Journal of Clinical Epidemiology 189 (2026) 112043

COMMENTARY

‘‘Identifying variables that independently predict…’’ is not a well-defined research task

John B. Carlin[a,b,*]

[a]*Clinical Epidemiology & Biostatistics Unit (CEBU), Murdoch Children’s Research Institute, Melbourne, Australia*
[b]*CEBU, University of Melbourne, Melbourne, Australia*

Accepted 31 October 2025; Published online 5 November 2025

Recent developments in the methodology of epidemiological research have emphasized the importance of achieving clarity of purpose by classifying research questions into one of three types: descriptive, predictive, and causal. Dyer’s paper in a recent issue of the journal [1] makes a welcome contribution to this literature, highlighting several specific ‘‘mistakes’’ that are commonly made in the design and reporting of research data analysis, despite increasing recognition of the value of the ‘‘three tasks’’ classification. In related work, a colleague and I reviewed the continuing high prevalence of poorly conceived statistical analyses (especially using multivariable regression), provided a detailed outline of how these problems can be traced back to the way in which statistical methods are taught, with an emphasis on models and techniques ahead of purposes, and outlined a way forward based on reforming biostatistics teaching programs [2].

Although many of Dyer’s points are well taken (and are welcome even if largely reiterating concerns expressed elsewhere), their overall coherence is reduced by the author’s desire to recognize a variant of the prediction research task that he describes as ‘‘identifying variables that predict some health state’’. Specifically, I believe his ‘‘mistake 1’’ is largely misconceived. He emphasizes correctly that the concept of confounding has no role in addressing descriptive and predictive research questions but then seeks to defend a role for regression adjustment in a variant of prediction research that is described as ‘‘prognostic factor research’’. In a rejoinder [3] to invited commentaries on our ‘‘call for reform’’ paper, we point out that prognostic factor research does not generally address a well-posed research question. It seeks to identify (independent) prognostic factors while failing first to provide a clear definition of a prognostic factor.

Following the lead of others (especially an influential overview paper from the PROGRESS guidelines group [4]), Dyer provides the general definition that a prognostic factor is a variable that is ‘‘associated with a future health outcome among people with a particular disease or health status’’. Taken at face value, however, this would allow many variables to be defined as prognostic factors, as truly null bivariate associations are uncommon. The field of prognostic factor research thus promotes the idea that attention should focus on factors for which their ‘‘prognostic value over established prognostic factors’’ can be established. However, there is no clarity around exactly what this means and how it can be defined and then evaluated. Instead, in a tradition that seems to have originated with a great pioneer of biostatistical methods in practice, Doug Altman [5], researchers turn to the use of multivariable regression for assessing the evidence that a putative predictive factor demonstrates an *independent* effect, over and above other predictors.

The difficulty is that this approach appears to tie the definition of prognostic factor to a null hypothesis about a coefficient in an assumed regression model. This in turn makes it very difficult to discern an inherent meaning in the concept, unless the proposed model is assumed to represent the ‘‘true’’ data generating process, which is highly implausible in practice [2]. Ultimately, we end up with the circularity that the definition of a prognostic factor in practice is that it is a factor that has been declared to have a statistically significantly coefficient in a multivariable regression model (‘‘adjusting’’ for other predictors) in a currently available dataset. Thus the definition becomes inseparable from the method for determining whether the factor meets the definition, that is, whether it has an ‘‘independent’’ effect… Furthermore, this method may clearly return different results depending on which other predictors have been included, details of model specification (especially for continuous variables, but also with

Funding: This research received no external funding.

* Corresponding author. Clinical Epidemiology & Biostatistics Unit, Murdoch Children’s Research Institute, Flemington Road, Parkville, Victoria 3052, Australia.

E-mail addresses: jbcarlin@unimelb.edu.au; john.carlin@mcri.edu.au.

https://doi.org/10.1016/j.jclinepi.2025.112043

respect to interactions), statistical significance threshold, and, of course, sample size.

These difficulties may be illustrated in the brief example on prognosis in amyotrophic lateral sclerosis that Dyer uses. He cites an extensive systematic review [6] that examines evidence on a wide range of potentially prognostic factors, observing that because a clinical indicator (bulbar onset) is more likely to be seen at diagnosis in older (rather than younger) patients, then it was important ''to confirm that bulbar onset is predictive of poorer prognosis over and above age in adjusted analyses''. This implies that a rather simple regression model could provide a definitive answer to an inherently complex and ill-defined question. In fact, the results of such an analysis would depend on many analytic decisions around model specification (scale of outcome, nature of age relationship, presence or absence of interactions, including across populations), not to mention the key issue of sample size. In a small study, with an ambitious model, the ''effect'' of interest might well appear to be ''nonsignificant''. It is interesting that the authors of the review only conclude, sensibly, that ''there is a general consensus that older age and bulbar onset are negatively related to amyotrophic lateral sclerosis outcome, but the complex relationship between age, female gender, and bulbar onset remains to be clarified.'' This example highlights the difficulties that arise from conflating the complexities of complex biological processes with the relative simplicity of multivariable linear regression models.

