Journal of Clinical Epidemiology 109 (2019) 142

Journal of Clinical Epidemiology

LETTER TO THE EDITOR

A glimpse of the difference between predictive modeling and classification modeling

Whittle et al.'s article, ''Measurement Error and Timing of Predictor Values for Multivariable Risk Prediction Models are Poorly Reported'' (May 2018), reviews a variety of articles and raises concerns regarding the reliability of covariates used in prediction modeling. This is indeed an important topic because often covariates do not necessarily represent their true values.

Prediction and classification modeling share similar characteristics. There are problems in both domains in the attempt to identify associations between covariates and outcomes. However, there are substantial differences between the two sets of problems. Prediction modeling relies on defining specific points in time (such as a patient's discharge dates or certain procedure dates). Then a variety of covariates of the patient's historical medical profile are pulled to serve as an input for a machine-learning algorithm to predict an outcome of interest (e.g., mortality, readmission). By contrast, classification modeling problems commonly focus on assessing the risk for the patient to be associated with a disease or to rule out the disease. Such models do not necessarily rely on defining any disease index dates or baselines.

As Whittle et al. pointed out correctly, our nonalcoholic fatty liver disease (NAFLD) model [1] was not developed with the intention to be used at a specific time but rather to identify large-scale longitudinal cohorts. By design, we considered the entire lifetime of a patient; our intention was to develop a disease classification model, not a prediction model.

Medical publications that focus on disease classifications using health records are rare. A few, however, have been published—for instance, models that classify rheumatoid arthritis, Crohn's disease, and ulcerative colitis. Similarly, our NAFLD article describes a classification model, which should not be characterized as a prediction model. Our NAFLD algorithm is a pioneering attempt to use health records to identify patients at a high risk for NAFLD. Subsequently, our algorithm formed the basis for studies published in high-impact journals. For instance, using our algorithm, researchers from the Cleveland Clinic validated our discovery regarding the interplay between cardiovascular risk and liver disease in NAFLD. Both studies were published in *The American Journal of Gastroenterology* [2,3]. Our article thus should not to be listed along with articles referred to by Whittle et al. as a ''prediction model,'' and definitely not as ''poor.''

Uri Kartoun
Center for Computational Health
IBM Research
Cambridge, MA, USA
Center for Computational Health, IBM Research,
75 Binney St., Cambridge, MA 02142, USA.
Tel.: 857-500-2425.
E-mail address: *uri.kartoun@ibm.com*

References

[1] Corey KE, Kartoun U, Zheng H, Shaw SY. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Dig Dis Sci 2016;61:913–9.

[2] Corey KE, Kartoun U, Zheng H, Chung RT, Shaw SY. Using an electronic medical records database to identify nontraditional cardiovascular risk factors in nonalcoholic fatty liver disease. Am J Gastroenterol 2016;111(5):671–6.

[3] Mehta N, Singh T, Lopez R, Alkhouri N. The heart age is increased in patients with nonalcoholic fatty liver disease and correlates with fibrosis and hepatocyte ballooning. Am J Gastroenterol 2016; 111(12):1853–4.

https://doi.org/10.1016/j.jclinepi.2019.01.001

DOI of original article: https://doi.org/10.1016/j.jclinepi.2018.05.008.

Funding statement: The author received honoraria and travel funding from The American Association for the Study of Liver Diseases (October 2017).

Conflict of interest statement: The author has declared that no competing interests exist. The author confirms that the commercial affiliation with IBM does not alter his adherence to all Journal of Clinical Epidemiology policies.