Building a better model: abandon kitchen sink regression

Stefan Kuhle ,[1,2] Mary Margaret Brown,[3] Sanja Stanojevic[4]

[1]Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), Johannes Gutenberg University Mainz, Mainz, Germany
[2]Departments of Pediatrics and Obstetrics and Gynaecology, Dalhousie University, Halifax, Nova Scotia, Canada
[3]Department of Mathematics and Statistics, University of New Brunswick Saint John, Saint John, New Brunswick, Canada
[4]Department of Community Health and Epidemiology, Dalhousie University, Halifax, Nova Scotia, Canada

Correspondence to
Dr Stefan Kuhle;
stefan.kuhle@uni-mainz.de

Received 4 October 2023
Accepted 23 November 2023
Published Online First
6 December 2023

To cite: Kuhle S, Brown MM, Stanojevic S. *Arch Dis Child Fetal Neonatal Ed* 2024;**109**:F574–F579.

ABSTRACT

This paper critically examines 'kitchen sink regression', a practice characterised by the manual or automated selection of variables for a multivariable regression model based on p values or model-based information criteria. We highlight the pitfalls of this method, using examples from perinatal/neonatal medicine, and propose more robust alternatives. The concept of directed acyclic graphs (DAGs) is introduced as a tool for describing and analysing causal relationships. We highlight five key issues with 'kitchen sink regression': (1) the disregard for the directionality of variable relationships, (2) the lack of a meaningful causal interpretation of effect estimates from these models, (3) the inflated alpha error rate due to multiple testing, (4) the risk of overfitting and model instability and (5) the disregard for content expertise in model building. We advocate for the use of DAGs to guide variable selection for models that aim to examine associations between a putative risk factor and an outcome and emphasise the need for a more thoughtful and informed use of regression models in medical research.

INTRODUCTION

Kitchen sink regression is a pejorative term to describe a multivariable regression procedure where all available variables ('the kitchen sink') that may or may not be associated with an outcome of interest are entered into a model, followed by some form of automated or manual variable selection strategy based on p values or a model-based information criterion. The variables remaining in the model are then interpreted as 'independent risk factors' for the outcome of interest. The procedure, together with the availability of easy-to-use statistical programmes and the continued over-reliance on p values,[1] has likely contributed to the production of contradictory evidence and medical research waste.[2] The approach has serious issues both from a statistical and epidemiological standpoint,[3] and many journals have discouraged its use in submitted manuscripts,[4 5] but it still is pervasive in the literature.[6 7] The problem is further compounded by the uncritical use of estimates from these studies in systematic reviews and meta-analyses. The objectives of this review are to highlight the issues with kitchen sink regression using examples from neonatal medicine, and to offer alternatives to this method.

TO EXPLAIN OR TO PREDICT: MAKE UP YOUR MIND!

The first step towards valid inference is to be clear about the research question that is to be addressed: should a condition or the course of an illness be *predicted*, or is the aim to *explain* the occurrence of an outcome by examining the association between an exposure and a disease? Both approaches have different underlying assumptions, require, in part, different methodology and use different measures to describe their findings.[8] Unfortunately, these two approaches are often conflated in clinical research.[9]

If the aim is to predict future events, like the presence or the risk of a disease (screening, diagnostic testing) or the course of a disease (prognostic modelling), confounding is not a concern—any information that is available at the time of prediction can be used. The drawback, of course, is that intervening on a predictor will not necessarily affect the outcome. For example, the presence of several café au lait spots in an infant may predict neurofibromatosis, but surgically removing them will not influence the course of the disease. Automated variable selection procedures may be acceptable for predictive modelling under certain circumstances, but they should be used with great caution.[10] Most importantly, it cannot be stressed enough that the estimates from a prediction model have no meaningful causal interpretation. Unfortunately, with the current hype around big data, machine learning and artificial intelligence, we can expect many more papers using causal language to incorrectly tout predictors as 'risk factors'.

