Behavioral Sciences and the Law
Behav. Sci. Law **31**: 8–22 (2013)
Published online 13 February 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/bsl.2052

Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer

Jay P. Singh, Ph.D.*,†,‡

The predictive validity of violence risk assessments can be divided into two components: calibration and discrimination. The most common performance indicator used to measure the predictive validity of structured risk assessments, the area under the receiver operating characteristic curve (AUC), measures the latter component but not the former. As it does not capture how well a risk assessment tool's predictions of risk agree with actual observed risk, the AUC provides an incomplete portrayal of predictive validity. This primer provides an overview of calibration and discrimination performance indicators that measure global performance, performance in identifying higher-risk groups, and performance in identifying lower-risk groups. It is recommended that future research into the predictive validity of violence risk assessment tools includes a number of performance indicators that measure different facets of predictive validity and that the limitations of reported indicators be routinely explicated.

The field of violence risk assessment has grown rapidly over the past several decades, and with the development of a large number of structured assessment tools has come a literature seeking to establish the validity of their risk predictions. In the clinical context of structured violence risk assessment, *predictive validity* is the ability of total scores, probabilistic risk bins from actuarial instruments, or categorical risk judgments from structured professional judgment (SPJ) instruments to correctly assess the likelihood of violence. The assumption is that by accurately estimating violence risk, persons who could benefit most from the development of risk management plans and the provision of treatment resources can be identified, while those who pose the lowest risk can be screened out of interventions focused on violence risk reduction (Andrews & Bonta, 2010; Singh, Grann, Lichtenstein, Långström, & Fazel, 2012).

Broadly, the predictive validity of violence risk assessments can be divided into two components: calibration and discrimination. Calibration refers to how well a risk assessment tool's predictions of risk agree with actual observed risk, whereas discrimination refers to how well an instrument is able to separate those who went on to be violent from those who did not (Cook, 2007). Measuring only one of these components does not provide a complete picture of predictive validity. And as those performance indicators currently used in the risk assessment validation literature capture either calibration or discrimination but not both (Singh, Desmarais, &

*Correspondence to: Jay P. Singh, Ph.D., Department of Mental Health Law and Policy, University of South Florida, 13301 Bruce B. Downs Blvd., Tampa, FL 33612, U.S.A. E-mail: jaysingh@usf.edu
†Department of Mental Health Law and Policy, University of South Florida, 13301 Bruce B. Downs Blvd., Tampa, FL 33612, U.S.A.
‡Institute of Health Sciences, Molde University College, Molde, Norway.

Van Dorn, 2013),[1] it is imperative that authors of studies investigating predictive validity calculate and report more than one. A further dimension that needs to be taken into consideration when selecting performance indicators is whether they measure global accuracy (overall ability to identify both high risk and low risk groups), high risk accuracy (ability to identify high risk groups, specifically), or low risk accuracy (ability to identify low risk groups, specifically). The importance afforded to each form of performance will depend on the context in which assessments of risk are made.

The aim of the present article is to provide an overview of calibration and discrimination indicators that measure global, high risk, and low risk performance, including: sensitivity and specificity; the positive predictive value (PPV) and negative predictive value (NPV); the number needed to detain (NND) and number safely discharged (NSD); the diagnostic odds ratio (DOR) and logistic odds ratio (OR); the point-biserial correlation coefficient (r_{pb}); and the area under the curve (AUC). Each of these performance indicators captures a different dimension of predictive validity, and each has potential pitfalls that warrant recognition. The equations and operational definitions for these performance indicators are provided in Table 1, and, to assist in the conceptualization of the reviewed statistics, a simple flowchart has been developed (Figure 1).

SENSITIVITY AND SPECIFICITY

What Are They?

Sensitivity is a high risk discrimination index representing the proportion of violent individuals who were judged to be at high risk, whereas specificity is a low risk discrimination index representing the proportion of non-violent individuals who were judged to be at low risk (Altman & Bland, 1994a). Sensitivity and specificity are calculated using information available in a 2×2 contingency table, which organizes assessment and outcome information into counts of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) at a single cut-off threshold (Figure 2). TPs are individuals judged to be at high risk who went on to commit a violent act, TNs are individuals judged to be at low risk who did not go on to commit a violent act, FPs are individuals judged to be at high risk who did not go on to commit a violent act, and FNs are individuals judged to be at low risk who went on to commit a violent act. These four values can be computed using any software package that provides cross-tabulations between dichotomous variables. McNemar's (1947) χ^2 test can be used to compare differences in the sensitivities or specificities produced using risk assessment tools administered to the same individuals.

Potential Pitfalls?

Contrary to popular belief, sensitivity and specificity are influenced by the base rate of violence (Brenner & Gefeller, 1997; Li & Fine, 2011), making it a misconception that they are intrinsic properties of an instrument (Moons & Harrell, 2003). If a risk assessment instrument is reasonably accurate at discriminating between violent and non-violent groups, sensitivity tends to increase (and specificity decrease) as base rate increases,

[1] The likelihood statistics and Brier score are examples of performance measures that take into consideration both calibration and discrimination abilities.

