Psychological Assessment

Methodological Advances in Statistical Prediction

Howard N. Garb and James M. Wood

Online First Publication, March 11, 2019. http://dx.doi.org/10.1037/pas0000673

CITATION
Garb, H. N., & Wood, J. M. (2019, March 11). Methodological Advances in Statistical Prediction. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000673

American Psychological Association

Psychological Assessment

© 2019 American Psychological Association
1040-3590/19/$12.00

http://dx.doi.org/10.1037/pas0000673

Methodological Advances in Statistical Prediction

Howard N. Garb
Joint Base San Antonio–Lackland, San Antonio, Texas

James M. Wood
University of Texas at El Paso

Thirty years ago, Dawes, Faust, and Meehl (1989) argued that mental health professionals should routinely use statistical prediction rules to describe and diagnose clients, predict behaviors, and formulate treatment plans. Subsequent research has supported their claim that statistical prediction performs well when compared to clinical judgment. However, many of the things we thought we knew about statistical prediction have changed. The purpose of this literature review is to describe methodological advances in statistical prediction. Three broad areas are covered. First, while statistical prediction rules are valuable for criterion-referenced assessment (e.g., predicting violence, recidivism, treatment outcomes), they are valuable only for some norm-referenced assessment tasks (e.g., diagnosis but not describing personality and psychopathology). Second, statistical prediction is particularly prominent for the prediction of violence and criminal recidivism. Results from this area will be used to describe the validity of traditional clinical judgment, structured professional judgment, and statistical prediction. The results support the use of both structured professional judgment and statistical prediction. The effect of allowing professionals to override statistical predictions consistently led to lower validity. Third, issues in building statistical prediction rules are described, including the assignment of weights to predictors, the emergence of new statistical analyses (e.g., machine learning), and the role of theory. As research has progressed, statistical prediction has become one of the most exciting areas of psychological assessment.

Public Significance Statement
Advances in statistical prediction will allow us to better predict a range of behaviors and events including violence, criminal recidivism, onset of psychosis, and psychotherapy failure.

Keywords: risk assessment, statistical prediction, actuarial prediction, machine learning, clinical judgment

The purpose of this literature review is to describe methodological advances in statistical prediction. Statistical predictions can be made using logistic regression, Cox regression, hierarchical linear modeling, machine learning algorithms, and unit weight linear rules, as well as other types of statistical analyses.

In a landmark article in the journal *Science*, Dawes, Faust, and Meehl (1989) compared the validity of clinical judgment (judgments made by professionals) with the validity of statistical prediction. With a sample of about 100 studies drawn from the social sciences, they found that the validity of statistical prediction equaled or exceeded the validity of clinical judgment in almost every case that had been examined. Tasks ranged from predicting college grades to predicting responses to electroconvulsive therapy. Similar results were obtained in later meta-analyses. Grove, Zald, Lebow, Snitz, and Nelson (2000) reviewed studies in psychology and medicine, while Ægisdóttir et al. (2006) reviewed studies in which a psychological or mental health judgment was made. Grove et al. (2000) found that on average mechanical prediction (including statistical prediction) was about 10% more accurate than clinical judgment, and Ægisdóttir et al. (2006) found a 13% advantage in accuracy for statistical prediction.

Dawes et al. (1989) asserted that statistical prediction rules can be used for virtually any type of task including diagnosis, description, prediction, and treatment planning. Agreeing with this claim, Grove et al. (2000) found that, “Superiority for mechanical-prediction techniques was consistent, regardless of the judgment task, types of judges, judges’ amount of experience, or other types of data being combined” (p. 19). Similarly, Ægisdóttir et al. (2006) found that across different judgment tasks, statistical prediction was consistently as accurate as, or more accurate than, clinical judgment.

Another important issue is whether clinicians should override statistical predictions. That is, should they combine statistical predictions with other information or should they use statistical prediction by itself? Dawes et al. (1989) argued that statistical rules should be used without making any adjustments, even when clinical judges have

Howard N. Garb, Reid Medical Clinic, Joint Base San Antonio–Lackland, San Antonio, Texas; James M. Wood, Department of Psychology, University of Texas at El Paso.

The views expressed are those of the authors and do not reflect the official views or policy of the Department of Defense or its components.

Correspondence concerning this article should be addressed to Howard N. Garb, Reid Medical Clinic, Joint Base San Antonio–Lackland, 1940 Carswell Avenue, San Antonio, TX 78236-1500. E-mail: howard.n.garb.civ@mail.mil

extra information. They also argued that gaining access to additional information did nothing to close the gap between statistical prediction and clinical judgment. Grove et al. (2000) found that "whether the judges had more data or equal amounts of data relative to the mechanical formula made little difference in the relative superiority of mechanical prediction" (p. 24). Ægisdóttir et al. (2006) found that (a) increasing the amount of information available to clinicians decreased the validity of their judgments and (b) making statistical predictions available to clinicians did not improve their accuracy.

Dawes et al. (1989) expressed the hope that statistical prediction rules would become as common, and as valuable, as psychological tests. They even argued that it is "irrational" for professionals who identify themselves as scientific to not use statistical prediction rules (p. 1673). They also recommended that mental health professionals use statistical prediction rules even when criterion scores are unavailable.

Good criterion information is oftentimes unavailable, making it difficult to derive and validate a statistical prediction rule (Garb, 1998, 2005). For example, to describe personality and psychopathology, one could evaluate a statistical prediction rule by using a psychological test as a criterion. Unfortunately, this would involve evaluating a statistical prediction rule with assessment information. One would not be able to learn if a statistical prediction rule is more or less valid than the psychological test. Hence, being able to construct statistical prediction rules when good criterion data are unavailable could be an important advance for psychological assessment.

Dawes et al. (1989) explained how to do this (also see Dawes & Corrigan, 1974; Dawes, 1979). Instead of using criterion data to derive differential weights, weights can be obtained by some other method, for example, by assigning equal weights to each predictor variable. Though the weights for equal weight linear rules are not empirically derived using data, they were thought by Dawes et al. (1989) to be as good as regression rules, partly because differential weight prediction rules may not generalize well across diverse settings and samples. Since equal weight linear rules have done well in studies when criterion information was available, one can argue that they can be expected to do well when it is not available. Even though the predictions themselves would never be evaluated, Dawes et al. argued for their use because the *methodological approach* is supported by empirical evidence. This provocative argument was still being made 20 years later by Vrieze and Grove (2009): "Equal weights eliminates the need for expensive criterion-variable data sets and extensive research to design statistical algorithms for mechanical prediction" (p. 526).

The claims made by Dawes et al. (1989) will be examined as we describe methodological advances in statistical prediction. To describe methodological advances, three broad areas will be addressed: (a) the value of statistical prediction for criterion-referenced and norm-referenced assessment; (b) the validity of clinical judgment, structured professional judgment, and statistical prediction in the context of predicting violence and criminal recidivism; and (c) strategies for building statistical prediction rules.

Criterion-Referenced and Norm-Referenced Assessment

Criterion-referenced assessment is designed to predict events and outcomes, and *norm-referenced assessment* is designed to measure constructs such as personality traits and diagnoses. This terminology can be adapted to describe statistical prediction rules (Helmus & Babchishin, 2017). Although Dawes et al. (1989) argued that statistical prediction should be used for all judgment tasks, the distinction between criterion-referenced prediction rules and norm-referenced prediction rules will allow us to think about when statistical prediction rules are likely to be of value.

