NCCD | National Council on Crime & Delinquency

Developing and Validating Risk Assessment Instruments for Justice Agencies

Chris Baird, Chair, NCCD Board of Directors

A Question of Evidence, Part Two
Brief 4 of 6
February 2017

Articles on risk assessments from the last few decades include numerous claims that these systems are actuarial models. In truth, most are not. True actuarial systems are developed in the following manner: Once a study cohort is identified, a wide range of variables is selected for inclusion in data analysis. Factors analyzed may be based on prior research, theory, speculation, or simple curiosity. A dependent (or outcome) variable—some measure of recidivism over a specific period of time—is also identified. Statistical analysis is then conducted to determine which factors are related to recidivism. Those not related are eliminated. The objective is to determine which combination of risk factors most accurately classifies the cohort into different levels of risk based on actual outcomes.

If a study cohort is large enough, it is divided into construction and validation samples. This is important, because the best results obtained are almost always attained with construction samples. Once an instrument is developed, testing it on a second, or validation, sample provides a better estimate of accuracy in actual practice.

If the study cohort is not sufficient to create two samples, the instrument may be implemented in the agency for which it was developed. Then, to determine how well the risk tool works in actual practice, a prospective validation should be conducted using new cases. There are both advantages and disadvantages to prospective validation. One advantage is that it provides data not only from a separate sample, but from a different time period as well. The principal disadvantage is that prospective validations take much longer to complete.

In essence, actuarial systems are produced by data analysis. However, most systems currently in use were not developed through analysis, but rather constructed by researchers or clinicians. Because developers cite prior research and theory in factor selection, these systems contain variables thought to be related to recidivism (Vincent, Guy, & Grisso, 2013). However, subsequent validation studies have shown that many factors contained in these instruments are not related to continued criminal or delinquent activity. As noted earlier, Flores, Travis, and Latessa (2004) found this to be true of the YLS; Austin and colleagues found it true of the LSI-R (2003). In a comprehensive study of instruments used in juvenile justice, NCCD found that several instruments, including the YLS/CMI, PACT, COMPAS-Youth, and the YASI, all contain factors with little or no relationship to recidivism (Baird et al., 2013). Table 1 lists factors from the analysis of PACT that demonstrated little or no correlation with recidivism.

Table 1: Correlations for Selected PACT Risk Factors and Recidivism for Probationers in Florida

Risk Factor	Correlation
Prior Weapon Referrals	0.00
Prior Felonies Against Persons	0.02
Escapes	0.00*
Commitment Orders/One Day or More	0.03
Gender	0.04
History of Mental Health Issues	0.04

*Actual value is 0.002.

Factors not related to recidivism introduce substantial "noise" and dilute the relationship between overall risk scores and outcomes. The result of including such factors is, that while most of these instruments contain enough real correlates with recidivism and demonstrate a modest ability to classify cases, accuracy could be improved by removing factors unrelated to outcomes. In many instances, the level of improvement attained using true actuarial techniques is substantial (Baird et al., 2013). Researchers need to return to true actuarial development methods to ensure optimal classification of cases. If cases are not accurately classified, all other goals of case management may be seriously compromised.

The research field has also generally abandoned the most important means for evaluating the validity of risk assessment instruments. As noted by Gottfredson and Snyder (2005), two measures should be used to establish the validity and utility of risk assessment systems: (1) the degree of discrimination observed between recidivism rates for cases at different levels of risk and (2) the distribution of cases throughout the risk levels. An earlier National Institute of Corrections publication (Baird, 1991) stipulated the same criteria for evaluating the efficacy of risk assessment instruments. Silver and Banks (1998) not only identified these criteria as critical but actually developed a summary statistic that assesses how well a cohort is partitioned into different risk groups and the extent to which group outcomes vary from the base rate for the entire cohort. Their work was predicated on the position that distribution and the level of discrimination attained are critical to understanding the power of any system. While measures of specificity, sensitivity, association, and false positives/false negatives are useful, they simply are general measures of validity that do not accurately convey the utility of a system in everyday decision making. Yet claims of validity are frequently based solely on these measures. Most analyses published between 1995 and 2010 did not report recidivism rates for different risk groups.

In recent years, risk assessment validity has been based almost exclusively on two measures: the AUC (area under the curve) or simple correlations between risk scores and recidivism. Both are general measures of validity and, while useful measures, do not take into account two factors that are enormously important: the overall recidivism rate for the study cohort and the distribution of cases across risk levels.

