Evidence for Risk Estimate Precision: Implications for Individual Risk Communication

Grant T. Harris
Queen’s University and University of Toronto

Christopher T. Lowenkamp
Administrative Office of the US Courts, Probation and Pretrial Services Office and
University of Missouri Kansas City

N. Zoe Hilton
University of Toronto and Waypoint Research Institute

Version
Post-print/accepted manuscript

Citation (published version)

Publisher’s Statement
This is the peer reviewed version of the following article: Evidence for Risk Estimate Precision: Implications for Individual Risk Communication, which has been published in final form at DOI: 10.1002/bsl.2158. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

How to cite TSpace items

Always cite the published version, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the author manuscript from TSpace because you cannot access the published version, then cite the TSpace version in addition to the published version using the permanent URI (handle) found on the record page.

This article was made openly accessible by U of T Faculty. Please tell us how this access benefits you. Your story matters.
Evidence for Risk Estimate Precision: Implications for Individual Risk Communication

Grant T. Harris
Queen’s University and University of Toronto

Christopher T. Lowenkamp
Administrative Office of the US Courts, Probation and Pretrial Services Office and
University of Missouri Kansas City

N. Zoe Hilton
University of Toronto and Waypoint Research Institute


Author Note

Grant T. Harris, Department of Psychology, Queen’s University; and Department of Psychiatry University of Toronto. Christopher T. Lowenkamp, Department of Criminal Justice and Criminology, University of Missouri Kansas City; and Administrative Office of the US Courts, Probation and Pretrial Services Office. N. Zoe Hilton, Department of Psychiatry, University of Toronto; and Waypoint Research Institute, Penetanguishene.

We would like to thank Sonja Dey for assistance with manuscript preparation.

Correspondence concerning this article should be addressed to Christopher Lowenkamp, One Columbus Circle, N.E., Room 4-328, Washington, DC, 20544. Email contact: christopher_lowenkamp@ao.uscourts.gov.
Abstract

Actuarial risk assessment instruments using well established predictor variables measured at the individual level (e.g., age, criminal history, psychopathy) discriminate well between recidivists and nonrecidivists across diverse samples. Data indicating the relative risk of recidivism can inform policy decisions about allocating resources according to risk within a correctional system, consistent with the first of the Risk, Need, and Responsivity (RNR) principles. Evidence for the precision of absolute risk as applied to an individual based on scores from many samples, however, has proven challenging. In this paper, we present a study examining the association of actuarial risk estimate precision with sample size using the Post Conviction Risk Assessment (PCRA; Lowenkamp et al., 2013), in samples of up to 26,642 offenders. Results indicate that the precision of individual estimates can be demonstrated with sufficient sample size. We believe that the implications of absolute risk for the communication of an individual offender’s risks have been poorly understood. We argue that the purpose of individual-level risk communication is to ensure the effective application of policy, which requires matching a new case to aggregate data. We illustrate how an offender’s risk might thus be communicated, and conclude that this function is distinct from management of an individual’s criminogenic needs and identification of effective and suitable treatments.

Key Words: actuarial, offender, precision, risk communication, violence
Evidence for Risk Estimate Precision: Implications for Individual Risk Communication

In the development of actuarial risk assessment instruments, outcome data from large samples are used to identify independent variables or predictors that best distinguish offenders who meet the criteria for recidivism from those who do not. These variables are combined, sometimes using weights based on their predictive validity, to yield a numerical score for each individual case. Particular scores or ranges are then used to define groups expected to exhibit usefully distinct rates of recidivism, often known as risk categories. This expectation is then tested in validation studies using new samples. To the extent that predictors are measured at the individual level and exhibit robust, replicable associations with recidivism, they can be used to assign a new individual to one of the existing risk categories. In order for these categories to be used to communicate the individual’s risk of recidivism, certain criteria must be met, the most important of which is demonstration of the actuarial instrument’s discriminative or predictive validity. The discriminative validity of several actuarial instruments is well established. For example, there have been more than 70 tests of the Static-99/2002 family of actuarial tools for the assessment of the risk of sexual recidivism, involving more than 20,000 cases, overall yielding a moderate predictive effect (Hanson & Morton-Bourgon, 2009). There have been a similar number of tests of the VRAG/SORAG actuarial system for the assessment of the risk of violent recidivism, involving more than 15,000 cases, overall yielding a large effect (Harris, Rice, Quinsey, & Cormier, 2015). A concern has arisen, however, that these empirical findings reflect the performance of actuarial assessment for groups but not for individuals. Performance needs to be demonstrated for individuals because forensic decisions are about individual offenders (Hart & Cooke, 2013; Hart, Michie, & Cooke, 2007). In this article, we analyze original data from a very large sample to examine the extent to which the precision of individual
estimates depends on sample size, and consider the implications with respect to group versus individual prediction. We also discuss the somewhat related issue of absolute risk in the context of the principles of offender intervention.

**Predictive Validity of Actuarial Risk Assessment Instruments**

Scales using only a few items reliably predict recidivism among violent and sexual offenders (e.g., Campbell, French, & Gendreau, 2009; Hanson & Morton-Bourgon, 2009; Harris, Rice, & Quinsey, 2010). The twelve-item *Violence Risk Appraisal Guide* (VRAG; Harris, Rice, & Quinsey, 1993; Harris et al., 2015) and fourteen-item *Sex Offender Risk Appraisal Guide* (SORAG; Harris et al., 2015) have been replicated in independent samples from different countries, of men and women, bearing different ethnic composition, followed up for a range of time periods from a few months to more than twenty years, and having different inclusion criteria (e.g., sex offenders, forensic patients, developmentally delayed clients, civil psychiatric patients, domestic offenders, and so on; Harris et al., 2010, 2015). Actuarial instruments are widely used across the United States, such as the *Post Conviction Risk Assessment* (PCRA) which has been shown to discriminate among those with and without arrest for any new offense (Lowenkamp, Johnson, Holsinger, VanBenschoten, & Robinson, 2013).