DOI: 10.1002/bsl

Table 1. Equations and operational definitions of performance indicators in violence risk assessment research

Performance indicator	Equation	Operational definition
Sensitivity (Sens)	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	The proportion of those who engaged in an antisocial act who were judged to be at high risk
Specificity (Spec)	$\frac{\mathrm{TN}}{\mathrm{TN+FP}}$	The proportion of those who did not engage in an antisocial act who were judged to be at low risk
Positive predictive value (PPV)	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	The proportion of those judged to be at high risk who did go on to engage in an antisocial act
Negative predictive value (NPV)	$\frac{\mathrm{TN}}{\mathrm{TN+FN}}$	The proportion of those judged to be at low risk who did not go on to engage in an antisocial act
Number needed to detain (NND)	$\frac{1}{PPV}$	The number of individuals judged to be at high risk who would need to be detained to prevent a single antisocial act
Number safely discharged (NSD)	$\left(\frac{1}{1-\mathrm{NPV}}\right)-1$	The number of individuals judged to be at low risk who could be discharged before a single antisocial act
Diagnostic odds ratio (DOR)	$\frac{\mathrm{TP}\times\mathrm{TN}}{\mathrm{FP}\times\mathrm{FN}}$	The ratio of the odds of a high risk classification in those who engaged in antisocial act to the odds of a high risk classification in those who did not
Logistic odds ratio (OR)	$\frac{e^{a+bX}}{e^{a+bX}+1}$	The ratio of the odds of a lower risk classification in those who did not engage in an antisocial act to the odds of a higher classification in those who did
Point-biserial correlation coefficient (r_{pb})	$\frac{\bar{x}_{\mathrm{AntisocialClassification}}-\bar{x}_{\mathrm{NotAntisocialClassification}}}{\sqrt{\frac{1}{n_{\mathrm{Total}}}\sum_{x_{\mathrm{Min}}}^{x_{\mathrm{Max}}}(x-\bar{x})^2}}\times\sqrt{\frac{n_{\mathrm{Antisocial}}\times n_{\mathrm{NotAntisocial}}}{n_{\mathrm{Total}}^2-n_{\mathrm{Total}}}}$	The direction and strength of the association between risk classification and having engaged in an antisocial act or not
Area under the curve (AUC)	$\frac{1}{2}\sum_{T_{\mathrm{Min}}}^{T_{\mathrm{Max}}}(Sens_{T_i-1}+Sens_{T_i})\times(Spec_{T_i-1}-Spec_{T_i})$	The probability that a randomly selected individual who engaged in an antisocial act received a higher risk classification than a randomly selected individual who did not

Note: TP, number of true positives; FN, number of false negatives; TN, number of true negatives; FP, number of false positives; T, cut-off threshold; n, number of participants; $\bar{x}$, mean score; x, individual score; X, independent variable value; a, constant; b, slope.

although there is no exact functional relationship between prevalence and either performance indicator (Kraemer & Gibbons, 2009). A second potential pitfall of sensitivity and specificity is that they assume a single cut-off threshold on a risk assessment tool. This is problematic, as commonly used risk instruments including the Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 2006), the Level of Service

DOI: 10.1002/bsl

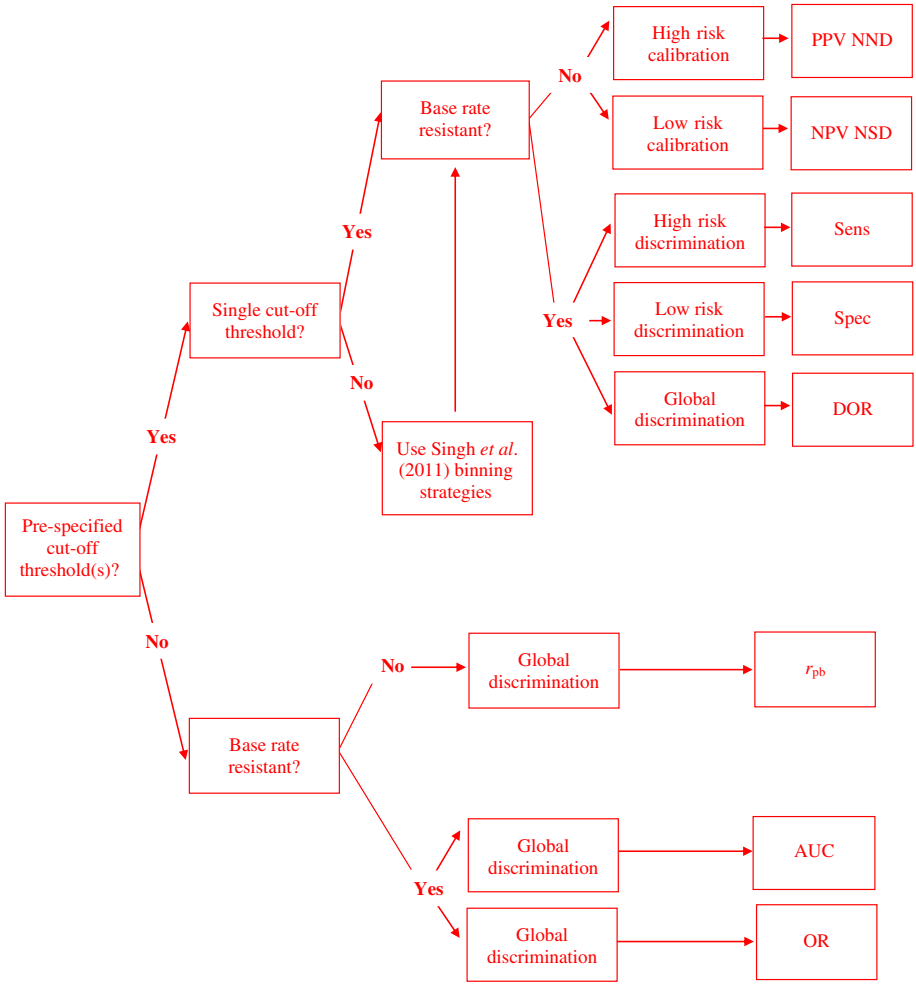

Figure 1. Performance indicator flowchart. PPV, positive predictive value; NPV, negative predictive value; NND, number needed to detain; NSD, number safely discharged; DOR, diagnostic odds ratio; OR, logistic odds ratio; r_{pb}, point-biserial correlation coefficient; AUC, area under the curve; Sens, sensitivity; Spec, specificity.