A further concern with Dyer's discussion of prognostic factor studies is that there is a contradiction between elements of the advice given under ''mistake 1'' and under ''mistake 2''. Under the former heading, he recommends that studies ''present an adjusted estimate'', to ''assess evidence of […] the extent to which the candidate prognostic factor may be a predictor over and above the established prognostic factors'', implying that the regression coefficient in question might be of interest, beyond testing the corresponding null hypothesis. Yet under ''mistake 2'' he advises, quite reasonably in my view, that ''the individual coefficients (in a multivariable prediction model) are not particularly meaningful''.

Related concerns apply to Dyer's discussion of regression adjustment in descriptive epidemiology. The examples cited do not clearly explain the potential role for regression models. If a researcher wishes to ''describe COVID-19 mortality rates […] according to age groups, sex, or ethnicity'', then they should do just that, that is, present an appropriate cross-tabulation. A regression model does not assist with this, unless, for example, the researcher is interested in describing sex differences in a hypothetical population in which the age distribution does not differ between the sexes, for which purpose a regression model might provide an avenue for model-based standardization of differences between the sexes. In another (sketch) example, the author refers to describing average patient trajectories over time, for the *estimation* of which a linear mixed-effects model might be useful. Examining the difference between men and women might be performed in a number of ways, with the proposed approach of introducing the variable sex ''into the model'' one that might be reflexive for many statisticians but in fact relies on strong assumptions—in particular that the age/time trajectories in the outcome measure are parallel between the two sexes, in the scale chosen for the outcome.

Here and elsewhere, the author exhibits some of the same tendencies that we documented in our article [2], whereby statisticians (and others, often leaning on what they have learned from statisticians) tend to jump reflexively to the fitting of regression models as if they are an omnibus tool for answering many questions, even in the case of questions that have not been well specified. In this regard, expressions such as ''independent effect'' and ''accounting for additional variables'' (last sentence in the paragraph on descriptive epidemiology) should be avoided, because they do not have a clear meaning. These considerations underlie the importance of a structured approach to question specification and analysis planning, following our ''roadmap'' concept [2]. This requires beginning with absolute clarity of specification of the research question, within one of the three categories, including defining estimands (target parameters) of interest, followed by consideration of assumptions that are necessary for the data to yield valid answers. At the final step, an analysis plan is specified, often introducing further assumptions in the form of parametric models to facilitate the estimation of target parameters. Following this approach provides a good recipe for avoiding the sorts of mistakes that Dyer discusses.

In summary, I do not believe that studies aiming to ''identify'' independent predictors or ''prognostic factors'' are addressing well-defined research questions. Indeed, beyond the issues already raised, there is a broader question of the extent to which it is ever sensible to frame a research question as if it could be answered dichotomously, as in ''is this an (independent) prognostic factor?'' Prediction questions, which include prognosis, are those that involve the development of a model or algorithm to provide predictions of outcomes using available variables that are potential predictors. Some variables may have greater predictive value than others, but this should be assessed by comparing the predictive value of the model or algorithm with and without the use of that variable, not by examining its ''independent effect'' in a multivariable regression model. More broadly, debates on whether to ''adjust'' or not for certain variables in a regression model can only be answered by situating the analysis within a sharply defined research question and a sharply defined rationale for specifying a regression model in the first place.

CRediT authorship contribution statement

John B. Carlin: Conceptualization, Methodology, Writing − original draft.

Declaration of competing interest

There are no competing interests for any other author.

Acknowledgments

The author thanks Margarita Moreno-Betancur for helpful comments on a draft and related discussions.

Data availability

No data were used for the research described in the article.

References

[1] Dyer B. The distinction between causal, predictive, and descriptive research − there is still room for improvement. [e-pub ahead of print]. J Clin Epidemiol 2025. https://doi.org/10.1016/j.jclinepi.2025.111960.

[2] Carlin JB, Moreno-Betancur M. On the uses and abuses of regression models: a call for reform of statistical practice and teaching. Stat Med 2025;44(13-14):e10244. https://doi.org/10.1002/sim.10244.

[3] Carlin JB, Moreno-Betancur M. Rejoinder to commentaries on: on the uses and abuses of regression models: a call for reform of statistical practice and teaching. Stat Med 2025;44(13-14):e70065. https://doi.org/10.1002/sim.70065.

[4] Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis research strategy (PROGRESS) 2: prognostic factor research. PLoS Med 2013;10(2):e1001380. https://doi.org/10.1371/journal.pmed.1001380.

[5] Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 1998;52(1):289−303. https://doi.org/10.1023/A:1006193704132.

[6] Chiò A, Logroscino G, Hardiman O, Swingler R, Mitchell D, Beghi E, et al. Prognostic factors in ALS: a critical review. Amyotroph Lateral Scler 2009;10(5−6):310−23. https://doi.org/10.3109/17482960802566824.