Law Hum Behav (2010) 34:275–281
DOI 10.1007/s10979-010-9227-3

COMMENTARY

Individual Confidence Intervals Do Not Inform Decision-Makers About the Accuracy of Risk Assessment Evaluations

R. Karl Hanson · Philip D. Howard

Published online: 17 June 2010

Abstract Some recent articles have proposed that the confidence interval for the predicted outcome of a single case can be used to describe the predictive accuracy of risk assessments (Hart et al. Br J Psychiat 190:60–65, 2007b; Cooke and Michie, Law Hum Behav 2009). Given that the confidence intervals for an individual prediction are very large, Cooke and colleagues have questioned the wisdom of applying recidivism rates estimated from group data to single cases. In this article, we argue that the confidence intervals for the recidivism outcome predicted for a single case will range between zero to one (i.e., be uninformative) when the outcome is dichotomous and the predicted probability is between .05 and .95. This is true by definition and limits the utility of using individual confidence intervals to measure predictive accuracy. Consequently, other quality indicators (many of which are non-quantitative) are needed to determine the accuracy and error of risk evaluations.

Keywords Risk assessment · Prediction · Confidence intervals

Author note The views expressed are those of the authors and not necessarily those of Public Safety Canada, the National Offender Management Service or the University of Birmingham.

R. K. Hanson (✉)
Corrections Research, Public Safety Canada, 340 Laurier Ave., West, Ottawa, ON K1A 0P8, Canada
e-mail: karl.hanson@ps-sp.gc.ca

P. D. Howard
National Offender Management Service, England and Wales, UK

P. D. Howard
University of Birmingham, Birmingham, UK

Psychological assessments of recidivism risk can have serious consequences for the individuals assessed and their potential victims. It is important, therefore, that evaluators specify the error and accuracy with which such evaluations are made. Some recent articles have proposed that the confidence interval for the predicted recidivism rates of a single case can be used to describe the predictive accuracy of risk evaluations (Hart, Cooke, & Michie, 2007b; Cooke & Michie, 2009). Given that these individual confidence intervals are very large, some commentators have questioned the wisdom of applying recidivism rates estimated from group data to individual cases (Jailer's Dilemma, 2007). In this article, we argue that individual confidence intervals provide little information concerning the accuracy of a risk assessment. When the outcome is dichotomous, the confidence intervals for recidivism prediction will almost always range from zero to one (i.e., be uninformative). Consequently, other indicators of predictive accuracy are needed, many of which are non-quantitative. When these indicators provide satisfactory results, and estimates are reported responsibly, the application of group-based recidivism rates to individual cases will be appropriate.

The Nature of Empirical Probability Estimates

In order to understanding the empirical probability estimates generated by actuarial risk assessments, it is important to distinguish them from other common forms of probability, such as the probabilities associated with diagnostic uncertainty (Akobeng, 2006a, b) or classical probabilities associated with games of chance (Hald, 1990). In classical probability theory, the probability of a particular outcome is defined as the proportion of all

possible events that include that outcome divided by all possible outcomes. For example, the probability of drawing a spade from a fair deck of cards is ¼: 13 (spades) divided by 52 cards in total. These types of probabilities are exact, true by definition, and proved by logic (mathematics). In applied risk assessment, however, the number of possible conditions is not known and is too large to specify in advance. In such situations, the probability of the outcome cannot be exactly known. It can, however, be estimated from empirical evidence.

Psychological assessments of recidivism risk are different from assessments of disorders (diagnosis) because the outcome of interest is inherently uncertain. If a physician uses a diagnostic test to identify cancer, the diagnosis will be true or false depending on whether the cancer is actually present. Even when the true state-of-affairs is determined in fact (assuming an omniscient view of current affairs), there can be uncertainty about the diagnosis. Consequently, diagnoses are routinely expressed as probabilities (e.g., based on your test results and the base rate among 50-year-old males, there is a 92% chance that the tumor is cancerous, see Akobeng, 2006b). With a determined state-of-affairs, it is appropriate to discuss the accuracy of the test in terms of sensitivities, specificities, and areas under the receiving operator characteristics curve (AUC under ROC, Akobeng, 2006a, 2007).

