Public Safety Risk Assessment CLEARINGHOUSE

Validation of Risk Assessment Tools

What is Validation?

Risk assessment tools can inform criminal justice decisions, such as whether to release or detain pretrial defendants awaiting trial and what level of supervision to impose upon probationers. By providing an empirical and objective assessment of an individual's risk of recidivism, these tools can improve upon traditional practice of criminal justice decision-making driven by professional judgment of criminal justice stakeholders. For this approach to be effective, however, the risk assessment tool must be as accurate as possible in its prediction of an individual's recidivism risk. It is also important to ensure that the risk assessment tool is legally and practically sound, and does not exacerbate existing disparities in criminal justice outcomes.

The process of determining how well a tool performs at predicting risk is called **validation**, and a risk assessment's performance is referred to as predictive validity. There are other types of validity, and they are typically discussed in the context of how individual risk factors are conceptualized and measured (see "Standards for Educational and Psychological Testing" for more information). In the context of evaluating the performance of risk assessment instruments, however, predictive validity is the primary measure of a tool's performance.

How to Validate Risk Assessment Tools?

For criminal justice risk assessments, validation tests whether a tool's estimated risk for an individual corresponds to actual behavior. This requires additional data against which the tool's prediction can be tested. Depending on the source of these data to be used for validation (or the mode of collecting such data), there are two general approaches to validating the performance of risk assessment tools. As illustrated below, this is also closely related to how the tool is developed and implemented.

First, validation can be done at the time of tool development. Once the tool developer identifies a population of interest and existing data about that population, a portion of the data can be used to devise a risk prediction model (i.e., the development sample) and the remainder of the data for validation tests. For example, if the case characteristics (e.g., charge information and criminal history) and outcomes (e.g., failure to appear and new arrest) of 1,000 pretrial defendants can be measured historically, the tool developer may randomly select 500 of them to develop a risk prediction model. The developer then tests the performance of the model using data on the remaining 500 pretrial defendants (i.e., the validation sample). In this approach, data can be partitioned into two splits or multiple splits. When multiple splits are used, the developer repeatedly examines one of the splits at a time and averages results from the multiple validation tests. The two splits method, known as hold-out validation, is commonly employed in criminal justice applications largely because it is easier to implement. But, it is important to recognize that (1) the performance of validation tests is sensitive to test settings and the data used and (2) examining validation in a few different ways is therefore highly recommended as opposed to relying solely on a single method or a single data run.

Second, historical data may not be readily available. The tool developer must then collect new data. For example, a justice agency may adopt an instrument that has been developed elsewhere without

any local customization. The agency can collect relevant information on the population of interest as they come through the justice system and conduct risk assessment on them. In this approach, the data needed for validation would be collected *prospectively* as the agency needs to follow their cases over time to observe their recidivism outcome. In principle, the instrument can be said to be valid to the extent that those identified as high-risk recidivate at a higher rate than those identified as low-risk. It is important to note that the agency runs the risk of misusing the instrument in this approach if the instrument was developed on a significantly different sample of individuals.

Example: Suppose you want to know how to buy corn seeds that are likely to give you the best yield performance. You are given 200 seeds to figure out what size, color, and shape of seeds are most likely to germinate and grow into healthy plants. You randomly pick 100 seeds and plant each of them to see which ones successfully grow on your property. You determine that mid-sized, uniformly round, bright yellow seeds are the most likely to grow in your local environment. This is your prediction model: medium, round, bright yellow seeds are most likely to thrive. To test your model, you try the remaining 100 seeds and see how often your guide leads you to find a seed that successfully grows. If the guide holds up well enough, your seed buying guide will be said to be "valid."

But, what does "well enough" mean? Suppose our buying guide helped us accurately identify successful seeds 50 out of 100 times. Some may say the performance of our buying guide is not good, equivalent to that of flipping a coin. However, if the proportion of seeds that could grow into healthy plants is only 10 percent in the universe of all cone seeds, a 50 percent chance of identifying a potentially successful seed should be considered quite high. In other words, the predictive performance of risk assessment instruments should be evaluated and interpreted within the context of their use.

There are several statistical metrics that assess the performance of risk assessment tools. It is important to consider the context in which the tool is used when using such metrics. Equally important is to assess the tool's performance in various ways to ensure the tool reliably performs to the expectations of criminal justice stakeholders, as well as scientific standards. Relying solely on one particular diagnostic metric or one aspect of performance may not lead to a balanced, reliable evaluation of the tool's performance.

What Are the Key Metrics in Validation?

When considering which actuarial risk assessment tool to adopt, practitioners should ask questions about how the tool's validity has been assessed and how it can be improved. Here are some of the key questions regarding predictive validity that can be explored without too much technical detail:

- How well does a tool separate those who experience an outcome of interest (i.e., recidivism) from those who do not? This criterion is called "discrimination," and is most typically used to evaluate the performance of risk assessment tools. The Area Under the Curve (AUC) is one of the most widely used metrics. The AUC of 1.0 indicates perfect discrimination and the AUC of 0.5 indicates no discrimination.

- How accurately does the tool predict the likelihood of such an outcome? This criterion is called "calibration." For example, if we predict a 40% risk of recidivism for an individual released from a correctional facility, the observed frequency of recidivism should be approximately 40 out of 100 individuals with such a prediction. A graphical assessment of calibration can be easily implemented with predictions on the x-axis and the outcome on the y-axis. A perfectly calibrated instrument should yield a 45-degree line.

- How frequently does the tool inaccurately predict a low-risk individual to be at high-risk (i.e., false positive errors) and vice versa (i.e., false negative errors)?

- How sensitive are validation results to different test settings (i.e., different samples, methods)? How does the tool perform across subgroups by race, ethnicity, and gender?

Why Is Revalidation Important?

Agencies should plan to monitor and, if possible, improve the performance of tools on a periodic basis. This can be achieved through revising the tool's scoring algorithm or updating the list of risk predictors used. It is important to note that the performance of risk assessment tools may change over time and differ across populations. As test settings in which a risk assessment tool was developed change (e.g., characteristics of criminal justice populations, criminal justice interventions), the tool should be reevaluated and recalibrated for optimal performance. If a justice agency frequently adopts or updates their practice and experiences notable changes in the characteristics of their criminal justice populations, it would be important to reassess the performance of risk assessment tools at short intervals. The more accurately a tool predicts risk, the more effectively criminal justice interventions can be assigned, and the safer a community will ultimately be.

The **Public Safety Risk Assessment Clearinghouse** is a one-stop resource that provides practitioners and policymakers with up-to-date and objective information about risk assessment as well as training and technical assistance on its use.

For more information, please visit: https://psrac.bja.ojp.gov/

Suggested citation for this publication:

KiDeuk Kim (2017). *Validation of risk assessment tools. (Policy Brief Number 2017-04)*. Washington, DC: The Public Safety Risk Assessment Clearinghouse.

This project was supported by Grant No. 2015-ZB-BX-K004 awarded by the Bureau of Justice Assistance. 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 U.S. Department of Justice.