		Outcome: Violent	Outcome: Not violent
Assessment	High risk	True positive	False positive
	Low risk	False negative	True negative

Figure 2. 2 x 2 contingency table.

Inventory-Revised (LSI-R; Andrews & Bonta, 1995), and the Historical, Clinical, Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) were designed to use more than a single threshold. Finally, sensitivity and specificity are limited by their retrospective orientation, resulting in a lack of direct relevance to clinical decision-making (Guggenmoos-Holzmann & van Houwelingen, 2000).

DOI: 10.1002/bsl

POSITIVE AND NEGATIVE PREDICTIVE VALUES

What Are They?

Two performance indicators with arguably greater relevance to clinical decision-making are the PPV and NPV. The PPV is the proportion of those judged to be at high risk of committing a violent act who go on to do so, whereas the NPV is the proportion of those judged to be at low risk who do not (Altman & Bland, 1994b). These calibration indices capture more the usefulness of risk assessments in practice, as they emphasize the prospective prediction of violent outcomes. The information needed to calculate PPV and NPV can be found in 2×2 contingency tables. Differences in PPVs or NPVs produced using overlapping samples can be measured using Smith's (1951) χ^2 test, Bennett's (1972) χ^2 test, Leisenring, Alonzo, and Pepe's (2000) Wald test, or Wang, Davis, and Soong's (2006) Wald test.

Potential Pitfalls?

Both PPV and NPV are base rate-dependent and vary depending on the population, time at risk, and outcome of interest. Though this arguably increases their clinical validity and immediacy (McQuay & Moore, 1997; Pinson & Gray, 2003), it has been argued that such variation makes them difficult to compare across studies (Mossman, 1994a). The latter argument has been used to suggest that sensitivity and specificity (or their derivative, the AUC) will be superior to PPV and NPV when describing predictive validity unless the base rate of violence will always be the same whenever a risk assessment tool is used (Mossman, 1994b). However, comparing performance indicators that measure calibration and discrimination is akin to comparing apples and oranges: both are useful in their own right for different purposes. A second potential pitfall is that the PPV and NPV also rely on the use of a single cut-off threshold, limiting their usefulness in risk assessment schemes with more than two risk categories.

NUMBER NEEDED TO DETAIN AND NUMBER SAFELY DISCHARGED

What Are They?

Two comparatively new calibration performance indicators in the violence risk assessment literature are the NND and NSD. The NND calculates the number of individuals judged by a risk assessment tool to be at high risk of committing a violent act who would need to be detained in order to prevent a single incident of violence from occurring in the community (Fleminger, 1997). The NSD calculates the number of individuals judged to be at low risk who could be discharged prior to a single violent incident occurring in the community (Fazel, Singh, Doll, & Grann, 2012). These prospectively-oriented statistics are useful in that they simulate clinical decision-making by providing estimates of the number of individuals who may be either unnecessarily detained or discharged prior to necessary risk reduction when relying on the results of a risk assessment tool. The indicators are based on the number needed to treat

DOI: 10.1002/bsl

(NNT) statistic, which has gained wide acceptance over the past decade as a measure of treatment effect in the medical literature (Cook & Sackett, 1995). To calculate the NND and NSD, outcome information from a 2×2 contingency table is required. Current statistical packages do not calculate these performance indicators, though they are simple to compute manually (Table 1).

Potential Pitfalls?

The NND and NSD are limited in that their interpretation is a moral rather than a statistical matter. For example, some may consider the unnecessary detention of, say, five people to prevent the violent behavior of a sixth an appropriate measure to ensure public safety, whereas others may feel that the civil rights of those five unnecessarily detained individuals are of greater importance. Thus, the NND and NSD are difficult indicators for which to establish standardized guidelines for interpretation. A second potential pitfall of the NND and NSD is that, similar to the PPV and NPV, both are base rate-dependent. A third potential pitfall is that the NND and NSD rely on the use of a single cut-off threshold, though formulae for the statistical conversion from alternative performance indicators to the NNT may be adapted to help address this (Furukawa, 1999; Hilton, Reid, & Paratz, 2006; Kraemer & Kupfer, 2006).

DIAGNOSTIC AND LOGISTIC ODDS RATIOS

What Are They?

The DOR and logistic OR are global discrimination indices. The DOR is the ratio of the odds of a high risk classification in the violent group (i.e., the odds of a true positive) relative to the odds of a high risk classification in the non-violent group (i.e., the odds of a false positive; Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003). In cases where investigators wish to adjust these odds for potential confounding variables or to use continuous (e.g., total score on a risk assessment tool) as opposed to dichotomous independent variables, logistic regression may be used (Grimes & Schulz, 2002). Illustratively, an investigator might use logistic regression (and the resulting OR) to compute the odds of violent conviction adjusting statistically for age, sex, and criminal history. DORs are calculated using information available in 2×2 contingency tables, whereas logistic regression can be carried out in most general-purpose statistical packages, including SPSS (SPSS Inc, 2012), STATA (StataCorp., 2011), and SAS (SAS Institute Inc, 2012). Both DORs and logistic ORs are resistant to changes in the base rate of violence and range from 0 to infinity. As researchers and many clinicians are familiar with the concept of an odds ratio, they may also be easier for non-specialists to comprehend than more statistically complex performance indicators (Farrington & Loeber, 2000). Differences between odds ratios may be tested using Breslow and Day's (1987) χ^2 test of heterogeneity.

Potential Pitfalls?