Statistical prediction is often successful when used to predict specific behaviors and outcomes. For example, criterion-referenced prediction rules currently play a central role in forensic settings for the prediction of violence and criminal recidivism (Monahan & Skeem, 2016). In mental health settings, in addition to predicting violence, they have been useful or at least showed promise for predicting (a) onset of psychosis (Cannon et al., 2008), (b) course of mental disorders (Kessler et al., 2016), (c) psychotherapy failure (Boswell, Kraus, Miller, & Lambert, 2015), and (d) suicide attempts and suicides (Kessler et al., 2015; Walsh, Ribeiro, & Franklin, 2017). In military settings, statistical prediction has been used to identify trainees who are likely to have poor mental health or behavioral outcomes, so that these trainees can be interviewed and recommendations and referrals can be made (Garb, Wood, & Baker, 2018). For all of these tasks, one can expect statistical prediction to be more successful than clinical judgment.

Another example of criterion-referenced statistical prediction is treatment selection (Cohen & DeRubeis, 2018). The Personalized Advantage Index (DeRubeis et al., 2014) can be used for treatment selection when (a) more than one treatment is under consideration, (b) comparative outcome data are available, and (c) pretreatment variables are related to outcomes across treatment interventions. The Veterans Health Administration has also proposed using statistical prediction to select treatments. They are encouraging research that would target patients with the highest probabilities of responding to interventions (Kessler et al., 2017, p. 6).

Norm-referenced statistical prediction rules can improve clinical practice by being able to predict semistructured and structured interview diagnoses. Semistructured and structured interview diagnoses are more reliable and valid than unstructured clinical diagnoses (e.g., Andreas, Theisen, Mestel, Koch, & Schulz, 2009; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009; Widiger & Lowe, 2010; Zimmerman, 1994). However, because they are time-consuming, semistructured and structured interviews are infrequently used in clinical practice. If statistical prediction rules could accurately predict semistructured or structured interview diagnoses and if they were relatively easy to use, then this would help to improve clinical practice.

In a study on predicting semistructured interview diagnoses of bipolar disorder in a pediatric sample (Youngstrom, Halverson, Youngstrom, Lindhiem, & Findling, 2018), different statistical prediction rules were compared. Included were Bayesian algorithms varying in complexity, logistic regression models varying in complexity, and supervised least absolute shrinkage and selection operation regression model. The least absolute shrinkage and selection operation model was the most complex as it can handle a large number of predictors (even having more predictors than cases), interactions, and nonlinear transformations. All of the statistical prediction rules were more valid than traditional clinical judgment. Complex models degraded rapidly when they were derived in one type of clinic (e.g., an academic clinic) and cross-validated in another (e.g., a community clinic). The naïve Bayesian approach and other relatively simple models performed well on cross-validation. They were thought to be the most promising rules for upgrading clinical practice. The use of probability nomograms, a clinician-friendly procedure for using the Bayesian approach, has recently been described by Youngstrom and Van Meter (in press).

The case for using norm-referenced statistical prediction rules to describe personality and mental status is less compelling. Dawes et al. (1989) recommended that equal weight linear rules (or similar rules) be used when criterion scores are unavailable. To do so would involve having a clinician identify different predictors. Equal weights would then be assigned to these predictors. Item scores would be summed to make a prediction or rating. Of course, this is already done for many scales. For example, on the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), a value of 0 to 3 is assigned for each item and these values are added to calculate the total score.

However, when the aim is to assess dimensions of personality or psychopathology, it would make little sense to turn to equal weight linear rules rather than to the psychometric approach that is commonly used for test construction. A prototypical example of the psychometric approach is described in a classic article by Clark and Watson (1995), which recommended that scale construction should involve such strategies as (a) creating an overinclusive and theory-based pool of items, (b) administering these items along with variables that measure closely related constructs to a heterogeneous sample of individuals representing the entire range of the target population, and (c) selecting items that are unidimensional, usually on the basis of factor analysis. Psychometric strategies such as these have been widely adopted and proven highly effective for generating valid and useful measures. With its widespread acceptance and many successful applications, the psychometric approach to test construction, rather than statistical prediction, is clearly the better option for assessing personality and facets of psychopathology (e.g., symptoms, mental status).

In conclusion, criterion-referenced statistical prediction rules are now being used successfully for a wide range of tasks, and norm-referenced statistical prediction rules have been shown to be important for diagnosis. However, advances in the description of personality and psychopathology are more likely to come from psychometrics than statistical prediction.

Predicting Violence and Criminal Recidivism

Risk assessment is a broader term than risk prediction. It refers to risk management, risk decision making, and risk communication, in addition to risk prediction (Heilbrun, 2003). In forensic contexts, risk management often involves finding ways to reduce or manage an individual's risk for violence (Heilbrun, 2009). In other contexts, risk management may refer to reducing an individual's risk for self-harm or for becoming psychotic. Many forensic psychologists have argued that a risk assessment scale should allow us to not only predict an outcome, but also provide information that is helpful for risk management (Bonta, 1996). The focus of this review is on prediction, but we recognize the importance of other issues.

Studies have evaluated the predictive validity of more than 120 assessment instruments for evaluating the risk of harming others (e.g., Singh, Serper, Reinharth, & Fazel, 2011). Being able to predict violence in mental health settings and in the community is of obvious importance, but much of the research has been conducted in forensic settings. Along with statistical prediction rules, structured professional judgment tools that provide guidelines for risk assessment have also been developed. In forensic settings, these instruments are considered to be a requisite component of the assessment process (Williams, Wormith, Bonta, & Sitarenios, 2017). Results will be described for (a) structured professional judgment, (b) statistical prediction rules, (c) comparisons to unstructured professional judgment (traditional clinical judgment), and (d) overriding statistical prediction.

To describe the results, area under the curve (AUC), a statistic taken from signal detection theory (Macmillan & Creelman, 1991), will be used. Values for the AUC can range from .00 (perfect inaccurate discrimination) to .50 (chance level of discrimination) to 1.00 (perfect discrimination). For example, for the prediction of violence, an AUC value of .67 indicates that there is a 67% likelihood that a randomly selected individual who later behaves violently obtained a higher score than a randomly selected individual who did not later behave violently. The major advantage of using the AUC to report results is that it is relatively independent of base rates across studies. Other performance statistics are also important, but for clarity of presentation, AUC values will be the preferred statistic in this article.

Structured Professional Judgment

Structured professional judgment tools provide guidelines for clinicians and other professionals who are engaged in risk assessment and management. Some guidelines have been created solely on the basis of clinical practice, while others are based on a combination of research, theory, and clinical practice. Structured professional judgment tools can be used to assess individuals in correctional and forensic mental health settings as well as individuals in other mental health settings and in the community. For example, structured professional judgment has been used to help clinicians predict aggressive behavior among men with schizophrenia living in the community (Michel et al., 2013). However, structured measures are much more commonly used in forensic settings to predict future violent criminal offenses.

