Data leakage in health outcomes prediction with machine learning

Applications of machine learning algorithms to predict the incidence of health outcomes have an enormous potential to improve clinical practice and lower healthcare costs [1]. Dr. Ye and colleagues’ recent article on hypertension incidence prediction on JMIR adds to this literature [2], but its potential gain in knowledge and applicability is hindered by a major flaw.

The objective of the study was to “develop and validate prospectively a risk prediction model of incident essential hypertension within the following year”. The authors follow good prediction protocols by applying a high-performing machine learning algorithm (XGBoost) and by validating the results on unseen data from the following year. The algorithm attained a very high area under the curve (AUC) of 0.870 for incidence prediction of hypertension on the following year.

The authors follow this impressive result by commenting on some of the most important predictive variables, such as demographic features, diagnosed chronic diseases and mental illness. The ranking of the variables that were most important for the predictive performance of hypertension, however, is included in the Multimedia Appendix and does not have any of these on the top. Of the six most important variables, five were: Lisinopril, Hydrochlorothiazide, Enalapril maleate, Amlodipine besylate and Losartan potassium. All of these are popular antihypertensive drugs.

Data leakage occurs when one or more features used to train the algorithm has hidden within itself the result of the outcome, and is considered one of the most frequent mistakes in machine learning [3]. This is different from predictive importance as it usually comes after the outcome, i.e. it is a consequence of the outcome that is being predicted and not the other way around.

A classic example from machine learning textbooks is the inclusion of the ID number of the patient as a predictor. While this should not have predictive importance if randomly assigned, it is frequent that patients coming from the same hospital have similar ID numbers in multicenter datasets. In the case of cancer prediction, for example, machine learning algorithms will learn that similar ID numbers that come from oncology hospitals have a higher probability of cancer.

By including the use of antihypertensive drugs as predictors for hypertension incidence on the following year, the authors open the possibility that the machine learning algorithm will focus on predicting who already had hypertension, but just did not have this information on the medical record at baseline.

While this would work for a prediction competition [4], it is not of particular scientific or clinical interest. In the case of the latter, just one variable (the use of hypertension drug) is sufficient for physicians to infer the presence of hypertension, while for the former, the knowledge of this being a highly predictable event (as measured by the AUC) is severely impaired.

Reporting the predictive performance of specific datasets for different diseases is an important new field in epidemiology and data science. The authors can still contribute to this literature by providing the new AUC of the prediction after addressing the data leakage issue.
References