A potential pitfall of the DOR arises from the number and distribution of violent cases across the cells of the 2×2 contingency table. Sample DORs can be biased and

DOI: 10.1002/bsl

unstable when the number of TPs, FPs, TNs, or FNs is small, resulting in a large standard error. Further, the DOR is undefined when either the number of FP or FN cases is zero, though this may be addressed by adding a constant of +1 to each cell (Higgins, Deeks, & Altman, 2008). A second potential pitfall of the DOR is that it is inferior to Youden's (1950) $\mathcal{J}$ index[2] for identifying the cut-off threshold that balances the sensitivity and specificity of a risk assessment tool. When both DOR and Youden's $\mathcal{J}$ are calculated for each cut-off threshold on a risk assessment tool, using the DOR results in a threshold nearer the extremes of the boundary range (Böhning, Holling, & Patilea, 2011). A related third potential pitfall arises from over-generalizing the DOR for a risk assessment tool that has multiple potential cut-off thresholds, as the DOR for different thresholds may vary considerably. Under such circumstances, logistic regression represents a more appropriate methodology for estimating odds ratios. However, logistic regression is limited by its inability to distinguish between risk assessment tools with high true positive rates and those with high true negative rates (Pepe, Janes, Longton, Leisenring, & Newcomb, 2004). While it has also been argued that calibration-based risk ratios (RRs) measuring relative risk are more clinically meaningful than odds ratios (Grimes & Schulz, 2008), the OR approximates the RR when the base rate of violence is low (Robbins, Chao, & Fonseca, 2002). Thus, the use of logistic regression may be more appropriate when violence or sexual offending are the outcomes of interest, but less so when general recidivism is being predicted. Guidelines for estimating relative risk from odds ratios are provided by Davies, Crombie, & Tavakoli (1998).

POINT-BISERIAL CORRELATION COEFFICIENT

What Is It?

The point-biserial correlation coefficient (r_{pb}) is a global discrimination index measuring the direction and strength of association between a continuous variable and a dichotomous variable (Das Gupta, 1960). In the context of violence risk assessment, the continuous variable is usually the total score on a structured instrument and the dichotomous variable is whether violence was perpetrated. The square of r_{pb} estimates the percentage of variance that is shared between the continuous and dichotomous variables. Although r_{pb} has also been used to investigate the association between violence and both actuarial risk bins and SPJ risk judgments, these categorical estimates are ordinal rather than continuous in nature. The r_{pb} coefficient can be computed by any software package that has a module for correlational analysis (e.g., SPSS, STATA, SAS). To test for differences in r_{pb}, Fisher's (1924) z-test can be used for independent samples, Steiger's (1980) z-test for overlapping samples, and Pearson and Filon's z-test for non-overlapping but correlated samples (Raghunathan, Rosenthal, & Rubin, 1996).

[2] Youden's $\mathcal{J}$ is equal to (sensitivity + specificity) – 1. The threshold that produces the highest $\mathcal{J}$ value is equivalent to the point of inflection on a ROC curve.

DOI: 10.1002/bsl

Potential Pitfalls?

The r_{pb} coefficient does not differentiate between TP and TN rates overall or for specific cut-off thresholds of a risk assessment tool. At most, r_{pb} provides an index of the association between risk assessments and a dichotomous outcome, but even in this capacity r_{pb} is easily misunderstood. Though r_{pb} is constrained to values that fall between –1.00 and +1.00, the range of possible values for r_{pb} may fall within a much narrower range depending on the prevalence of the dichotomous outcome (Breaugh, 2003). The further the base rate of violence deviates from 50%, the more constrained the possible values of r_{pb} (Nunnally, 1978). This issue is particularly important when predicting rare events. For example, in the case where only 5% of a given sample is violent, the maximum possible r_{pb} is 0.47. In the more extreme case where only 1% of a sample is violent, the maximum possible r_{pb} is 0.27. When the base rate of violent behavior is particularly low or high, it is advisable to compare the r_{pb} that has been found in the sample with the maximum possible r_{pb} achievable. Consider the case in which investigators obtain an r_{pb} of 0.20 in a sample with a base rate of violence of 1%. Were they to mistakenly assume that the maximum value of r_{pb} in this situation was +1.00, they might draw the conclusion that the risk assessment tool under investigation had relatively weak discriminative abilities, when in fact the r_{pb} obtained was close to the maximum value that could be obtained given the base rate. Tables of maximal r_{pb} at different base rates are available in the works of Lord and Novick (1968) and Gradstein (1986).

A further pitfall is that inferential tests of statistical significance for the r_{pb} coefficient are imprecise. The standard method in testing the significance of r_{pb} follows the same procedures as for the Pearson product–moment correlation coefficient, which measures the association between two continuous variables. However, these procedures are based on the linear regression model, the assumptions of which are not satisfied with r_{pb}. Some investigators have therefore recommended using maximum likelihood estimate methodology to test the significance of the r_{pb} coefficient (Pampel, 2000). However, it is important to keep in mind that maximum likelihood estimates are likely to be imprecise with samples of fewer than 500 participants (Long, 1997). Thus, unless maximum likelihood tests of significance are used with large samples, significance tests for r_{pb} should be regarded as approximations.

AREA UNDER THE CURVE

What Is It?

The AUC is a global discrimination index that is equal to the probability that a randomly selected violent individual received a higher risk classification (i.e., higher total score, actuarial risk bin, or SPJ risk judgment) than a randomly selected non-violent individual (Altman & Bland, 1994c). The AUC provides an index of an instrument's true positive rate (sensitivity) and false positive rate (1 – specificity) across cut-off thresholds. Although it has been argued that summarizing the receiver operating characteristic (ROC) curve, a plot of TP and FP rates across a risk assessment tool's cut-off thresholds, into a single number loses valuable information (Obuchowski, 2005), the AUC's reliance on estimates of sensitivity and specificity makes it resistant to changes in the base rate of violence. Apart from its cut-off independence and base rate resistance, the most

DOI: 10.1002/bsl

commonly provided justification as to why the AUC is used to measure the discrimination validity of violence risk assessments is because it has become the *de facto* standard in the field (Singh et al., 2013).