Professionals using structured professional judgment tools do not typically make precise predictions (e.g., probability ratings). Instead, they usually make item ratings and final ratings. Final ratings are made to indicate low, moderate, or high risk for violence. To make these risk ratings, professionals are supposed to consider the items making up the structured professional judgment tool, but they are free to also use available information that may not be reflected in their item ratings.

Total scores based on item ratings are not typically calculated by professionals, but they have been calculated for studies by research investigators. Most studies on structured judgment have examined the relation between these total scores and outcomes, rather than the relation between professionals' final ratings and outcomes. Nevertheless, a substantial number of studies on structured judgment have examined both total scores and professionals' final ratings.

A recent meta-analysis by Chevalier (2017) identified 22 structured professional judgment measures designed to help professionals evaluate risk for violence. These measures include the Brief Spousal Assault Form for the Evaluation of Risk (Kropp, Hart, &

Belfrage, 2005); the Early Assessment Risk List for boys (Augimeri, Koegl, Webster, & Levene, 2001) and for girls (Levene et al., 2001); the Estimate of Risk of Adolescent Sexual Offense Recidivism (Worling & Curwen, 2001); the Historic, Clinical, Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997); HCR-20–Version 3 (Douglas et al., 2014); the Spousal Assault Risk Assessment Guide (Kropp & Hart, 2000); the Structured Assessment for Violence Risk in Youth (Borum, Bartel, & Forth, 2003); the Short-Term Assessment of Risk and Treatability (Webster, Martin, Brink, Nicholls, & Middleton, 2004) and its adolescent version (Nicholls, Viljoen, Cruise, Desmarais, & Webster, 2010); and the Sexual Violence Risk-20 (Boer, Hart, Kropp, & Webster, 1997). Studies were included in the meta-analysis when (a) one or more of these structured professional judgment measures was administered and (b) results for predictive validity were reported.

The Chevalier (2017) meta-analysis included data from 69 samples that were collected in 60 studies with 10,871 participants (Chevalier, 2017). Mean weighted AUC values were .70 for research investigators' total scores and .70 for professionals' final ratings. The results did not vary by the specific measure used or by type of outcome (e.g., aggressive behavior, sexual recidivism). Higher AUC values were obtained for research studies (.71) than field studies (.65).

Chevalier (2017) also found that the addition of clinicians' final ratings to total scores in logistic regression equations consistently led to a statistically significant increment in validity. Thus, even after controlling for total scores, the clinicians' final ratings accounted for additional variability in the outcomes. These results suggest that validity may be highest when clinical judgments, along with other information, are entered into a statistical prediction rule.

Statistical Prediction Rules

Monahan and Skeem (2016) recently concluded that, among the many well-validated statistical and structured professional judgment instruments designed to predict criminal recidivism, "There is no compelling evidence that one validated tool forecasts recidivism better than another" (p. 500). Supporting this conclusion, Yang, Wong, and Coid (2010) evaluated seven commonly used risk assessment tools along with two instruments that were designed to measure psychopathy. They found that they all predicted violence at about the same moderate level of validity. The seven risk assessment tools were the: General Statistical Information for Recidivism (GSIR; Bonta, Harman, Hann, & Cormier, 1996), HCR-20 (Webster et al., 1997), Level of Service Inventory (LSI) and its revised version (LSI-R; Andrews & Bonta, 1995), Offender Group Reconviction Scale (Copas & Marshall, 1998), Risk Matrix 2000 for Violence (Thornton, 2007), Violence Risk Assessment Guide (VRAG; Harris, Rice, & Quinsey, 1993), and the Violence Risk Scale (Wong & Gordon, 2006). The two measures of psychopathy were the Psychopathy Checklist–Revised (PCL-R; Hare, 2003) and its screening version (Salekin, Rogers, & Sewell, 1996). Data were collected in 28 studies. Sample size ranged from 6,348 to 7,221 for different risk assessment tools and from 34 to 1,650 by study. A within-subject design was used, meaning that studies were included in the meta-analysis only if more than one risk assessment tool was administered to the same sample of subjects and the same outcome variable was used for all subjects. For most of the prediction and measurement instruments and their accompanying scales, AUC values ranged from .65 to .71. Yang et al. (2010) concluded that

> If the intention is only to predict future violence, then the 9 tools are essentially interchangeable; the selection of which tool to use in practice should depend on what other functions the tool can perform rather than on its efficacy in predicting violence. (p. 740)

In a more recent meta-analysis, Williams et al. (2017) reanalyzed data from Singh, Grann, and Fazel (2011). Results for the following nine risk assessment tools were included in their meta-analysis: Structured Assessment of Violence Risk in Youth (Borum et al., 2003), HCR-20 (Webster et al., 1997); LSI-R (Andrews & Bonta, 1995); VRAG (Quinsey, Harris, Rice, & Cormier, 2006); Spousal Assault Risk Assessment Guide (Kropp & Hart, 2000); PCL-R (Hare, 2003); Sex Offender Risk Appraisal Guide (Quinsey et al., 2006); *STATIC-99 Coding Rules, Revised, 2003* (Harris, Phenix, Thornton, & Hanson, 2003); and Sexual Violence Risk–20 (Boer et al., 1997). Data were from 88 independent samples in 68 studies with 25,980 participants. Studies were included in the meta-analysis when (a) one or more of the risk assessment tools was administered to a sample of subjects and (b) results for predictive validity were described. Williams et al. (2017) calculated confidence intervals for the estimates of predictive validity for each risk assessment tool. The mean size of overlap across the confidence intervals for the risk assessment tools was about 50%. Williams et al. (2017) concluded that offender risk measures are more alike than different in their validity for predicting violence.

One reason why statistical prediction rules and measures of psychopathy may achieve similar levels of validity is because they are measuring similar factors. In an innovative study by Kroner, Mills, and Reddon (2005), "coffee can" statistical prediction rules were constructed and compared to the PCL-R (Hare, 1991), the LSI-R (Andrews & Bonta, 1995), the VRAG (Harris et al., 1993), and the GSIR (Nuffield, 1982). To construct the coffee can prediction rules, every item in the PCL-R, LSI-R, VRAG, and GSIR was individually written on a separate card. All of the cards were then placed in a large empty coffee can. Four "coffee can prediction rules" were created by randomly selecting cards from the can, with 13 cards selected for each rule. Another item, "number of prior incarcerations," was added to all four of the "coffee can prediction rules." This was done because the PCL-R, LSI-R, VRAG, and GSIR all contain at least one item related to prior offending or prior criminal behavior. Kroner et al. (2005) found that the PCL-R, LSI-R, VRAG, and GSIR did not predict criminal convictions and revocations of parole any better than the coffee can measures. Predictions were made for 206 offenders after their release from prison. A factor analysis of the coffee can items yielded four factors: (a) criminal history, (b) persistent criminal lifestyle, (c) antisocial personality, and (d) alcohol/mental health issues. Because all of the statistical prediction rules that were examined by Kroner et al. (2005) contained items that measure the same common factors, they achieved comparable levels of validity.