In contrast to diagnoses, risk assessments estimate the likelihood of an event that has not yet happened, and may never happen. There are inherently stochastic, and the future outcome can only be estimated with a certain probability. Consequently, empirically based risk assessments involve two separate steps. The first step identifies the risk relevant characteristics of the individual being assessed (i.e., "diagnostic" accuracy). The second step estimates the likelihood that individuals with these characteristics will have the outcome of interest. This logic is routinely used, for example, in the estimation of insurance risk (e.g., young men pay more than older women for car insurance) and the identification and communication of health risks (e.g., overweight smokers are at increased risk for cancer).

In agreement with Hart et al. (2007b), we believe that the accuracy of empirical probabilities estimates require justification. We disagree, however, about the utility of the methods they propose for evaluating the accuracy of risk assessments for individuals. When the outcome is inherently uncertain, different evidence is required to establish the value of an assessment method than when the outcome of interest is a determined state-of-affairs. Being "right" cannot involve certainty that a patient or offender will or will not be violent in the future. Instead, justification of accuracy of a risk assessment procedure involves evidence that the estimated probability is sufficiently credible that it can be used for the decision at hand. The following evidence would be relevant in determining the credibility of a risk assessment:

1. Are the factors considered in the evaluation actually related to the outcome? This can be established by follow-up studies, such as those summarized in previous meta-analyses of the risk factors for sexual offenders (e.g., Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005) or mentally disordered offenders (Bonta, Law & Hanson, 1998).
2. Were risk relevant characteristics neglected? This is determined by comparing the known risk factors with the risk factors assessed in a specific evaluation. Professional judgment is required to determine the conceptual overlap between related risk factors (e.g., non-contact sexual offences and self-reported paraphilic interests).
3. What is the accuracy of the evaluation in identifying risk factors? This can be expressed in terms of reliability (for specific tests) or in terms of reliability and validity for evaluations based on inferred characteristics (e.g., sexual deviance inferred from self-report and offence history).
4. How well is the assessment procedure able to distinguish between recidivists and non-recidivists? This can be indexed, for example, by AUC values from follow-up studies by evaluators using the same assessment procedures. Assessment procedures that attain only modest AUC values are likely to be deficient in some or all of the above three respects. Low AUC values can also arise, however, when the sample is relatively homogeneous (low variability on the predictor variables) or when the outcome is inherently difficult to predict. In recidivism studies using criminal records, for example, considerable measurement error would be expected in the outcome variable given that a substantial proportion of true recidivists are not detected by the criminal justice system.
5. What is the expected recidivism rate for an individual with a specific set of risk factors? Evidence in support of expected recidivism rates can be gathered from follow-up studies of individuals with these same characteristics (the same class of offender). Evidence increases given large sample sizes and similar findings across diverse settings. Inference is strengthened given credible studies focusing on samples that closely resemble the features and circumstances of the individual being assessed.

Two recent articles (Cooke & Michie, 2009; Hart et al., 2007b) have criticized the use of actuarial risk assessment instruments (ARAIs) to provide answers to question 5 above. If these criticisms are upheld, the

question, "What is the expected recidivism rate for an individual…?" becomes meaningless, as in most circumstances an individual's expected recidivism rate has such wide confidence intervals that the estimate is rendered useless. If true, this would be a serious challenge to the applicability of any empirically based risk procedure to any individual for anything, and some authors have interpreted their results as proof of the impossibility of predictions for individuals (Jailer's Dilemma, 2007). Not all commentators agreed. Mossman and Sellke (2007) as well as Harris, Rice, and Quinsey (2007; see also Harris & Rice, 2007) were unconvinced by Hart et al.'s argument, citing problems with the specific statistics used and the general conclusions. In this article, we provide additional arguments as to why Hart et al.'s criticisms of ARAIs are misplaced, as well as suggestions on how predictions from ARAIs for individual offenders can be used responsibly.

