COMMENTARY

Five Opportunities for Innovation in Violence Risk Assessment Research

Jay P. Singh

Global Institute of Forensic Research, Reston, Virginia, and Molde University College

Structured risk assessment instruments are frequently used to aid in a number of medico-legal decisions. With hundreds of available risk assessment tools and a steadily growing literature on their utility, it is important to examine possibilities for research innovation. The present article provides an overview of 5 recommendations for innovation in risk assessment research: (a) move beyond focusing on the idiographic-versus-nomothetic controversy and start focusing on the nomothetic-versus-nomothetic controversy, (b) move beyond discrimination validity, (c) move beyond using comparisons against chance, (d) move beyond using rules-of-thumb to justify claims of predictive validity, and (e) move beyond assuming incidents of recidivism occurred after index offenses.

Keywords: violence, risk assessment, forensic, crime, innovation

Over the past 30 years, the development of structured instruments to aid in the evaluation of violence (including sexual violence) risk has become a cottage industry. There are now literally hundreds of available risk assessment tools and a large research literature suggesting their superior reliability and predictive validity over unstructured clinical judgments (Singh, Serper, Reinhart, & Fazel, 2011). With over 40 systematic reviews and meta-analyses on the topic (Singh & Fazel, 2010), and an average of 17 new risk assessment articles being published each month (Alliance for International Risk Research, 2013), it is important to take a step back from this fast-moving train and examine possibilities for research innovation. The present article provides an overview of five opportunities for innovation in risk assessment research, inspired by the criticisms brought forth by DeClue and Zavodny (2014), in this issue.

Whereas DeClue and Zavodny focus their arguments on a single actuarial risk assessment tool, the Static-99R, this report aims to provide recommendations applicable to research on any structured instrument.

Recommendation 1: Move Beyond the Idiographic-Versus-Nomothetic Controversy and Start Focusing on the Nomothetic-Versus-Nomothetic Controversy

Over the past 7 years, much research and commentary has focused on the issue of whether actuarial risk assessment tools can reliably apply group-based estimates of recidivism risk to individuals (e.g., Cooke & Michie, 2010; DeMatteo, Batastini, Foster, & Hunt, 2010; Hart, Michie, & Cooke, 2007; Scurich, Monahan, & John, 2012). This has come to be known as the idiographic-versus-nomothetic controversy. Although the discussion of ecological fallacies is important in principle, what seems to have been forgotten is that there is only a pressing need for research in this area if practitioners are actually assigning group-based probabilities to individuals (e.g., “Mr. X has a Y% probability of recidivating in Z years”). But only a minority of practitioners is doing so.
Indeed, according to recent international survey evidence (Singh, 2013b), less than a quarter of clinicians communicate risk in this manner. Abandoning actuarial risk assessment tools based solely on the argument that group-based estimates are not applicable to individuals is a slippery slope. By this logic, most research findings should be disregarded, as the preponderance of scientific investigations attempts to make conclusions about individuals based on group-level findings. Consider the public health problems that would result from disregarding widely used medical screening tests such as mammograms or prostatic-specific antigen tests due to their use of normative cut-off thresholds to inform decisions about individuals.

In the end, what is worrying in the actuarial risk assessment literature is not so much that group-based estimates cannot be applied to individuals, but rather that group-based estimates appear to not be applicable to groups. Two meta-analyses in the past year have found that rates of recidivism in groups judged to be at “high risk” by both sex offender (Singh, Fazel, Gueorguieva, & Buchanan, 2013) and violence risk assessment tools (Singh, Fazel, Gueorguieva, & Buchanan, 2014) vary so much that practitioners simply cannot assume manual-based probabilistic estimates of recidivism risk to be accurate. These large reviews found that sexual recidivism rates in high-risk groups were higher in younger samples assessed by actuarial instruments when conviction was the study outcome, and violent recidivism rates in high-risk groups were higher in samples with fewer men where an structured professional judgment (SPJ) tool had been administered to a population with an overall higher recidivism rate. And recent primary research has confirmed that, even when replication studies match the sample and design characteristics of normative investigations closely and follow manual-based protocols exactly, group-based recidivism rates still do not hold (Rossegger, Endrass, Gerth, & Singh, 2014; Rossegger, Gerth, Singh, & Endrass, 2013).