Related results were found for measures of *static and dynamic risk factors* for violent criminal recidivism. Static risk factors will not change and cannot be treated. Examples are gender of victims and a client's number of prior offenses. Dynamic risk factors can be addressed by treatment. Examples are sexual preoccupation, deviant sexual interests, and impulsive tendencies. An emphasis on measuring dynamic risk factors has been widely praised as leading to a new generation of statistical prediction rules (Bonta, 1996). Yet, in a meta-analysis (Van den Berg et al., 2018), the addition of dynamic risk measures to static risk measures resulted in only a small increase in validity for predicting recidivism for sexual offenses. For example, for the prediction of sexual recidivism, a Cox hazard ratio of 1.08 was obtained (19 studies, 13 unique samples, $N = 3{,}747$). Dynamic risk measures may be valuable even if they are able to prevent only a small number of sexual offenses, but, as in the Kroner et al. (2005) study, it has been surprisingly difficult to improve predictions. Although a dynamic risk factor (e.g., antisocial cognition) can help clarify the effects of a static risk factor (e.g., criminal history) by way of mediation (the effect of criminal history on recidivism may be mediated by antisocial cognition), it may add only a small increment in predictive validity (Walters, 2017).

As noted in the preceding paragraphs, different statistical prediction rules may reach a limit beyond which it is difficult to improve. This can occur when a criterion is imperfect. Seto (2005) described the "noise" in the criterion scores when predicting recidivism:

> . . . some recidivists are not detected by police, there is jurisdictional variation in the likelihood of laying charges or obtaining convictions, and there is jurisdictional variation in plea bargaining . . . it is very unlikely that prediction of recidivism will ever obtain the very high AUCs (e.g., .95 to 1.00) that are possible for those biomedical tests that can be quickly verified by an accepted standard, such as tissue biopsy. (p. 165)

There are many additional reasons why it can be difficult to improve validity. Meehl (1978) famously wrote about the slow progress of psychology. His main recommendation, that there be less reliance on statistical significance testing of the null hypothesis, has largely been accepted. However, he also discussed reasons why it is difficult to improve theory and prediction in psychology. For example, the sheer number of historical causal influences that affect behavior can be long and difficult to detect. Meehl (1978) observed that

> Every thoughtful clinician realizes that the standard life history that one finds in a medical chart is, from the standpoint of thorough causal comprehension, so thin and spotty and selective as to border on the ludicrous. But there is also what I would view as an important causal source of movement in one rather than another direction of divergent causality, namely, inner events, such as fantasies, resolutions, shifts in cognitive structure, that the patient may or may not report and that he or she may later be unable to recall. (p. 810)

Thus, a difficulty in improving upon predictions of violence and recidivism may be due, in part, to individuals being unable to recall and report their changing cognitive processes.

Comparisons to Unstructured Professional Judgment

Results on the comparison of statistical prediction, structured professional judgment, and unstructured professional judgment will be described. The results support the use of both structured professional judgment and statistical prediction.

A highly cited meta-analysis by Mossman (1994) found that clinicians' unstructured predictions of violence have a medium level of validity. For the short-term prediction of violence, with results from six studies, the average AUC was .69. Long-term predictions of violence, based on seven studies, were about as accurate, with an AUC value of .64. For all unstructured clinical judgments (including medium-term predictions), based on the results from 17 studies, the average AUC was .67. Mossman (1994) also reported results for statistical predictions. Based on the results from 14 studies, the average AUC for statistical prediction rules was .71.

In a related study (Vogel, Ruiter, Hildebrand, Bos, & van de Ven, 2004), predictions of violent recidivism and general recidivism were made for 120 forensic psychiatric patients. The base rate for violent recidivism was 36%, and the base rate for general recidivism was 52%. For the prediction of violent recidivism, AUC values were .68 for unstructured clinical judgment, .79 for structured professional judgment, and .82 for a statistical prediction rule (total score on the HCR-20; Webster et al., 1997). For the prediction of general recidivism, AUC values were .63 for unstructured judgments, .66 for structured professional judgment, and .70 for the statistical prediction rule (total score on the HCR-20).

Predictions for sex offenders were evaluated in a meta-analysis of studies on sex offenders that used data from 118 studies (Hanson & Morton-Bourgon, 2009). Predictions were made for (a) sexual offense recidivism, (b) violent (including sexual) offense recidivism, and (c) any recidivism. Effect sizes (d) were reported rather than values for AUC. For the prediction of sex offense recidivism, unstructured clinical judgment ($d = 0.42$) and structured professional judgment ($d = 0.46$) were about equally accurate while statistical prediction was substantially more accurate ($d = .67$). For the prediction of violent offense (including sexual offense) recidivism, unstructured clinical judgment ($d = 0.22$) was less accurate than structured professional judgment ($d = 0.31$) and statistical prediction ($d = 0.51$). For predicting recidivism for any reason, unstructured clinical judgment ($d = 0.11$) had lower validity than structured professional judgment ($d = 0.26$) and statistical prediction ($d = 0.52$). The authors observed that if the unstructured clinical judgments, structured professional judgments, and statistical predictions had been made for the same samples, this would have allowed for more precise comparisons.

Overriding Statistical Predictions

There are several good reasons why a professional might feel justified in overriding a statistical prediction. For example, a professional may want to override a statistical prediction because of (a) changing environmental factors since the statistical prediction rule was first derived and cross-validated, (b) protective factors that are present in a particular client, and (c) unique personal characteristics that are not captured by a statistical prediction rule (Childs, Frick, Ryals, Lingonblad, & Villio, 2014).

A number of studies on the prediction of sexual criminal recidivism have found that allowing professionals (clinicians, probation officers, classification officers) to adjust statistical predictions led to lower validity, though sometimes the decrement in validity was not statistically significant (Guay & Parent, 2018; Hanson, Helmus, & Harris, 2015; Schmidt, Sinclair, & Thomasdóttir, 2016; Storey, Watt, Jackson, & Hart, 2012; Wormith, Hogg, & Guzzo, 2012; also see Hanson & Morton-Bourgon, 2009, who cited three unpublished studies that also reported negative findings for statistical overriding). For example, although the manuals for the Youth Level of Service/Case Management Inventory (Hoge & Andrews, 2006) and Youth Level of Service/Case Management Inventory 2.0 (Hoge & Andrews, 2011) state that an override of the statistical predictions should occur only in rare circumstances, Schmidt et al. (2016) reported that the override feature was used to adjust statistical predictions for 74% of 204 sexual offenders and 41.6% of 185 nonsexual offenders. The use of the override feature led to worse predictive validity. A dramatic drop in validity was reported by Wormith et al. (2012) for the use of the Level of Service/Case Management Inventory (Andrews, Bonta, & Wormith, 2004). When risk level was adjusted by personnel, predictive validity fell from $r = 33$ to $r = .02$ for 669 sex offenders and from $r = .36$ to $r = .14$ for 3,694 nonsexual offenders. When probation officers were trained to override statistical predictions in only one case out of 20 when working with the Level of Service/Case Management Inventory (Andrews et al., 2004), the rate of overrides was only 6.5% (Guay & Parent, 2018). However, the overrides still decreased validity.