Hart, Cooke, and Michie (2007b)

Hart et al. (2007b) make a distinction between the confidence intervals for group estimates, and the confidence interval for the probability associated with an individual member of the group. In an argument based on Wilson's (1927) formula, they claim that the confidence intervals for the individual probabilities are so large as to be useless. To fully understand their argument, it is useful to examine the statistical procedures involved.

In most introductory statistics texts, the variance of a proportion is typically defined as

$$\hat{p}(1-\hat{p})/n \tag{1}$$

and the 95% confidence interval is given as

$$\hat{p} \pm Z\sqrt{\hat{p}(1-\hat{p})/n} \tag{2}$$

In these equations, $\hat{p}$ is the observed probability (recidivists/total), n is the total sample size used to estimate $\hat{p}$, and Z is the value from the normal distribution corresponding to the desired confidence interval (typically, 1.96 for 95% CI).

This estimate of the variance (and CI) is not ideal and various alternatives have been proposed (Agresti & Coull, 1998). One problem with Eq. 2 is that the variance is zero under two quite different conditions: (a) when the sample size is infinitely large and (b) when the sample size is sufficiently small that all the occurrences are of the same type (i.e., all recidivists or all non-recidivists). Consequently, Wilson (1927) proposed a new approach, which is reported as Formula 2 from Agresti and Coull (1998):

$$\left(\hat{p} + \frac{Z^2}{2n} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p}) + Z^2/4n}{n}}\right) \Big/ \left(1 + Z^2/n\right). \tag{3}$$

This formula may seem confusing, but, as described by Agresti and Caffo (2000), it is essentially the standard formula (Eq. 2) with a constant added to the top and bottom. Adding the constant makes it possible to create confidence intervals from only one observation. Adding this constant tilts the estimates toward ½ in the absence of additional information. When Z^2 (i.e., $[1.96]^2 = 3.84$) is replaced by 4, the formula for the 95% C.I. simplifies to

$$\tilde{p} \pm Z\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} \tag{4}$$

where $\tilde{n} = n + 4$, and $\tilde{p} = (\text{recidivists} + 2)/(n + 4)$ (Agresti & Caffo, 2000). The decision to add the constant (approximately two recidivists and two non-recidivists) to the formula is based on the assumption that, given an absence of information, the best estimate of a proportion is closer to ½ than to zero (or one). In this way, the constant functions like a Bayesian prior estimate.

The purpose of Eq. 4 is to estimate the range in which the true recidivism rate for the class is likely to fall based on the observed sample from that class. The true value for the class (or group) should lie within the 95% CI of the observed proportion 19 times out of 20 (Wilson, 1927). With large sample sizes, the confidence intervals are small; the confidence intervals bound zero and one when the sample size decreases to zero. Agresti and Caffo (2000) specifically mention that, unlike the standard procedure (Eq. 1), Wilson's (1927) formula can be completed with no observations:

> In fact, we note in closing (and with tongue in cheek) that the adjusted intervals… have the advantage that, as in Bayesian methods, one can do an analysis without having any data. In the single-sample case, the adjusted sample then has $\tilde{p} = 2/4$, and the 95% confidence interval is $.5 \pm ((.5)(.5)/4)^{1/2}$, or [0, 1]….[This analysis is] uninformative, as one would hope from a frequentist approach with no data. No one will get into too much trouble using them! (Agresti & Caffo, 2000, p. 288).

As Mossman and Sellke (2007) point out, however, the way Hart et al. (2007b) used Wilson's (1927) formula did, in fact, get them into trouble: "The 95% CIS for 'individual risk' piles nonsense on top of meaninglessness… With '1' in place of '*n*,' the formulae just do not mean anything" (Mossman & Sellke, 2007). The Wilson formulae are intended to bound the confidence interval for the group estimates of the proportion of recidivists. They cannot provide meaningful information about single observations, which can only take the values zero and one.

