Measuring Risk Assessment Tool Performance

Emily Tiry and KiDeuk Kim

March 2021

Criminal justice stakeholders are increasingly using risk assessment tools to inform their decisionmaking, with the intention that these tools will help them achieve specific goals. Risk assessment tools support decisionmaking by classifying people into groups based on their risk of reoffending so that stakeholders can identify the appropriate levels of supervision and/or treatment. However, these tools need to be validated periodically to ensure that they continue to work well over time. During validation, the interpretation of the tool's performance metrics and the decision about which metrics should carry the most weight often depends on the context in which the risk assessment is used. For example, validation of a pretrial risk assessment tool might focus on different metrics from validation of a tool to predict sexual recidivism. In this brief, we describe the commonly used performance metrics and what factors to consider when interpreting them.

Types of Performance Metrics

Three main types of performance metrics used for evaluating the predictive performance of risk assessment tools include accuracy, calibration, and discrimination. Each type of metric provides information about a different aspect of predictive performance.

Accuracy measures how frequently a tool correctly predicts the outcome of interest (for example, recidivism). In other words, accuracy measures how often someone who is predicted to be high risk

reoffends, and vice versa. The opposite version of this concept is also used and is referred to as the "error rate." Practitioners are often interested in knowing the *type* of mistakes the tool might make. For example, they might want to know how often a tool inaccurately predicts that someone who does not reoffend scores as high risk ("false positive" errors) and how often someone who does reoffend is labeled low risk ("false negative" errors). This is explained in more detail in the next section.

Rather than grouping people into categories of high and low risk, **calibration** typically measures performance by comparing the predicted probability of an outcome with the observed probability of the outcome among a target population. As an example, say 40 out of a group of 100 people will reoffend. For that group, a risk assessment tool with good calibration would return an average predicted probability of reoffending of roughly 40 percent. Calibration is commonly measured using the Brier score, which averages the differences between the predicted probabilities and the actual outcomes (Brier 1950). A low Brier score means the tool is well calibrated (the lowest possible score is zero).

Whereas calibration measures how well a tool captures absolute risk, **discrimination** provides a measure of how well a tool captures relative risk. In other words, discrimination measures how well a tool separates people at high and low risk. A common measure of discrimination is the area under the curve (AUC) score, which describes the chance that a randomly sampled person who reoffends will have a higher risk score than a randomly sampled person who does not reoffend. An AUC score of 0.5 means that the tool is no better at deciding who is high risk than flipping a coin. An AUC score between 0.5 and 0.7 is considered poor, between 0.7 and 0.8 is acceptable, between 0.8 and 0.9 is excellent, and above 0.9 is outstanding (Hosmer, Lemeshow, and Sturdivant 2013).

How to Interpret Performance Metrics

The interpretation of the performance metrics for a risk assessment tool and the weight placed on each metric should vary depending on the context in which it is applied. In particular, there are two important characteristics of the context that stakeholders should consider: how often the outcome of interest occurs, and the costs of getting the prediction wrong. Let's use the prediction of sexual reoffending as an example of how to use these characteristics to interpret the performance metrics.

Frequency of Outcome

First, consider how often sexual reoffending occurs. Results of past studies have shown that the rate of sexual reoffending is relatively low—typically less than 10 percent are rearrested for another sex offense within one year (Langan, Schmitt, and Durose 2003; Sample and Bray 2003). Suppose that 10 out of 100 people will sexually reoffend within one year. This is an example of an "unbalanced" sample—there are many more people who will not reoffend than people who will. It is more difficult to correctly predict reoffending in these cases, both because there are fewer examples of reoffending to learn from, and because you can get good overall accuracy simply by always guessing the more common category. Suppose it is your job to guess whether people will reoffend or not in the next year. If you always guess that they will not reoffend, you will be correct 90 percent of the time. In this context, your overall

accuracy is good, but you would not be correctly identifying any of the people who actually will reoffend. To ensure that you are doing a good job in identifying that group, you might want to look at more specific accuracy measures such as the "sensitivity" of the tool—that is, the percentage of people who do reoffend who are correctly predicted as high risk.