Building Statistical Prediction Rules

How should we go about building statistical prediction rules and risk assessment instruments? A study by Grann and Långström (2007) illustrates several important issues. To predict violent recidivism in a sample of 404 violent offenders diagnosed with either a personality disorder or schizophrenia, they used 10 items of the Historical subscale of the Historical, Clinical, Risk 20 (HCR-20; Webster et al., 1997). They found that an equal weight linear rule did as well as differential weight rules and better than a neural network procedure that attempted to use complicated relations among predictors to make predictions. They concluded that theory building should be emphasized more than the development of complex statistical prediction models. We will discuss (a) the performance of equal weight linear rules, (b) machine learning (which includes neural network procedures), and (c) the role of theory for selecting and weighing predictors.

Assignment of Weights to Predictors

One advantage of using equal weights is convenience. To get a total score or prediction, one can simply add the item scores. Interestingly, many risk assessment scales weigh items equally (e.g., Violence Risk Scale, Wong & Gordon, 2006; Level of Service/Case Management Inventory, Andrews et al., 2004; Static-99R, Hanson & Thornton, 1999). These measures have been evaluated using criterion scores, and they received at least moderate empirical support.

Another advantage for using equal weights is that they can be more robust than differential weights (Grann & Långström, 2007). If a derivation sample size is too small, then assigning differential weights to predictors may capitalize on random variance. Overfitting will lead to shrinkage in predictive validity in cross-validation samples. Thus, a differential weight linear rule may outperform an equal weight linear rule on a derivation sample, but not when the rules are cross-validated. In other words, the results may not generalize to other samples.

An example of the overfitting of a differential weight linear rule was given by Helmus and Thornton (2015). They analyzed data for 19 samples with Static-99R (Hanson & Thornton, 1999) item data ($N = 7{,}461$) and eight samples with Static-2002R (Hanson, Helmus, & Thornton, 2010) item data ($N = 2{,}951$). For roughly half of the Static-99R and Static-2002R items, the relation between items and criterion varied significantly across samples in ways that could not be explained by sampling error. When differential weights vary across samples, there are a number of options that investigators may want to consider including assigning equal weights.

Dawes et al. (1989) recommended that equal weights be used even when criterion scores are unavailable. This recommendation can be problematic, however, if a predictor is not monotonically related to a criterion. As noted by Wainer (1976), equal weights perform well when: "(a) All predictor variables are oriented properly (if you don't know what direction the criterion variable lies with respect to a predictor, that predictor shouldn't be used); and (b) the predictor variables are intercorrelated positively" (p. 213). If one has access to criterion scores, then one can check to make sure that a predictor variable is monotonically related to a criterion. One cannot do this without criterion scores.

The assumption that a predictor is monotonically related to an event or outcome can sometimes be wrong. For example, when conducting a study on mental health screening with United States Air Force trainees, the authors of this article and their colleagues thought that excessive alcohol use would be monotonically related to failure to complete term of service. Failure to complete term of service is an important issue for the United States Air Force because so many active duty enlisted personnel are discharged before they complete their first four years. In a sample of 89,032 active duty personnel, the relation between excessive drinking and discharge rate was not monotonic (Garb, 2013). The highest discharge rates were obtained for personnel who drank excessively *more* than once a week (five or more drinks in a sitting). Surprisingly, personnel who never drank excessively had higher discharge rates than personnel who drank excessively once a month or once a week.

Another problem with using equal weight linear rules when criterion data are unavailable is that clinicians will not know the validity of their own predictions and prediction rules. In the fields of clinical and forensic psychology, we did not find studies on the use of unit weight linear rules when criterion data are unavailable. Indeed, because one cannot evaluate their validity, it is unclear how researchers could present results for them.

Machine Learning

Machine learning is a branch of artificial intelligence that allows computers to learn by discovering patterns in empirical data. The computers are not explicitly programmed to find those patterns based on a priori knowledge. A strength of machine learning is that it can combine enormous numbers of predictors in nonlinear and highly interactive ways. Overfitting a model may be difficult to prevent because so many parameters are being fit to the data. However, machine learning is designed for the analysis of high-dimensional data with hundreds or thousands of predictors and relatively few cases, and statistical procedures have been developed to prevent overfitting. For example, in a study discussed below, Kessler et al. (2015) used the machine learning method *elastic net* that penalizes overfitting.

Many different statistical techniques are used in machine learning, for example, artificial neural networks, decision tree learning, and support vector machines. Machine learning is closely related to statistics, and the distinction between traditional statistical analyses and machine learning can be fuzzy (Duwe & Kim, 2017). For example, in the machine learning field, it is not uncommon to see an analysis like logistic regression referred to as a machine learning algorithm.

The "black box" metaphor refers to being unable to see what a machine learning model is doing. In general, machine learning programs are not designed to offer an explanation for their predictions. This can make it difficult to understand the basis for the predictions. Different machine learning algorithms differ in transparency, so the "black box" metaphor is not descriptive of all of machine learning. A lack of transparency in decision-making can raise ethical issues, especially in legal settings. Offenders may feel that they have a right to know what variables caused them to be seen as being at risk for violence or recidivism, and psychologists in court may have to defend their not knowing how information was combined. When prediction rules have similar levels of validity, there are important reasons why psychologists should favor the rules that have greater transparency.[1]

Machine learning will likely have a major impact in the field of medicine (Obermeyer & Emanuel, 2016), although its influence on mental health practice is more difficult to predict. Machine learning is already being used for judgment tasks in radiology and pathology (e.g., to provide a second reading of a mammogram). Radiologists and pathologists focus largely on interpreting medical images, so by digitizing a medical image, one is making available to a machine learning program most of the information that a radiologist and pathologist would have. Criterion data are often available to radiologists and pathologists (e.g., whether a tumor is benign or malign), though it is less often available in other areas of medicine.

In mental health and forensic settings, the value of machine learning may depend on the complexity of the data. When using information from a single risk assessment tool, results on the value of machine learning have been mixed (e.g., Berk & Bleich, 2013; Hamilton, Neuilly, Lee, & Barnoski, 2015; Liu, Yang, Ramsay, Li, & Coid, 2011; Tollenaar & van der Heijden, 2013). To learn whether newer machine learning analyses perform better than older methods, Duwe and Kim (2017) compared 12 statistical methods: (a) simple logistic regression, (b) logistic regression with nonlinear and interaction terms, (c) regularized logistic regression, (d) decision trees, (e) naïve Bayes, (f) artificial neural networks, (g) support vector machines, (h) bagged trees, (i) random forests, (j) LogitBoost, (k) MultiBoosting, and (l) logistic model trees. Offenders released from prison between 2003 and 2008 made up the training data set, which was used to develop the models. The test data set, made up of offenders released from prison in 2009 and 2010, was used to evaluate the prediction models. The total sample consisted of 27,772 offenders. Predictions were made for five different types of recidivism, with base rates ranging from 1% to 47%. Results were analyzed separately for male and female offenders, so it was also possible to compare the different statistical analyses across varying levels of sample size. Depending on the type of recidivism that was predicted, the number of prediction variables ranged from 10 to 55. The difference between the best and worst statistical techniques on the test data set was modest, with the newer machine learning algorithms generally performing better (e.g., overall average AUC for LogitBoosting = .78 vs. overall average AUC for decision trees = .73). The authors concluded that machine learning algorithms should be considered when developing a risk assessment instrument, but they acknowledged that the use of machine learning in criminal justice is in its infancy.

