RISK CLASSIFICATION AND JUVENILE DISPOSITIONS: WHAT IS THE STATE OF THE ART?

*Peter R. Jones, David R. Schwartz, Ira M. Schwartz, Zoran Obradovic, and Joseph Jupin**

I. INTRODUCTION

The past two decades have witnessed increased concern with the problem of serious and violent juvenile crime.[1] The juvenile justice and child welfare public policy response has varied across the states, with some adopting a get-tough stance involving increased supervision, incapacitation, and waiver, and others looking to improve the effectiveness of rehabilitative and preventive strategies.[2] Increased attention has focused on the ability of both the juvenile justice and child welfare systems to make the most appropriate decisions on placement. There is some consensus that the most appropriate decision involves an informed matching of youth needs with system responses.[3] Of course, this requires an ability to appropriately assess youth needs for control, supervision, and intervention, and the availability of a comprehensive continuum of care from which decision makers can select. Increasingly, these assessments of youth needs involve the classification of juvenile risk.[4]

The methods of risk assessment currently used in juvenile justice and child welfare cover an enormous spectrum, from highly informal "clinical" decisions to sophisticated actuarial risk assessments that reflect best practice for the field.[5]

---

* Peter R. Jones, Vice Provost of Undergraduate Studies and Professor of Criminal Justice, Temple University; David R. Schwartz, President and CEO, Q-linx Corp.; Ira M. Schwartz, President and CEO, Jewish Federation of Greater Philadelphia; Zoran Obradovic, Director, Information Science and Technology Center, Temple University; and Joseph Jupin, Information Science and Technology Center, Temple University.

1. Ilyse Grinberg et al., *Adolescents at Risk for Violence: An Initial Validation of the Life Challenges Questionnaire and Risk Assessment Index*, 40 ADOLESCENCE 573, 573-74 (2005).

2. *See* OFFICE OF JUVENILE JUSTICE & DELINQUENCY PREVENTION, COMPREHENSIVE STRATEGY FOR SERIOUS, VIOLENT, AND CHRONIC JUVENILE OFFENDERS 7 (1998) (discussing states' responses in attempting to fulfill responsibilities of juvenile justice system).

3. *See* TED PALMER, THE RE-EMERGENCE OF CORRECTIONAL INTERVENTION 35-39 (1992) (discussing studies that show effectiveness of treatments where specific programs are chosen depending on certain characteristics of an offender).

4. *See* KELLY HANNAH-MOFFAT & PAULA MAURUTTO, DEP'T OF JUSTICE CAN., YOUTH RISK/NEED ASSESSMENT: AN OVERVIEW OF ISSUES AND PRACTICES § 1.0, at 1 (2003), http://www.justice.gc.ca/en/ps/rs/ rep/2003/rr03yj-4/rr03yj-4_1.html (noting that classifying risks and needs of youths using "new actuarial and quasi-actuarial techniques" is growing trend among researchers).

5. For a review of literature dealing with methods of risk assessment currently used in juvenile justice and child welfare, see generally Tim Brennan, *Classification: An Overview of Selected Methodological Issues*, *in* PREDICTION & CLASSIFICATION: CRIMINAL JUSTICE DECISION MAKING 201

Research studies have repeatedly demonstrated that empirically derived and validated risk assessment tools can efficiently estimate the risk of reoffending or maltreatment and substantially improve upon clinical risk assessment performed by individual caseworkers or even by clinically based assessment tools. Consequently, actuarial risk assessment is increasingly supplanting clinical practices that are widely perceived as being overly subjective and discretionary. Many jurisdictions have worked with agencies such as the National Council on Crime and Delinquency ("NCCD") and the National Institute of Corrections ("NIC") to develop, introduce, or validate actuarial risk assessment tools.[6] Those that use actuarial models are generally convinced of the models' superior performance, pointing to attributes of objectivity, nonarbitrariness, efficiency, and consistency as the main benefits.[7] Unfortunately, there is only a very limited understanding of what constitutes best practice in risk assessment, and consequently, risk assessment in the field is often woefully short of what should be considered acceptable.

(Don M. Gottfredson & Michael Tonry eds., 1987) (calling for upgrade in quality of classification methods by integrating interdisciplinary methodological frameworks); Stephen D. Gottfredson, *Prediction: An Overview of Selected Methodological Issues*, *in* PREDICTION & CLASSIFICATION: CRIMINAL JUSTICE DECISION MAKING, *supra*, at 21 (discussing why actuarial approach methods are superior and describing methods of analysis); Roger Tarling & John A. Perry, *Statistical Methods in Criminological Prediction*, *in* PREDICTION IN CRIMINOLOGY 210 (David P. Farrington & Roger Tarling eds., 1985) (comparing various statistical predictive techniques); Leslie T. Wilkins, *The Politics of Prediction*, *in* PREDICTION IN CRIMINOLOGY, *supra*, at 34 (discussing superiority of statistical over clinical analyses and moral implications of predictive assessments); Christopher Baird et al., *Risk Assessment in Child Protective Services: Consensus and Actuarial Model Reliability*, 78 CHILD WELFARE 723 (1999) (comparing actuarial-based systems with consensus-based systems); Robyn M. Dawes et al., *Clinical Versus Actuarial Judgment*, 243 SCI. 1668 (1989) (finding actuarial methods superior at predicting human behavior over clinical methods); Robert D. Hoge, *Standardized Instruments for Assessing Risk and Need in Youthful Offenders*, 29 CRIM. JUST. & BEHAV. 380 (2002) (comparing three standardized risk assessment measures and explaining why they are preferred over clinical methods). *See also* Peter R. Jones, *Risk Prediction in Criminal Justice*, *in* CHOOSING CORRECTIONAL OPTIONS THAT WORK 33, 35 (Alan T. Harland ed., 1995) (stating that evidence points to success of statistical over clinical prediction methods); D.A. Andrews et al., *Does Correctional Treatment Work? A Clinically Relevant and Psychologically Informed Meta-Analysis*, 28 CRIMINOLOGY 369, 377-80 (1990) (providing examples of statistical analyses); Christopher Baird & Dennis Wagner, *The Relative Validity of Actuarial- and Consensus-Based Risk Assessment Systems*, 22 CHILD. & YOUTH SERVS. REV. 839, 867 (2000) (finding actuarial-based systems more accurate than consensus-based systems).

6. According to the NCCD website, the organization has collaborated with over seventy-five state and local jurisdictions as well as with the United States Department of Justice. More information about the NCCD can be found at The National Council on Crime and Delinquency, http://www.nccd-crc.org/nccd/n_index_main.html (last visited Jan. 2, 2007).

7. *See* Baird & Wagner, *supra* note 5, at 842-43 (discussing accuracy of actuarial risk assessment systems); Dawes et al., *supra* note 5, at 1673 (determining that actuarial method of predicting human behavior is superior to clinical method because its accuracy can be more easily monitored); Jones, *supra* note 5, at 35 (stating that evidence points to success of statistical over clinical prediction methods); David R. Schwartz et al., *Computational Intelligence Techniques for Risk Assessment and Decision Support*, 26 CHILD. & YOUTH SERVS. REV. 1081, 1082-83 (2004) (suggesting that actuarial risk assessment methods are more reliable than consensus or clinical methods).

Researchers in the field of risk assessment understand that the superior performance of actuarial over clinical risk assessment is to be earned rather than simply assumed.[8] Developing ineffective actuarial assessment tools by using poor and inappropriate methodological or statistical techniques that produce risk assessments with little or no predictive validity is quite possible. Such tools not only possess none of the implied advantages of actuarial assessment but have the danger of "laundering" poor and inappropriate assessments as scientific, objective results. Poorly developed or inappropriately employed actuarial risk assessment is damaging because it contributes to poor decision making and gives the decision maker an unwarranted optimism and faith about the validity and utility of the estimates of risk that buttress their decisions. The consequences of poor risk assessment in juvenile justice and child welfare are serious and far reaching, leading to inappropriate or inequitable decisions that significantly undermine the chances of successful treatment in any given case. Poor risk assessment can place a child at risk of abuse or neglect, justify removal from a home, or place in the community a youth at significant risk of continued delinquency.

Given these potential problems, the increased use of risk assessment tools raises concern about the issue of standards of quality and appropriateness. No clear set of minimum standards by which risk assessment tools may be evaluated exists. Studies have compared clinical and empirical approaches and demonstrated the greater effectiveness and validity of the latter type.[9] Other, more recent studies have tested the validity of existing actuarial assessment tools in different settings—sometimes finding that they work well and on occasion showing that risk assessments do not travel well.[10] There are to our knowledge, however, no studies that compare different types of actuarial risk assessment to ascertain the variability and the predictive validity of the classifications.

Absent clear guidelines about best practice or minimal standards, the superiority of actuarial over clinical approaches is assumed but rarely demonstrated. This absence of standards may explain why there is very little case law that challenges the admissibility of actuarial risk assessments as a scientific basis for decision making. The paucity of case law is somewhat surprising—aside

---

8. *See* Dawes et al., *supra* note 5, at 1673 (explaining that although actuarial methods have great benefits, "quality controls" should be used to ensure that actuarial method is appropriately working).