The frequency of the outcome of interest also has important implications for a tool's calibration. Particularly in samples like this one where the reoffending rate is relatively low, there is a danger of overestimating the probability of reoffending, which could lead to unnecessary incarceration or supervision as well as expending resources that could be better used elsewhere. For that reason, it is important to monitor whether a tool is underestimating or overestimating risk in your population. As for discrimination, one benefit of using the AUC score to measure a tool's performance is that AUC is not as affected by how often the outcome of interest occurs (or, the "base rate") as the other types of metrics. It focuses more on measuring how well the tool can tell the difference between people who are high risk and people who are low risk, but not necessarily how well it measures the probability of reoffending. For this reason, AUC might be a good metric for stakeholders to consider if they are using the tool to determine how to prioritize and target a fixed amount of resources.

Cost of Incorrect Predictions

The other consideration in interpreting a performance metric is the cost of incorrect predictions. The tool could make two types of mistakes: it could predict that someone who would not reoffend is high risk, and it could predict that someone who would reoffend is low risk. In many applications, the cost of one type of mistake is greater than the cost of the other. Using the example of sexual offending: if you were to guess that no one would reoffend, you would be correct 90 percent of the time, but in 10 percent of cases a new sex offense would occur that might have been prevented if those cases had been deemed high risk. These new offenses would impose costs to the victims, to the criminal justice system, and so on. A jurisdiction might be willing to accept more false positives—people who are deemed high risk but would not have reoffended—to prevent more false negative predictions and to potentially prevent additional sex offenses from occurring. In other situations, such as a pretrial risk assessment tool that predicts the risk of failing to appear in court, jurisdictions might choose to use a tool that more often makes the opposite error. False negatives—or predicting that someone who will not show up for their court date is at low risk—might be more acceptable as to avoid the costs of unnecessary incarceration.

As mentioned above, inaccurate calibration can overestimate the probability of reoffending, which could lead to unnecessary treatment, supervision, or incarceration and impose not only budgetary costs but also human costs. Conversely, underestimating risk has the potential for significant costs, as the example above demonstrates. On the other hand, discrimination metrics—such as AUC—are not as helpful for understanding how accurate a tool's predictions are, so they are not as useful when considering the costs of misclassification.

Other Considerations

In addition to these statistical measures of risk assessment performance, stakeholders should consider the equity of the tool's performance across important subgroups such as race and ethnicity. So far, the field lacks widespread agreement about what constitutes an "equitable" versus "biased" risk assessment tool. Several groups of experts have proposed competing measures that are constructed in such a way that it is impossible for a tool to do well on all of them at the same time.[1] As with the statistical performance measures described above, the specific context in which the tool will be applied should help drive which metric is given the most weight (Roberts Freeman, Hu, and Jannetta 2021).

Conclusion

Whether choosing a new actuarial risk assessment tool or validating one that has already been implemented, practitioners should carefully consider their overall goals for implementing the tool, as well as the context in which they will be implementing it. In order to ensure the best possible performance within a jurisdiction, practitioners should interpret and weigh a tool's performance metrics with these factors in mind. In addition, a tool's performance might change as the local rules, laws, and populations in a jurisdiction undergo shifts that should be taken into account. This makes it important to revalidate a tool so that it can be modified, if necessary, to continue to perform as it is expected.

Note

[1] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *Washington Post*, October 17, 2016, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas.

References

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Hosmer, David W., Jr., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Langan, Patrick A., Erica L. Schmitt, and Matthew R. Durose. 2003. *Recidivism of Sex Offenders Released from Prison in 1994*. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Roberts Freeman, Kelly, Cathy Hu, and Jesse Jannetta. 2021. *Racial Equity in Criminal Justice Risk Assessment*. Washington, DC: Urban Institute.

Sample, Lisa L., and Timothy M. Bray. 2003. "Are Sex Offenders Dangerous?" *Criminology & Public Policy* 3 (1): 59–82.

About the Authors

Emily Tiry is a research associate in the Urban Institute's Justice Policy Center, where her current research focuses on developing and validating risk assessment tools, criminal court case processing, and measuring sexual misconduct and assault in the business context.

KiDeuk Kim is a senior fellow in the Justice Policy Center, where he leads multidisciplinary research teams to examine issues related to criminal justice interventions and policies. He has extensive research experience and national expertise in actuarial decisionmaking in criminal justice, criminal case processing, and policy evaluations.

Acknowledgments

This brief was supported by Grant No. 2015-ZB-BX-K004 awarded by the Bureau of Justice Assistance. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. The Bureau of Justice Assistance is a component of the Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.