Conducting ROC curve analysis in order to obtain the AUC requires two pieces of information: (1) risk scores (e.g., each participant's total score on the VRAG), actuarial risk bins (e.g., into which of the five risk bins of the LSI-R each participant is classified), or categorical risk judgments from SPJ instruments (e.g., whether each individual is judged to be at low, moderate, or high risk on the HCR-20); and (2) a dichotomous outcome (e.g., violent conviction vs. no violent conviction upon follow-up). General-purpose statistical packages such as SPSS, STATA, and SAS have pre-installed ROC curve analysis modules. Third-party software such as ROCKFIT (Metz, Herman, & Roe, 1998), LABROC4 (Metz, Herman, & Shen, 1998), and PROPROC (Pan & Metz, 1997) have also been developed to conduct ROC curve analysis and calculate AUCs. Research suggests these programs produce similar estimates, although each has been found to have its shortcomings when applying ROC curve methodology (Stephan, Wesseling, Schink, & Jung, 2003). Tables have been published to assist in the conversion between r_{pb}, Cohen's *d*, and the AUC for comparative purposes in the violence risk assessment literature (Rice & Harris, 1995).

Statistical tests have been developed to assess differences in AUCs. The two most commonly used tests in the risk assessment literature are the parametric method of Hanley and McNeil (1983) and the non-parametric method of DeLong, DeLong, and Clarke-Pearson (1988). However, the developers of the former test recommended the use of the latter more than 15 years ago (Hanley & Hajian-Tilaki, 1997). If statistical tests of differences are not used, it has been suggested that one can compare the 95% confidence intervals of AUCs calculated using the same outcome to see whether there is evidence of overlap (Rugge, 2006) or whether the confidence interval contains the chance AUC value of 0.50 (Helmus & Hanson, 2007). However, such comparisons should be made keeping in mind that comparing 95% confidence intervals is not the statistical equivalent of an $\alpha = 0.05$ test of differences, a common misconception in the behavioral and medical sciences (Belia, Fidler, Williams, & Cumming, 2005).

Potential Pitfalls?

The AUC is commonly misinterpreted as measuring the calibration accuracy of risk assessment tools such that higher AUCs mean more accurate prospective prediction. To demonstrate why this is not the case, consider the following example: Let us imagine that we are conducting a study investigating the predictive validity of assessments made using the Static-99 (Hanson & Thornton, 1999), a commonly used actuarial instrument designed to predict the likelihood of recidivism in adult male sex offenders. According to the Static-99 manual, scores range from +0 to +12, with scores of +6 and above denoting that a sex offender is at high risk of recidivism. Say we recruit 100 sex offenders about to be released into the community, follow them for a pre-specified period of time, and then use criminal records to find out whether they were convicted of a subsequent sexual offense. Now say we find that only one of the 100 sex offenders recidivated, and that this individual had a Static-99 score of +5, whereas the 99 non-recidivists had scores from +0 to +4. In other words, no one was judged to be at high risk of recidivism. However, this exact situation would result in a perfect AUC of 1.00. So does the AUC really capture a tool's ability to accurately predict who will

DOI: 10.1002/bsl

offend in the future? Arguably not. In addition, the AUC does not take into consideration that actuarial tools such as the Static-99 were never meant to be used in this associative manner – scores or risk bins are supposed to be cross-referenced with statistical tables published by tool authors in order to convert total scores or risk categories into probabilistic estimates of future violence risk. Rather than measuring whether a tool accurately predicts future violence, the AUC serves as a rank sum measure of discrimination (Steyerberg et al., 2010), the statistical equivalent of the probability that a blindfolded clinician presented with two hats, one filled with the risk assessment results of individuals who were violent and one filled with the results of individuals who were not, would rummage around both and randomly pull out a higher risk classification from the former. Thus, AUCs of 1.00 do not represent perfect prediction, but rather perfect discrimination, and statistically significant AUCs are not a sign that an instrument has the ability to identify high risk individuals, specifically.

A second potential pitfall of the AUC is that it offers base rate resistant estimates of predictive validity in a field where clinical utility is inherently dependent upon base rates. ROC curve analysis was developed as a diagnostic methodology as opposed to a prognostic methodology (Cook, 2008). That is, ROC curve analysis and the AUC answer the question, "Could an adverse event that has already occurred have been predicted?" This is not the situation that professionals working in mental health and correctional settings find themselves faced with on a daily basis, needing rather to answer the question, "Will the prediction that I have made come true in the future?" These are two fundamentally different questions. While the former can be at least partially answered by calculating an instrument's sensitivity and specificity, which are resistant to (though not independent of) changes in the prevalence of an adverse outcome, the latter is functionally dependent on how often an adverse event such as violence occurs in a population. If, for instance, 99% of residents judged to be at high risk in the town of Dangerville go on to be violent, but only 1% of residents judged to be at high risk in the town of Peaceton go on to be violent, it makes little sense to not take such base rate information into consideration when making practical decisions about whether risk estimates made using a given assessment tool should be used to aid in decisions concerning public protection, individual liberty, and resource allocation.