If, on the other hand, our aim is to understand the aetiology of a condition by examining the association of a putative risk factor and a disease, confounding and bias need to be minimised to make causal inferences. Ultimately, if a causal effect is identified, it can become the target of an intervention that will then change the outcome. For example, if the aim is to estimate the effect of smoking during pregnancy on neonatal mortality, relevant confounders of the association (such as maternal age and socioeconomic status) have to be identified and controlled for in our analysis. If an association between smoking and neonatal mortality is found, a behavioural intervention to reduce smoking in pregnant women could then be designed.

DIRECTED ACYCLIC GRAPHS: A PRIMER

To better illustrate the problems with data-driven, hypothesis testing-based approaches to variable selection, the language of causal inference, directed acyclic graphs (DAG), will be briefly introduced.

Causal effects are directional (eg, cigarette smoking causes lung cancer), but statistics does not have the language to express this directionality. The standard regression equation $Y=a+bX$ applies if X

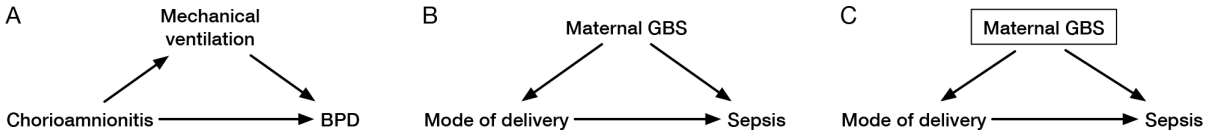


Figure 1 Directed acyclic graph representation of mediation (A), confounding (B) and conditioning on a confounder (C). BPD, bronchopulmonary dysplasia; GBS, group B streptococcus.

causes *Y*, but it may also apply if *Y* causes *X*, or if a third variable *C* (ie, a confounder) causes both *X* and *Y*. Valid causal inferences cannot be made when directionality is ignored, since modern definitions of confounding and bias are based on the (hypothesised) direction of effects.[11] Motivated by this shortcoming of classical statistics, the computer scientist Judea Pearl developed DAGs (or causal graphs as they are also referred to) to describe causal relationships.[12]

DAGs are structural representations of the researcher's hypothesised relationships between measured or unmeasured variables that affect an association of interest. A set of rules can then be used to identify variables that confound the association. DAGs can also help identify selection and measurement bias. They provide an intuitive, visual representation of a complex problem that makes the researcher's assumptions about variable relationships transparent. It is important to note that relationships in a DAG are entirely based on a priori knowledge and are generally not tested in the data, nor are associations between variables in a DAG estimated.[13–15]

A DAG is drawn by connecting variables that are assumed to be related (based on content expertise, existing data or best guess) with single-headed arrows pointing from the presumed cause to the presumed effect. Variables that are deemed to be unrelated are not connected with an arrow. Most importantly, DAGs should be based on *all* variables that are relevant to the relationship of interest, not just the variables that the researchers have in their dataset. Because the relationships between variables in a DAG are indicated by directional arrows, the graph is 'directed'. A DAG may not contain a forward path of arrows that form a closed loop, it is 'acyclic'.

The DAG in figure 1A says that *chorioamnionitis* leads to *mechanical ventilation* which in turn leads to *bronchopulmonary dysplasia* (*BPD*). In other words, mechanical ventilation is mediating part of the effect of chorioamnionitis on BPD. There is also an arrow from *chorioamnionitis* to *BPD*, indicating a direct effect (eg, the infection itself damaging the fetal/neonatal lung tissue).