In a response to Mossman and Sellke (2007), Hart, Cooke, and Michie (2007a) acknowledged the problems with their application of the Wilson formula, but go onto blame the test developers:

> …They also criticized us for using an ad hoc procedure to estimate the margin of error for individual risk estimates, which they opined served only to 'pile nonsense on top of meaninglessness.' We must plead guilty to some of the charges leveled by Mossman and Sellke —indeed, we pled guilty in our paper, acknowledging the conceptual and statistical problems with the approach we used. In our defence, we claimed duress: Because developers used inappropriate statistical methods to construct ARAIs, we could not use appropriate methods to evaluate them (Hart et al., 2007a, p. 561).

According to Hart et al. (2007a), logistic regression is an appropriate method for calculating the confidence intervals associated with specific scores. In a subsequent presentation, Hart (2008) used logistic regression to estimate recidivism probabilities and their associated confidence intervals using numeric scores from the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997). He fitted an equation with four predictor variables (the four domains from the SVR-20) and presented the confidence intervals for two different outputs. The first output was the complete set of distinct predicted values associated with each combination of scores observed in this data set; the second output examined the predicted values after they were collapsed into only two groups (60 low or moderate risk offenders versus 30 high risk offenders). As expected, the confidence intervals for the grouped data were narrower than the confidence intervals for any specific probability (any specific score pattern). As well, the confidence intervals for the probabilities associated with specific scores were very wide. Hart (2008) interpreted these findings to support his earlier conclusions that the width of confidence intervals increases as sample size decreases, resulting in virtually useless confidence intervals when applied to a single case.

We believe that his findings do not support his conclusions. The finding that the grouped data had smaller confidence intervals than the ungrouped data is uncontroversial, given that *group* confidence intervals would be expected to decrease as the sample size of the groups increases. The finding of wide confidence intervals for specific scores, however, is not a necessary feature of logistic regression analyses. In Hart (2008), the large confidence intervals can be attributed to the small sample size (16 recidivists, 74 non-recidivists) and four predictor variables, resulting in many unique combinations of scores being populated by a single case. When large sample sizes are used (2,000+), logistic regression provides narrow confidence intervals for specific scores (see Hanson, Helmus, & Thornton, 2010).

Cooke and Michie (2009)

In a more recent publication, Cooke and Michie (2009) argue that even with large sample sizes, logistic regression will not provide sufficient accuracy for prediction of the behavior of individual offenders. Cooke and Michie make a distinction between the confidence interval for the group estimate, and the prediction region for an individual case. This is standard practice for linear regression, in which the prediction region is the expected range for the difference between the predicted value and the observed value (e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996, §2.5). Cooke and Michie (2009) use the example of height regressed on weight to demonstrate the unreliability of applying linear regression to the individual. The standard deviations are such that a male estimated to weigh 80 kg could weigh anything between 60 and 100 kg. This application of linear regression is uncontroversial. Linear regression is a standard method for generating a point estimate as well as an estimate of the amount of residual variability given a predictor with a known relationship to the quantity predicted. With a normally distributed, linear variable (such as weight), it is informative to provide numerical estimates of the extent to which the observed value would be expected to be different from the predicted value.

Cooke and Michie then apply the same logic to logits (logged odds ratios), and find that the prediction region ranges almost from zero to one. We argue that this is not an empirical finding; instead, it is the consequence of the dichotomous nature of the outcome variable. The apparent variation across scores can be attributed to the approximations inherent in fitting a logistic function to a specific set of data.

Confidence Intervals for Dichotomous Predictions

We agree that the confidence interval for the individual's risk estimate will be close to [0, 1] for most applied contexts, but disagree with Hart et al. on this implications of this fact. A confidence interval of [0, 1] is a consequence of having only two possible outcomes, and says almost nothing about the predictive accuracy or utility of a prediction instrument.