Additional pitfalls of the AUC that have been pointed out in the recent statistical literature include:

- Small sample sizes ($n < 200$) result in large inaccuracies in the estimated population parameters underlying ROC analysis (Hanczar et al., 2010).
- Adding or removing risk or protective factors that significantly effect how many individuals are classified into the correct risk categories (e.g., actuarial risk bins or SPJ risk judgments) of a risk assessment tool has minimal effect on the AUC (Cook, 2007, 2008). Removing weaker predictors, which would result in greater parsimony, also has little effect on AUC values (Royston, Moons, Altman, & Vergouwe, 2009).
- The parameterization of the AUC currently used in the violence risk assessment literature does not take time at risk into account. Alternative models have been recently proposed to take stochastic information into account (Chambless & Diao, 2006; Heagerty & Zheng, 2005), though these have not made their way into the mainstream yet.
- Though intuitively analogous to likelihood tests, comparing AUCs to test for evidence of incremental validity when additional information is added to a model

DOI: 10.1002/bsl

composed of established predictors (e.g., testing whether AUCs increase when adding clinical judgments to total scores) is unlikely to produce valid findings. Evidence suggests this is because comparing AUCs is only of use when differentiating between models that have no discriminative utility and those that do (Marzban, 2004; Vickers, Cronin, & Begg, 2011; Ware, 2006). Alternatives to comparisons of AUCs for prognostic prediction have been shown to be more statistically powerful and produce more valid results in this regard (Pencina, D'Agostino, D'Agostino, & Vasan, 2008).

- More than any other performance indicator reported in the violence risk assessment literature, the AUC is interpreted according to benchmarks as to what constitutes a small, moderate, or large magnitude effect size (Singh et al., 2013). However, these benchmarks were never intended to be used to evaluate the performance of predictive models (Mossman, 2013), and there is considerable variation in rules-of-thumb, suggesting that caution is warranted when using them.
- A common counterpoint to the difficulty in predicting low base rate behaviors, especially severe forms of violence such as sexual offending (Vrieze & Grove, 2008), is that risk assessment tools predict violence better than simply "betting the base rate". While this could be true, the appropriate way to test such a hypothesis is not using the AUC. A chance AUC of 0.50 does not represent "betting the base rate," and an AUC significantly higher than chance does not represent an instrument's ability to prognostically predict the likelihood of violence better than the base rate. Recently developed methodology should aid researchers interested in establishing this property (Moskowitz & Pepe, 2004; Pencina et al., 2008).
- The inflection point of the ROC curve is commonly identified as the "optimal" cut-off threshold, because it is where an instrument balances sensitivity and specificity. However, once the practical considerations of misclassification costs and prevalence are added, and sensitivity and specificity pairs on the ROC curve are weighted accordingly, the "optimal" cut-off threshold can change dramatically (Perkins & Schisterman, 2006).

Supplementing the AUC

Reporting only the AUC, as do over half of violence risk assessment validation studies (Singh et al., 2013), does not provide adequate evidence of a risk assessment tool's predictive validity. The AUC measures discrimination but not calibration, meaning that it paints but half the picture. This said, available calibration performance indicators that could be used to describe an instrument's performance in identifying higher- versus lower-risk groups (e.g., PPV, NPV, NND, NSD) depend on a single cut-off threshold, which many modern risk assessment tools lack. While strategies have been developed to combine actuarial risk bins and SPJ risk judgments to use these performance indicators (Singh, Grann, & Fazel, 2011), what would arguably be more useful is a set of cut-off independent calibration indicators that measure high and low risk performance separately and could be reported alongside the discrimination-based and more global AUC. Such indicators are overdue though, a global index, the net reclassification index (NRI; Pencina et al., 2008) is already available. Novel graphical techniques such as test validation plots (Neller & Frederick, 2013), predictiveness curves (Pepe et al., 2008),

DOI: 10.1002/bsl

and predictive ROC curves (Shiu & Gatsonis, 2008) may also prove useful in portraying the calibration component of predictive validity across cut-off thresholds.

CONCLUDING REMARKS

It is difficult to believe that it has been over 30 years since Monahan (1981) seminally reviewed the extant dangerousness prediction literature using indices derived from 2×2 contingency tables such as sensitivity, specificity, and the predictive values. It is perhaps even more difficult to believe that it has been approximately 20 years since Hart, Webster, and Menzies (1993) questioned this approach in light of the conceptual shift from the dichotomous construct of dangerousness to the continuous construct of risk, since Mossman (1994a) published his influential article introducing ROC curve analysis and the AUC to the literature, and since Rice and Harris (1995) recommended that these become the new standard for the measurement of predictive validity in the field of violence risk assessment. Although a number of performance indicators are available to researchers, the use of ROC curve analysis and the AUC has become ubiquitous in studies attempting to establish predictive validity. Expert opinion is divided as to whether this has been a positive development (Douglas, Otto, Desmarais, & Borum, 2012; Kroner, 2007; Sjöstedt & Grann, 2002; Szmukler, 2012; Urbaniok, Rinne, Held, Rossegger, & Endrass, 2008). Regardless, as the AUC captures the discrimination but not the calibration component of predictive validity, reporting only this performance indicator does not supply sufficient evidence of predictive utility. Future research into the predictive validity of violence risk assessment tools should include a number of performance indicators that measure different facets of predictive validity, and the limitations of reported indicators should be routinely explicated. Providing more comprehensive statistical descriptions of tool performance has the potential to help give researchers, clinicians, and policymakers a clearer picture of whether structured assessment instruments may be useful in practice.

ACKNOWLEDGMENTS

The author sincerely thanks Drs. Joshua Wallace, Helen Doll, and Beom Lee for their statistical guidance, and Drs. Eva Kimonis and Patrick Kennealy for their valued proofreading.

REFERENCES

Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1. Sensitivity and specificity. *British Medical Journal*, *308*, 1552. DOI:10.1136/bmj.308.6943.1552.

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *British Medical Journal*, *309*, 102. DOI:10.1136/bmj.309.6947.102.

Altman, D. G., & Bland, J. M. (1994c). Diagnostic tests 3: Receiver operating characteristic plots. *British Medical Journal*, *309*, 188. DOI:10.1136/bmj.309.6948.188.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, & Law*, *16*, 39–55. DOI:10.1037/a0018362.