9. For a discussion of studies that have demonstrated the greater effectiveness of empirical approaches over clinical approaches, see generally John Monahan, *The Scientific Status of Research on Clinical and Actuarial Predictions of Violence*, *in* 2 SCIENCE IN THE LAW: SOCIAL AND BEHAVIORAL SCIENCE ISSUES § 2-2.1.1(2); (David L. Faigman et al. eds., 2002); Tarling & Perry, *supra* note 5; Dawes et al., *supra* note 5; William M. Grove & Paul E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedure: The Clinical-Statistical Controversy*, 2 PSYCHOL. PUB. POL'Y & L. 293 (1996).

10. For more recent reviews of evaluations of actuarial assessment tools, see generally Grinberg et al., *supra* note 1; Shane R. Jimerson et al., *The Santa Barbara Assets and Risks Assessment to Predict Recidivism Among Male and Female Juveniles: An Investigation of Inter-rater Reliability and Predictive Validity*, 27 EDUC. & TREATMENT CHILD. 353 (2004); Cornelis Stadtland et al., *Risk Assessment and Prediction of Violent and Sexual Recidivism in Sex Offenders: Long-term Predictive Validity of Four Risk Assessment Instruments*, 16 J. FORENSIC PSYCHIATRY & PSYCHOL. 92 (2005).

from challenges to child welfare placement decisions or juvenile delinquency detention decisions based on risk, there are potential claims of immunity from liability in cases where juveniles remain in or are removed from the community based on a risk score.

This Article addresses the lacuna in the actuarial risk assessment literature by comparing the performance of three actuarial risk assessment tools that represent three different levels of sophistication.[11] It examines the extent to which juveniles are placed in the same risk category irrespective of the method used.[12] The Article also examines the ability of each method to accurately predict reoffending.[13] For both these issues, it reviews performance across key demographic criteria such as gender and race.[14]

The study utilizes a rich and unique data source as the basis for developing the actuarial risk assessment classifications. The first assessment classifies juveniles using a long-established and frequently validated generic risk assessment method—the Wisconsin Delinquency Risk Assessment Tool. Although now very dated and obsolete, the tool is currently used either in its original or revised version in many settings across the United States.[15] The second approach utilizes a customized risk assessment tool developed specifically for the juvenile sample being studied.[16] It uses configural analysis techniques that have been employed previously in risk assessment research and offers several advantages over more traditional regression-based approaches. The third approach also develops a customized actuarial risk assessment for the juvenile sample studied, using an extremely powerful and sophisticated neural networks approach.[17] By comparing the results of each of the approaches, it is possible to establish the extent to which risk classification remains—as earlier studies suggested—stable across methodologies. If the results indicate different classifications or predictive validities across assessment types, however, it undermines the confidence currently placed in actuarial risk assessment. It also begs two additional questions. First, how should minimal standards of performance for actuarial risk assessment in the juvenile justice and child welfare

11. See *infra* Parts III.B-D for a comparison of the performance of the Wisconsin Delinquency Risk Assessment, the customized model using configural analysis, and the customized model using neural networks.

12. See *infra* Part III.E for an examination of results of the comparison in relation to juvenile risk placement.

13. See *infra* Part III.F for an examination of the accuracy of each method in predicting reoffenders.

14. See *infra* Part III.A for an overview of the demographic data used in these assessments.

15. Seventy-three agencies representing forty-four states and twenty-four localities, collectively supervising well over half that nation's probationers and parolees, reported that just over one-third of agencies used the Wisconsin instrument or an instrument based on the Wisconsin model. U.S. DEP'T OF JUSTICE, NAT'L INST. OF CORR., TOPICS IN COMMUNITY CORRECTIONS: OFFENDER ASSESSMENT 4 (2003).

16. See *infra* Part III.C for an analysis of the customized model using configural analysis.

17. See *infra* Part III.D for an analysis of the customized model using neural networks.

fields be established? Second, who will provide the impetus for establishing and maintaining such standards?

A. Risk Assessment in Juvenile Justice and Child Welfare

The development and utilization of formal risk assessment tools in juvenile justice and child welfare decision making has been a relatively recent but increasingly widespread phenomenon.[18] The types of assessment tools currently used in juvenile justice and child welfare fall into one of two broad categories—clinically based tools relying on the professional judgment and expertise of clinicians, and empirically derived actuarial tools that identify statistically significant predictors of the risk outcome and create a scoring mechanism that allows for classification of the youth.

Clinical assessment tools tend to perform poorly in comparison to empirical tools largely because they allow for widely differential selection and weighting of information about the subject.[19] Factors that intuitively and theoretically appear to be associated with risk may in fact have little or no predictive validity. Similarly, two or more factors assumed to have an independent impact on risk may be highly intercorrelated so that their combined inclusion effectively biases the risk assessment in one particular direction. From a purely methodological perspective, therefore, one reason why the research repeatedly demonstrates the greater predictive validity of empirical tools is that hypothesized predictors of risk have to be both statistically significant and theoretically independent.[20]

Recognition of the superiority of actuarial approaches to risk assessment has prompted state and local jurisdictions, as well as many public and private agencies, to integrate standardized risk assessment into the decision-making stages of juvenile justice and child welfare processes.[21] This process has been aided and facilitated by research and technical assistance programs involving agencies such as the Office of Juvenile Justice and Delinquency Prevention ("OJJDP"), the NIC, and the NCCD. One example of the process is the widely used Wisconsin Delinquency Risk Assessment Tool. This actuarial-based risk assessment tool, developed by the Wisconsin Department of Juvenile Corrections with assistance from NCCD in the late 1970s, has been revised and

---

18. *See* Baird & Wagner, *supra* note 5, at 840 (stating that, as of 1996, at least thirty-eight out of fifty-four U.S. states and territories used formal risk assessment tools; twenty-six implemented tools after 1987).

19. *Id.* at 842; Dawes et al., *supra* note 5, at 1671-72; Grove & Meehl, *supra* note 9, at 293-95.

20. *See* PAUL E. MEEHL, CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE 3-4, 136 (1954) (labeling clinical methods as hypothesizing predictors of risk as opposed to actuarial methods that use statistics and concluding that use of actuarial method is "unavoidable" because clinical methods are untrustworthy); Baird et al., *supra* note 5, at 743 (finding that actuarial methods are more consistent and reliable than consensus, or clinical, methods); Dawes et al., *supra* note 5, at 1671 (finding that actuarial methods are more accurate than clinical judgment methods because clinical methods rely on independent judgments that can dramatically fluctuate and are not statistically significant).

21. Baird & Wagner, *supra* note 5, at 839-40.

revalidated.[22] One version or other of the assessment tool is found in current use in many juvenile justice agencies nationally. The original Wisconsin Risk Assessment Tool attempts to measure the risk of recidivism through an objective assessment of eight factors determined to be statistically significant predictors of reoffending.[23] Based on the total scores across all items, a juvenile is classified into a risk category. The original and the revised Wisconsin Risk Assessment Tool are examples of established risk tools that are adopted or adapted for use nationwide. Other well-known and similar types of assessment include the Washington State Juvenile Court Risk Assessment, Orange County Risk Assessment, Missouri Risk Assessment Scale, Global Risk Assessment Device, and from Canada, the Youth Level of Service/Case Management Inventory ("YLS/CMI"). A body of research validates these assessment tools[24]—particularly in Canadian applications[25]—but the fact remains that in the majority of settings where these tools are used, their predictive validity is assumed rather than demonstrated.

The growth in use of actuarial risk assessment is fueled by the fact that it is considered "good practice" in juvenile justice and child welfare. OJJDP's Comprehensive Strategy for Serious, Violent, and Chronic Juvenile Offenders states that the use of risk and needs assessment tools is a critical component of an effective juvenile justice system.[26] Unfortunately, the statements supporting the use of empirical risk assessment are broadly philosophical rather than specifically methodological, and they do not make clear that the potential benefits and advantages of the approach are contingent on the integrity and quality of both the development and implementation of the tool. Most likely, many jurisdictions as well as juvenile justice and child welfare programs are using forms of actuarial risk assessment that are simply not valid for their setting or population. In these cases, it is a mistake to assume any "good practice"

---

22. *See* Jones, *supra* note 5, at 43 (discussing programs developed through assistance from the NIC); Baird & Wagner, *supra* note 5, at 842 (stating that NCCD has conducted actuarial research in at least seven states since 1989). *See generally* OFFICE OF JUVENILE JUSTICE & DELINQUENCY PREVENTION, *supra* note 2 (introducing its strategy for dealing with varying degrees of juvenile offenders).

23. See *infra* Part III.B for a list of the eight factors.

24. *See generally* Baird & Wagner, *supra* note 5 (comparing Washington Risk Assessment Matrix with California Family Assessment Factor Analysis and Michigan Structured Decision Making System's Family Risk Assessment of Abuse and Neglect); Peter R. Jones et al., *Identifying Chronic Juvenile Offenders*, 18 JUST. Q. 479 (2001) (describing study identifying chronic juvenile offenders in Orange County, California).

