to inferior results compared to backward elimination
Likelihood	The probability of the data to be observed under a given model
Maximum likelihood	Statistical estimation techniques which sets unknown model parameters (e.g., regression coefficients) to those values which are such that the observed data are most plausibly explained
Multicollinearity	Almost perfect correlation of explanatory variables. Causes ambiguity in estimation of regression coefficients and selection of variables. Can be a problem if regression coefficients should be interpretable, that is, in etiologic research contexts
Prognostic models	Statistical models used to prognosticate a patient's outcome (e.g., graft loss or death) by the values of some variables available at the time point at which this prediction is made
Univariable prefiltering (bivariable analysis)	Each candidate variable is evaluated in its association with the outcome. Only significant variables are entered in a multivariable model. Often used by researchers, but should be avoided
Variable, independent or dependent	Independent variable: a variable considered as predicting or explaining patient outcome. Also termed predictor or explanatory variable, respectively. Dependent variable: the variable representing the outcome under study, for example, occurrence of rejections, patient survival, or graft survival

International in 2015, 49 applied multivariable regression modeling, and variable selection was used in 30 (65%) of those 49 articles. However, it is not commonly known that there hardly exists any statistical theory which justifies the use of these techniques. This means that quantities such as regression coefficients, hazard ratios or odds ratios, *P*-values or confidence intervals may suffer from systematic biases if variable selection was applied, and usually, the magnitude or direction of these biases is unpredictable [15]. This is in sharp contrast to the simplicity of application of variable selection approaches, being ready to use in statistical standard software such as IBM SPSS [16] or SAS software [17].

The popularity of variable selection approaches is based on five myths, that is, "believes" lacking theoretic foundation. Before discussing these myths in this review, it should be noted that no variable selection approach can guard against general errors in setting up a statistical modeling problem, for example, putting the "cause" as outcome variable and the "effect" as independent, adjusting effects for later outcomes, or the like. Moreover, it is assumed here that, without looking into the data, a set of candidate predictors has already been preselected by clinical expertise, for example, a prior belief that those predictors could be related to the outcome. Availability in a data set alone is not the basis for their consideration for a model.

Myth 1: "The number of variables in a model should be reduced until there are 10 events per variables." *No!*

Simulation studies have revealed that multivariable models become very unstable with too low events-per-variable (EPV) ratios. Current recommendations suggest that a minimum of 5–15 EPV should be available, depending on context [15,18]. However, practitioners often overlook

that this recommendation refers to *a priori* fixed models which do not result from earlier selection [15]. If variable selection is considered, the rule should consider the number of candidate variables with which the selection process is initialized, and probably even much higher values such as 50 EPV are needed to obtain approximately stable results, as the selection adds another source of uncertainty to estimation [19]. If the number of candidate variables seems to be too large, background knowledge obtained from analyses of former studies or from theoretical considerations should be applied to prefilter variables in order to meet the EPV-rule requirements of a problem. Causal diagrams such as directed acyclic graphs can also help in discarding candidate variables before statistically analyzing the data [20,21].

Myth 2: "Only variables with proven univariable-model significance should be included in a model." *No!*

While it is true that regression coefficients are often larger in univariable models than in multivariable ones, also the opposite may occur, in particular if some variables (with all positive effects on the outcome) are negatively correlated. Moreover, univariable prefiltering, sometimes also referred to as "bivariable analysis," does not add stability to the selection process as it is based on stochastic quantities, and can lead to overlooking important adjustment variables needed for control in an etiologic model. Although univariable prefiltering is traceable and easy to do with standard software, one should better completely forget about it as it is neither a prerequisite nor providing any benefits when building multivariable models [22].

Myth 3: "Insignificant effects should be eliminated from a model." *No!*

Eliminating a variable from a model means to put its regression coefficient to exactly zero – even if the likeliest value for it, given the data, is different. In this way, one is moving away from a maximum likelihood solution (which has theoretical foundation) and reports a model which is suboptimal by intention. Eliminating weak effects may also be dangerous as in etiologic studies, bias could result from falsely omitting an important confounder. This is because regression coefficients generally depend on which other variables are in a model, and consequently, they change their value if one of the other variables are omitted from a model. This "change-in-estimate" [23] can be positive or negative, that is, away from or toward zero. Hence, it may happen, that after eliminating a potential confounder another adjustment variable's coefficients moves closer to zero, changing from "significant" to "nonsignificant" and hence leading to the elimination of that variable in a later step. However, despite its usual detrimental effects on bias, elimination of very weak predictors from a model can sometimes decrease the variance (uncertainty) of the remaining regression coefficients. Dunkler *et al.* [23] have proposed "augmented backward elimination," a selection algorithm which leaves insignificant effects in a model, if their elimination would cause a change in the estimate of another variable. Thus, their proposal extends pure "significance"-based sequential elimination of variables ("backward elimination") and is of particular interest in etiologic modeling.