More traditional statistical prediction rules may be as valid as machine learning models when data are less complex (e.g., when using items from a single measure as predictors). When using more complex data (e.g., electronic health records), preliminary results on the value of machine learning have been impressive (Kessler et al., 2015; Walsh et al., 2017). For example, to predict suicides, Kessler et al. (2015) used data from 38 U.S. Army and Department of Defense administrative data systems, including sociodemographic data (e.g., recent job loss), criminal justice data (e.g., violent crime victimization or perpetration), measures of registered weapons, and pharmacy and medical data (e.g., quality of care, prior suicidal behaviors). Because the data are longitudinal, it is possible for machine learning to use information about the effects of time on complex data structures (e.g., complex feedback loops). Predictions were made for soldiers who were hospitalized for the treatment of a psychiatric disorder (N = 40,820). Within one year of hospitalization, 68 (0.17%) of the soldiers committed suicide. Bivariate associations were calculated between 421 of the predictors and suicides. Results were statistically significant for 131 of the predictors. The statistical prediction rule, which was created using a three step analysis that included features of machine learning, had an AUC value = .84. This is a high level of validity, but we will have to wait to learn how the rule does when cross-validated in new samples.

Role of Theory for Building Statistical Prediction Rules

The role of theory in statistical prediction is not the same as the role of theory in psychological measurement. A statistical prediction rule does not have to be a measure of a construct, and thus one does not need to conduct a factor analysis of the variables included in the statistical prediction rule. However, statistical prediction is stronger when it has a theoretical base, in particular, when something is known about the root causes of the behaviors or events that are being predicted (Silver, 2012). As an example from outside the area of psychology, statistical prediction rules about global warming are widely accepted not only because of their predictive validity, but *more importantly* because of their basis in scientific theory. Atmospheric concentrations of greenhouse gases (e.g., carbon dioxide) are increasing as a result of human activity, and these increases will enhance the greenhouse effect resulting in increased warming. This theory is well supported by scientific evidence.

[1] A useful alternative to machine learning is the "information-theoretic" approach (Burnham & Anderson, 2003). It argues for the active role of the researcher in helping to build and interpret parsimonious and logical prediction models.

For the field of psychological assessment, there is general agreement that theory can help professionals identify "candidate variables." These are variables that *may* have predictive power and therefore deserve to be studied and evaluated by researchers while they are developing a new statistical prediction rule. For example, for more than 20 years, the statistical prediction of violence and criminal recidivism has been informed by theory, with the introduction of instruments such as the LSI-R (Andrews & Bonta, 1995). This was a departure from earlier risk assessment procedures like the VRAG (Harris et al., 1993) that used items that were known to have predictive validity without consideration of their theoretical value.

One could argue that attempts to improve predictive validity using theory have failed. For the prediction of violence and recidivism, the use of prediction rules that are informed by theory has not added to validity, as the most widely used statistical prediction rules are all thought to be close to each other in validity (Kroner et al., 2005; Monahan & Skeem, 2016; Williams et al., 2017; Yang et al., 2010). The inclusion of items that reflect dynamic factors was a theoretical advance, but did not lead to improved predictive validity (Van den Berg et al., 2018). The biggest advance in predictive validity may come with the use of machine learning, and it does not rely on theory. However, it would be a mistake to conclude that advances in theory have not occurred. Advances in theory can occur *without* an advance in predictive validity. Also, by using theory to help build statistical prediction rules, one can provide information that is theoretically meaningful to the professionals who will use the statistical prediction rules. This information can be explored in follow-up interviews, and it may help with case management and communication, even if it does not lead to improved prediction. Finally, the development of causal models may help us build new statistical prediction rules. Walters (2017) has argued in favor of improving prediction by conducting mediation analyses to help us understand the causal relations among variables.

There may be another role for theory in building a statistical prediction rule. Some researchers have made compelling arguments that professional judgment should be used to discard variables from a prediction rule, even when those variables have been empirically shown to have predictive power. Specifically, Duwe and Kim (2017) argued that items should be considered for inclusion in a statistical prediction rule only when the direction of their impact is consistent with existing theory. As a hypothetical example, if a researcher was developing a statistical rule to predict violence and found that the beta weight for "number of DUIs" was negative (that is, more DUIs predicted *less* risk of violence), the researcher might be justified in removing the DUI item from the statistical prediction rule even before conducting cross-validation analyses. Although this view represents a significant departure from the way statistical prediction rules have been derived in the past, it does have merit.

Generalizability of Findings

A problem with the use of risk measures in forensic assessment is that they generally perform more poorly in the field than in the lab (Edens & Boccaccini, 2017). One reason this may occur is because of an allegiance effect (the most favorable results for an instrument are often reported by the investigators who created the measure; Blair, Marcus, & Boccaccini, 2008; Singh, Grann, & Fazel, 2013), but another reason is because results are sometimes not robust across samples. For example, a regression rule may accurately describe the variance in a sample, and may accurately describe variance in a validation sample drawn from the same population, but the magnitude of relations among predictors and a criterion may be different in a new sample at a new location. This would lead to lower validity. If equal weight linear rules are used, the drop-off from the lab to the field may not be as large.

Machine learning models can be based on thousands of predictors and hundreds of interactions (e.g., Kessler et al., 2015) so there is a possibility for problems with generalizability. However, machine learning normally includes analyses that are intended to control for overfitting. Machine learning may accurately capture very complex relations in a data set, but those relations may differ in populations at other clinics or may change over time at the same local clinic. Under these circumstances, one may obtain different distributions of scores on the thousands of predictors. For example, in the Kessler et al. (2015) study, a machine learning model was used to make predictions for 40,820 psychiatric inpatients, 68 of whom committed suicide, using data from 38 U.S. Army and Department of Defense administrative data systems. Machine learning may have accounted for true variance in this sample, but when they seek to replicate their results there may not be patients who obtained the same scores on all of the predictor variables. Individuals who commit suicide in a new sample may be described by a different set of predictor values and interactions. Machine learning may be able to capture true variance in both samples, yet a model may be valid in one sample but not another. If such a problem arises, it will be especially perplexing because of the nontransparency of machine learning.

Generalizability may depend on the type of statistical prediction rule that is being used. An equal weight linear rule may be equally valid across different settings (e.g., across prison, community, and hospital settings—even across countries). In contrast, findings for a machine learning model may not be generalizable, not because the machine learning model is incorrectly modeling interactions in a data set but because those interactions do not occur across settings. To improve generalizability, advances in theory will also be important.

Approaches to Statistical Prediction

Many of the claims made by Dawes et al. (1989) have been supported. Statistical prediction rules perform as well as, and often better than, traditional clinical judgment. Adjustments made by professionals to statistical predictions based on their having more information available often leads to a decrease in validity. Equal weight linear rules continue to do well in many studies.

There have also been many changes. We can now recognize that statistical prediction rules do not do well in all judgment tasks. Specifically, psychometric methods are preferable to statistical prediction rules for the description of personality and psychopathology. Structured professional judgment, largely unknown 30 years ago, performs as well as statistical prediction. Machine learning, also largely unknown 30 years ago, can be expected to achieve higher levels of validity when data are complex. However, if a data set contains results for a single psychological assessment instrument, the incremental validity of using machine learning rather than a traditional statistical analysis may be modest. Finally, the importance of theory for statistical prediction has become more widely recognized.