25. *See* Hoge, *supra* note 5, at 390-91 (explaining YLS/CMI method). *See generally* HANNAH-MOFFAT & MAURUTTO, *supra* note 4, § 4.0, at 9-13 (reviewing and critiquing risk assessment procedures in Canada, including YLS/CMI); David P. Cole & Glenn Angus, *Using Pre-Sentence Reports to Evaluate and Respond to Risk*, 47 CRIM. L.Q. 302 (2003) (describing benefits of presentence reports in risk assessment); Ivan Zinger, *Actuarial Risk Assessment and Human Rights: A Commentary*, 46 CANADIAN J. CRIMINOLOGY & CRIM. JUST. 607, 609 (2004) (commenting on accuracy of third-generation actuarial risk assessment techniques, particularly Canada's LSI-R, in predicting recidivism).

26. OFFICE OF JUVENILE JUSTICE & DELINQUENCY PREVENTION, *supra* note 2, at 13-14.

benefits accrue, and it is possible that serious errors of classification are being made.

B. *Evaluating Risk Assessments*

The existing methodological literature on risk assessment tends to divide into three categories. There are studies that compare clinical and empirical risk instruments in terms of their predictive validity, studies that validate existing actuarial risk assessments, and studies that compare the performance of different forms of empirically based risk assessment.[27]

Almost all studies in the first category reach the conclusion that actuarial risk outperforms clinical assessment tools.[28] A review by Professors Don Gottfredson and Stephen Gottfredson of such comparisons concluded that "[i]n virtually every decision-making situation for which the issue has been studied, it has been found that statistically developed predictive devices outperform human judgments."[29]

In addition to outperforming clinical decision making, actuarial assessment tools have the distinct advantage of providing an explicit basis for the risk decision.[30] Unlike clinical risk assessment, actuarial models do not require one to accept classification decisions on faith.[31] Professors Eileen Gambrill and Aron Shlonsky warn against over confidence in actuarial models within the field of child welfare, pointing out that "actuarial models are rarely able to predict reabuse at acceptable levels of sensitivity (correctly classifying those children who will be reabused)."[32] They further note that "[a]lthough actuarial models

---

27. For examples of each type of study, see generally D.A. Andrews et al., *The Recent Past and Near Future of Risk and/or Need Assessment*, 52 CRIME & DELINQ. 7 (2006) (comparing performance of different types of empirical methods); Baird & Wagner, *supra* note 5 (comparing predictive validity of actuarial- and consensus-based or clinical methods); Grinberg et al., *supra* note 1 (validating existing actuarial method).

28. *See* Stephen D. Gottfredson & Don M. Gottfredson, *Accuracy of Prediction Models*, *in* 2 CRIMINAL CAREERS AND "CAREER CRIMINALS" 212, 247 (Alfred Blumstein et al. eds., 1986) (explaining that various studies have determined statistical predictive devices are more accurate than clinical methods); Jones, *supra* note 5, at 35-36 (suggesting that all available evidence regards statistical predictive devices over clinical devices); Baird & Wagner, *supra* note 5, at 842-43 (noting that research has shown superiority of actuarial method over clinical method); Dawes et al., *supra* note 5, at 1673 (reviewing research that propels theory that actuarial assessment is as accurate or more accurate at prediction than clinical assessment).

29. Gottfredson & Gottfredson, *supra* note 28, at 247.

30. *See* Monahan, *supra* note 9, § 2-2.1.1(2), at 102-03 (discussing actuarial study indicating that substance abuse, prior arrests for violent crimes, and young age were significant predictors of future violence); Tarling & Perry, *supra* note 5, at 212-13 (explaining that variables used to predict reconviction can remain relatively constant); Dawes et al., *supra* note 5, at 1671 (stating that actuarial tools are effective in identifying significant variables that contributed to a decision); Grove & Meehl, *supra* note 9, at 317 (stating that actuarial data allows most influential variables to be identified).

31. *See* Gottfredson & Gottfredson, *supra* note 28, at 247-48 (citing various studies in which results from actuarial devices were superior to individual human's judgment).

32. Eileen Gambrill & Aron Shlonsky, *Risk Assessment in Context*, 22 CHILD. & YOUTH SERVS. REV. 813, 825 (2000).

tend to be the best predictors of future maltreatment, they are far from perfect."[33]

The research literature on validation of existing assessment tools is rapidly expanding.[34] Unfortunately, government or contract researchers, rather than independent researchers with no stake in the performance of the instruments, produced many of these studies.[35]

The literature on the comparative performance of different actuarial approaches suggests that the type of method makes little difference.[36] There are several reasons why this is possible. The first is that theory in the field of juvenile justice and child welfare is still developing when compared with applications of prediction studies in fields such as medicine, econometrics, and engineering. The second is the constraint of poor data. Existing limitations on the range of available measures added to problems of reliability and validity of the data and therefore limited the ability of more sophisticated statistical approaches to achieve their potential. Professor Peter Jones argues that "[w]ithout better and different data we simply cannot improve on the basic analytic approaches of the past."[37] Similarly, Professor Stephen Gottfredson warns that limited and generally poor quality data combined with the highly random nature of delinquent behavior ensures that prediction research will rarely explain more than fifteen to twenty percent of the outcome variance and may never do better than thirty percent.[38] Professors Peter Schmidt and Ann Witte concur and caution against overly optimistic goals for prediction studies in the field of delinquency, pointing to the fact that even in disciplines with well-developed, specific theories and relatively accurate data, prediction instruments struggle to explain more than half the variation in the outcome measure.[39] A third reason

33. *Id.* at 826.

34. For a review of such literature, see JESS MCDONALD & JOHN GOAD, ILL. DEP'T OF CHILDREN & FAMILY SERVS., ILLINOIS CHILD ENDANGERMENT RISK ASSESSMENT PROTOCOL 10 (2002), *available at* http://www.state.il.us/dcfs/docs/cerap2002.pdf (concluding that Child Endangerment Risk Assessment Protocol reduced recurrence rates of maltreatment for at-risk children); Grinberg et al., *supra* note 1, at 598 (finding "initial support" for validity of The Life-Challenges Questionnaire and Risk Assessment Index as a risk assessment measure); Jimerson et al., *supra* note 10, at 370 (concluding that Santa Barbara Assets and Risks Assessment provides enhanced assessment of juveniles for both males and females); Stadtland et al., *supra* note 10, at 105-06 (finding that various assessment tools efficiently predicted violent nonsexual, noncontact sexual, and contact sexual recidivism).

35. *See, e.g.*, HANNAH-MOFFAT & MAURUTTO, *supra* note 4, § 4.0, at 9-10, § 6.1, at 27 (noting that government and contract researchers conducted many of these studies and suggesting that independent researchers should be conducting these studies to ensure lack of bias).

36. Tarling & Perry, *supra* note 5, at 211 ("On the basis of her results and those of other prediction studies that she reviewed, Simon argued that no method was greatly superior to any other and concluded that 'from this examination of statistical methods for combining data . . . in practice all of them work about equally well.'" (quoting FRANCES H. SIMON, HOME OFFICE RESEARCH STUDY NO. 7, PREDICTION METHODS IN CRIMINOLOGY: INCLUDING A PREDICTION STUDY OF YOUNG MEN ON PROBATION 154 (1971)).

37. Jones, *supra* note 5, at 43-44.

38. Gottfredson, *supra* note 5, at 24-25.

39. PETER SCHMIDT & ANN DRYDEN WITTE, PREDICTING RECIDIVISM USING SURVIVAL

why researchers expect little variation across statistical approaches is that the lack of differentiation is more apparent during the validation than the construction of a risk instrument so that apparent gains in the creation of an assessment tool will disappear upon testing. Professor Roger Tarling and Analyst John Perry warned of this when they completed their review of seven different statistical approaches by concluding that "no method is consistently better than any other in validation samples."[40] A fourth reason is that the methods being compared are often basic, regression-based techniques that share similar assumptions about the data.[41]

Most of the comparative studies use traditional, statistical methodologies, and it is rare to find contributions from nontraditional disciplines such as engineering or bioinformatics.[42] For this reason, it is premature to assume that the type of actuarial assessment tool is of little importance. The present study employs three different actuarial assessment tools that are founded upon widely different mathematical platforms and vary enormously in their range of sophistication and power. If these approaches yield divergent results for the same sample of juveniles—in terms of both classification and predictive validity—then we are forced beyond the generic label of actuarial assessment to determine the quality of the tool and the results it yields.

II. DEVELOPING RISK ASSESSMENTS

Before examining in more detail the three approaches compared in the present study, it is worth outlining some methodological issues that affect the quality of any actuarial risk assessment study. These issues include sample size, quality and range of available data, and the statistical basis of the assessment tool itself.

A. Sample Size

Sample size is important for several reasons. It affects the ability of the assessment tool to identify important patterns among subgroups defined by factors such as race and gender. For example, if the etiology of offending for females is different from that of males, then small delinquent samples that are predominantly male will not yield enough female cases for their specific patterns of offending to be recognized. The resulting assessment tool will likely perform poorly for females.

MODELS 13-14 (1988).