Figure 1B shows a DAG with confounding: *maternal group B streptococcus (GBS) status* is a cause of both the exposure (*mode of delivery*) and the outcome (*neonatal sepsis*). Note that this definition of a confounder differs from the one often found in older textbooks, which just requires a non-directional association with the exposure; the latter definition is less precise and should not be used. Our DAG illustrates in an intuitive fashion how maternal GBS status confounds the association between mode of delivery and neonatal sepsis: besides the hypothesised and yet to be tested effect of *mode of delivery* on *neonatal sepsis*, there is a second pathway connecting *mode of delivery* and *sepsis* through the variable *maternal GBS status*. That pathway runs against the direction of the arrowhead and is therefore considered a non-causal path. This so-called 'backdoor path' creates a non-causal association between *mode of delivery* and *sepsis*, even if there were no direct association between the two variables. To obtain an unbiased estimate for the association between the two variables, this path needs to be blocked. A path in a DAG can be blocked by conditioning on a variable that is on the path. 'Conditioning on' refers to stratification, restriction or adjustment—all familiar and mathematically identical approaches to dealing with confounding. A variable that has been conditioned on has a box drawn around it by convention (figure 1C). Now the backdoor path through *maternal GBS status* is blocked, and an analysis of neonatal sepsis on mode of delivery conditional on maternal GBS status will yield the unbiased effect estimate for the association between the two variables (assuming that there are no other confounding variables).

Finally, there is one other arrangement of three variables in a DAG. In the DAG in figure 2A, both *intrauterine infection* and *congenital anomalies* are causally related to *gestational age at birth*. Because the two arrowheads collide in *gestational age at birth*, the variable is called a 'collider'. By contrast to mediation and confounding, the term *collider* is new to the epidemiological literature, although its effects have been described previously as Berkson's bias and in Simpson's paradox,[16 17] among others. Two variables that are connected through a collider (here: *intrauterine infection* and *congenital anomalies* through *gestational age at birth*) are *not* correlated through this pathway. However, if the collider *gestational age* is conditioned on, *intrauterine infection* and *congenital anomalies* will become correlated. For example, if the sample is restricted to preterm infants, and an infant in the sample does not have congenital anomalies, then their preterm birth is more likely to be due to an intrauterine infection. Conversely, if they had no intrauterine infection, then their preterm birth is more likely to be due to congenital anomalies.

Collider-stratification bias is responsible for a number of paradoxical, spurious associations that haunt the literature, such as the false observation that smoking reduces mortality among low birthweight infants, where conditioning on birth weight creates a pathway from *smoking* to *mortality* through all other (unmeasured) common causes of low birth weight (*U*) and *mortality* (eg, congenital anomalies) (figure 3).[18] More recently, collider bias has been implicated in the aetiology of paradoxical protective associations between smoking and severity of COVID-19.[19]

With this foundation laid, the shortcomings of hypothesis testing-driven models for making causal inference will now be discussed.

Issue 1: directionality of relationships is ignored

If we do not know the mechanism that generated our data or can at least make assumptions about it, it is impossible to determine if a variable in a model is a confounder, a mediator or a collider. For example, in figure 1B, content expertise suggests that maternal GBS status is a confounder of the association between mode of delivery and sepsis. However, if a hypothesis test determines that maternal GBS status should be excluded from the model, because its p value is too large (which may happen if the cut-offs are set very tight, like 0.1 or lower), the estimate for the association between mode of delivery and sepsis will be biased, because a non-significant variable can still confound an

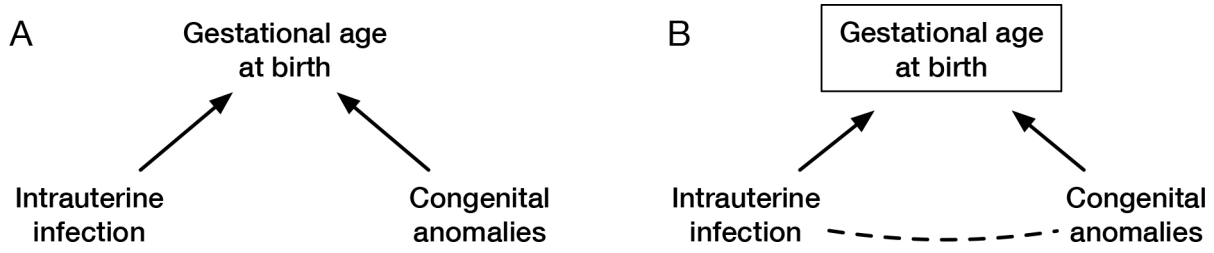


Figure 2 Directed acyclic graph representation of a collider. Intrauterine infection and congenital anomalies are both causally associated with gestational age at birth, making it a collider (A). Conditioning on the collider (eg, by restricting the sample to preterm infants), as indicated by the box around it (B), will create a spurious association between the two causes: if intrauterine infection did not cause the preterm birth of an infant in the sample, congenital anomalies become more likely as the cause of preterm birth, and vice versa.