The accepted standard for quantifying the predictive accuracy of actuarial scores or categories to discriminate between recidivists and nonrecidivists is the Receiver Operating Characteristic Area under the Curve (ROC area, or AUC; e.g., Hanson & Howard, 2010; Mossman, 1994; Rice & Harris, 2005). The ROC area entails the comparison of offenders recorded either as recidivists or nonrecidivists in a given sample at several possible cut scores, each of which defines a unique pair of sensitivity and specificity statistics. A plot of these pairs captures the size of the predictive effect. This statistic can also be compared across studies,
allowing for aggregated findings to yield a single average ROC area statistic. For example, the VRAG/SORAG system yields an average ROC area under the curve exceeding .72 for studies of violent recidivism in the community, the outcome for which it was designed (Harris et al., 2010; Harris et al., 2015). Importantly, the ROC area indexes relative predictive accuracy, but does not address the absolute recidivism rates associated with each score or category.

**Individual Confidence Intervals**

Assessing predictive accuracy using ROC area requires a group of cases at each score or in each category, some of whom meet the criteria for recidivism and some of whom do not. A question has arisen as to how much information about the risk of recidivism is actually conferred by an actuarial score when a user is attempting to assess the risk of a single new case. Hart and Cooke (2013) concluded that actuarial risk assessment “cannot be used to estimate the specific probability or absolute likelihood that an individual person will commit violence” (Hart & Cooke, 2013, p.95). Hart and colleagues (Hart & Cooke, 2013; Hart et al., 2007) used frequentist reasoning to equate the absolute probability of recidivism among a group of individuals to a single new individual being assessed who receives the same actuarial assessment score. They then calculated intervals around a point estimate for an individual and demonstrate non-equivalence with estimates derived from group data. This approach has created a good deal of confusion in the relevant clinical and legal literature, being used, for example, to argue that risk assessment tools cannot discriminate among offenders with respect to recidivism any better than informal assessment can (e.g., Starr, 2014).

In the present article, our aim is not to offer a point-by-point criticism of the approach of using confidence intervals to determine the precision of individual risk estimates, which has been done elsewhere (e.g., Hanson & Howard, 2010; Harris et al., 2008; Mossman & Sellke, 2007).
Instead, we compare the results of such precision tests as a function of sample size. This issue is best examined with large samples, so we began with a data set of 26,642 offenders whose outcome was known, and repeated the test in four sub-samples of decreasing size to a minimum of 90 cases.

**Method**

**Sample**

The cases for this study were drawn from the United States Probation database system. All offenders under federal supervision (probation or a term of supervised release) are tracked in this system. Probation officers assess and reassess offenders on their caseloads over time using the Post Conviction Risk Assessment (PCRA; Johnson, VanBenschoten, Robinson, & Lowenkamp, 2011; Lowenkamp, Johnson, Holsinger, VanBenschoten, & Robinson, 2013) and the results of these assessments are maintained in the system. Cases in the present study were selected if the offender was assessed by a probation officer between November 1, 2009 and October 1, 2013, and outcomes for violent recidivism were known. These cases were not included in any previously reported research on the PCRA. These criteria were met by 132,102 offenders of whom 4% were violent recidivists. To avoid problems with statistical prediction associated with low base rates, we then selected all 4691 offenders who recidivated with a post-assessment arrest for a violent offence and added 21,946 randomly selected cases so that the proportion of recidivists was 18% of the sample, to equal that reported by Hart and Cooke (2013).

**Independent Variable Measures**

The database system contained information about PCRA scores taken from federal presentence reports, risk assessment reports, criminal record checks, and records held in an
offender pretrial and probation tracking system. The PCRA scores were completed by probation officers after reviewing official records and interviewing the offender. Each of the fifteen items is rated and scored by officers and then summed to create an overall total risk score. PCRA scores range from 0 to 18 and are used to categorize individuals into low (0 to 5), low-moderate (6 to 9), moderate (10 to 12), and high-risk categories (13 to 18). Individual risk communication is normally based on the probability of recidivism; for example, in the high risk category, “47% of offenders have their supervision revoked and 30% are rearrested within the first 190 days from their initial assessment” (United States Courts Administrative Office, 2012).

Discriminative validity of the PCRA was demonstrated by Lowenkamp et al. (2013) as PCRA categories predicted the occurrence of subsequent arrest for any offense in post-construction validation samples with ROC areas of .71 to .78. Furthermore, other work in progress with the present sample has demonstrated the PCRA's discriminative validity for arrest for violent behavior with ROC areas of .76 to .77 (Lowenkamp, Holsinger, and Cohen, 2014). The latter finding supported our current use of violent recidivism as the dependent measure.

In the present study, we did not attempt to validate the total PCRA score or its four risk categories. Instead, we aimed to demonstrate the precision of estimates for actuarial assessment derived from the PCRA domains. The PCRA items are grouped into five domains. The logistic regression model used to calculate predicted probabilities included the scores on these five domains rather than the total PCRA score; however, we used the total PCRA score to later calculate both predicted probabilities and actual failure rates that are associated with a PCRA score. The criminal history domain includes five items (pertaining to previous arrests, violent offending, offending during supervised release, institutional adjustment, and age at intake to supervision) with a possible range of 0 to 9 points. The education and employment domain
includes four items (highest level of education, employment, jobs in the past year, and general work history) with a possible range of 0 to 3 points. The substance use domain includes five items (substance use causing disruptions to work/school/home, hazardous substance use, legal problems, continuing use despite problems, and current alcohol or drug problem) with a possible range of 0 to 2 points. The social network domain includes six items (marital status, current living situation, lack of familial support, stability of the family situation, criminogenic peer networks, and lack of prosocial support systems) with a possible range of 0 to 3 points. The cognitions domain includes two items (antisocial attitudes/values and attitude toward supervision or change) with a possible range of 0 to 1 points.

