Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Thornley, J Biom Biostat 2012, 3:7
DOI: 10.4172/2155-6180.1000e115

Editorial Open Access

Causation and Statistical Prediction: Perfect Strangers or Bedfellows?

Simon Thornley*

Professional Teaching Fellow, Epidemiology and Biostatistics, University of Auckland, New Zealand

A quiet revolution is occurring in biostatistics and epidemiology from the field of computer science. Directed acyclic graphs (DAGs), developed and refined by Judea Pearl to address issues of causation, have led to new insights into old epidemiological concepts [1]. His ideas shed new light on confounding, selection bias, and strategies for variable selection in regression analysis. Pearl's thinking has de-emphasised the use of purely statistical tools (such as hypothesis testing or information measures), in favour of reasoning about the causal influence between variables, and conducting statistical analyses in the light of prior knowledge and informed scientific understanding of the subject under study.

DAGs are used to support the development of plausible causal pathways for the risk of disease. In contrast, building regression models does not typically facilitate this process; the model provides a structure, and variables are added or taken from the structure according to statistical indices, such as *p*-values. DAGs provide a framework for deciding what variables should be included. Models generally have two purposes in epidemiology: to elicit the effects of exposures and for prediction. Often these purposes are conflicting–a confounder one attempts to eliminate from a causal model may be an important component in a prediction model - but in either case DAGs seem to be useful. For eliciting the effects of exposure, they help decide what variables it is sensible to adjust for, and for prediction, they may avoid optimism, or overfitting of statistical models. Traditionally, statistical approaches to avoiding optimism have included use of cross-validation and boot-strap resampling, which do not, necessarily, take into account the causal considerations [2].

From a common sense perspective, it seems that causation and prediction are linked. If we understand the causes of an event, surely that would help to predict its behaviour. From our (colleagues and myself) research into cardiovascular disease, we believe that causal thinking can help in the development of accurate prediction models. One of Pearl's key concepts that seem to be relevant is that of colliders. If a variable, which has many unobserved strong influences, is included in a prediction model then adjusting for the variable may introduce the effects of the unobserved factors that, otherwise, would not affect the outcome. Through the effects of colliders, Pearl shows that inclusion of such variables in prediction regression models is not only questionable by statistical considerations, but also causal considerations.

An example from Cardiovascular Disease (CVD) risk prediction research may help to clarify. Many CVD risk prediction models include drug treatment, anti-hypertensive and statin therapy at enrolment [3,4]. Typically, the estimated effect is positive, treatment apparently increases, rather than reduces the risk of cardiovascular disease. This is counter-intuitive and at odds with published meta-analyses of randomised-controlled-trial data, which suggest that treatment with these drugs results in consistently reduced risks of disease [5-7]. As treatment variables are likely to be related to unobserved causal influences, such as propensity to take treatment (and seek preventive healthcare), unrecorded prognostic factors, and the doctor's leaning towards treating risk factors for disease, treatment may be considered, in Pearl's term, a collider and should not be included in a model (Figure 1 shows a DAG drawn from our understanding of the interaction between variables, both observed and unobserved). For more details, we refer readers to the original paper [8]. Since publishing the paper,

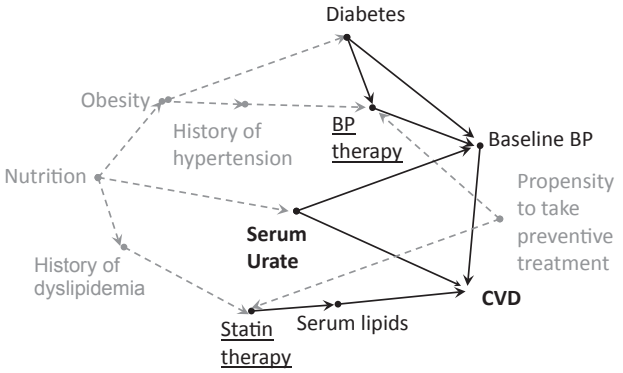

Figure 1: A directed acyclic graph (DAG), characterising the direction of selected influences* on risk of developing cardiovascular disease.
BP- Blood pressure; CVD-Cardiovascular disease.
*Dark font variables are those which are observed, whereas grey font represent unobserved. Drug treatment variables, which we consider colliders, which introduce the influence of unobserved variables not directly linked with the outcome, are underlined.