association. Similarly, a mediator (eg, *mechanical ventilation* in figure 1A) or a collider that are entered into or kept in a model because their coefficient is statistically significant will bias the effect estimate of the parent variable. This problem is illustrated by Simpson's paradox, where the interpretation of the results changes, depending on the causal structure of the underlying data.[17] This causal structure, however, will only become clear through a DAG developed using content expertise.

Issue 2: effect estimates often have no meaningful causal interpretation

There is a plethora of scientific literature describing 'independent risk factors' for a condition, as indicated by the presence of a variable in a multivariable regression model. These estimates are then interpreted and compared with one another. This practice is perhaps the most pervasive misuse of multivariable models and has led to misinterpretation of data and associations.

Consider the DAG in figure 4A for the association between maternal smoking during pregnancy and the risk of preterm birth. Based on the output from a regression model for preterm birth on maternal smoking, adjusting for the potential confounding factors maternal age and socioeconomic status (figure 4B), we conclude that maternal smoking during pregnancy is associated with a 2.5 times increased risk for preterm birth, controlling for age and socioeconomic status. Assuming the DAG is complete and there is no measurement error, this estimate represents the *total* effect of maternal smoking on preterm birth. Many researchers will interpret the estimates for the other variables in the model in the same fashion. For example, lower socioeconomic status is associated with a 1.5-fold increased risk for preterm birth, controlling for maternal smoking and age. However, this interpretation is incorrect because the coefficients in the regression model represent different effects. If we rearrange the position of the variables in the DAG so that socioeconomic status and preterm birth are now placed at the bottom (figure 4C), we can see that maternal smoking is a mediator of this association. Therefore, the effect estimate for socioeconomic status does not represent the *total* effect on preterm birth, but only the *direct* effect; the indirect effect mediated by maternal smoking has been removed by adjusting for it in the model. This phenomenon of incorrectly interpreting and comparing the coefficients in a regression table has been dubbed the '*Table 2 fallacy*' (in reference to Table 2 in a publication often containing the regression table output).[20] Researchers should recognise that each coefficient must be interpreted within the context of its corresponding variable, taking into account the specific relationships and potential confounding and mediating factors involved. In the example above, the correct estimate for the effect of socioeconomic status on preterm birth can only be obtained from a regression model that does not contain the mediator maternal smoking, only the confounder maternal age (figure 4D); the resulting effect estimate is considerably larger (relative risk 1.8 vs 1.5). It is conceivable from this simple example that in a model with many more variables, selected by an algorithm with no consideration for the directionality of effects, any meaningful interpretation of the effect estimates will be impossible.

Issue 3: increased alpha error rate due to multiple testing

The most significant issue from a statistical point is the multiple testing that is required to build a model with stepwise regression.