**Dependent Variable (Recidivism)**

The dependent variable in this study was arrest for a violent offense, rather than any arrest, in order to be more closely comparable with Hart & Cooke (2013) who used arrest for sexual violence. Our definition of violence followed that of the National Crime Information Center (NCIC) and included robbery, homicide, assault (domestic or nondomestic), stalking, and kidnapping and excluded sex offenses. Arrest for violent offenses occurring any time after the assessment date were identified using criminal history records or rap sheets gathered through the NCIC and Access to Law Enforcement System databases. Recidivism was defined dichotomously as at least one arrest for a violent offense, versus no arrest for a violent offense, after administration of the PCRA and up to October 15, 2013. The mean follow up length was 569 days with a standard deviation of 265 days.

**Procedure and Statistical Analysis**

We followed the same methodological approach as Hart and Cooke (2013). We copied the syntax they reported and modified it to accommodate our variable names and customization
of graphs (Figure 1). As such, our analyses followed an identical process. First, an actuarial risk assessment instrument (ARAI) to predict arrest for a violent offense was developed using the five domains on the PCRA using logistic regression (Figure 1, line 1 of syntax). The regression model produced the following equation:

\[
\text{ARAI} = -3.87 + \text{Criminal History Score} \times 0.39 + \text{Employment/Education} \times 0.06 \\
+ \text{Drugs/Alcohol} \times 0.06 + \text{Social Networks} \times 0.18 + \text{Cognitions} \times 0.27
\]

Once an actuarial instrument algorithm was generated, the generated instrument scores and standard error were calculated. ROC analyses were conducted for each sample reported below (Figure 1, lines 2-7). The next step involved estimating group risk and margins of error (Figure 1, lines 8-11). A high-risk group (the upper third of cases ordered on the probability of failure) and a low risk group (the lower two thirds of the cases) were created. Using these groups, equality of proportions tests on recidivism were carried out as were point estimates and confidence intervals for odds ratios (the odds of failure for the high risk group is 4.8 times that of the low risk group). The probabilities of failure, upper and lower 95% confidence interval values and the upper and lower bounds of the confidence interval were then calculated (Figure 1, lines 12-15). In addition to summarizing the predicted probabilities of failure and the confidence interval, we ran \( t \)-tests to identify significant differences (see Table 4). Finally, a series of graphs was generated to plot the probability of failure – the individual risk estimate – and the 95% confidence interval of this estimate (Figure 1, lines 19-20). Given the large number of cases presented in some of the graphs we elected to use \texttt{rcap} rather than \texttt{rcapsym}, which generates a range plot with spikes capped with marker symbols. We also added syntax that added reference lines to the graphs as well as adjusted the placement and size of graph titles.
We repeated this syntax to estimate a logistic regression model predicting arrest for a violent offense for the overall sample and sub-samples of decreasing size. The first sample contained all those offenders for whom we had complete data on the independent variables of interest. The second through fifth subsamples were merely randomly selected cases from the larger dataset containing cases with complete data. That is, we generated a dataset with fifty percent of the whole sample, five percent of the whole sample, 500 cases, and 90 cases. Following this process we generated five charts, one on each of these subsamples, to demonstrate the effect of sample size on the size of the confidence interval.

Hart and Cooke’s (2013) results were reported for a single sample of 90 offenders. Although we used only one sample of each size to generate the exact results reported below, the syntax was run several times for each sample size (as well as several times using a base sample of 132,102 offenders without the stratified sampling for recidivists used in the present procedure). We did so without setting a seed and obtained very close estimates and numbers, leading to identical conclusions.

Consistent with Hart and Cooke (2013), we used a variable follow up for the logistic regression analyses, which enabled us not only to replicate their method but also to maximize sample size which was essential to our ability to test the precision of risk estimates in large samples. It is important to note that, as with Hart and Cooke (2013), the purpose of the present study was not to test the discriminative validity of the PCRA as a tool to estimate the probability of future violence; this validity has been demonstrated elsewhere (Lowenkamp et al., 2013) and it is used here for illustrative purposes only.

Results
Table 1 presents the descriptive statistics on the demographic characteristics of the sample, including the distribution of PCRA risk categories. The average overall score on the PCRA was 6.81 (out of a possible 18 points). Average scores for each of the PCRA domains are also shown in Table 1 and indicate that most of the risk points are attributable to criminal history. The follow-up period for recidivism was variable with an average follow-up time of 569 days (median = 566) with over 73% of the sample being followed for one year or more. In this time, 18% of the sample was rearrested for a violent offense during the follow-up period (as per sample construction) and 28% was rearrested for any offense. The PCRA yielded an ROC area under the curve of .73 for any arrest in the total sample.

The descriptive statistics for the PCRA are presented in Table 2 for each sample used in these analyses. Given that the 50%, 5%, 500-case and 90-case samples were randomly drawn from the larger dataset, it is expected that the mean, minimum, maximum values would be fairly similar. The mean logit values from the logistic regression models incorporating the PCRA domains range in value from -2.01 to -1.82 across the samples, and the minimum and maximum values range from -4.43 to -3.79 and +0.35 to +1.04, respectively. The mean standard errors associated with the PCRA vary considerably according to the size of the sample, $r (n=5) = -.68$, from 0.01 for the full sample and 0.74 for the sample of 90 cases, and the minimum and maximum values of the standard error increase accordingly. Where ROC areas are shown in Table 2, we also calculated Harrell's C Index (an alternative to the ROC area under the curve when follow-up times vary; Harrell, Lee, & Mark, 1996) for the PCRA’s prediction of violent arrest in the whole sample and in each subsample, which produced similar results. The PCRA achieved discriminative predictive validity with a large effect size (over .71; Rice & Harris, 2005) in all samples.
The group estimates of risk and margins of error are presented in Table 3. We divided the sample based on the syntax and method that Hart and Cooke (2013) used. That is, we divided the sample into three equal groups based on the PCRA score and then the lower two groups were combined into the low risk group and the upper group (the top 1/3 of the cases on the PCRA) became the high risk group. This group division was created each time the syntax was run on a sample. As indicated in Table 3, the failure rates for the high-risk group range from 31% to 37% across the different samples; and for the low-risk group, 8% to 10%. The 95% confidence intervals show slight variability about the estimates for the larger samples and larger variability for the smaller samples, particularly for the high-risk group. The failure rates for the high-risk group are 3 to 4 times greater than the failure rates for the corresponding low-risk group. For each sample, the probability of failure for the high-risk group significantly differs for the probability of failure for the low risk group (see Table 3 note for $Z$ and $p$ values).