The AUC has become particularly popular in recent years. Supporters cite the fact that this measure does not consider base rates or the distribution of cases across risk levels as strengths of the AUC, noting that this allows for easy comparisons of results

across systems. These "strengths," however, are in reality serious weaknesses. The AUC represents the chance that a true positive (i.e., a recidivist) selected at random will have a higher risk score than a true negative (a non-recidivist) also randomly selected. However, there are many scenarios where a high AUC can be attained for a system that produces low levels of discrimination and has extremely limited utility. This is especially true when there are few true positives (i.e., rates of recidivism are low). Hence, when one instrument clearly produces a higher level of discrimination between risk levels than another, AUC values for the two scales can be similar. This allows supporters of risk instruments that are based on prior research and theory to insist their instruments are as accurate as actuarial models.

This is precisely what a group of respondents did to challenge the conclusions of a recent study of instruments used in the juvenile justice field (Baird et al., 2013). Interestingly, these reviewers made no claim that the later-generation instruments were better, only that they were equivalent in terms of predictive accuracy, a step back from earlier claims made by Andrews, Bonta, and Wormith (2006). Their view is that tools with similar AUCs should produce approximately equal classification results; it is only a matter of selecting the proper cut points. There is little evidence, however, that this is true. In the study cited above, different cut points were used for all of the tools analyzed in an effort to optimize classification results. Improvement was noted for only one system, which was an anomaly due to a scoring system that produced a very narrow range of scores and the fact that two risk factors accounted for virtually all of the discrimination attained.

To understand how misleading the assumption that similar AUCs translate into equal classification tools is, consider one of the examples used to support this argument. AUCs for a risk assessment instrument used in Georgia and an actuarial instrument developed using data from the same jurisdiction were .64 and .67, respectively. This fact, combined with similar comparisons from other jurisdictions, led the respondents to conclude: "Fundamentally, this study provides evidence that tools that differ in their length, format, and foci can achieve similar levels of predictive utility." However, other measures of predictive utility clearly demonstrate that sole reliance on the AUC is problematic. Table 2 compares discrimination results as well as distribution across risk levels for the same two assessments. The original tool placed only 1% of cases in the high-risk category: In effect the system identified only two risk categories and placed nearly nine of every 10 youth at the lowest risk level. The actuarial system produced a much better distribution across risk levels and effectively separated cases into low-, moderate-, and high-risk categories. As a result, the DIFR statistic developed by Silver and Banks to measure the power of a risk prediction model was much higher (.61 vs. .40) for the actuarial system.

In sum, despite the instruments' similar AUC values, classification results from the actuarial system are clearly superior. The actuarial system has far greater utility, despite producing only a slightly higher AUC. This group of respondents appear to have given no consideration to other measures of predictive utility in reaching their conclusion.

The level of discrimination attained by different tools simply cannot be ignored. The primary purpose of risk assessment is to assign cases to different risk levels. Either implicitly or explicitly, the assigned risk level plays a role in case decision making, ranging from assigning a supervision level in the community to informing a decision to incarcerate a young person. Given the importance of these decisions, risk assessment systems must optimize differences in outcomes observed for cases at different risk levels. It is clear that assessments with similar AUC values often produce very different classification results.

The standard for measuring the efficacy of a risk assessment model should be the level of discrimination attained between outcomes for cases at each risk level. AUC values may be helpful in scale construction, but they fail to accurately convey how well a risk model operates in actual practice.

Table 2: Comparing Assessments With Similar AUCs

Risk Level	Actuarial Risk Instrument		Original Risk Instrument	
	% at Level	Recidivism Rate	% at Level	Recidivism Rate
Low	32%	17.0%	88%	25.3%
Moderate	44%	37.1%	11%	52.4%
High	24%	49.1%	1%	57.5%

References

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52*(1).

Baird, C. (1991). *Validating risk assessment instruments used in community corrections*. Washington, DC: National Institute of Corrections.

Baird, C., Healy, T., Johnson, K., Bogie, A., Wicke Dankert, E., & Scharenbroch, C. (2013). *A comparison of risk assessment instruments in juvenile justice*. Madison, WI: National Council on Crime and Delinquency.

Gottfredson, D. M., & Snyder, H. N. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts*. Washington, DC: National Center for Juvenile Justice, Office of Juvenile Justice and Delinquency Prevention.

Silver, E., & Banks, S. (1998). *Calibrating the potency of violence risk classification models: The dispersion index for risk (DIFR)*. Washington, DC: American Society of Criminology.

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). *Risk assessment in juvenile justice: A guidebook for implementation*. Chicago, IL: John T. and Catherine D. MacArthur Foundation.