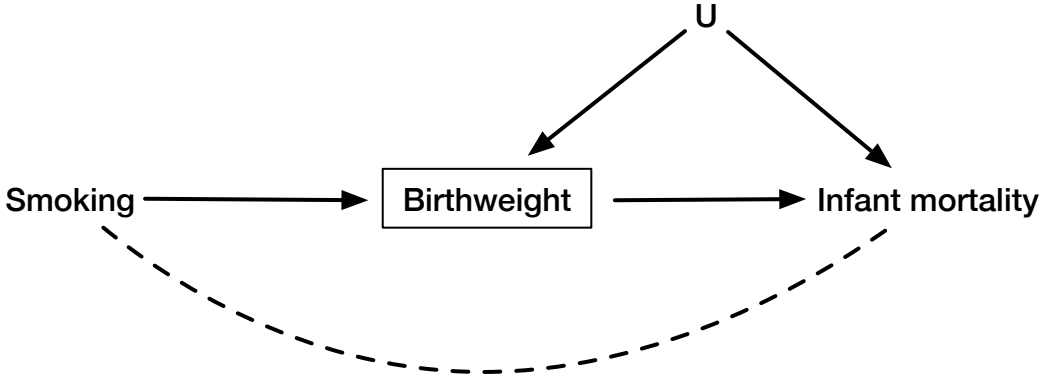
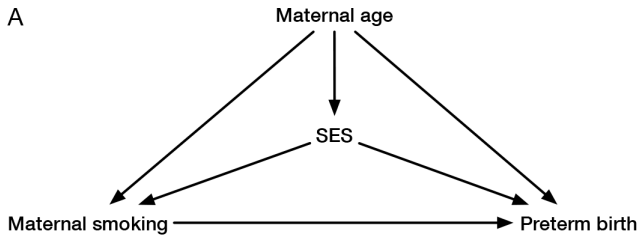
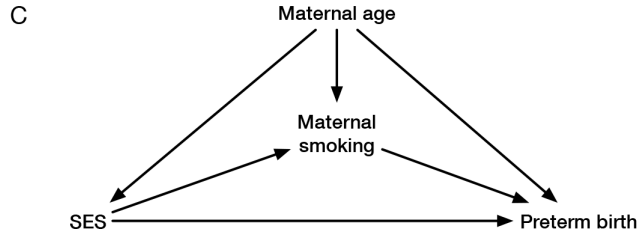


Figure 3 Directed acyclic graph representation of the 'birthweight paradox'. When examining the association between *smoking during pregnancy* and *infant mortality*, conditioning on the collider *birth weight* (eg, by restricting the sample to low birthweight infants) will induce a spurious, negative association between *smoking* and *infant mortality* through unmeasured common causes of birth weight and mortality (*U*) such as intrauterine infection or congenital anomalies.[18]



B

Variable	Relative risk (95% CI)
Smoking	2.5 (1.9, 3.2)
Low SES	1.5 (1.1, 1.8)
Age <20 years	1.2 (0.9, 1.6)



D

Variable	Relative risk (95% CI)
Low SES	1.8 (1.4, 2.3)
Age <20 years	1.4 (1.1, 1.8)

Figure 4 Directed acyclic graph representation (A) and regression output (B) for the association between *maternal smoking during pregnancy* and *preterm birth* with confounding by socioeconomic status (*SES*) and *maternal age*. When *SES*, the confounder of the original association, is treated as the exposure, *maternal smoking* becomes a mediator of this association (C). The effect estimate for *SES* from the original model can therefore not be interpreted as the total effect of *SES* on *preterm birth*, because it is inappropriately adjusted for the mediator *maternal smoking*. The correct estimate for the total effect of *SES* on *preterm birth* comes from a model that does not contain the mediator *maternal smoking*, only the confounder *maternal age* (D).

The stepwise algorithm performs a large number of significance tests that are hidden from the user by the software. By the time the algorithm produces the final model, the alpha error rate is already well above 5% due to multiple testing. As a result, the p values for each coefficient cannot be compared against the standard 0.05 significance cut-off; they must be compared against a much lower cut-off, but that cut-off is difficult to determine. So that 'significant' risk factor that the model suggests may not be significant when the multiple testing is taken into account. The same issue arises when the researcher performs a series of single-variable regressions to select covariates for their regression model; this is simply forward stepwise regression performed manually.