Andrews, D. A., & Bonta, J. (1995). LSI-R: The Level of Service Inventory – Revised. Toronto, ON: Multi-Health Systems.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396. DOI:10.1037/1082-989X.10.4.389.
Bennett, B. M. (1972). On comparisons of sensitivity, specificity and predictive value of a number of diagnostic procedures. *Biometrics*, *28*, 793–800.
Böhning, D., Holling, H., & Patilea, V. (2011). A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical Methods in Medical Research*, *20*, 541–550. DOI:10.1177/0962280210374532.
Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management*, *29*, 79–97. DOI:10.1177/014920630302900106.
Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine*, *16*, 981–991. DOI:10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N.
Breslow, N., & Day, N. (1987). *Statistical methods in cancer research*. Lyon: International Agency for Research on Cancer.
Chambless, L. E., & Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, *25*, 3474–3486. DOI:10.1002/sim.2299.
Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clinical Chemistry*, *54*, 17–23. DOI:10.1373/clinchem.2007.096529.
Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*, 928–935. DOI:10.1161/CIRCULATIONAHA.106.672402.
Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: A clinically useful measure of treatment effect. *British Medical Journal*, *310*, 452–454. DOI:10.1136/bmj.310.6977.452.
Das Gupta, S. (1960). Point biserial correlation coefficient and its generalization. *Psychometrika*, *25*, 393–408. DOI:10.1007/BF02289756.
Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal*, *316*, 989. DOI:10.1136/bmj.316.7136.989.
DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.
Douglas, K. S., Otto, R. K., Desmarais, S. L., & Borum, R. (2012). Clinical forensic psychology. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, volume 2: Research methods in psychology*. (pp. 213–244) Hoboken, NJ: John Wiley & Sons.
Farrington, D. P., & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour & Mental Health*, *10*, 100–122. DOI:10.1002/cbm.349.
Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal*, *345*, e4692. DOI:10.1136/bmj.e4692.
Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematics*, *2*, 805–813.
Fleminger, S. (1997). Number needed to detain. *British Journal of Psychiatry*, *171*, 287. DOI:10.1192/bjp.171.3.287a.
Furukawa, T. A. (1999). From effect size into number needed to treat. *Lancet*, *353*, 1680.
Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, *56*, 1129–1135. DOI:10.1016/S0895-4356(03)00177-X.
Gradstein, M. (1986). Maximal correlation between normal and dichotomous variables. *Journal of Educational Statistics*, *11*, 259–261. DOI:10.3102/10769986011004259.
Grimes, D. A. & Schulz, K. F. (2008). Making sense of odds and odds ratios. *Obstetrics & Gynecology*, *111*, 423–426. DOI:10.1097/01.AOG.0000297304.32187.5d.
Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet*, *359*, 248–252. DOI:10.1016/S0140-6736(02)07451-2.
Guggenmoos-Holzmann, I., & van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity. *Statistics in Medicine*, *19*, 1783–1792. DOI:10.1002/1097-0258(20000715)19:13<1783::AID-SIM497>3.0.CO;2-B.
Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*, 822–830. DOI:10.1093/bioinformatics/btq037.
Hanley, J. A., & Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of areas under receiver operating characteristic curves: An update. *Academic Radiology*, *4*, 49–58. DOI:10.1016/S1076-6332(97)80161-4.
Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.
Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders* (User Report 99–02). Ottawa, ON: Department of the Solicitor General of Canada.
Hart, S. D., Webster, C. D., & Menzies, R. J. (1993). A note on portraying the accuracy of violence predictions. *Law & Human Behavior*, *17*, 695–700. DOI:10.1007/BF01044690.
Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*, 92–105. DOI:10.1111/j.0006-341X.2005.030814.x.

Helmus, L., & Hanson, R. K. (2007). Predictive validity of the Static-99 and Static-2002 for sex offenders on community supervision. *Sexual Offender Treatment*, *2*, 1–14.
Higgins, J., Deeks, J., & Altman, D. G. (2008). Special topics in statistics. In Higgins J. & Green S. (Eds.), *Cochrane handbook for systematic reviews of interventions 5.0.0*. London: Wiley.
Hilton, D. J., Reid, C. M., & Paratz, J. (2006). An under-used yet easily understood statistic: The number needed to treat (NNT). *Physiotherapy*, *92*, 240–246.
Kraemer, H. C., & Gibbons, R. D. (2009). Where do we go wrong in assessing risk factors, diagnostic and prognostic tests? The problems of two-by-two association. *Psychiatric Annals*, *39*, 711–718. DOI:10.3928/00485713-20090625-05.
Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*, 990–996. DOI:10.1016/j.biopsych.2005.09.014.
Kroner, D. G. (2007). Issues in violent risk assessment: Lessons learned and future directions. *Journal of Interpersonal Violence*, *20*, 231–235. DOI:10.1177/0886260504267743.
Leisenring, W., Alonzo, T. A., & Pepe, M. S. (2000). Comparison of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*, *56*, 345–351. DOI:10.1111/j.0006-341X.2000.00345.x.
Li, J., & Fine, J. P. (2011). Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics*, *4*, 1–13. DOI:10.1093/biostatistics/kxr008.
Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather & Forecasting*, *19*, 1106–1114. DOI:10.1175/825.1.
McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.
McQuay, H. J., & Moore, A. (1997). Using numerical results from systematic reviews in clinical practice. *Annals of Internal Medicine*, *126*, 712–720.
Metz, C. E., Herman, B. A., & Shen, J. H. (1998). Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Statistics in Medicine*, *17*, 1033–1053. DOI:10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z.
Metz, C. E., Herman, B. A., & Roe, C. A. (1998). Statistical comparison of two ROC curve estimates obtained from partially-paired datasets. *Medical Decision Making*, *18*, 110–121. DOI:10.1177/0272989X9801800118.
Monahan, D. (1981). *The clinical prediction of violent behavior*. Rockville, MD: US Department of Health and Human Services.
Moons, K. G., & Harrell, F. E. (2003). Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology*, *10*, 670–672.
Moskowitz, C. S., & Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, *5*, 113–127.
Mossman, D. (1994a). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting & Clinical Psychology*, *62*, 783–792. DOI:10.1037/0022-006X.62.4.783.
Mossman, D. (1994b). Further comments on portraying the accuracy of violence predictions. *Law & Human Behavior*, *18*, 587–593. DOI:10.1007/BF01499177.
Mossman, D. (2013). Evaluating risk assessments using receiver operating characteristic analysis: Rationale, advantages, insights, and limitations. *Behavioral Sciences & the Law*, *31*, 23–39.
Neller, D. J., & Frederick, R. I. (2013). Classification accuracy of actuarial risk assessment instruments. *Behavioral Sciences & the Law*, *31*, 141–153.
Nunnally, J. C. (1978). Psychometric theory (2nd ed). New York: McGraw-Hill.
Obuchowski, N. A. (2005). Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Academic Radiology*, *12*, 1198–1204. DOI:10.1016/j.acrs.2005.05.013.
Pampel, F. C. (2000). Logistic regression: A primer. Thousand Oaks, CA: Sage.
Pan, X., & Metz, C. E. (1997). The "proper" binormal model: Parametric ROC curve estimation with degenerate data. *Academic Radiology*, *4*, 380–389. DOI:10.1016/S1076-6332(97)80121-3.
Pencina, M. J., D'Agostino, R. B., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, *27*, 157–172.
Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., & Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, *167*, 362–368. DOI:10.1093/aje/kwm305.
Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, *159*, 882–890. DOI:10.1093/aje/kwh101.
Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *7*, 670–675. DOI:10.1093/aje/kwj063.