Ranges of individual level probability of failure and 95% confidence intervals are presented in Table 4. The range for the low-risk group is 1% to 22% with a mean that ranges from 9% to 10% across the samples. The individual level probabilities for the high-risk group range from 20% to 74% with a mean from 32% to 36% across the samples. The values reported for the CIs in Table 4 are, conceptually, the prediction intervals referred to by Hart et al. (2007). The width of the intervals should not cause alarm as these are the range of values that can occur for the 95% CI about a predicted probability for all individuals and do not represent the 95% CI about the mean failure rate for a given group.

Figure 2 provides a visual display of the data contained in Table 4 with greater granularity as these figures graph the 95% CI as a function of individual risk estimates for all of the cases. In each of the figures there are two reference lines. The horizontal reference line
marks the mean failure rate for the entire sample. The vertical reference line is positioned at the cutoff between the low and high risk group. Figure 2a presents the individual 95% CIs for the whole sample (N = 26,642). There are several important aspects of this figure to note. First, it is the case that almost all of the high-risk cases (those to the right of the vertical reference line) have a risk estimate and 95% CI far above the base rate for the group. Second, most low risk cases have a risk estimate and 95% CIs that exclude the base rate. Third, while the confidence intervals for contiguous cases do overlap, within each risk group, there are cases within each group whose associated risk estimates and 95% CIs are distinct from other cases within the same risk group. In the overwhelming majority of cases, it is possible to say that the subjects’ predicted risk of failure is greater or lower than the base rate. In the remaining figures (2b-2e) the confidence intervals get wider. As the number of cases in the sample decreases, the likelihood that case probabilities and confidence intervals are distinct from other cases within the same risk category decreases. Further, as the sample size decreases, more of the confidence intervals include the base rate. The sample with 90 cases produces a figure (Figure 2e) that is nearly identical to that reported by Hart & Cooke (2013) figure, also based on 90 cases.

Figure 3 displays the failure rates and confidence intervals about those failure rates for each PCRA score, in the whole sample. In most instances, a group’s failure rate is significantly different from the failure rate of another group that scores only two or three points apart.

Figure 4 shows the predicted probabilities as a function of the score on the PCRA, with the median and distribution of cases within each score illustrated by a box and whisker plot. In most instances the interquartile ranges of scores two points apart do not overlap, and in some instances, the interquartile ranges of contiguous scores do not overlap. This provides a visual presentation of prediction interval such as Hart, Michie, and Cooke (2007) described. The low
density at high scores indicates the small number of cases with such scores on the PCRA (however, one could design a risk assessment to yield categories based on proportion of cases that ensures less “thinning” at extreme scores).

**Discussion**

In the present study, we demonstrated that, despite fairly consistent and high discriminative accuracy, confidence intervals around probability estimates widened as sample size decreased. The same syntax as previously used to conclude that an actuarial risk assessment had no precision in risk estimates, and therefore no validity, was here used to demonstrate that this “precision” was dependent on sample size and could easily be observed in sufficiently large samples. We conclude that the accuracy of actuarial risk assessment cannot be disproven using the frequentist reasoning and related analysis of confidence intervals. Concurring with Hanson and Howard (2010), we suggest these confidence intervals around estimates for small or single-case samples have no bearing on individual risk communication.

We argue that the focus on absolute risk of recidivism in individual level risk communication is sometimes misdirected. Social and criminal justice policies are almost always about offenders’ risk of actual repeated criminal offending, whereas actuarial assessments identify the risk of detected recidivism (a subset of all actual recidivism). Such absolute estimates can be derived from actuarial systems. Although there is evidence that the rank ordering of offenders by actuarial methods based on detected recidivism is very close to the rank ordering that would be observed if all recidivistic conduct could be measured (Farrington, 1985), the rates of actual and detected recidivism differ, and do so as a function of actuarial scores.

**On Improving Absolute Risk Estimates**
It has been proposed that score-by-score recidivism rates can be easily corrected using Bayes’ Theorem together with the overall base rate in any replication sample or population (Mossman, 2006), but the accuracy of such corrections cannot be assumed (Harris & Rice, 2007, 2013). When base rates increase, observed rates for individual categories also generally increase, but not as implied by simple application of Bayes. The only apparent exceptions to this are corrections due to variations in the duration of opportunity which do appear to follow Bayes fairly well. Importantly, the base rate of actual recidivistic conduct (as opposed to detected recidivism) is almost always unknown, obviating such a simple Bayesian solution. Furthermore, the ways observed rates under-estimate the actual behavior of interest could vary by institution, jurisdiction, temporal cohort, operational definition of recidivism, follow-up duration, and offender subpopulation, and so on. There is some research on how detected rates vary as a function of some of study features (referred to as calibration; Helmus, 2009). For example, when there has been a close match between such features in replication studies and the original construction, we reported very close correspondence and successful replication of observed category-by-category rates (Harris, Rice, & Cormier, 2002; Harris et al., 2003), and some independent studies permit the conclusion of no significant difference between norms and observed rates in new samples (e.g., Kroner, Stadtland, Eidt, & Nedopil, 2007; Mills, Jones, & Kroner, 2005; Tengstrom, 2001). There is no research yet, however, on how the difference between actual recidivism and detected recidivism varies as a function of other factors such as follow-up duration. Where such factors are very different, category-by-category rates of observed recidivism differ (e.g., Looman, 2006; Rossegger, Gerth, Singh, & Endrass, 2013; Snowden, Gray, Taylor, & MacCulloch, 2007).
The more crucial issue for the present purposes pertaining to absolute rates of recidivism, however, is how these study features might be related to the difference between detected and total recidivism. Overall, interpretation of absolute risk requires knowledge of jurisdiction-specific conditions, plus information on the difference between actual and observed rates. Interpretation and communication of individual risk assessment scores likewise require a jurisdictional or policy-level perspective.