Myth 4: "The reported *P*-value quantifies the type I error of a variable being falsely selected." *No!*

First, while the probability of a type I error mainly depends on the significance level, a *P*-value is a result of data collection and analysis and quantifies the plausibility of the observed data under the null hypothesis. Therefore, the *P*-value does not quantify the type I error [24]. Second, after a sequence of elimination or selection steps, standard software reports *P*-values only from the finally estimated model. Any quantities from this last model are unreliable as they do not "remember" which steps have been performed before. Therefore, *P*-values are biased low (as only those *P*-values are reported which fall below a certain threshold), and confidence intervals are often too narrow, claiming, for example, a confidence level of 95% while they actually cover the true value with a much lower probability [15]. On the other hand, there is also the danger of false elimination of variables, the possibility of which is not quantified at all by just reporting the final model of a variable selection procedure. To overcome these problems, statisticians have argued in favor of using resampling techniques or relative AIC- or Bayes factor-based approaches to explore alternative models and their likelihood to fit the data, and to use averages over competing models instead of just selecting one ultimate model [25]. Such analyses may provide valuable insight in how stable models are and how many and which competing models would be selected how often. Resampling can also be used to quantify selection probabilities of variables or pairs of correlated, competitive variables [26]. Unfortunately, with the exception of SAS/PROC GLMSELECT software [17], we do not know of any implementations of these very useful approaches in standard

Transplant International 2017; 30: 6–10

software. For example, while simple bootstrap resampling is indeed implemented in IBM SPSS software [16], its application is restricted to a model with a fixed set of independent variables rather than to evaluating the stability of variable selection across resamples.

Myth 5: "Variable selection simplifies analysis." *No!*

While a smaller model may be easier to use and – at first glance – to report, there are many problems to be solved when variable selection techniques are considered. First, an appropriate variable selection method has to be selected for the problem at hand. Statisticians have recommended backward elimination as the most reliable one among those that can be easily achieved with standard software [27]. Second, an often arbitrary choice has to be made about the selection parameter, that is, the significance level to decide whether an effect should be retained in a model. While smaller values such as 0.05 or 0.01 are only recommended for very large sample sizes (EPV of 100 or above), in the vast majority of applications, a value of 0.2 or 0.157 (corresponding to selection based on AIC) or even 0.5 (resulting in very mild selection) will be a better choice. Third, as selection is a "hard decision" but often based on vague quantities, investigations on model stability should accompany any applied variable selection to justify the decision for the model finally reported or at least to quantify the uncertainty related to the selection of the variables. This has to be done with resampling methods, which, until robust implementations are available in standard software, are still cumbersome to implement. Such evaluations are also needed (and even more computationally demanding) for best subset searches, that is, letting the computer evaluate all different models that can be thought of using a given set of candidate predictors. Similar stability investigations should be carried out if modern variable selection methods such as the LASSO [28] or boosting [29] are used.

By way of conclusion, we see that five myths have obscured the problems of variable selection that have been identified in recent decades by statisticians. While variable selection methods seem simple to use and handy to build multivariable models, issues such as selection uncertainty or bias in the reported quantities have too often been overlooked by practitioners. Before using variable selection techniques one should critically reflect whether such methods are needed in a particular study at all, and if yes, whether there is enough data available to justify elimination or inclusion of variables in a model just by "letting the data speak." By contrast, expert background knowledge, for example, formalized by directed acyclic graphs [20,21], is usually a much better guide to robust multivariable models. At least, such knowledge (and not data-driven methods!) should be used to restrict the number of candidate variables competing for selection to a number compatible with published EPV rules. In line with many other statisticians, we think that for prognostic models backward elimination with a selection criterion of 0.157 and without a preceding univariable prefiltering is a good starter, but sometimes other choices for the selection criterion may be more appropriate. For etiologic models, "augmented backward elimination" [23] preceded by a careful preselection based on assumptions on the causal roles of variables [20,21] is a reasonable approach. Whenever investigators decide to use statistical variable selection approaches, they should use them with care and should add sensitivity and robustness analyses. Ideally, such analyses are conducted using resampling techniques, but often it will already be helpful if robustness of basic conclusions of a study is demonstrated by comparing the main results with those obtained after eliminating some variables from the main model or including additional ones. The key messages of this review are summarized in Box 2.

Box 2. Key points to remember

1. The five myths presented here are all misconceptions of variable selection.
2. Often there is no scientific reason to perform variable selection. In particular, variable selection methods require a much larger sample size than estimation of a multivariable model with a fixed set of predictors based on clinical expertise.
3. If a researcher needs to perform variable selection, and sample size is large enough and the candidate predictors have been carefully selected based on prior knowledge, then backward elimination with a *P*-value criterion of 0.157 is a good choice for prognostic models, and augmented backward elimination for etiologic models.
4. Variable selection should always be accompanied by sensitivity analyses to avoid wrong conclusions.

