Multivariable Risk Prediction Models

It’s All about the Performance

In this issue of Anesthesiology, Dalton et al. report on the development, validation, and recalibration of a prediction model with the fundamental aim to provide an accurate tool for risk adjustment to compare health outcomes (inhospital mortality) across hospitals. With the increasing availability of observational, registry, and hospital audit data sets, there has been a surge in developing new multivariable risk prediction models (risk scores, diagnostic, and prognostic models). Risk prediction models can be used not only to aid in a diagnostic or prognostic setting, but also as a risk-adjustment tool in a monitoring and auditing context. One of the key differences between models for diagnosis or prognosis and risk adjustment for hospital profiling is the former type of models is for clinical use on individual patients, whereas the latter is typically retrospective for groups of patients.

The purpose of the risk adjustment is to try and create a level playing field for comparing healthcare providers so that hospitals with different case-mix can be fairly compared. Risk adjustment helps to disentangle variations in patient outcomes attributable to patient characteristics that are generally not under the control of the treating clinicians or healthcare provider from aspects such as quality of care, as well as unmeasured case-mix factors and any data errors.

Using a subset of 10 million records from an impressively large registry containing 24 million inpatient discharges, Dalton et al. developed a prediction model using information present-on-admission model, namely, International Classification of Disease, 9th Revision, Clinical Modification, diagnosis and procedural codes in addition to age and gender. Using the elastic net method to derive their model, which shrinks regression coefficients toward zero in an attempt to avoid overfitting, the final model included a colossal 1976 regression coefficients. Characterizing the predictive ability of a risk prediction model are concepts of discrimination and calibration. Discrimination is the ability of the model to differentiate between individuals who do and do not die, i.e., those who ultimately go on to die should generally have higher predicted risks compared to those who do not. Calibration reflects the closeness predictions to observed outcomes. Both components are important, yet calibration is frequently overlooked. In their analysis, Dalton et al. emphasize the importance of having a sufficiently well-calibrated model. Despite the impressively large dataset, with a large number of deaths that drive the effective sample size in prediction studies (although interestingly only the event rate for the 2009 validation cohort is reported), the calibration of the present-on-admission model on the internal validation data (random 20% of the original dataset) was surprisingly quite poor. It remains unclear what contributed to the poor calibration, whether it was the inclusion of nearly 2,000 predictors or whether there was any difference in the case-mix of these 2,000 predictors between the original development and initial validation cohorts; although given the large datasets, random splitting will tend to produce two very similar datasets, except by chance. The elastic net approach to develop the model should also be questioned; typically such an approach is used when the number of predictors exceeds the number of individuals in the study (such as in genomics), which clearly was not the case here. Thus, it would be useful to have a clearer understanding to what extent the elastic net approach is useful and examine subsequent performance properties in both internal and external validation cohorts. If use of the elastic net produces a poorly calibrated model on very similar data used to develop the model, then while recalibration could fix this, it seems far from ideal. We agree with Dalton et al. that models should indeed undergo periodic recalibration to reflect...
temporal changes in case-mix (predictors and outcomes), yet in practice this disappointingly rarely happens, and thus requiring users to constantly recalibrate the model before using it seriously hampers its usefulness. Furthermore, the flexible recalibration approach used requires sufficient understanding of the actual methodology, and whether this can be translated into reality remains to be seen.

In their analysis, Dalton et al.\(^1\) compare their present-on-admission model to their recently developed Risk Stratification Index that does not distinguish between factors known on admission and those occurring during hospitalization, thus severely limiting its usefulness.\(^8\) This raises a couple of critical issues, namely, the importance for authors to compare newly developed risk prediction models against existing models (or current practice), and also for reviewers, editors, and readers to question why a new risk prediction model is needed and for authors to fully justify why a new model is needed.

The large majority of the hundreds of risk prediction models published each year, regardless of their purpose, are not used and it is arguable in many instances whether there was any intention in doing so. Possible reasons contributing to the failure of uptake are likely to be concerns about the methods, opaque reporting, lack of independent validation, face validity, and ultimately lack of trust in these models due to a combination of these factors (and more). When contemplating deriving yet another new model, for which there are often preexisting models for the same outcome, authors should be strongly encouraged to conduct systematic reviews, either as a separate article or embedded within the same article.\(^9\) If a new model is warranted, then authors should, providing there is sufficient data available on predictors, conduct head-to-head comparisons against existing models to provide readers and potential users with adequate information to make a clear objective judgment about the usefulness of all the models.\(^10\)–\(^12\)

After recalibrating their model,\(^13\) Dalton et al.\(^1\) carried out a temporal validation using data from 2009. The resulting performance appears excellent with good calibration and remarkably high discrimination (c-statistic = 0.958). However, before considering using the model, and indeed any prediction model, an external validation of the model should be carried out, again comparing to existing competing prediction models. Ideally this should be done by truly independent investigators using a completely separate contemporary or more recent (collected) dataset from that used to derive the model (different hospitals).

Finally, it is important the authors provide full and transparent reporting of all details used to derive and validate their models.\(^14\) Numerous systematic reviews evaluating the methodological quality of level of reporting of risk prediction models have all found deficiencies.\(^7\)\(^9\) Only when all the relevant information on quality of the data, statistical methodology, and model performance details are reported can readers and potential users make an informed evaluation. Expert-led consensus-based guidelines are currently being developed to aid editors, reviewers, authors, and readers of risk prediction models.\(^14\) While in this instance providing a table of characteristics is clearly infeasible for nearly 2,000 predictors, some basic clinical information on the 2004–2008 and 2009 cohorts would be useful to observe whether there were any noticeable differences. Also noted earlier, while the number of deaths in the study will be clearly large, most studies are not this large, so providing adequate information on the number of deaths in all cohorts used to derive and validate a model is required.

**References**