**Policy-Level Risk Communication**

The Risk-Need-Responsivity principles of criminal justice practice and offender intervention (e.g., Andrews & Bonta, 2010) state as a first fundamental principle that intervention (i.e., treatment, custody, supervision) should be apportioned in accordance with risk. That is, the highest risk individuals within any given decision juncture receive the most intensive and extensive intervention available at that juncture. Risk principle decisions are best implemented at the policy level. For example, a policy decision could assign a community disposition to any offender convicted of a first-time no-contact sex offense and scoring in the lowest two risk categories of the VRAG-R (Rice, Harris, & Lang, 2013); the most extensive treatment services might be assigned to all contact sex offenders scoring in the highest three categories. Any individual offender meeting these criteria would be managed according to the policy. Those assigned to intervention would undergo further assessment of criminogenic needs (i.e., measureable aspects of offenders or circumstances empirically related to the commission of crime and recidivism)—the Need principle. The most empirically efficacious interventions are delivered as appropriate to each individual’s specific strengths and deficits (the Responsivity Principle). In this way, two offenders in the same category of actuarial risk might be managed quite differently, depending on further individual-level assessment. To implement the risk
principle, the distribution of actuarial scores for all relevant cases in the population is used in conjunction with knowledge of the resources available in the jurisdiction or agency’s clientele to allocate these resources. The type of evidence necessary for valid use of a tool in this manner is that demonstrating relative predictive validity, as evidenced by statistics such as the ROC area.

Knowing an individual offender’s percentile rank with respect to risk, rather than expected proportions of detected recidivism, is relevant to allocating resources within a forensic or correctional system. If all other cases score higher than a given offender, then that offender should receive few available intervention resources (if any). On the other hand, if all relevant cases score lower than the given offender, then that individual should receive proportionately more of the available intervention resources. Thus, we suggest that expert risk communication involves more than accurate actuarial assessment; it should also include information about where the individual’s score falls within the relevant population on the same assessment tool. Where no data are available on the distribution of actuarial scores in any particular agency’s population, the appropriate course would be to rely on the tool’s normative data. Good risk communication about an individual offender, we suggest, also includes useful information on the available custody, supervision, and treatment resources (and ideally the established magnitude of their causal effects on violent behavior) so as to permit the decision-maker to apportion those resources most appropriately. In Table 5, we illustrate some tentative rubrics for expert communication of violence risk that incorporate this suggestion.

Once the available extent of resources is identified, the task of risk assessment and communication leads into the assessment and management of an individual’s criminogenic needs. Psychological assessments for identifying needs such as psychopathy and substance abuse have been extensively validated and are suited to individual-level decisions about
appropriate treatments. There is good empirical evidence on the individual differences that are and are not associated with violent recidivism – the criminogenic needs of the second RNR principle (for a recent summary and example, see Skeem, Winter, Kennealy, Louder, & Tatar, 2014). Evidence on the effectiveness of clinical interventions to cause reductions in criminal behavior among offenders released to the community is also developing, but practitioners can have much less confidence in the extent to which an offender will successfully respond to any given intervention (the third element of RNR). Thus, effective policy and practice should address risk communication and offender management separately. We readily acknowledge that our suggestions require knowledge not normally possessed by experts who conduct risk assessment and communication; our point in making them is to illustrate the need for effective risk communication being processed at the level of policy.

**When Do Absolute Risk and the Assessment of Individual Risk Matter?**

Some decisions might truly require an estimate of each individual’s absolute risk of recidivism. Excellent accuracy of a risk assessment tool would allow one to expect that, if an offender’s score on the tool falls in category A, and category A was associated with X% of cases engaging in recidivism, then a group of new offenders falling within category A would exhibit that same X% likelihood. An assessor could then communicate that an individual offender with a score in category A had an estimated likelihood of recidivism equal to X%.

On the other hand, even if the tool had large and replicated discriminative accuracy for the prediction of recidivism, users might still be unable to estimate expected rates (e.g., Helmus, Hanson, Thornton, Babchishin, & Harris, 2012). That is, the rate of detected recidivism as a function of actuarial scores or categories can exhibit considerable unexplained variability (i.e., poor calibration), as well as not reflecting all actual recidivism. Some forensic decisions,
though, implicitly or explicitly refer to absolute estimates, such as denying bail to an accused person who is more likely than not to recidivate, or discharging a forensic patient unless she exhibits “substantial” risk. In our experience, judges sometimes expect experts to provide unqualified estimates of absolute risk of recidivism (rather than estimates of detected recidivism within a specified opportunity). We suggest that this use of actuarial methods to render individual assessments is invalid.

Even so, there is the prospect that such actuarial data could eventually be useful in making individual absolute probability estimates. This prospect depends on future empirical work; for example, studies of the ratio of detected to total recidivism as a function of actuarial score, the shape of the recidivism hazard functions for each actuarial category, and the specific effects (independent of actuarial score) of treatment on recidivism. In this context, it could be statistically or formally logical to conceive of the probability of recidivism for an individual offender as distinct from the proportion of recidivists reported for that individual’s actuarial risk category in prior follow-up studies. This difference is the distinction between Bayesian and frequentist conceptualizations of “probability,” respectively (Harris & Rice, 2007, 2013; Mossman & Peng, 2014; Scurich & John, 2012). Until the requisite empirical work permits a Bayesian approach to individual assessment, we propose that frequentist statements are valid; e.g., “This man’s actuarial score places him at the 75th percentile among offenders previously adjudicated to have committed a violent offense, and in a subgroup of offenders of whom 62% are expected to meet the operational criteria for violent recidivism within 12 years of opportunity,” or “Based on his actuarial score at the 80th percentile, this man’s likelihood of meeting the operational definition for violent recidivism within the next five years of opportunity is estimated to be 58% (and, 19 times out of 20, between 51% and 70%).” We look forward to a
day when sufficient research has been conducted to permit valid Bayesian statements of the form, “Based on all the available empirical research on violent recidivism among offenders previously adjudicated to have committed a violent offense, together with this man’s actuarial score, he has an estimated 52% probability of committing any criminally violent act within the rest of his life.”