Funding

The authors have declared no funding.

Conflict of interest

The authors have declared no conflicts of interest.

Acknowledgements

We acknowledge Isabell Gläser's and Thomas Hillebrand's help with conducting the systematic review of the use of variable selection methods in *Transplant International*. Furthermore, we would like to thank two anonymous reviewers for helpful comments on a previous version of this manuscript.

REFERENCES

1. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Testing for causality and prognosis: etiological and prognostic models. *Kidney Int* 2008; **74**: 1512.
2. Von Düring ME, Jenssen T, Bollerslev J, *et al.* Visceral fat is better related to impaired glucose metabolism than body mass index after kidney transplantation. *Transpl Int* 2015; **28**: 1162.
3. Bhat M, Hathcock M, Kremers WK, *et al.* Portal vein encasement predicts neoadjuvant therapy response in liver transplantation for perihilar cholangiocarcinoma protocol. *Transpl Int* 2015; **28**: 1383.
4. Pianta TJ, Peake PW, Pickering JW, Kelleher M, Buckley NA, Endre ZH. Evaluation of biomarkers of cell cycle arrest and inflammation in prediction of dialysis or recovery after kidney transplantation. *Transpl Int* 2015; **28**: 1392.
5. Rompianesi G, Montalti R, Cautero N, *et al.* Neurological complications after liver transplantation as a consequence of immunosuppression: univariate and multivariate analysis of risk factors. *Transpl Int* 2015; **28**: 864.
6. Zijlstra LE, Constantinescu AA, Manintveld O, *et al.* Improved long-term survival in Dutch heart transplant patients despite increasing donor age: the Rotterdam experience. *Transpl Int* 2015; **28**: 962.
7. Fernández-Ruiz M, Arias M, Campistol JM, *et al.* Cytomegalovirus prevention strategies in seropositive kidney transplant recipients: an insight into current clinical practice. *Transpl Int* 2015; **28**: 1042.
8. Martinez-Selles M, Almenar L, Paniagua-Martin MJ, *et al.* Donor/recipient sex mismatch and survival after heart transplantation: only an issue in male recipients? An analysis of the Spanish Heart Transplantation Registry. *Transpl Int* 2015; **28**: 305.
9. Rodríguez-Perálvarez M, García-Caparrõs C, Tsochatzis E, *et al.* Lack of agreement for defining 'clinical suspicion of rejection' in liver transplantation: a model to select candidates for liver biopsy. *Transpl Int* 2015; **28**: 455.
10. Pezawas T, Grimm M, Ristl R, *et al.* Primary preventive cardioverter-defibrillator implantation (Pro-ICD) in patients awaiting heart transplantation. A prospective, randomized, controlled 12-year follow-up study. *Transpl Int* 2015; **28**: 34.
11. Prasad GVR, Huang M, Silver SA, *et al.* Metabolic syndrome definitions and components in predicting major adverse cardiovascular events after kidney transplantation. *Transpl Int* 2015; **28**: 79.
12. Tripon S, Francoz C, Albuquerque A, *et al.* Interactions between virus-related factors and post-transplant ascites in patients with hepatitis C and no cirrhosis: role of cryoglobulinemia. *Transpl Int* 2015; **28**: 162.
13. Somers J, Ruttens D, Verleden SE, *et al.* A decade of extended-criteria lung donors in a single center: was it justified? *Transpl Int* 2015; **28**: 170.
14. Nagai S, Mangus RS, Anderson E, *et al.* Post-transplant persistent lymphopenia is a strong predictor of late survival in isolated intestine and multivisceral transplantation. *Transpl Int* 2015; **28**: 1195–204.
15. Harrell FEJ. *Regression Modeling Strategies*, 2nd edn. Switzerland: Springer, 2015.
16. IBM Corp. *IBM Statistics for Windows*, 23.0 edn. New York, NY: IBM Corp, 2013.
17. SAS Institute Inc. *SAS/STAT*, 9.4 edn. Cary, NC: SAS Institute Inc., 2012.
18. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007; **165**: 710.
19. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer, 2009.
20. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37.
21. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics* 2011; **67**: 1406.
22. Sung 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**: 907.
23. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PLoS One* 2014; **9**: e113677. doi:10.1371/journal.pone.0113677
24. Goodman S. A dirty dozen: twelve P-value misconceptions. *Semin Hematol* 2008; **45**: 135.
25. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*, 2nd edn. New York, NY: Springer, 2002.
26. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* 1992; **11**: 2093.
27. Royston P, Sauerbrei W. *Multivariable Model-Building – A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Chichester: John Wiley & Sons Ltd, 2008.
28. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 1996; **58**: 267.
29. Breiman L. Arcing classifiers. *Ann Stat* 1998; **26**: 801.