This nomothetic-versus-nomothetic quandary is more concerning than the idiographic-versus-nomothetic debate, as a main benefit of actuarial instruments over alternative risk assessment approaches is the production of probabilistic estimates of group-based recidivism. And if these estimates do not hold for different risk groups or risk scores (cf. Helmus, Thornton, Hanson, Babchishin, & Harris, 2012), then it could be argued that the actuarial approach has a difficult road ahead in forensic mental health.

**Recommendation 2: Move Beyond Discrimination Validity**

In the context of structured risk assessment, predictive validity refers to the ability of an instrument to predict the likelihood of an adverse outcome (Singh & Petrila, 2013). This form of construct validity can be divided into two equally important components: discrimination and calibration (for a primer, see Singh, 2013a). Consistent with a diagnostic approach, discrimination refers to the ability to retroactively detect whether recidivists were judged to have been at higher risk than nonrecidivists. Consistent with a prognostic approach, calibration refers to the ability to prospectively predict whether individuals judged to be at higher risk are more likely to recidivate. Given that risk assessment in practice relies inherently upon conditions of uncertainty, the calibration component of predictive validity is perhaps more relevant. But calibration performance indicators are reported in less than a third of risk assessment predictive validity studies, with the general reasoning being that discrimination indicators are less influenced by outcome prevalence rates and, hence, easier to compare across studies (Singh, Desmarais, & Van Dorn, 2013). This said, both discrimination and calibration validity should arguably be established before implementing an instrument in practice.

As Bayesian thought is currently in vogue in the field (e.g., Beauregard & Mieczkowski, 2009; Harris & Rice, 2013; Wollert, 2006), it may be that the time has come to explore advancing the statistical methodology used in the field to take into consideration the issue of outcome prevalence rather than to systematically sweep this issue under the proverbial rug. For example, instead of providing readers with receiver operating characteristic (ROC) curves, predictiveness curves could be constructed (Pepe et al., 2008). Rather than publishing areas under the ROC curve (AUC), probabilistic AUCs could be reported (Shiu & Gatsonis, 2008). Should researchers wish to combine both discrimination and calibration in a single performance indicator, the Brier score and its var-
ious decompositions could be calculated (Rubach, 2010). There is mixed evidence of utility for “fast and frugal” approaches (Gigerenzer & Goldstein, 1996, p. 650) such as decision trees and neural network models (Caulkins, Cohen, Gorr, & Wei, 1996; Grann & Långström, 2007).

**Recommendation 3: Move Beyond Using Comparisons Against Chance**

Risk assessment researchers have a number of statistical methods at their disposal when measuring predictive validity. Many of these methodologies—including correlation, regression, and ROC curve analysis—incorporate null hypothesis significance testing (NHST). NHST is a statistical method by which the likelihood of a research hypothesis being “true” is evaluated (Carver, 1978). The conventional null hypothesis in the behavioral sciences is chance. However, there are no practitioners whose alternative to using a risk assessment tool is simply flipping a coin (and, if there are, please alert me so malpractice charges can be brought forth). As a field, we cannot be content to rely on methods of statistical testing that lack practical meaning.

The ideal comparison against which to compare the performance of risk assessment tools would be unstructured clinical judgments. However, to recommend that case-control study designs be used when clinical guidelines suggest that structured assessments are preferable to unstructured evaluations would be unethical. An alternative would be to use routinely available pieces of information that would likely drive unstructured clinical judgments as a proxy. Recent research has shown that simple models composed of just three pieces of routinely available information (age, sex, and criminal history) produce rates of discrimination validity similar to widely used risk assessment scales (Fries, Rossegger, Endrass, & Singh, 2013; Buchanan & Leese, 2006; Fazel, Singh, Doll, & Grann, 2012). This transition would mean that the p values produced by commonly used statistical packages, which assume a null hypothesis of chance, would no longer be reported. Rather, predictive validity estimates produced by tools versus routinely available demographic information would be compared using established statistical tests such as the Steiger (1980) z test or the Pearson and Filon (1898) z test for differences in correlations, the Breslow and Day (1987) χ² test for differences in odds ratios, or the DeLong, DeLong, and Clarke-Pearson (1988) χ² test for differences in AUCs. In interpreting the findings of these tests of differences, it should be remembered that statistical significance—no matter how it is calculated—does not necessarily mean practical importance, and smaller p values do not necessarily mean stronger relationships (Gliner, Leech, & Morgan, 2002).