**Using Actuarial Scores to Communicate Individual Risk**

Our study demonstrates that individual offender risk communication can rely on actuarial assessment. We conclude that the belief that imprecision of risk estimates in previous studies limits individual risk communication is erroneous. New cases can be matched to aggregate data, permitting the effective application of policy through individual-level risk communication.

We argue that the individual’s relative risk – that is, the offender’s rank order with respect to risk in a given population – is more informative than the observed or expected proportion of cases with the same score detected as having met the criteria for recidivism. Actuarial data are most informative where they are embedded in policy, and less so when front line decisions by police, courts, or treatment providers focus on the interpretation of individual absolute probabilities of recidivism. The function of risk communication should be treated distinctly from management of an individual’s criminogenic needs and identification of effective and suitable treatments, and the role of expert risk communication should be to help decision makers effectively apply an existing policy to an individual offender.
References


Table 1

Sample Demographics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21,946</td>
<td>82</td>
</tr>
<tr>
<td>Female</td>
<td>4,716</td>
<td>18</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>620</td>
<td>2</td>
</tr>
<tr>
<td>Black</td>
<td>10,628</td>
<td>40</td>
</tr>
<tr>
<td>Native American</td>
<td>615</td>
<td>2</td>
</tr>
<tr>
<td>White</td>
<td>14,552</td>
<td>55</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Unknown</td>
<td>85</td>
<td>&lt;1</td>
</tr>
<tr>
<td><strong>PCRA Risk Category</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>9,739</td>
<td>37</td>
</tr>
<tr>
<td>Low/Moderate</td>
<td>10,764</td>
<td>40</td>
</tr>
<tr>
<td>Moderate</td>
<td>4,679</td>
<td>18</td>
</tr>
<tr>
<td>High</td>
<td>1,480</td>
<td>6</td>
</tr>
<tr>
<td><strong>Violent Reoffense</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>21,971</td>
<td>82</td>
</tr>
<tr>
<td>Yes</td>
<td>4,691</td>
<td>18</td>
</tr>
<tr>
<td><strong>Any Reoffense</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>19,150</td>
<td>72</td>
</tr>
<tr>
<td>Yes</td>
<td>7,512</td>
<td>28</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>38.50</td>
<td>11.29</td>
</tr>
<tr>
<td>Criminal History Score</td>
<td>4.28</td>
<td>2.39</td>
</tr>
<tr>
<td>Education/Employment Score</td>
<td>1.05</td>
<td>0.98</td>
</tr>
<tr>
<td>Drugs and Alcohol Score</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>Social Networks Score</td>
<td>1.11</td>
<td>0.80</td>
</tr>
<tr>
<td>Cognitions Score</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>Total Score</td>
<td>6.81</td>
<td>3.50</td>
</tr>
<tr>
<td>Follow Up Time (days)</td>
<td>569.89</td>
<td>264.98</td>
</tr>
</tbody>
</table>
Table 2

Generated Actuarial Risk Assessment Instrument (ARAI) Score Means and Predictive Accuracies (ROC Area) for Arrest for Violent Offense

<table>
<thead>
<tr>
<th>Sample</th>
<th>Measure</th>
<th>N</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>ARAI</td>
<td>26,662</td>
<td>-1.87</td>
<td>-3.87</td>
<td>0.80</td>
<td>1.05</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>SE ARAI</td>
<td>26,662</td>
<td>0.04</td>
<td>0.02</td>
<td>0.09</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>ARAI</td>
<td>13,331</td>
<td>-1.82</td>
<td>-3.79</td>
<td>0.77</td>
<td>1.02</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>SE ARAI</td>
<td>13,331</td>
<td>0.06</td>
<td>0.03</td>
<td>0.13</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>ARAI</td>
<td>1,333</td>
<td>-1.92</td>
<td>-4.06</td>
<td>0.35</td>
<td>1.07</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>SE ARAI</td>
<td>1,333</td>
<td>0.19</td>
<td>0.10</td>
<td>0.39</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>ARAI</td>
<td>500</td>
<td>-2.01</td>
<td>-4.43</td>
<td>0.54</td>
<td>1.12</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td>SE ARAI</td>
<td>500</td>
<td>0.32</td>
<td>0.16</td>
<td>0.68</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>ARAI</td>
<td>90</td>
<td>-1.88</td>
<td>-4.17</td>
<td>1.04</td>
<td>1.11</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>SE ARAI</td>
<td>90</td>
<td>0.74</td>
<td>0.43</td>
<td>1.19</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

Note. Recidivism rates in each sample range from 16% to 18%.
Table 3

*Group Failure Rates and 95% Confidence Intervals*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Measure</th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>High Risk</td>
<td>9,106</td>
<td>0.330</td>
<td>0.005</td>
<td>0.319, 0.339</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>17,556</td>
<td>0.096</td>
<td>0.022</td>
<td>0.092, 0.101</td>
</tr>
<tr>
<td>50%</td>
<td>High Risk</td>
<td>4,580</td>
<td>0.333</td>
<td>0.007</td>
<td>0.319, 0.346</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>8,751</td>
<td>0.099</td>
<td>0.003</td>
<td>0.093, 0.106</td>
</tr>
<tr>
<td>5%</td>
<td>High Risk</td>
<td>445</td>
<td>0.330</td>
<td>0.022</td>
<td>0.287, 0.374</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>888</td>
<td>0.091</td>
<td>0.010</td>
<td>0.072, 0.110</td>
</tr>
<tr>
<td>500</td>
<td>High Risk</td>
<td>169</td>
<td>0.308</td>
<td>0.036</td>
<td>0.238, 0.377</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>331</td>
<td>0.091</td>
<td>0.060</td>
<td>0.060, 0.122</td>
</tr>
<tr>
<td>90%</td>
<td>High Risk</td>
<td>30</td>
<td>0.367</td>
<td>0.088</td>
<td>0.194, 0.539</td>
</tr>
<tr>
<td></td>
<td>Low Risk</td>
<td>60</td>
<td>0.083</td>
<td>0.036</td>
<td>0.013, 0.153</td>
</tr>
</tbody>
</table>