Issue 4: overfitting and model instability

When a model is fitted to a set of data, regression coefficients for each variable need to be estimated. The maximum likelihood algorithm will essentially solve an equation with p-1 unknowns (where p is the number of variables in the model) or 'degrees of freedom'. If a large number of variables are added to a model based on statistical significance, the model fit will be very good to near perfect. The problem is that such a model will be fit to the peculiarities of our data and will be near useless when applied to other datasets. Many predictors identified by a stepwise algorithm are simply noise variables.[21] The model may become sensitive to even small changes in the original data, because the statistical significance of a variable may have been strongly dependent on the presence of other variables in the model.[22] These models also may produce regression coefficients that are disproportionally large and give the appearance that some variables are particularly important, when in fact they are not.[3] Lastly, with more variables in the model, the variance of the estimates will increase, and therefore the power to detect true associations will decrease. Similar to the increased alpha error rate of the final model, the degrees of freedom that were used in the building of the model (even if the variables were not included in the model) still count against the p-1 available degrees of freedom, meaning that the fitted model from a stepwise regression will have fewer degrees of freedom than suggested in the regression table and will be overfitted.

Issue 5: common sense is left at the door

The strongest tool for building a useful and meaningful model is the researcher's understanding of and insight into the content area, which will allow them to appropriately preprocess the data, examine the data for outliers and non-linear relationships between a covariate and the outcome, identify errors and make an informed choice of all relevant variables to enter into the model. Without these human insights, the information we can gain from data is limited or even wrong.

BUILDING A BETTER MODEL

The first step to building a better model is to be clear from the outset what the aim of the model building is: prediction or explanation. If the former, then confounding is not a concern. However, to reiterate, the researcher must be aware that in this case, they must refrain from interpreting the magnitude or direction of the coefficients as they do not have a meaningful causal interpretation (other than a being a weight that is given to a variable in the prediction). Also, issue 3 (increased alpha error rate) and issue 4 (overfitting) still apply to prediction models.

If, on the other hand, explanation (measuring associations) is the goal, all the issues described above apply, and researchers should describe their assumptions about the relevant variable relationships in a DAG first. A DAG will by no means protect researchers from incorrectly specifying a model, but it will facilitate understanding of the variable relationships and make

the researchers' assumptions transparent to the reader. Tutorials[14 15 23 24] and online tools (https://dagitty.net/)[25] are available to help researchers to learn and use DAGs.

If a single exposure is of interest, the DAG should be used to identify the relevant adjustment set that is then used in the regression analysis. If there is doubt about some of the hypothesised relationships in the DAG, sensitivity analyses based on adjustment sets from alternative DAGs may be performed.

If researchers wish to examine several potential risk factors for an outcome, they should create one DAG that contains all relationships for all relevant variables. For each of the exposures, they then identify the respective adjustment sets from the DAG and run separate regression models for each exposure with the corresponding adjustment set.[20]

BUT WHAT ABOUT…? OTHER APPROACHES TO VARIABLE SELECTION AND MODEL BUILDING

The change-in-estimate method,[26] whereby a variable that changes the main effect by more than 5–20% (suggested cut-offs vary) after being added to the model is considered a 'confounder', is a popular approach to variable selection. This approach—by contrast to statistical significance-based approaches—examines what confounding really is about, namely: does a variable change the estimate for the association of interest? However, without considering the causal structure by drawing a DAG, this method may mistake a mediator or collider for a confounder, since their addition to the model would also change the estimate of the main effect. This approach also has limitations when used with logistic regression or Cox proportional hazards models: due to a statistical property of the OR and HR ('noncollapsibility'), estimates for these measures may change even when a variable that is not a confounder (but strongly related to the outcome) is added to the model.[27]

Shrinkage methods such as penalised maximum likelihood estimation,[28] Lasso or ridge regression[29] can help reduce the magnitude of the overly optimistic estimates. Penalised maximum likelihood estimation does so by applying a shrinkage factor to the coefficients in a regression model after fitting. Lasso shrinks coefficients towards zero during fitting; covariates that do not significantly improve the fit of the model are shrunk until they are forced out of the model entirely. These methods can be used to reduce the number of covariates in a model and avoid issue 3 (increased alpha error rate) and issue 4 (overfitting), but still require some preselection of covariates informed by content expertise and causal inference considerations.