The 95% confidence interval is defined as the interval within which 95% of the possible outcomes are expected to be found. The outcome whereby the individual either

reoffends or does not reoffend is an example of the Bernoulli process: that is, a binary variable (e.g., yes/no), which is a single case from a binomial distribution. The sole parameter of the Bernoulli distribution is p, the probability of a "positive" outcome or, in this case, recidivism. A binary variable can, therefore, only ever take the values 0 (e.g., does not reoffend) or 1 (e.g., reoffends). The 95% confidence interval around p (i.e., the interval covering 95% of the possible outcomes) is [0,1] whenever p is between .05 and .95. When p is .05 or lower, the interval is [0,0]. When p is .95 or higher, the interval is [1,1]. This is true *by definition* and does not need to be empirically established.

Implications of the 95% Confidence Interval

Although some will argue that confidence intervals ranging from zero to one render prediction impossible, a more plausible interpretation is that "individual" confidence intervals have no utility as measures of predictive accuracy for dichotomous outcomes. Imagine a test that identifies with certainty two groups of offenders. One group has a 94% chance of reoffending, and the other group has a 6% chance of reoffending. Further imagine that by using this test it is possible to demonstrate that all recidivism rates observed in previous studies can be explained by the relative proportion of high-risk and low-risk offenders. Such a test would revolutionize correctional practice, and the test developers would appear on the cover of Science and the New York Times. If judged on the basis of individual confidence intervals, however, this test would have limited utility for decision-making because the confidence intervals for the individual predictions overlap.

How then, should the accuracy of risk assessments be reported? We believe that one plausible approach to reporting predictive accuracy is to report the confidence interval for the estimate derived from the group that the individual most closely resembles (e.g., Hosmer & Lemeshow, 2000, §3.9). Evaluators will also need to make a separate, non-quantitative judgment concerning the extent to which the offender's risk is accurately reflected by the group data. This judgment will involve the quality of the research supporting the assessment procedure, as well as the extent to which the procedure takes into account all relevant risk factors.

For any single study, confidence intervals shrink as the sample size increases, and Cooke and Michie (2009) correctly showed from PCL-R/recidivism data that confidence intervals will also be narrower around the sample mean (where more data were available). As the sample size becomes very large, the width of the confidence intervals will approach zero. Confidence intervals are only valid, however, when all relevant risk factors have been included. When major risk factors have been neglected, the confidence intervals are too narrow.

Assume, for example, that a quality control tester examines coins from a bin containing a mixture of coins produced by two different machines, both of which produce biased coins: Machine 1 produces coins with 60% probability of heads, and Machine 2 produces coins with 40% probability of heads. Based on flipping 10,000 coins, the tester would be expected to observe a rate of close to 50% heads. Using Eq. 2 described above, the 95% confidence interval would be small (±.0098), and fail to identify meaningful variation in the rates. Consequently, small confidence intervals for a single study are also of limited utility for describing the stability of expected recidivism rates.

Evaluators have increased confidence in the empirical estimates when the same results are found across diverse settings and samples. One way of quantifying the stability of predicted recidivism rates is through meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009). Meta-analysis can estimate whether the amount of between-study variability is more than would be expected by chance (Cochrane's Q statistic), and to quantify the amount of variability. Furthermore, random-effects meta-analysis can explicitly include between-study variability into the prediction intervals (Borenstein et al., 2009, §17). Even with very large sample sizes, these prediction intervals could be large when there is real variability in the results across studies. For example, meta-analysis of 100 samples of 100 coins, each sample drawn entirely from one or other of the machines in the above example, would identify that the variance between samples was more than would be expected by chance.

Although it is easy to suggest factors neglected by a risk tool, such assertions gain credibility when it can be demonstrated that the external risk factor adds incrementally to risk prediction. ARAIs gain credibility when the incremental contribution of plausible, external risk factors has been examined and found to provide no new information. Note that this standard cannot be easily quantified, as it depends on the credibility of the external risk factors and the quality of the existing research evidence.

When there is meaningful variation in the recidivism rates across samples and settings, evaluators will need to select the norms that most closely resemble the case at hand. In all cases, evaluators will need to make judgments concerning the extent to which specific risk tools are appropriate for specific decisions with specific offenders, and judge the extent to which risk factors external to the actuarial tool should influence these decisions.