Note. *Z* = -47.4438, *p* < .001; ^b^ *Z* = -33.356, *p* < .001; ^c^ *Z* = -10.9338, *p* < .001; ^d^ *Z* = -6.2004, *p* < .001; ^e^ *Z* = -3.3142, *p* < .001. Low Risk = cases falling into the lowest two thirds of the whole sample on the PCRA; High Risk = cases falling into the top third of sample on the PCRA.
Table 4

Summary of Individual Level Probability and 95% Confidence Intervals Overall and by Risk Group

<table>
<thead>
<tr>
<th>Sample</th>
<th>Measure</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>p(Failure)</td>
<td>26,662</td>
<td>0.176</td>
<td>0.131</td>
<td>0.020</td>
<td>0.690</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>26,662</td>
<td>0.020</td>
<td>0.014</td>
<td>0.004</td>
<td>0.719</td>
</tr>
<tr>
<td>High</td>
<td>p(Failure)</td>
<td>9,106</td>
<td>0.329</td>
<td>0.088</td>
<td>0.222</td>
<td>0.690</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>9,106</td>
<td>0.035</td>
<td>0.013</td>
<td>0.150</td>
<td>0.072</td>
</tr>
<tr>
<td>Low</td>
<td>p(Failure)</td>
<td>17,556</td>
<td>0.097</td>
<td>0.586</td>
<td>0.020</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>17,556</td>
<td>0.013</td>
<td>0.007</td>
<td>0.004</td>
<td>0.047</td>
</tr>
<tr>
<td>50%</td>
<td>p(Failure)</td>
<td>13,331</td>
<td>0.180</td>
<td>0.129</td>
<td>0.022</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>13,331</td>
<td>0.029</td>
<td>0.019</td>
<td>0.007</td>
<td>0.101</td>
</tr>
<tr>
<td>High</td>
<td>p(Failure)</td>
<td>4,580</td>
<td>0.331</td>
<td>0.085</td>
<td>0.225</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>4,580</td>
<td>0.049</td>
<td>0.018</td>
<td>0.021</td>
<td>0.101</td>
</tr>
<tr>
<td>Low</td>
<td>p(Failure)</td>
<td>8,751</td>
<td>0.100</td>
<td>0.059</td>
<td>0.022</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>8,751</td>
<td>0.018</td>
<td>0.010</td>
<td>0.007</td>
<td>0.067</td>
</tr>
<tr>
<td>5%</td>
<td>p(Failure)</td>
<td>1,333</td>
<td>0.171</td>
<td>0.128</td>
<td>0.169</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>1,333</td>
<td>0.089</td>
<td>0.061</td>
<td>0.021</td>
<td>0.282</td>
</tr>
<tr>
<td>High</td>
<td>p(Failure)</td>
<td>445</td>
<td>0.325</td>
<td>0.789</td>
<td>0.223</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>445</td>
<td>0.154</td>
<td>0.052</td>
<td>0.069</td>
<td>0.282</td>
</tr>
<tr>
<td>Low</td>
<td>p(Failure)</td>
<td>888</td>
<td>0.094</td>
<td>0.002</td>
<td>0.017</td>
<td>0.221</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>888</td>
<td>0.057</td>
<td>0.033</td>
<td>0.021</td>
<td>0.200</td>
</tr>
<tr>
<td>500</td>
<td>p(Failure)</td>
<td>500</td>
<td>0.164</td>
<td>0.132</td>
<td>0.118</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>500</td>
<td>0.172</td>
<td>0.098</td>
<td>0.030</td>
<td>0.451</td>
</tr>
<tr>
<td>High</td>
<td>p(Failure)</td>
<td>169</td>
<td>0.318</td>
<td>0.096</td>
<td>0.202</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>169</td>
<td>0.239</td>
<td>0.089</td>
<td>0.107</td>
<td>0.451</td>
</tr>
<tr>
<td>Low</td>
<td>p(Failure)</td>
<td>331</td>
<td>0.085</td>
<td>0.056</td>
<td>0.012</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>331</td>
<td>0.092</td>
<td>0.558</td>
<td>0.030</td>
<td>0.332</td>
</tr>
<tr>
<td>90</td>
<td>p(Failure)</td>
<td>90</td>
<td>0.178</td>
<td>0.155</td>
<td>0.015</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>90</td>
<td>0.342</td>
<td>0.174</td>
<td>0.135</td>
<td>0.764</td>
</tr>
<tr>
<td>High</td>
<td>p(Failure)</td>
<td>30</td>
<td>0.356</td>
<td>0.142</td>
<td>0.191</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>30</td>
<td>0.525</td>
<td>0.148</td>
<td>0.270</td>
<td>0.764</td>
</tr>
<tr>
<td>Low</td>
<td>p(Failure)</td>
<td>60</td>
<td>0.089</td>
<td>0.046</td>
<td>0.015</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>60</td>
<td>0.251</td>
<td>0.097</td>
<td>0.135</td>
<td>0.510</td>
</tr>
</tbody>
</table>

Note (first t-value difference between means: second t-value difference between 95% CI): t = -2.60e+02, p < .001; b t = -1.80e+02, p < .001; c t = -1.30e+02, p < .001, d t = -59.211, p < .001; e t = -41.883, p < .001; f t = -34.207, p < .001; g t = -22.591, p < .001; h t = -13.344, p < .001; i t = -10.533, p < .001
Table 5