Lastly, it should be mentioned that machine learning methods are not valid tools for variable selection or causal inference unless they are used in very specific situations and approaches (eg, targeted maximum likelihood estimation or g-computation).[30] Machine learning methods are designed for predictions, and commonly can not produce effect estimates for individual variables. They have some theoretical advantages over conventional regression, but they are quite often not the magic bullet they are made out to be.[31]

CONCLUSIONS

We have described the key issues around the use of stepwise regression for model building and strongly discourage its use for explanatory (risk factor) research. The use of DAGs based on content expertise to guide variable selection for explanatory models is essential. Researchers should use prior knowledge about variable relationships and directionality in their model building instead of handing control over to an algorithm and hypothesis testing to decide which factors are 'relevant' based on their statistical significance.

Acknowledgements None of the insights presented in this review are new or based on our own work; credits are due to the great minds of the field like Judea Pearl, Frank Harrell, Sander Greenland, Miguel Hernan and others. All authors of this review have used hypothesis testing-based variable selection for explanatory models in the past (we were young and needed the p values!) but have since recognised the error in their ways.

Contributors SK, MMB and SS conceived the idea for the manuscript. SK wrote the initial manuscript draft, reviewed and revised the manuscript and approved the final manuscript as submitted. MMB and SS reviewed and revised the manuscript, and approved the final manuscript as submitted.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; internally peer reviewed.

ORCID iD

Stefan Kuhle http://orcid.org/0000-0001-9417-3727

REFERENCES

1 Greenland S, Senn SJ, Rothman KJ, *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337–50.
2 Ioannidis JPA, Greenland S, Hlatky MA, *et al*. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383:166–75.
3 Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004;66:411–21.
4 Thompson B. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. *Educ Psychol Meas* 1995;55:525–34.
5 Lederer DJ, Bell SC, Branson RD, *et al*. Control of confounding and reporting of results in causal inference studies. *Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals Ann Am Thorac Soc* 2019;16:22–8.
6 Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. *Eur J Epidemiol* 2009;24:733–6.
7 Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *Eur J Epidemiol* 2019;34:725–30.
8 Schooling CM, Jones HE. Clarifying questions about "risk factors": predictors versus explanation. *Emerg Themes Epidemiol* 2018;15:10.
9 Varga TV, Niss K, Estampador AC, *et al*. Association is not prediction: a landscape of confused reporting in diabetes - A systematic review. *Diabetes Res Clin Pract* 2020;170:108497.
10 Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:431–49.
11 Howards PP, Schisterman EF, Poole C, *et al*. "Toward a clearer definition of confounding" revisited with directed acyclic graphs. *Am J Epidemiol* 2012;176:506–11.
12 Pearl J. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 2009.
13 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37–48.
14 Williams TC, Bach CC, Matthiesen NB, *et al*. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatr Res* 2018;84:487–93.
15 Tennant PWG, Murray EJ, Arnold KF, *et al*. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol* 2021;50:620–32.
16 Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology* 2012;23:159–64.
17 Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. *Int J Epidemiol* 2011;40:780–5.
18 Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered *Am J Epidemiol* 2006;164:1115–20.
19 Griffith GJ, Morris TT, Tudball MJ, *et al*. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020;11:5749.
20 Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013;177:292–8.
21 Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Brit J Math & Statis* 1992;45:265–82.
22 Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 2004;57:1138–46.

23 Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *J Clin Epidemiol* 2022;142:264–7.
24 Bandoli G, Palmsten K, Flores KF, *et al*. Constructing causal diagrams for common perinatal outcomes: benefits, limitations and motivating examples with maternal antidepressant use in pregnancy. *Paediatr Perinat Epidemiol* 2016;30:521–8.
25 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;45:1887–94.
26 Greenland S, Mickey RM. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;130:1066.
27 Whitcomb BW, Naimi AI. Defining, quantifying, and interpreting "noncollapsibility" in epidemiologic studies of measures of "effect." *Am J Epidemiol* 2021;190:697–700.
28 Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer Science & Business Media, 2001.
29 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B* 2005;67:301–20.
30 Blakely T, Lynch J, Simons K, *et al*. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. *Int J Epidemiol* 2021;49:2058–64.
31 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.