The [0,1] Confidence Interval Is Not Unique to Actuarial Assessment

Despite the limitations of the analyses provided by Hart et al. (2007b) and Cooke and Michie (2009), they arrived at an (almost) correct answer concerning confidence intervals. Actuarial assessments will usually have a [0,1] confidence interval when applied to an individual. This feature is not unique, however, to actuarial risk assessment. It would apply to all predictions for any dichotomous outcome using any method. Advocates of structured professional judgement want to exempt SPJ from this criticism by not providing explicit recidivism estimates for nominal risk categories. We do not believe that they can.

Information about the probability of recidivism is inherent in the use of terms such as "low," "moderate," and "high" risk. When an evaluator states that an offender's risk is "low" based on a SPJ (or any) evaluation, the evaluator is communicating that the expected probability of recidivism for this offender is different from (less than) it would have been if the offender had been deemed "high" risk. The criticism of overlapping [0,1] confidence intervals applies whenever one of the nominal risk categories includes an implied probability of recidivism between 5 and 95%. Consider, for example, an SPJ tool such as HCR-20 (Webster, Douglas, Eaves, & Hart, 1997). When offenders are placed in the Low-risk category, does this imply that their probability of recidivism is always below 5%? When they are considered High, is the probability always over 95%? Unless both of these are the case, the confidence intervals for Low and High-risk ratings of an individual will overlap; if neither is the case, the confidence intervals will be identical. Medium risk classifications will certainly carry a [0,1] confidence interval for the individual.

Defenders of the SPJ approach could argue that these methods are not intended to assess absolute recidivism rates, but only broad categories of relative risk (i.e., low, moderate, high). Nevertheless, the logical consequence of their interpretation of the [0,1] confidence interval apply to all statements about the recidivism risk of individuals. Relative risk ratings would have little utility if their implied recidivism rates were identical. Therefore, if, despite our arguments above, the [0,1] confidence interval is still regarded as a flaw of estimates resulting from actuarial assessments, it must equally be regarded as a flaw of risk judgments from SPJ assessments.

Conclusion

Given that risk assessment in corrections and mental health has serious consequences for those evaluated and for public safety, it is important that the accuracy of such risk decisions are known. Some recent articles have proposed using the confidence interval for the outcome of an individual case as a measure of predictive accuracy. Such confidence intervals have limited utility, however, as they will range between zero and one whenever the estimated probability is between 5 and 95%. Although some will argue that this proves the impossibility of individual level prediction, our view is that individual confidence intervals are uninformative about the accuracy of the risk assessment procedure when the outcome is dichotomous.

We believe that group data can and should be used to estimating the recidivism risk of a particular case. Evaluators need to complete two tasks. The first task is the administration of the risk tool (i.e., scoring items, creating total scores, or summary risk judgments). This stage of the evaluation can be summarized in risk communications by phrases such as "Of offenders with similar age, gender, and criminal history, about two in ten are known to reoffend within 2 years of their date of release into the community," or "Previous studies have found that 20% of offenders with similar psychological characteristics as Mr. X were reconvicted of a sexual offence within 5 years" (for other options, see Babchishin & Hanson, 2009). Evaluators should also be clear in their reports that these are group data and the recidivism risk of Mr X. may be higher or lower based on factors not measured by this assessment procedure, or due to other limitations of the risk procedure.

The evaluation should not stop there, however. A second judgment is required concerning the credibility of the risk assessment procedure for addressing the recidivism risk of this specific individual. Statements concerning the credibility of the risk assessment procedure should be explicit when high-stakes evaluations are completed by experts with the training necessary to appreciate the relevant research literature. In the routine application of risk assessment tools, judgments concerning the credibility of the assessment process will often be implicit: they can be inferred by the evaluator's decisions to include or exclude specific risk assessment procedures, and in the weight accorded to particular measures in the overall risk evaluation. For many evaluators, the decision to use a particular tool is not their own. Correctional and mental health systems frequently specify the tools to be used—a decision that can be informed by consultations with experts, cost–benefit analyses, and the ability of the typical user to use the tool for specific purposes (e.g., triage into treatment, community supervision levels). In all cases, however, there needs to be a judgment that *this* measure is appropriate for *this* decision concerning *this* individual, whether the responsibility for this judgment is assumed by a centralized authority or by the end user.