40. David P. Farrington & Roger Tarling, *Criminological Prediction: The Way Forward*, *in* PREDICTION IN CRIMINOLOGY, *supra* note 5, at 264.

41. *See* Gottfredson, *supra* note 5, at 26-27 (explaining how statistical methods sometimes assume that results from one sample can explain results from similar sample).

42. For examples of studies of human behavior from nontraditional disciplines, see generally ALBERT NIGRIN, NEURAL NETWORKS FOR PATTERN RECOGNITION (1993) (studying frameworks of pattern recognition by neural networks); T.A. Arentze et al., *Using Decision Tree Induction Systems for Modeling Space-Time Behavior*, 32 GEOGRAPHICAL ANALYSIS 330 (2000) (discussing use of decision tree induction systems to predict an individual's activity and travel choices).

Roger Tarling and John Perry cautioned that most of the variation in the performance of different statistical approaches to assessment disappears upon validation.[43] Therefore, it is critical that any risk tool development includes validation—preferably on an independent, randomly selected sample. This requires that the available sample is divided into construction (training) and validation (testing) subsamples, with the actual proportions determined by the investigator. For small samples, this is a difficult trade-off because the benefits of validation are potentially outweighed by the decreased ability to identify significant predictors from the reduced sample.

Finally, the lower the base rate for the outcome—for example, violent offending and sexual offending can have base rates of less than five percent even for delinquents—the larger the sample necessary to develop reliable risk assessments. In a sample of five hundred delinquents, there may be only twenty-five individuals who commit a sexual offense, and this inevitably restricts the ability of any assessment device to reach its full potential. There is no required sample size for the development of a risk assessment model, but for the reasons outlined here, samples of less than five hundred should be treated with extreme caution.

B. Quality of Data

Empirical risk models require the analysis of numeric values. These numeric values signify objective measures such as age, gender, race, or the number of prior arrests. They also signify more subjective measures that reflect second-party assessments of behavior or personality. Reviews of risk assessments identify the use of highly subjective measures such as "could make better use of time," "[n]on-rewarding parental relations," "inconsistent parenting," "poor social skills," and "inadequate supervision."[44] For such measures, it is difficult if not impossible to determine the consistency of interpretation and measurement. Inevitably, measures vary along a spectrum from the value free to those that involve "substantial speculation and morally laden subjective assessments."[45] The reliability and validity of such data are dependent on quality of training by those making the assessments.

Data quality is affected significantly by the prevalence and source of missing data. At its simplest, the problem of missing data means values are unavailable for varying proportions of cases on different measures. A more complex issue involves the process that gives rise to the missing data. To the extent that the underlying process is biased in some way, and that bias is associated with specific attributes of the sample, then the resulting model will be inappropriately shaped by the missing data. An example will illustrate. In juvenile justice files, it is common to find information on whether a female has a child of her own. Comparison of different sources of such information—official files, staff reports,

43. Tarling & Perry, *supra* note 5, at 227.

44. HANNAH-MOFFAT & MAURUTTO, *supra* note 4, § 4.1, at 12-13.

45. *Id.*

and self-report—will often be highly correlated. The same is not true for males. Analyses for the data set used in the present study show that reports of delinquent males with children of their own increase quite significantly when one compares official file, staff report, and self-report sources. The availability of this information increases respectively with each of these sources.

For the investigator, there are several options to dealing with the missing values problem. One may remove variables with excessive proportions of missing data—the actual threshold usually being determined by the investigator. An alternative is to adopt one of several forms of missing value substitution—a process that can range from simplistic mean substitution to sophisticated missing data estimation algorithms that learn from patterns of data among the remaining good data.[46]

Despite its significance, the issue of missing data is treated as almost inconsequential in many studies.[47] It represents one of the most serious threats to the validity of any risk assessment, however, and represents an essential indicator of quality control. Recognition of the problem is dependent on the expertise of the researcher. In inexperienced hands, it is possible for an investigator to believe they are studying hundreds of cases with dozens of potential measures when in fact the effective sample is a small portion of the whole and several important predictors are missed through the lack of information.

C. Theoretical Understanding: How Should We Select and Measure the Predictors?

Scholarly interest in risk assessment within justice and welfare disciplines has a long history.[48] In the 1920s, Ernest Burgess completed his famous prediction study on parole decision making,[49] and subsequent work by Lloyd Ohlin[50] in the 1950s, as well as Don Gottfredson, Leslie Wilkins, and Peter

---

46. RODERICK J.A. LITTLE & DONALD B. RUBIN, STATISTICAL ANALYSIS WITH MISSING DATA 60-61 (1987).

47. *See generally* DEAN J. CHAMPION, MEASURING OFFENDER RISK: A CRIMINAL JUSTICE SOURCEBOOK (1994) (reviewing thoroughly a range of issues pertaining to development and application of risk assessment within criminal and juvenile justice without once referring to problems of missing data). *But see* DON M. GOTTFREDSON & HOWARD M. SNYDER, OFFICE OF JUVENILE JUSTICE & DELINQUENCY PREVENTION, NATIONAL CENTER FOR JUVENILE JUSTICE REPORT, THE MATHEMATICS OF RISK CLASSIFICATION: CHANGING DATA INTO VALID INSTRUMENTS FOR JUVENILE COURTS 6 (2005), *available at* http://www.ncjrs.gov/ html/ojjdp/209158 (dealing specifically with the mathematical aspects of risk assessment and stating that "missing data cause great problems").

48. *See* Jones, *supra* note 5, at 36 (detailing various studies attempting to predict criminality since the 1920s).

49. *See generally* Ernest W. Burgess, *Factors Determining Success or Failure on Parole*, *in* THE WORKINGS OF THE INDETERMINATE-SENTENCE LAW AND THE PAROLE SYSTEM IN ILLINOIS 205 (Andrew A. Bruce et al. eds., 1928) (discussing factors that may help determine whether an individual is more or less likely to violate his or her parole).

50. *See generally* LLOYD E. OHLIN, SELECTION FOR PAROLE: A MANUAL OF PAROLE PREDICTION (1951) (discussing factors that may help determine whether an individual is likely to

Hoffman in the late 1970s, further refined risk classification through work on "salient factor scores" and "parole guidelines."[51] The work of Steven and Eleanor Glueck is one of the best-known prediction studies in the history of delinquency[52] and was so influential that President Richard Nixon was advised that the Gluecks' Social Prediction Table enabled "9 out of 10 delinquents [to be] correctly identified at the age of 6."[53]

Review of the theoretical literature in the justice and welfare disciplines generates a very wide array of potential predictors. Unfortunately, most of these are not available among the usually constrained range of measures routinely maintained by programs or agencies. The list of potential predictors that can be found in a juvenile court or child welfare agency includes such factors as gender, race, age, family relations, living arrangements, degree of family involvement, substance abuse, and past offending record.[54]

Less commonly found are important predictors such as early problem behavior,[55] parenting and family management techniques,[56] family disruption and family size or structure,[57] parental or sibling criminality or delinquency,[58] delinquent peers,[59] alcohol abuse,[60] personality,[61] self-esteem,[62] attitudes,[63]

---

violate his or her parole and how to utilize these factors).

51. DON M. GOTTFREDSON ET AL., GUIDELINES FOR PAROLE & SENTENCING 17-36 (1978).

52. David P. Farrington & Roger Tarling, *Criminological Prediction: An Introduction*, *in* PREDICTION IN CRIMINOLOGY, *supra* note 5, at 7 (citing STEVEN GLUECK & ELEANOR T. GLUECK, UNRAVELING JUVENILE DELINQUENCY (1950)).

53. Farrington & Tarling, *supra* note 52, at 8.

54. For a discussion of potential predictors typically used in a juvenile court or welfare agency, see generally GWEN A. KUTZ & LOUIS E. MOORE, THE "8% PROBLEM": CHRONIC JUVENILE OFFENDER RECIDIVISM (1994); Charles W. Dean et al., *Criminal Propensities, Discrete Groups of Offenders, and Persistence in Crime*, 34 CRIMINOLOGY 547 (1996); Christy A. Visher et al., *Predicting the Recidivism of Serious Youthful Offenders Using Survival Models*, 29 CRIMINOLOGY 329 (1991).

55. *See* Sheila Mitchell & Peter Rosa, *Boyhood Behaviour Problems as Precursors of Criminality: A Fifteen-Year Follow-Up Study*, 22 J. CHILD PSYCHOL. & PSYCHIATRY & ALLIED DISCIPLINES 19, 33 (1981) (finding that a child is more likely to be convicted of crime later in life if he or she has previously exhibited "anti-social behavior (stealing, lying, destructiveness and wandering from home)").

56. *See* DAVID RILEY & MARGARET SHAW, HOME OFFICE RESEARCH STUDY NO. 83, PARENTAL SUPERVISION AND JUVENILE DELINQUENCY 1, 2 (1985) (noting that child offenders "tend to come from large families, to have parents with a criminal record, to have poor or erratic discipline at home, conflict between parents, and poor supervision by their parents").