Pinson, L., & Gray, G. (2003). Number needed to treat: An underused measure of treatment effect. *Psychiatric Services*, *54*, 145.
Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed). Washington, DC: American Psychological Association.
Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178–183. DOI:10.1037/1082-989X.1.2.178.
Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting & Clinical Psychology*, *63*, 737–748. DOI:10.1037/0022-006X.63.5.737.
Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC, Cohen's *d* and *r*. *Law & Human Behavior*, *29*, 615–620. DOI:10.1007/s10979-005-6832-7.
Robbins, A. S., Chao, S. Y. & Fonseca, V. P. (2002). What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes. *Annals of Epidemiology*, *12*, 452–454. DOI:10.1016/S1047-2797(01)00278-2.
Royston, P., Moons, K. G., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ*, *338*, 1373–1377. DOI:10.1136/bmj.b604.
Rugge, T. (2006). *Risk assessment of male aboriginal offenders: A 2006 perspective*. (User Report 2006-01). Ottawa: Public Safety Canada.
SAS Institute Inc. (2012). *SAS: Version 12.0*. Cary, NC: Author.
Shiu, S. Y., & Gatsonis, C. (2008). The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. *Philosophical Transactions of the Royal Society*, *366*, 2313–2333. DOI:10.1098/rsta.2008.0043.
Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: a second-order systematic review. *Behavioral Sciences & the Law*, *31*, 55–73.
Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, 499–513. DOI:10.1016/j.cpr.2010.11.
Singh, J. P., Grann, M., Lichtenstein, P., Långström, N., & Fazel, S. (2012). A novel approach to determining violence risk in schizophrenia: Developing a stepped strategy in 13,806 discharged patients. *PLoS ONE*, *7*, e31727. DOI:10.1371/journal.pone.0031727.
Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health*, *1*, 179–183. DOI:10.1080/14999013.2002.10471172.
Smith, C. A. B. (1951). A test for heterogeneity of proportions. *Annals of Eugenics*, *16*, 16–25. DOI:10.1111/j.1469-1809.1951.tb02455.x.
SPSS Inc. (2012). *SPSS for Windows: Release 21.0*. Chicago: Author.
StataCorp. (2011). *Stata statistical software: Release 12*. College Station, TX: Author.
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.
Stephan, C., Wesseling, S., Schink, T., & Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clinical Chemistry*, *49*, 433–439. DOI:10.1373/49.3.433.
Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*, 128–138. DOI:10.1097/EDE.0b013e3181c30fb2.
Szmukler, G. (2012). Risk assessment for suicide and violence is of extremely limited value in general psychiatric practice. *Australian & New Zealand Journal of Psychiatry*, *46*, 173–174. DOI:10.1177/0004867411432214.
Urbaniok, F., Rinne, T., Held, L., Rossegger, A., & Endrass, J. (2008). *Forensische Risikokalkulationen: Beurteilung der Anwendbarkeit und Validitat verschiedener Verfahren*. [Forensic risk assessment: Assessing the applicability and validity of different methods]. *Fortschritte der Neurologie-Psychiatrie*, *76*, 470–477. DOI:10.1055/s-2008-1038228.
Vickers, A. J., Cronin, A. M., & Begg, C. B. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology*, *11*, 1–7. DOI:10.1186/1471-2288-11-13.
Vrieze, S. I., & Grove, W. M. (2008). Predicting sex offender recidivism. I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law & Human Behavior*, *32*, 266–278. DOI:10.1007/s10979-007-9092-x.
Wang, W., Davis, C. S., & Soong, S. J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, *25*, 2215–2229. DOI:10.1002/sim.2332.
Ware, J. H. (2006). The limitations of risk factors as prognostic tools. *New England Journal of Medicine*, *355*, 2615–2617. DOI:10.1056/NEJMp068249.
Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR- 20: Assessing risk for violence. Version 2*. Burnaby, BC: Simon Fraser University, Mental Health, Law, and Policy Institute.
Youden, D. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35. DOI:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.