*Suggested Individual Risk Communication*

<table>
<thead>
<tr>
<th>Population: <em>Insanity Acquitter</em></th>
<th>Population: <em>Domestic Violence Arrestee</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Context: <em>Less Secure Custody</em></td>
<td>Decision Context: <em>Bail</em></td>
</tr>
<tr>
<td>Instrument: <em>VRAG-R</em></td>
<td>Instrument: <em>ODARA</em></td>
</tr>
<tr>
<td>Assessment: Mr. A is currently detained in a maximum security psychiatric facility. Based on information from Mr. A’s institutional file and associated records, his score on the <em>Violence Risk Appraisal Guide—Revised</em> (VRAG-R) is in the fourth of nine categories. Compared to all such cases served by this state’s Department of Mental Health, Mr. A’s score of -8 places him at the 35th percentile in this jurisdiction. No mitigating or exacerbating factors (exclusive of violence risk) have been identified in this case.</td>
<td>Assessment: Mr. B was arrested, charged with second degree battery of his wife, and taken into custody by city police this evening. Based on police records and the arresting officers’ investigation, Mr. A received a score of 9 on the <em>Ontario Domestic Assault Risk Assessment</em> (ODARA). Compared to all such cases dealt with in this jurisdiction, this score places Mr. D at the 98th percentile for the risk of domestic violence recidivism.</td>
</tr>
<tr>
<td>Policy Application: Examination of all the VRAG-R scores for all the inpatients of his maximum security psychiatric unit, plus the scores for all the state’s insanity acquitees with special attention to those on the waiting list for hospitalization in jail, indicates that occupancy of a maximum security bed by patients of Mr. A’s assessed risk means that higher risk individuals cannot receive optimal placement. Comparison with the risk scores of other residents of the medium security unit combined with a comprehensive measurement of the medium security unit’s personnel resources available for community supervision indicate that, at current staffing levels, patients of Mr. A’s assessed risk cannot be afforded community access—to do so would deny such access to persons of lower risk (or would provide community access with inadequate supervision).</td>
<td>Policy Application: Based on the court clearance rate for persons arrested for domestic violence over the past years, it can be estimated that detaining accused persons like Mr. B would require 290 days of custody. Analysis of ODARA scores for all the domestic violence cases handled by the city police over the past three years, in conjunction with the current policy for bed allocation in the city’s detention centers, indicate that maximal prevention of domestic violence recidivism while on bail is achieved by detaining all cases scoring 8 or higher on the ODARA. Adopting a lower cutoff means higher risk cases (higher than Mr. B’s assessed risk) must eventually be released, while adopting a higher cutoff means jail spaces that could be used to protect future victims will go unused unnecessarily.</td>
</tr>
<tr>
<td>Conclusion: Mr. A’s risk is commensurate in the present system with transfer to the state’s medium security unit.</td>
<td>Conclusion: Mr. B’s risk is commensurate with opposing release on bail.</td>
</tr>
</tbody>
</table>
Figure 1. Syntax

1. logit vioArrestPA ra_r_criminalhistory_total ra_r_educationemployment_total ra_r_drugsalcohol_total ra_r_socialnetworks_total ra_r_cognitions_total
2. predict ARAI, xb
3. predict se_ARAI, stdp
4. gen lb_ARAI=ARAI - invnormal(0.975)*se_ARAI
5. gen ub_ARAI=ARAI+invnormal(0.975)*se_ARAI
6. summarize ARAI se_ARAI
7. roctab vioArrestPA ARAI, graph summary
8 egen ARAIgroup= cut(ARAI), group(3) nolabel
9. recode ARAIgroup (0/1= 0)(2=1)
10. prtest vioArrestPA, by(ARAIgroup)
11. cc vioArrestPA ARAIgroup
12. gen prob_failure=exp(ARAI)/(1+ exp(ARAI))
13. gen lb_prob=exp(lb_ARAI)/(1+ exp(lb_ARAI))
14. gen ub_prob=exp(ub_ARAI)/(1+ exp(ub_ARAI))
15. gen CI_prob=ub_prob - lb_prob
16. summarize prob_failure CI_prob
17. ttest prob_failure, by(ARAIgroup)
18. ttest CI_prob, by(ARAIgroup)
19. sort prob_failure
20. twoway rcap lb_prob ub_prob prob_failure, ///
    xtitle("Probability of Failure", margin(0 5 0 0)) ///
    ytitle("Upper & Lower Bounds", margin(0 5 0 0)) ///
    yaxis(1 2) ylabel(0.0444 "Base Rate 0.044", axis(2) angle(0) labsize(2)) ///
    xaxis(1 2) xlabel(0.0404595 "Low Risk    High Risk", axis(2) angle(0) labsize(2) labgap(-5)) ///
    ytitle("", axis(2)) ///
    xtitle("", axis(2)) ///
    yline(0.0444, lstyle(refline) lwidth(thin) lcolor(red) lpattern(dash)) ///
    xline(0.0404595, lstyle(refline) lwidth(thin) lcolor(red) lpattern(dash)) ///
    lwidth(*1)
Figure 2. Plot of predicted individual probability of failure by lower and upper confidence interval bounds for a) $N = 26,662$ b) $N = 13,331$ c) $N = 1,333$ d) $N = 500$ e) $N = 90$. 

a)

b)
Figure 3. Mean actual failure rate and 95% confidence interval by PCRA score in the whole sample ($N = 26,642$)

Note. *Cases with scores of 17 and 18 are not included in this figure because neither cell included more than 20 cases.
Figure 4. Boxplot of predicted probability of failure and PCRA score in the whole sample ($N = 26,642$)

Note. Upper and lower bounds of boxes represent interquartile range; line within boxes represents median; upper and lower bounds of arms mark the upper and lower most adjacent values, respectively; circles represent outliers.