57. *See* DONALD J. WEST, DELINQUENCY: ITS ROOTS, CAREERS & PROSPECTS 29-30 (1982) (finding that juveniles are twice as likely to become delinquent if their parents have a low income, come from a large family, have a criminal record, or are "unsatisfactory" with "below-average intelligence").

58. *See* DAVID P. FARRINGTON, NAT'L INST. OF JUSTICE, FURTHER ANALYSES OF A LONGITUDINAL SURVEY OF CRIME AND DELINQUENCY: FINAL REPORT TO THE NATIONAL INSTITUTE OF JUSTICE 68-80 (1983) (discussing correlation between troubled youths and families with convicted parents).

59. Albert I. Reiss, Jr., *Co-Offender Influences on Criminal Careers*, *in* 2 CRIMINAL CAREERS AND "CAREER CRIMINALS," *supra* note 28, at 122-23.

60. *See* MICHAEL R. GOTTFREDSON, HOME OFFICE RESEARCH STUDY NO. 81, VICTIMS OF CRIME: THE DIMENSIONS OF RISK 14-15 (1984) (finding that self-reported drinking correlates with risk of personal victimization).

family bonding,[64] and school bonding.[65] Most of the available measures pertain to the individual rather than the family or the community. This unavailability is being seen as an increasing limitation as the literature continues to demonstrate the importance of predictors incorporating familial, peer, and environmental measures such as household violence,[66] gang membership,[67] and neighborhood.[68]

Finally, almost all the measures used in risk assessment are "static"—gender, race, number of prior arrests, etc.—and cannot change or decrease over time.[69] The recent risk assessment literature points to the importance of "dynamic" risk factors that indicate temporal trends in behavior and personality rather than snapshot measures frozen in time.[70] For example, a juvenile with poor school grades may behave differently if poor grades are the norm rather than a very recent development.

D. Methods of Developing a Risk Assessment Tool

The goal of most actuarial prediction studies is to select and combine a small number of predictors into a parsimonious model that will maximize predictive efficiency in terms of validity, cost, and utility. As noted above, that type of statistical approach has been found to have limited impact on overall

---

61. *See* WILLIAM MCCORD & JOAN MCCORD, THE PSYCHOPATH: AN ESSAY ON THE CRIMINAL MIND 8-17 (1964) (discussing personality traits that can characterize individual as psychopath, including being asocial, having uncontrollable desires, being impulsive, being aggressive, feeling little guilt, and not having capacity to develop lasting, loving relationships).

62. *See* David Thornton et al., *Pretreatment Self-Esteem and Posttreatment Sexual Recidivism*, 48 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 587, 587 (2004) (finding a correlation between low self-esteem and higher sexual recidivism).

63. *See* Andrews et al., *supra* note 27, at 7 (stating that treatment strategies for juvenile criminals should be matched to motivation of participants).

64. *See* Stephen A. Cernkovich & Peggy C. Giordano, *Family Relationships and Delinquency*, 25 CRIMINOLOGY 295, 297 (1987) (discussing that children in homes in which family members get along well with one another are less likely to have delinquency problems, even if parents are separated or divorced).

65. *See* Steven A. Cernkovich & Peggy C. Giordano, *School Bonding, Race and Delinquency*, 30 CRIMINOLOGY 261, 261 (1992) (noting that juvenile delinquency increases when juvenile's bond to school is low).

66. *See* Richard Dembo et al., *The Role of Family Factors, Physical Abuse and Sexual Victimization Experiences in High-Risk Youths' Alcohol and Other Drug Use and Delinquency: A Longitudinal Model*, 7 VIOLENCE & VICTIMS 245, 245 (1992) (finding significant relationship between physical abuse or sexual victimization and juvenile delinquency).

67. *See* Pamela K. Lattimore et al., *Risk of Death Among Serious Young Offenders*, 34 J. RES. CRIME & DELINQ. 187, 188 (1997) (noting that gang involvement, a factor that leads to victimization, may also lead to criminality).

68. Jones et al., *supra* note 24, at 480; Robert J. Sampson et al., *Assessing "Neighborhood Effects": Social Processes and New Directions in Research*, 28 ANN. REV. SOC. 443, 443-45 (2002).

69. *See* HANNAH-MOFFAT & MAURUTTO, *supra* note 4, § 2.0, at 3 (noting that risk assessments traditionally receiving the most attention stress "static historic factors, such as age, number and type of convictions, sexual offending, and relationship to victim").

70. *See id.* at 2-4 (discussing distinctiveness of "third-generation risk assessment" that takes into account both static and dynamic risk factors).

predictive validity. Francis Simon compared seven different statistical techniques and concluded that all worked equally well.[71] Subsequent criticism of Simon's analysis prompted Roger Tarling and John Perry to extend the examination of alternate methods by submitting the same data to analysis by two more sophisticated and appropriate prediction methods—Automatic Interaction Detection and Logistic Regression.[72] They concluded that both methods performed equally well but not significantly better than Simon's other approaches.[73] Steven Gottfredson and Michael Gottfredson completed a similar comparative study involving several analytic methods—linear additive models (OLS Multiple Regression and unweighted Burgess Method), clustering models (Predictive Attribute Analysis and Association Analysis), and multidimensional contingency table analysis.[74] They too concluded that "simpler and more easily understood and implemented statistical prediction devices may work as well as those based on more complex techniques."[75]

In summary, the juvenile justice and child welfare fields are experiencing significant development in the methods of actuarial risk prediction. There is slow but steady progress in the theoretical understanding of the behavior to be predicted—especially as the realm of measures expands to include the family, peer groups, and the community—but very limited development in the range or quality of available data. These weaknesses suggest that confidence in actuarial risk assessment is perhaps exceeding what is reasonable. We are in danger of creating a seriously misplaced and misleading aura of utility, accuracy, and validity among actuarial risk assessments in delinquency and child welfare. An international literature search conducted by Kelly Hannah-Moffat and Paula Maurutto revealed few published peer-reviewed academic studies of youth risk assessments, especially by investigators with no vested interest in the promotion of the tools.[76]

After several decades of research on actuarial risk assessment, there is need for the development of minimum standards and external regulation to ensure that these standards are met. To achieve the potential advantages of actuarial risk assessment we must do more than pay lip service to the process—we must ensure its validity and its integrity. As Leslie Wilkins reminds us, "[t]he ultimate

71. SIMON, *supra* note 36, at 154.

72. Simon used Predictive Attribute Analysis and OLS Multiple Regression in her work. *Id.* at 80, 91.

73. Tarling & Perry, *supra* note 5, at 214-26.

74. *See generally* STEVEN D. GOTTFREDSON & MICHAEL D. GOTTFREDSON, SCREENING FOR RISK: A COMPARISON OF METHODS 62-63 (1979) (determining analytic methods examined produced same degree of predictive efficiency).

75. Steven D. Gottfredson & Don M. Gottfredson, *Screening for Risk Among Parolees: Policy, Practice and Method*, *in* PREDICTION IN CRIMINOLOGY, *supra* note 5, at 54, 75; *see also* DON M. GOTTFREDSON ET AL., THE UTILIZATION OF EXPERIENCE IN PAROLE DECISION-MAKING: SUMMARY REPORT 18 (1974) (suggesting that less powerful and more simple prediction techniques may produce better predictive validity than more complex techniques given insufficient data).

76. HANNAH-MOFFAT & MAURUTTO, *supra* note 4, §4.0, at 10.

test of predictive methods is . . . neither the scientific nor the statistical nature of the exercises, but their honesty, rigor and moral underpinnings."[77]

III. CURRENT STUDY: COMPARING RISK ASSESSMENT TOOLS

The present study employs three assessment tools, each based on a different methodological approach. The first—the Wisconsin Risk Assessment Tool—was developed using multivariate logistic regression techniques.[78] The second approach employs configural analysis, an improved version of the automatic interaction detector method utilized by Tarling and Perry.[79] The third approach involves neural network analysis.

The comparative analysis poses a number of important questions. First, has the field progressed beyond the studies of Francis Simon,[80] Steven Gottfredson and Don Gottfredson,[81] and Roger Tarling and John Perry[82] to the point where it is possible to identify significant differences in the predictive validity of different statistical approaches? Second, do these differences translate into differential classifications of juveniles based on the type of assessment tool being used? For example, will a juvenile classified low risk by one model be classified high risk by another? Depending on these results, the analysis poses several secondary questions. What are appropriate minimal standards for actuarial risk classification? Where will the impetus for change be found? Ultimately, the study seeks to examine the extent to which actuarial tools deserve the methodological "high ground" they currently inhabit.

A. Data

The study utilizes data from the *ProDES* database of juvenile delinquents processed by Family Court in Philadelphia from 2000 to 2002.[83] The *ProDES* database comprises data on all juveniles whose Family Court disposition involved more than regular probation—either probation with the condition of attending a treatment program or placement in a state juvenile correctional facility. The dataset provides a wide array of measures including official records, staff assessments, and self-reported data. For this study, the available measures include the eight variables that comprised the original Wisconsin Risk

77. Wilkins, *supra* note 5, at 50.

78. S. CHRISTOPHER BAIRD ET AL., WIS. DIV. OF CORR., THE WISCONSIN CASE CLASSIFICATION/STAFF DEPLOYMENT PROJECT: A TWO YEAR FOLLOW-UP REPORT 9 (1979).