however, some empiric support for this concept has emerged from our exploration of CVD in a cohort of patients recruited from interactions with their family doctors, which we intend to publish in due course. In a directed acyclic graph estimated from the data – using a procedure implemented by Scutari [9] - drug treatment was conditionally dependent on a number of observed variables including age, sex, diabetes status, ethnic group and smoking status. With so many observed influences, it is likely that some unobserved variables also play a part in whether a subject opts to take treatment at enrolment, or not. Theoretically, if associations between the unobserved variables with the collider change over time, the information conveyed by drug treatment is likely to be unreliable. This is because of the induced conditional dependence resulting from the model structure (due to adjusting for the collider) between the unobserved variables and the disease outcome.

It is also possible that non-causal variables may upset risk prediction [10]. Staying with the topic of cardiovascular disease, if red wine drinking is not truly causal, but merely associated with other behaviors which reduce the risk of developing CVD, this variable may appear to be associated with disease and might be useful for prediction. If in the future, the association between red wine drinking and other

***Corresponding author:** Simon Thornley, Professional Teaching Fellow, Epidemiology and Biostatistics, University of Auckland, New Zealand, Tel: +64 9 3737599; Fax: +64 9 3737503; E-mail: s.thornley@auckland.ac.nz

Received October 25, 2012; **Accepted** October 27, 2012; **Published** November 02, 2012

Citation: Thornley S (2012) Causation and Statistical Prediction: Perfect Strangers or Bedfellows? J Biom Biostat 3:e115. doi:10.4172/2155-6180.1000e115

Copyright: © 2012 Thornley S. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

beneficial behaviours declines, or even reverses, for example if taxes or tariffs were lifted on red wine, and it became associated with adverse behaviours, then the previous information conveyed by this variable would disappear.

While the methods for doing so are not entirely clear, prediction is yet another area of biostatistics that is likely to benefit from causal considerations when selecting variables to predict disease events. The development of plausible causal models based on expert opinion and experimental, causal evidence is likely to lead to more reliable prediction models than using statistical methods alone.

Acknowledgements

The author thanks Associate Professor, Roger Marshall and Professor Rod Jackson for their valuable comments on drafts.

Funding

ST is funded by a Health Research Council of New Zealand Clinical Research Training Fellowship (reference number: 11/145).

References

1. Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press.
2. Harrell FE (2001) Regression Modeling Strategies. (1stedn), Springer, New York.
3. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, et al. (2007) Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 335: 136.
4. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, et al. (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 117: 743-753.
5. Taylor F, Ward K, Moore THM, Burke M, Davey SG, et al. (2011) Statins for the primary prevention of cardiovascular disease. The Cochrane Collaboration, John Wiley & Sons, Ltd.
6. Wiysonge CSU, Bradley HA, Mayosi BM, Maroney RT, Mbewu A, et al. (2009) Beta-blockers for hypertension. The Cochrane Collaboration, John Wiley & Sons, Ltd.
7. Wright JM, Musini VM (1996) First-line drugs for hypertension. Cochrane Database Syst Rev: John Wiley & Sons.
8. Thornley S, Marshall RJ, Jackson R, Gentles D, Dalbeth N, et al. (2012) Is serum urate causally associated with incident cardiovascular disease? Rheumatology (Oxford).
9. Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. J Stat Softw 55: 1-22.
10. Spirtes P (2010) Introduction to Causal Inference. J Mach Learn Res 11: 1643-1662.