Also different is how prediction rules are evaluated. When evaluating a statistical prediction rule, one will obviously want to weigh predictive validity, but other factors are also important. As noted above, generalizability may differ for different types of statistical rules. Also, depending on how a rule will be used, transparency can be important. This may be particularly true if professionals will be using the statistical predictions, for example, to conduct follow-up interviews or to testify in a legal setting.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341–382. http://dx.doi.org/10.1177/0011000005285875

Andreas, S., Theisen, P., Mestel, R., Koch, U., & Schulz, H. (2009). Validity of routine clinical *DSM–IV* diagnoses (Axis I/II) in inpatients with mental disorders. *Psychiatry Research, 170,* 252–255. http://dx.doi.org/10.1016/j.psychres.2008.09.009

Andrews, D. A., & Bonta, J. (1995). *Level of service inventory–Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The level of service/case management inventory*. Toronto, Ontario, Canada: Multi-Health Systems.

Augimeri, L. K., Koegl, C. J., Webster, C. D., & Levene, K. S. (2001). *Early assessment risk list for boys: EARL-20B, Version 2*. Toronto, Ontario, Canada: Earlscourt Child and Family Centre.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571. http://dx.doi.org/10.1001/archpsyc.1961.01710120031004

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior. *Criminology & Public Policy, 12,* 513–544. http://dx.doi.org/10.1111/1745-9133.12047

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15,* 346–360. http://dx.doi.org/10.1111/j.1468-2850.2008.00147.x

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20*. Burnaby, British Columbia, Canada: The Mental Health, Law, and Policy Institute, Simon Fraser University.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18–32). Thousand Oaks, CA: Sage.

Bonta, J., Harman, W. G., Hann, R. G., & Cormier, R. B. (1996). The prediction of recidivism among federally sentenced offenders: A revalidation of the SIR scale. *Canadian Journal of Criminology, 38,* 61–79.

Borum, R., Bartel, P., & Forth, A. (2003). *Manual for the Structured Assessment for Violence Risk in Youth (SAVRY)*. Odessa, FL: Psychological Assessment Resources.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25,* 6–19. http://dx.doi.org/10.1080/10503307.2013.817696

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin, Germany: Springer Science & Business Media.

Cannon, T. D., Cadenhead, K., Cornblatt, B., Woods, S. W., Addington, J., Walker, E., . . . Heinssen, R. (2008). Prediction of psychosis in youth at high clinical risk: A multisite longitudinal study in North America. *Archives of General Psychiatry, 65,* 28–37. http://dx.doi.org/10.1001/archgenpsychiatry.2007.3

Chevalier, C. S. (2017). *The association between structured professional judgment measure total scores and summary risk ratings: Implications for predictive validity* (Doctoral dissertation). Department of Psychology, Sam Houston State University, Huntsville, TX.

Childs, K., Frick, P. J., Ryals, J. S., Jr., Lingonblad, A., & Villio, M. J. (2014). A comparison of empirically based and structured professional judgment estimation of risk using the Structured Assessment of Violence Risk in Youth. *Youth Violence and Juvenile Justice, 12,* 40–57. http://dx.doi.org/10.1177/1541204013480368

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. http://dx.doi.org/10.1037/1040-3590.7.3.309

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology, 14,* 209–236. http://dx.doi.org/10.1146/annurev-clinpsy-050817-084746

Copas, J., & Marshall, P. (1998). The Offender Group Reconviction Scale: The statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society: Series A: Statistics in Society, 47C,* 159–171.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106. http://dx.doi.org/10.1037/h0037613

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations: A demonstration. *PLoS ONE, 9,* e83875. http://dx.doi.org/10.1371/journal.pone.0083875

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, version 3 (HCR-20V3): Development and overview. *International Journal of Forensic Mental Health, 13,* 93–108. http://dx.doi.org/10.1080/14999013.2014.906519

Duwe, G., & Kim, K. (2017). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review, 28,* 570–600. http://dx.doi.org/10.1177/0887403415604899

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment "out of the lab" and into "the real world": Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29,* 599–610. http://dx.doi.org/10.1037/pas0000475

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10299-000

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1,* 67–89. http://dx.doi.org/10.1146/annurev.clinpsy.1.102803.143810

Garb, H. N. (2013). [Lackland Behavioral Questionnaire and attrition data]. Unpublished raw data.

Garb, H. N., Wood, J. M., & Baker, M. (2018). The Lackland Behavioral Questionnaire: The use of biographical data and statistical prediction rules for public safety screening. *Psychological Assessment, 30,* 1039–1048. http://dx.doi.org/10.1037/pas0000542

Grann, M., & Långström, N. (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34,* 22–36. http://dx.doi.org/10.1177/0093854806290250

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk. *Criminal Justice and Behavior, 45,* 82–100. http://dx.doi.org/10.1177/0093854817719482

Hamilton, Z., Neuilly, M. A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology, 11,* 299–318. http://dx.doi.org/10.1007/s11292-014-9221-8

Hanson, R. K., Helmus, L., & Harris, A. J. R. (2015). Assessing the risk and needs of supervised sexual offenders: A prospective study using STABLE-2007, Static-99R, and Static-2002R. *Criminal Justice and Behavior, 42,* 1205–1224. http://dx.doi.org/10.1177/0093854815602094

Hanson, R. K., Helmus, L., & Thornton, D. (2010). Predicting recidivism amongst sexual offenders: A multi-site study of Static-2002. *Law and Human Behavior, 34,* 198–211. http://dx.doi.org/10.1007/s10979-009-9180-1

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1–21. http://dx.doi.org/10.1037/a0014421

Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex offenders* (User Report 99–02). Ottawa, Ontario: Department of the Solicitor General of Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist–Revised: Manual.* Toronto, Ontario, Canada: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Harris, A., Phenix, A., Thornton, D., & Hanson, R. K. (2003). *Static-99, coding Rules Revised, 2003.* Ottawa, Ontario: Solicitor General Canada.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315–335. http://dx.doi.org/10.1177/0093854893020004001

Heilbrun, K. (2003). Violence risk: From prediction to management. In D. Carson & R. Bull (Eds.), *Handbook of psychology in legal contexts* (2nd ed., pp. 127–142). Chichester, United Kingdom: Wiley.

Heilbrun, K. (2009). *Evaluation for risk of violence in adults.* New York, NY: Oxford University Press. http://dx.doi.org/10.1093/med:psych/9780195369816.001.0001

Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior, 44,* 8–25. http://dx.doi.org/10.1177/0093854816678898

Helmus, L. M., & Thornton, D. (2015). Stability and predictive and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism: A meta-analysis. *Criminal Justice and Behavior, 42,* 917–937. http://dx.doi.org/10.1177/0093854814568891

Hoge, R. D., & Andrews, D. A. (2006). *Youth Level of Service/Case Management Inventory (YLS/CMI) user's manual.* Toronto, Ontario, Canada: Multi-Heath Systems.

Hoge, R. D., & Andrews, D. A. (2011). *Youth Level of Service/case Management Inventory 2.0 (YLS/CMI 2.0): User's manual.* Toronto, Ontario, Canada: Multi-Health Systems.