79. Tarling & Perry, *supra* note 5, at 215-16.

80. *See* SIMON, *supra* note 36, at 72-134 (comparing several statistical techniques for their predictability).

81. *See* Gottfredson & Gottfredson, *supra* note 75, at 54 (comparing statistical efficiency of five methods for predicting parole risk).

82. *See* Tarling & Perry, *supra* note 5, at 212-26 (studying whether social and criminal histories of individuals were predictive of future criminal convictions using the Automatic Interaction Detection and Logistic Regression methods of analysis).

83. Peter R. Jones et al., *Evaluating Services to Delinquent Youth in Philadelphia: The* ProDES *Information System*, 59 J. PA. ASS'N PROBATION, PAROLE & CORRECTIONS 10, 10-13 (1999).

Assessment Tool and a wide range of additional measures pertaining to the juvenile, his or her family, his or her school, and his or her peers. All cases were screened for missing data on the Wisconsin Risk Assessment Tool measures and all cases missing these risk scores were removed.[84] The effective sample comprises 8,239 juveniles that are primarily male, African American, and between ages fourteen and seventeen (see Table 1). Additional attributes show almost forty percent of the juveniles in the sample have dispositions for personal offenses, twenty percent injured a victim, and twenty-three percent have two or more prior arrests. Approximately thirty percent of juveniles have a history of chronic drug abuse, and over twenty-two percent have IQ scores that classify them as "borderline" or "mentally deficient."

84. Removal of cases with missing data is not a problem in the present study since its function is solely to compare different methods of assessment for the same juveniles rather than build a risk assessment tool for use in the field.

TABLE 1
CHARACTERISTICS OF SAMPLE

VARIABLE	% IN SAMPLE	VARIABLE	% IN SAMPLE
Gender		**Type of Offense**	
Males	87.6	*Drugs*	24.6
Females	12.4	*Personal (e.g., assault)*	38.5
Race		*Property (e.g., theft)*	28.9
White	13.1	*Weapons*	6.0
Black	70.9	*Other*	1.9
Hispanic	14.2	**Injury to Victim**	
Asian	1.5	*No*	80.3
Other	0.3	*Yes*	19.7
Age			
10	0.2	**IQ**	
11	1.1	*Mentally Deficient*	6.7
12	2.7	*Borderline*	15.5
13	6.7	*Low Average*	16.1
14	13.1	*Average*	21.6
15	20.0	*High Average*	4.6
16	24.6	*Missing*	35.5
17	22.1	**Prior Arrests**	
18	7.9	*None*	51.2
19	1.3	*1*	25.8
20	0.2	*2*	12.1
		3 or more	10.9
History of Drug Abuse			
None	47.2		
Occasional	22.6		
Chronic	30.2		

All juveniles in the *ProDES* system are monitored during their program placement (involving a variable program length of stay—sample average thirty-two weeks) and for six months following program discharge. The outcome measure for the study is juvenile rearrests leading to new court petitions during the combined in-program and post-program periods. In total, almost thirty percent of the sample reoffended, though the figure varies by gender—31.9% for males and 12.7% for females.

B. *Modeling Risk I: The Wisconsin Risk Assessment Model*

The Wisconsin Risk Assessment Model was developed in the late 1970s and remains widely used in its original form, its revised form, or in some site-specific adaptation that reflects local needs and data availability. The assessment tool yields risk scores based on information derived from official records and interviews conducted by case management staff. The assessment should be completed within thirty days of a juvenile's assignment to a placement, and the *ProDES* system required this of all programs.

Risk classification is based on a total risk score calculated by simple summation of the eight risk items.[85] Though the risk assessment tool has undergone revision, the one employed here comprises the following items:

1. Age at first adjudication,
2. Prior delinquent behavior,
3. Prior institutional commitments of thirty-plus days,
4. Drug or chemical abuse,
5. Alcohol abuse,
6. Quality of parental control,
7. Evidence of school disciplinary problems, and
8. Quality of peer relationships.

The attributes of each item are weighted to reflect the magnitude of their association with recidivism. For example, "quality of parental control" involves a risk score of zero for "generally effective," two for "inconsistent," and four for "little or none." Other risk items have different weights to reflect their specific correlation to reoffending. The procedure for determining a juvenile's risk classification involves summation of total score and classification into low, low-medium, medium-high, and high risk using cutoff scores of six, twelve, and nineteen, respectively.[86]

The Wisconsin Risk Assessment Tool has been validated in different settings both in its original and various revised formats. Christopher Baird, Richard Heinz, and Brian Bemus reported on the validity of the original model, and using a three-level categorization based on total risk score, were able to predict rates of probation and parole revocations for a Wisconsin sample of over four thousand cases.[87] The overall base rate for revocations was eleven percent, and they identified rates of two, nine, and twenty-six percent for low-, medium-, and high-risk cases, respectively.[88] Professors Kevin Wright, Todd Clear, and Paul Dickson reported less favorable results for a New York sample of 366 probationers for which in-program information on a variety of indices of recidivism were available.[89]

---

85. Prorated total risk scores were calculated in cases involving any missing data.

86. Alternative cutoff points were evaluated but none provided improvement.

87. BAIRD ET AL., *supra* note 78, at 10 tbl.1, 21.

88. *Id.* at 10 tbl.1, 11 tbl.2.

89. K.N. Wright, T.R. Clear & P. Dickinson, *Universal Applicability of Probation Risk-*

The in-program "failure" rate, as defined by the above indicators, was thirty percent.[90] The New York study found no significant relationship between overall risk scores and recidivism, and only three of the eleven components of risk individually predicted failures at levels above chance. Other studies, such as Professors Eileen Gambrill and Aron Shlonsky's work in child welfare, have shown that actuarial risk assessments such as the Wisconsin Risk Assessment Tool do not perform adequately.[91]

The Wisconsin assessment-based risk classification of the juveniles in the *ProDES* sample is presented in Table 2. The cutoffs provided in the original study create four risk groups with almost fifteen percent of juveniles considered low risk, sixty-four percent low-medium or medium-high risk, and twenty-one percent high risk. The overall risk classification clearly varies by gender (see Table 3) and by race (see Table 4) with males and whites overrepresented in the high-risk category.

TABLE 2
WISCONSIN RISK CLASSIFICATION

Risk	%	N
Low	14.7	1215
Low-Medium	31.6	2603
Medium-High	32.4	2669
High	21.3	1752
Total	**100%**	**8239**

TABLE 3
WISCONSIN RISK CLASSIFICATION BY GENDER

Risk	Males		Females	
	%	N	%	N
Low	14.2	1022	18.9	193
Low-Medium	31.6	2277	31.8	326
Medium-High	32.7	2362	30.0	307
High	21.5	1554	19.3	198
Total	**100%**	**7215**	**100%**	**1024**

Chi-square 17.51; df = 3; $p < 0.001$

Assessment Instruments: A Critique, 22 CRIMINOLOGY 113-34 (1984).

90. *Id.*

91. Gambrill & Shlonsky, *supra* note 32, at 825-26.

TABLE 4
WISCONSIN RISK CLASSIFICATION BY RACE

Risk	White		Non-White	
	%	N	%	N
Low	11.3	122	15.3	1091
Low-Medium	25.1	270	32.6	2329
Medium-High	33.7	363	32.2	2306
High	30.0	323	20.0	1428
Total	**100%**	**1078**	**100%**	**7154**

Chi-square 71.25; df = 3; $p < 0.001$

C. *Modeling Risk II: Customized Model Using Configural Analysis*

The second approach to assessing risk involves the use of configural analysis on the Philadelphia sample itself. The method uses tree induction algorithms or nonparametric classification procedures to subdivide a sample into significant groupings on the basis of specific statistical and theoretical conditions.[92] The predictor variables used in the model are related to the outcome variable (reoffending) by a set of rules. As Professor Irene Casas notes,

> These methods do not require a functional form specification, thus allowing situations to be revealed by the data with minimum interference of the analyst. [Such] [d]ecision tree induction systems are under the supervised learning category techniques, which have as [their] purpose producing general rules from externally supplied examples.[93]

Configural analysis techniques are rarely found in the juvenile justice or child welfare literature, but they are more commonly used in other social science and business disciplines. They are often used when an analyst seeks to identify the optimal predictor from a wide range of competing options. They can be used as an alternative or as an adjunct to traditional regression-based techniques and offer considerable advantages when interested in identifying complex interactions among predictors.

The research reported here makes use of a segmentation technique called Answer Tree that is available within the software package, Statistical Package for the Social Scientist.[94]

---

92. Arentze et al., *supra* note 42, at 336-38. A tree induction algorithm, or nonparametric classification procedure, is a "discrete choice model" that predicts an individual's behavior and choices. *Id.* at 330. These algorithms "assume that individuals compare a number of choice alternatives on relevant attributes and choose the alternative that maximizes some measure of utility." *Id.*

93. Irene Casas, *Evaluating the Importance of Accessibility to Congestion Response Using a GIS-Based Travel Simulator*, 5 J. GEOGRAPHICAL SYS. 109, 119 (2003).