Although numbers and empirical data should inform the judgment concerning the credibility of the risk assessment procedure, the judgment is fundamentally qualitative. The point of the current article is that confidence intervals for individual cases do not inform decision-makers about this second step. Instead, we argue that quality of risk assessments should be judged by other criteria, such as their ability to distinguish between recidivists and non-recidivists, the stability of the observed recidivism rates across samples, the degree to which variables neglected by the risk assessment scheme add incrementally to predictive accuracy, and the utility of the risk assessment procedure to inform applied decisions, including whether the procedure can be applied affordably and reliably in practice.

References

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician, 54*, 280–288.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119–126. doi:10.1214/ss/1009213286.

Akobeng, A. K. (2006a). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Pædiactrica, 96*, 338–341. doi:10.1111/j.1651-2227.2006.00180.x.

Akobeng, A. K. (2006b). Understanding diagnostic tests 2: Likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Pædiactrica, 96*, 487–491. doi:10.1111/j.1651-2227.2006.00179.x.

Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Pædiactrica, 96*, 644–647. doi:10.1111/j.1651-2227.2006.00178.x.

Babchishin, K. M., & Hanson, R. K. (2009). Improving our talk: Moving beyond the "low", "moderate", and "high" typology of risk communication. *Crime Scene, 16*(1), 11–14. Retrieved from http://www.cpa.ca/sections/criminaljustice/publications.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence*. Vancouver, BC: The British Columbia Institute Against Family Violence.

Bonta, J., Law, M., & Hanson, R. K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123*, 123–142. doi:0033-2909/98/S3.00.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Cooke, D. J., & Michie, C. (2009). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. Advance online publication. *Law and Human Behavior.* doi:10.1007/s10979-009-9176-x.

Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York: Wiley.

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*, 348–362. doi:0022-006X/98/S3.00.

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.

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*, 1154–1163. doi:10.1037/0022-006X.73.6.1154.

Harris, G. T., & Rice, M. E. (2007a). Characterizing the value of actuarial risk assessments. *Criminal Justice and Behavior, 34*, 1638–1658. doi:10.1177/0093854807307029.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (2007, August 15). Abandoning evidence-based risk appraisal in forensic psychiatry: Comments on Hart et al. *British Journal of Psychiatry* (electronic letter). Retrieved March 29, 2010 from http://bjp.rcpsych.org/cgi/eletters/190/49/s60#5674.

Hart, S. D. (2008, March). Group vs. individual risk: Precision of estimates. In J. Dvoskin (Chair), *Thinking clearly about the accuracy of actuarial risk assessment instruments.* Symposium presented at the Annual Meeting of the American Psychology-Law Society (Division 41 of the American Psychological Association), Jacksonville, FL.

Hart, S. D., Cooke, D. J., & Michie, C. (2007a). Margins of error for individual risk estimates: Large, unknown and incalculable. *British Journal of Psychiatry* (electronic letter). Retrieved March 29, 2010, from http://bjp.rcpsych.org/cgi/eletters/190/49/s60#5674.

Hart, S. D., Cooke, D. J., & Michie, C. (2007b). Precision of actuarial risk assessment instruments: Evaluating the "margins of error" of group v. individual predictions of violence. *British Journal of Psychiatry, 190*, 60–65. doi:10.1192/bjp.190.5.s60.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Jailers Dilemma (2007, June 23). *Economist, 383*(8534), 90.

Mossman, D., & Sellke, T. M. (2007, July 5). Avoiding errors about "margins of error". *British Journal of Psychiatry* (electronic letter). Retrieved March 29, 2010, from http://bjp.rcpsych.org/cgi/eletters/190/49/s60#5674.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

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.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*, 209–212.