Kessler, R. C., Hwang, I., Hoffmire, C. A., McCarthy, J. F., Petukhova, M. V., Rosellini, A. J., . . . Bossarte, R. M. (2017). Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *International Journal of Methods in Psychiatric Research, 26,* e1575. http://dx.doi.org/10.1002/mpr.1575

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., . . . Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry, 21,* 1366–1371. http://dx.doi.org/10.1038/mp.2015.198

Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., . . . the Army STARRS Collaborators. (2015). Predicting suicides after psychiatric hospitalization in U.S. Army soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Journal of the American Medical Association Psychiatry, 72,* 49–57. http://dx.doi.org/10.1001/jamapsychiatry.2014.1754

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry, 28,* 360–374. http://dx.doi.org/10.1016/j.ijlp.2004.01.011

Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior, 24,* 101–118. http://dx.doi.org/10.1023/A:1005430904495

Kropp, P. R., Hart, S. D., & Belfrage, H. (2005). *Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER): User manual.* Vancouver, British Columbia, Canada: ProActive ReSolutions.

Levene, K. S., Augimeri, L. K., Pepler, D. J., Walsh, M. M., Webster, C. D., & Koegl, C. J. (2001). *Early Assessment Risk List for Girls: EARL-21G, Version 1* (Consultation ed.). Toronto, Ontario, Canada: Earlscourt Child and Family Centre.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology, 27,* 547–573. http://dx.doi.org/10.1007/s10940-011-9137-7

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* United Kingdom: Cambridge University Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Müller-Isberner, R., . . . Hodgins, S. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. *International Journal of Forensic Mental Health, 12,* 1–13. http://dx.doi.org/10.1080/14999013.2012.760182

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12,* 489–513. http://dx.doi.org/10.1146/annurev-clinpsy-021815-092945

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62,* 783–792. http://dx.doi.org/10.1037/0022-006X.62.4.783

Nicholls, T. L., Viljoen, J. L., Cruise, K. R., Desmarais, S. L., & Webster, C. D. (2010). *Short-Term Assessment of Risk and Treatability: Adolescent Version (START: AV; Abbreviated manual).* Coquitlam, Canada: BC Mental Health and Addiction Services.

Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines.* Ottawa, Ontario: Communication Division, Solicitor General of Canada.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375,* 1216–1219. http://dx.doi.org/10.1056/NEJMp1606181

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11367-000

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18,* 169–184. http://dx.doi.org/10.1002/mpr.289

Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist-Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice, 3,* 203–215. http://dx.doi.org/10.1111/j.1468-2850.1996.tb00071.x

Schmidt, F., Sinclair, S. M., & Thomasdóttir, S. (2016). Predictive validity of the youth level of service/case management inventory with youth who have committed sexual and non-sexual offenses: The utility of professional override. *Criminal Justice and Behavior, 43,* 413–430. http://dx.doi.org/10.1177/0093854815603389

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17,* 156–167. http://dx.doi.org/10.1037/1040-3590.17.2.156

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—But some don't*. New York, NY: Penguin.

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, 31,* 499–513. http://dx.doi.org/10.1016/j.cpr.2010.11.009

Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS ONE, 8,* e72484. http://dx.doi.org/10.1371/journal.pone.0072484

Singh, J. P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other psychiatric disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin, 37,* 899–912. http://dx.doi.org/10.1093/schbul/sbr093

Storey, J. E., Watt, K. A., Jackson, K. J., & Hart, S. D. (2012). Utilization and implications of the Static-99 in practice. *Sexual Abuse, 24,* 289–302. http://dx.doi.org/10.1177/1079063211423943

Thornton, D. (2007). *Scoring guide for Risk Matrix 2000.9/SVC*. Retrieved from https://www.birmingham.ac.uk/Documents/collegeles/psych/RM2000scoringinstructions.pdf

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society Series A: Statistics in Society, 176,* 565–584. http://dx.doi.org/10.1111/j.1467-985X.2012.01056.x

van den Berg, J. W., Smid, W., Schepers, K., Wever, E., van Beek, D., Janssen, E., & Gijs, L. (2018). The predictive properties of dynamic sex offender risk assessment instruments: A meta-analysis. *Psychological Assessment, 30,* 179–191. http://dx.doi.org/10.1037/pas0000454

Vogel, V. D., Ruiter, C. D., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health, 3,* 149–165. http://dx.doi.org/10.1080/14999013.2004.10471204

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology, Research and Practice, 40,* 525–531. http://dx.doi.org/10.1037/a0014693

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213–217. http://dx.doi.org/10.1037/0033-2909.83.2.213

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5,* 457–469. http://dx.doi.org/10.1177/2167702617691560

Walters, G. D. (2017). Beyond dustbowl empiricism: The need for theory in recidivism prediction research and its potential realization in causal mediation analysis. *Criminal Justice and Behavior, 44,* 40–58. http://dx.doi.org/10.1177/0093854816677566

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence (Version 2)*. Burnaby, Canada: Simon Fraser University, Mental Health, Law, and Policy Institute.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Middleton, C. (2004). *Manual for the short term assessment of risk and treatability (START): Version 1.0*. (Consultation ed.). Port Coquitlam, British Columbia, Canada: Forensic Psychiatric Services Commission and St. Joseph's Healthcare.

Widiger, T. A., & Lowe, J. R. (2010). Personality disorders. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (2nd ed., pp. 571–605). New York, NY: Guilford Press.

Williams, K. M., Wormith, J. S., Bonta, J., & Sitarenios, G. (2017). The use of meta-analysis to compare and select offender risk instruments: A commentary on Singh, Grann, and Fazel (2011). *International Journal of Forensic Mental Health, 16,* 1–15. http://dx.doi.org/10.1080/14999013.2016.1255280

Wong, S. C., & Gordon, A. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology, Public Policy, and Law, 12,* 279–309. http://dx.doi.org/10.1037/1076-8971.12.3.279

Worling, J. R., & Curwen, T. (2001). *Estimate of risk of adolescent sexual offense recidivism (ERASOR; Version 2.0)*. Toronto, Canada: Ontario Ministry of Community and Social Services.

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39,* 1511–1538. http://dx.doi.org/10.1177/0093854812455741

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136,* 740–767. http://dx.doi.org/10.1037/a0020473

Youngstrom, E. A., Halverson, T. F., Youngstrom, J. K., Lindhiem, O., & Findling, R. L. (2018). Evidence-based assessment from simple clinical judgments to statistical learning: Evaluating a range of options using pediatric bipolar disorder as a diagnostic challenge. *Clinical Psychological Science, 6,* 243–265. http://dx.doi.org/10.1177/2167702617741845

Youngstrom, E. A., & Van Meter, A. R. (in press). Working smarter, not harder: Comparing evidence based assessment to the conventional routine assessment process. In S. Dimidjian (Ed.), *Evidence-based practice in action*. New York, NY: Guilford Press.

Zimmerman, M. (1994). Diagnosing personality disorders. A review of issues and research methods. *Archives of General Psychiatry, 51,* 225–245. http://dx.doi.org/10.1001/archpsyc.1994.03950030061006

Received December 7, 2017
Revision received September 29, 2018
Accepted October 3, 2018 ■