94. *See* R. GNANADESIKAN, METHODS FOR STATISTICAL DATA ANALYSIS OF MULTIVARIATE OBSERVATIONS 333 (1997) (stating that since the late 1970s researchers have widely used the

> Answer Tree is a set of statistical algorithms, designed to select patterns of variables [that] are especially predictive of a pre-defined criterion variable. The methods are . . . ideal for use in relatively large samples, where the cases are characterized by many variables and where a multitude of different . . . variable patterns are expected to occur.[95]

Published risk assessment research involving this technique is found in many disciplines, including the health sciences,[96] marketing,[97] psychology,[98] geography and regional planning,[99] criminal justice,[100] and education.[101] From an analytic perspective, Answer Tree begins by dividing the initial sample (preferably a randomly selected construction subsample) into smaller first-order subgroups, which seek to maximize the variability of the criterion variable. Each first-order subgroup is then analyzed independently to identify the optimal predictor that maximizes the variability of the criterion variable for that subgroup. Each new category is similarly subdivided until one of several "stopping rules" are met. This process produces a tree-like structure (dendogram) with the initial sample at the top and a series of "branches" extending downward to the terminal subgroups. Each subdivision is based on the analyst's selection of the predictor variable that is considered the best predictor of the criterion variable for that sample. In the "exhaustive chaid" algorithm used in this study, the selection of the optimal predictor is based on a combination of the analyst's assessment of the quality of the predictor variables involved and their probability scores or effect sizes (Cohen's *d*) based on chi-squared tests. At each selection point, Answer Tree arranges the potential predictors by effect size and the analyst is free to select that predictor that best meets his requirements.[102] Model development ends when all remaining subgroups fail to meet analyst specified stopping rules based on either a minimal subgroup size or probability level.[103]

In the current study, Answer Tree repeatedly divides the initial construction sample into mutually exclusive and exhaustive subgroups on the basis of the

---

Statistical Package for the Social Scientist to perform data analysis).

95. Emma C. Smith & Klaus Grawe, *What Makes Psychotherapy Sessions Productive? A New Approach to Bridging the Gap Between Process Research and Practice*, 10 CLINICAL PSYCHOL. & PSYCHOTHERAPY 275, 278 (2003).

96. John Welte et al., *Risk Factors for Pathological Gambling*, 29 ADDICTIVE BEHAV. 323, 329 (2004).

97. Jedid-Jah Jonker et al., *Joint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-term Profitability*, 27 EXPERT SYSTEMS WITH APPLICATIONS 159, 159-60 (2004).

98. Smith & Grawe, *supra* note 95, at 279.

99. *See* Casas, *supra* note 93, at 119-24 (using the CHAID method of risk assessment research for geographical planning of transportation system).

100. Jones et al., *supra* note 24, at 490.

101. Bennie R. Grobler et al., *The Chaid-Technique and the Relationship Between School Effectiveness and Various Independent Variables*, 30 INT'L STUD. EDUC. ADMIN. 49, 49-50 (2002).

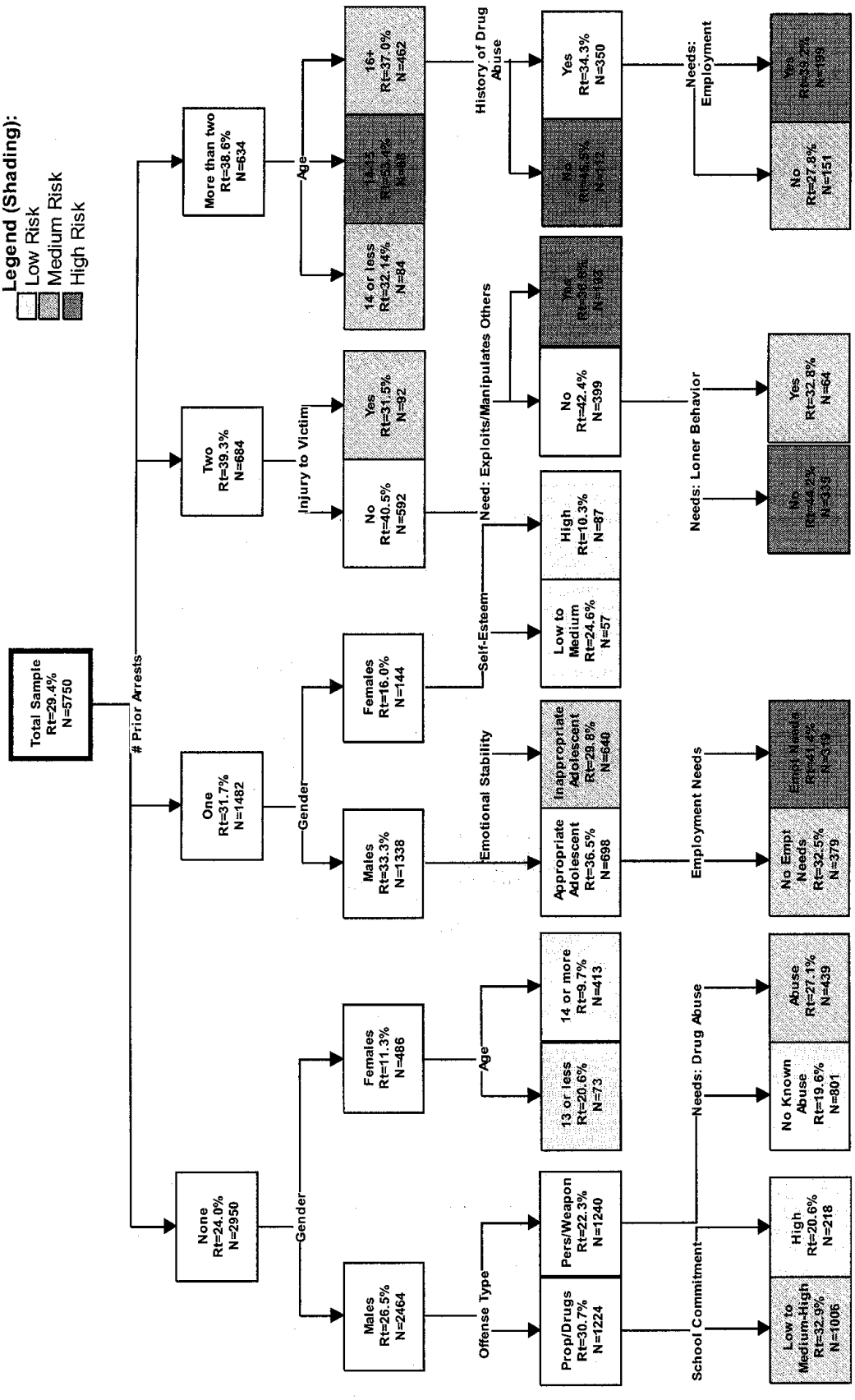
102. Smith & Grawe, *supra* note 95, at 278.

103. *Id.* at 279.

criterion measure of reoffending.[104] As a result, the predictors identified do not always apply to the entire sample (as would be the case with more common regression-based approaches). For example, the number of prior arrests is identified as the most significant initial predictor (see results for construction sample in Figure 1). The model subsequently searches for the next best predictor among the subgroups with no prior arrests, one prior, two priors, or more than two priors. For the first two of these subgroups, the second predictor is gender, for those with two priors the next predictor is "injury to victim" during commission of the current offense, and for those with more than two prior arrests the next predictor is the juvenile's age. The model adds predictors of different types to each of these subgroups, creating "branches" of different length, size, and composition for different types of juveniles.

104. Jay Magidson, *The CHAID Approach to Segmentation Modeling: CHi-squared Automatic Interaction Detection*, *in* ADVANCED METHODS OF MARKETING RESEARCH 118, 156-57 (Richard P. Bagozzi ed., 1994) (discussing that CHAID segmentation algorithm is superior to the previously used models when desired result is clusters with differing criterion).

Figure 1 Dendogram of Predictors of Rearrest



Each of the individual variables that comprise the Wisconsin Risk Assessment Tool was available to the configural analysis for inclusion in the model, but none was selected. Interestingly, several of the variables included in the Wisconsin needs assessment—a set of variables not considered to have predictive criminogenic value—were selected and included in the configural model. These "need" based variables included staff assessments of whether the juvenile was manipulative of others, was a loner, had employment needs, or had a drug or alcohol abuse problem.

The resulting risk classification is presented in Table 5, and the relationship of risk to gender and race is presented in Tables 6 and 7, respectively. When compared with the Wisconsin classification, it is evident that the configural analysis—which has three risk groups—places far more juveniles in the low-risk group and slightly fewer in the high-risk group. The distributions also vary significantly by gender and race. In the configural model, very few females are placed in the high-risk category, and the proportion of high-risk whites and non-whites are more comparable than is found in the Wisconsin classification.

