Crystal Ball or Magic8 Ball? Reply Hazy, Try Again

A. James O'Malley, PhD^{1,2} and Sandra L. Wong, MD MS^{1,3}

¹The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Lebanon, NH; ²The Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH; ³Department of Surgery, Dartmouth-Hitchcock Medical Center, Lebanon, NH

In this issue of *Annals of Surgical Oncology*, Zabor et al.¹ present a comparison of three online melanoma prediction tools. By applying data inputs from a single validation set, the authors demonstrate variability in prediction of individual survival across the tools. The findings highlight important distinctions not only in how the tools were developed, but also in the potential for the tools to contain inherent bias, the breadth of the tool's use or applicability, and the utility of a tool for its intended population (e.g., its ability to correctly discriminate patients with different levels of risk). The article also emphasizes the importance of understandable information representation as it applies to data interpretation and risk communication for health care providers and patients. We applaud the authors for highlighting several practical concerns about prediction tools and their unregulated availability, including the lack of documentation on their assumptions and performance.

In the age of data science, a surge of statistical techniques and the ever-increasing availability of computer software have led to the development of multiple prediction tools for a target (the outcome being predicted). The fact that different tools can yield substantially different predictions may be off-putting to clinicians who become frustrated with inconsistencies or may lead to misuse of the tools (e.g., selective adoption of a tool to support the use of a favored treatment). In the first instance, a clinicians struggling to incorporate tool-based predictions and other decision tools into their practice might give up trying to do

so. This highlights the need for predictive tools and other instruments to be accompanied by a clear statement of the assumptions and contexts under which they were developed, including the target population. Such information enables users to select the predictive tool most suited to their needs based on relevant medical or scientific reasons. Furthermore, if comfortable with multiple tools that each appear to be valid, a savvy user might combine the predictions into an overall prediction.

Predictive tools tend to be most relevant to the population whose data were used for training of the tools, and their utility for other populations typically is a decreasing function of how much extrapolation is needed. If different tools contain different predictors, it might be that none of them captures the true data-generating process. However, as British statistician George Box famously stated, "All models are wrong, but some are useful." Subject matter experts should always be consulted for assessment of the expected bias, if any, from the absence of a key predictor in a tool. This should occur both when the tool is designed, with suspected important key predictors reported by the developers together with the tool's performance, and when it is being considered for use.

The tools considered by Zabor et al.¹ involved different sub-populations of patients. Not surprisingly, the outputs from the tools differed significantly. It is concerning that a sizeable number of patients were omitted from re-evaluation of the tool. If patients with missing data differ from those with no missing values for the outcome or the predictors, the tool is likely to be biased in the sense of misrepresenting the population of interest as a potentially unusual subgroup with no missing data. This problem can be overcome by using multiple imputation before estimation of the tool. Appropriate handling of missing data should be encouraged when tools are developed. Therefore, although it is commendable that Zabor et al.¹ used inverse

censoring weights to deal with censored outcomes, it is to be regretted that they did not also appropriately handle the missing covariates. Because of the selective tendency in the development of predictive tools, the uncertainty in a prediction is almost universally under-acknowledged. The article describes the importance of testing a tool on data different from that used to train the tool. This leads to the discussion of the tool's calibration.

GENERAL ADVICE FOR TOOL CONSTRUCTION

A predictive tool should (1) fit the data on which it is trained, (2) make intuitive sense, and (3) predict unseen data nearly as well as it fits the training data. Standard model fitting and assessment methods in statistics (e.g., residual analysis) apply to the first factor. Subject matter experts are key to engage for establishment of the second factor. They provide a "sanity check." The tool developer should not rest until the reason for any nonintuitive features of the tool have been understood. One common mistake is to include a posttreatment variable in the model. Because such a variable may not be known at the point of deciding when the tool is intended to be used in practice, this may render the tool useless. The closer a predictive tool is to capturing the true data-generating mechanisms (i.e., to being a causal model), the more stable its performance is likely to be.

We focus in more detail on the third factor because it is the most involved step from the standpoint of a tool developer and contains key pieces of information that should be reported with the tool. Notably, all the factors in the following discussion pertain to the test data (the data withheld from the tool during its development) or an external set of data. The following three tests were proposed by D'Agostino et al.² as a basis for evaluating a predictive tool.

Effect Comparison

Traditionally, tools were based on regression models, which allow the key components of the tool, the regression coefficients, to be compared between the training and test samples. If the estimated regression coefficients are similar (e.g., their standardized difference is small, ideally less than 0.1), a developer can be assured that any model-building efforts that occurred in the tool development phase were not erroneously affected by data dredging. In other words, the tool appears stable and not a result of random idiosyncrasies in the data.

Recently, machine learning, neural networks, random forests, and other algorithmic-based methods have been used to develop predictive procedures. Although such

computationally heavy methods hold promise from the standpoint of predictive accuracy, their complexity makes them appear as black boxes, and parsing out the contribution of individual terms appears challenging. The assessment of intuitive sense and effect comparison between the training and test samples are elusive and unused elements of tool construction.

Calibration

Calibration measures the extent that the predictive outcomes match those of the actual outcomes. The test data are placed into "bins" or groups of similar predictive risk, and then for each such bin, the actual result is plotted against the predictive result using the tool. A perfect 45° line represents perfect calibration. Deviations from this nadir can be tested for statistical significance using a chi-square test (Hosmer–Lemeshow test).

Discrimination

The ability of the model to correctly discriminate cases with larger values of the outcome from those with smaller values typically is measured using the c-statistic or area under the receiver operating characteristic (ROC) curve.³ This is determined using both the model developed on the training data and the model developed on the test data. The difference represents the extent to which using the tool on new data compromises the performance (discriminative ability) of the tool—a penalty for real-world use. A big difference suggests that the predictive tool may suffer from overfitting. That is, it includes terms of no predictive value, which, therefore, only add noise to the predictions.

Cancer prediction tools have many uses. Foremost is their use for patient education and treatment decision making. Aside from perfunctory disclaimers that prediction tools should not be used as a substitute for medical advice, diagnosis, or treatment of any health condition or problem, most prediction tools do not come with "instructions" that would ideally spell out assumptions, applicability, and performance of the tool. No prediction tool can be expected to perform perfectly, and the health care provider is the one who must combine the prediction with expert clinical judgment. The relative weight to be placed on the tool's prediction versus the expert's opinion could be informed by registries or other tracking methods that collect these variables, together with treatment decisions and outcomes over time. This level of re-calibration could take prediction and prediction tools to the next level.

REFERENCES

1. Zabor EC, Coit D, Gershenwald JE, et al. Variability in predictions from online tools: a demonstration using Internet-based melanoma predictors. *Ann Surg Oncol*. 2018. <https://doi.org/10.1245/s10434-018-6370-4>.
2. D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P; CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*. 2001;286:180–7.
3. Zou KH, O'Malley AJ, Mauri L. Receiver operating characteristic (ROC) analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115:654–7.