**Recommendation 4: Move Beyond Using Rules-of-Thumb to Justify Claims of Predictive Validity**

There is little question that the AUC is the most commonly used performance indicator in predictive validity studies of risk assessment tools. In their systematic review of predictive validity methodology, Singh, Desmarais, and Van Dorn (2013) found that AUCs were frequently labeled as small, moderate, or large in terms of their magnitude. The usefulness of such rules-of-thumb is predicated upon there being some agreement in the field as to what ranges of AUCs constitute small, moderate, and large effect sizes. But there is no such agreement. Both the Singh review and an overview by Mossman (2013)—the clinical researcher who introduced the field to ROC curve analysis in the 1990s—have concluded that there is too much variability in rules-of-thumb for them to be practically useful. Rather than relying on rules-of-thumb, AUCs should be interpreted for readers using a standardized definition (e.g., “The probability that a randomly selected recidivist had a higher risk classification than a randomly selected nonrecidivist was X%”). Surprisingly, such a straightforward interpretation is provided in a minority of studies. Also, it is important to note that risk assessment tools that perform well at identifying recidivists but poorly at identifying nonrecidivists can produce the same AUCs as tools that perform well at identifying nonrecidivists but poorly at identifying recidivists (cf. Singh, Grann, Lichtenstein, Långström, & Fazel, 2012). Hence, statements that an instrument should be used in practice solely because the AUCs it produces are comparable with those of other instruments should be read with due caution.
**Recommendation 5: Move Beyond Assuming Incidents of Recidivism Occurred After Index Offenses**

The use of criminal registers to detect incidents of recidivism is commonplace in predictive validity studies of risk assessment tools (Singh, Grann, & Fazel, 2011). Though much has been written on the potential drawbacks of relying solely on register-based outcomes like charges or convictions (e.g., Davies, Clarke, & Duggan, 2004; Monahan et al., 2001), comparatively little attention has been paid to the closely related issue of pseudorecidivism. Pseudorecidivism refers to new charges or convictions handed down after an index offense precipitated by incidents occurring before that index offense (Quinsey, Harris, Rice, & Cormier, 2006).

Consider the following situation: Mr. X is charged with threatening his father-in-law with a knife and is released on bail. The next day he gets into an argument at a local bar, a confrontation that ends in Mr. X murdering the bartender. He is summarily convicted of second-degree murder and incarcerated in the state penitentiary. Upon admission to the facility, a risk assessment tool is administered with hopes of establishing the likelihood of future harmful behavior. Six months later, Mr. X goes to court to face the charge of threatening his father-in-law and is convicted. Is this new conviction an act of recidivism? No—the new conviction was for an incident that occurred before the index offense. But if researchers interested in the predictive validity of the administered risk assessment tool rely solely upon conviction outcomes without cross-referencing date-of-offense information, then their findings concerning the utility of the tool will be biased. It is important to control for this issue of pseudorecidivism to ensure that outcomes rates are as accurate as possible.

**Concluding Remarks**

The most important aspect of risk assessment research can be summed up in a single word: transparency. Consumers of research—be they other researchers, practitioners, or policymakers—need to know exactly how an instrument was tested and on whom in order to judge whether findings are applicable to their particular context. Consensus-based guidelines for what should be routinely reported in risk assessment research have recently been established (Singh, Yang, Mulvey, & the RAGEE Group, 2014), offering a promising way forward. With high hopes for the field, what must be advocated when it comes to structured risk assessment is neither nihilistic cynicism nor unfettered optimism, but rather objective and respectful scientific caution. Risk assessment tools are not panaceas, but they do represent a significant improvement in psychometric performance over unstructured approaches.

**References**


INNOVATION IN RISK ASSESSMENT

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of individual users and is not to be disseminated broadly.


Received January 10, 2014
Revision received June 12, 2014
Accepted September 2, 2014

---

**Correction to Storey and Hart (2014)**


http://dx.doi.org/10.1037/tam0000022