TABLE 5
CONFIGURAL RISK CLASSIFICATION

Risk	%	N
Low	28.5	2349
Medium	53.0	4369
High	18.5	1521
Total	**100%**	**8239**

TABLE 6
CONFIGURAL RISK CLASSIFICATION BY GENDER

Risk	Males		Females	
	%	N	%	N
Low	19.7	1420	90.7	929
Medium	59.8	4313	5.5	56
High	20.5	1482	3.8	39
Total	**100%**	**7215**	**100%**	**1024**

Chi-square 2222.131; df = 2; $p < 0.001$

TABLE 7
CONFIGURAL RISK CLASSIFICATION BY RACE

Risk	White		Non-White	
	%	N	%	N
Low	26.5	286	28.8	2060
Medium	57.7	622	52.3	3743
High	15.8	170	18.9	1351
Total	**100%**	**1078**	**100%**	**7154**

Chi-square 11.72; df = 2; $p < 0.01$

D. Modeling Risk III: Customized Model Using Neural Networks

The third approach uses artificial neural networks "because of their ability to learn from databases and build predictive models that can evolve and update themselves using random selections of case file data."[105] The potential of this type of statistical technique has been demonstrated in the child welfare arena where the final trained network was able to successfully categorize 89.6% of the cases in the testing population.[106]

Neural networks offer an appropriate statistical approach to risk assessment because they "are not organized around rules or programming, but learn the underlying behaviors of a model by analyzing its input and output values and adjusting the weights between neuron layers. In this respect the behavior of a neural network is trained using supervised or unsupervised techniques."[107] This type of "soft computing" methodology is widely used in research and practice by other fields because it is highly adaptive and robust. For example, "[n]eural networks have performed well as risk assessment and classification tools . . . in the medical decision-making literature."[108] Felipe Atienza developed a network to correctly classify 123 of 132 patients,[109] and Richard Orr trained a network to estimate mortality following cardiac surgery to the point where he achieved 91.5% accuracy for the training set and 92.3% accuracy for the validation set.[110] Though the prediction of health outcomes is quite different from the prediction of actual human behavior, the fact remains that the existing actuarial models used to predict human behavior are an adaptation of the health and life

105. Schwartz et al., *supra* note 7, at 1082.

106. *Id.* at 1091.

107. EARL COX, FUZZY LOGIC FOR BUSINESS AND INDUSTRY 582 (1995).

108. Schwartz et al., *supra* note 7, at 1084.

109. Felipe Atienza et al., *Risk Stratification in Heart Failure Using Artificial Neural Networks*, PROC. AMIA ANN. SYMP. 32, 34 (2000), *available at* http://www.amia.org/pubs/proceedings/symposia/2000/ D200367.pdf.

110. Richard K. Orr, *Use of Probabilistic Neural Network to Estimate the Risk of Mortality After Cardiac Surgery*, 17 MED. DECISION MAKING 178, 178 (1997).

insurance models developed in the past.[111] Child welfare literature discusses neural networks, but with the exception of David Schwartz, they have not been implemented in the field.[112] David Marshall and Diana English compared "logistic and linear multiple regression to neural networks using child protective service data from the State of Washington's risk assessment model. . . . [and] concluded that the neural network produced superior prediction and classification results."[113]

For the present study, the neural network analysis was run on the same *ProDES* database as produced the Wisconsin and configural analysis results. Though the neural network could in theory work with a wide array of potential predictor variables, it was decided, for purposes of this analysis, to restrict the model to those variables that had previously been identified as predictors in the configural analysis. The utilization of the configural model to constrain the array of potential predictors for neural networks analysis is an unusual but appropriate application of a preliminary noise reduction strategy. Indirectly, it also enhances direct comparison with the results of the configural model.

Classification of the dataset was undertaken using a multilayer perceptron ("MLP"), which is a feed-forward neural network model trained using the back-propagation algorithm. Such models are very powerful and suitable for learning highly complex concepts given sufficiently large data in training time that scales linearly with data size.[114] The MLPs were trained with seventy percent of the sample and tested with a thirty percent validation sample that was disjointed and randomly generated. This methodology is used for predictive purposes to test the accuracy of the model on unseen instances. For purposes of this study, the analysis was performed with a general purpose machine learning suite with mostly default parameters. Training MLPs is a computationally expensive and complex process that requires a large number of trials to identify the optimal set of parameters to produce the most accurate model. The training data is processed hundreds to thousands or more times to obtain the final model. The model was initially run on the entire sample, but because of the importance of gender differences with regard to reoffending, separate models were also run for the male and female sample subsets.

111. Schwartz et al., *supra* note 7, at 1084.

112. For a discussion of Schwartz's work with neural networks, see *supra* note 7, at 1081.

113. Schwartz et al, *supra* note 7, at 1092; *see also* David B. Marshall & Diana J. English, *Neural Network Modeling of Risk Assessment in Child Protective Services*, 5 PSYCHOL. METHODS 102, 103 (2000) (concluding that under certain conditions linear decision making is superior to clinical judgments).

114. NIGRIN, *supra* note 42, at 107-08.

TABLE 8
NEURAL NETWORK RISK CLASSIFICATION

Risk	%	N
Low	65.8	5424
High	34.2	2815
Total	**100%**	**8239**

TABLE 9
NEURAL NETWORK RISK CLASSIFICATION BY GENDER

Risk	Males		Females	
	%	N	%	N
Low	62.8	4532	87.1	892
High	37.2	2683	12.9	132
Total	**100%**	**7215**	**100%**	**1024**

Chi-square 235.32; df = 1; $p < 0.001$

TABLE 10
NEURAL NETWORK RISK CLASSIFICATION BY RACE

Risk	White		Non-White	
	%	N	%	N
Low	72.8	785	64.8	4634
High	27.2	293	35.2	2520
Total	**100%**	**1078**	**100%**	**7154**

Chi-square 26.96; df = 1; $p < 0.001$

E. The Risk Classifications Compared

In theory, even though the three risk assessment tools combine different types of measures in different ways, it is possible that the same juveniles will be identified as high or low risk by all three methods. Table 11 compares Wisconsin and configural classifications. If one combines the "low-medium" and "medium-high" categories of the Wisconsin Tool, the Wisconsin Tool classifies about twenty-five percent of the configural classification's low-risk youth as low risk, it classifies sixty-four percent of medium-risk juveniles as medium risk, and it classifies twenty-five percent of high-risk juveniles as high risk. There are also instances where the two classifications are at odds—ten percent of the configural classification's low-risk youth are assessed high risk by the Wisconsin

classification and, similarly, ten percent of the configural high-risk youth are assessed low risk by the Wisconsin classification.

TABLE 11
WISCONSIN BY CONFIGURAL RISK CLASSIFICATION

Wisconsin Risk Classification	Configural Risk Classification			
	Low Risk	Medium Risk	High Risk	Total
Low	25.4%	10.6%	10.2%	14.7%
Low-Medium	40.8%	28.2%	27.2%	31.6%
Medium-High	23.5%	35.3%	37.9%	32.4%
High	10.3%	25.9%	24.7%	21.3%
Total	**100%**	**100%**	**100%**	**100%**

Comparing the neural network with the Wisconsin classification indicates even less agreement (see Table 12). Of the juveniles assessed low risk by neural networks, only fifteen percent were low risk in the Wisconsin classification and twenty-one percent were assessed high risk. Similarly, of the juveniles assessed high risk by neural networks, twenty-one percent were also assessed high risk by the Wisconsin analysis, and over twelve percent were assessed to be low risk.

TABLE 12
WISCONSIN BY NEURAL NETWORK RISK CLASSIFICATION

Wisconsin Risk Classification	Neural Network Risk Classification		
	Low Risk	High Risk	Total
Low	15.9%	12.5%	14.7%
Low-Medium	31.4%	32.0%	31.6%
Medium-High	31.3%	34.5%	32.4%
High	21.4%	21.0%	21.3%
Total	**100%**	**100%**	**100%**

Comparing the third pair of assessments—configural and neural networks—shows greater agreement (see Table 13). Of the juveniles assessed low risk by neural networks, thirty-five percent were also low risk in the configural analysis; of those assessed high risk by neural networks, twenty-seven percent were similarly assessed by the configural model.

These results provide definitive proof that type of assessment does matter. Table 14 focuses solely on juveniles classified as high risk by each of the assessment tools. Over half the juveniles in the sample were assessed high risk by

one or other of the tools, and yet less than two percent were assessed high risk by all three tools.

TABLE 13
CONFIGURAL BY NEURAL NETWORK RISK CLASSIFICATION

Configural Risk Classification	Neural Network Risk Classification		
	Low Risk	**High Risk**	**Total**
Low	34.7%	16.6%	28.5%
Medium	51.3%	56.3%	53.0%
High	14.0%	27.0%	18.5%
Total	**100%**	**100%**	**100%**

TABLE 14
COMPARING RISK CLASSIFICATIONS: HIGH RISK JUVENILES

Classification Agreement	N	%	% Reoffend
None High	4664	56.6	3.0
One High	2185	26.5	59.1
Two High	1221	14.8	69.2
Three High	169	2.1	91.1
Total	**8239**	**100%**	**29.5%**

F. Predictive Validity

The risk assessments can be tested by comparing the predicted risk classification with actual reoffending patterns. The results for the Wisconsin Risk Assessment Tool show that 24.4% of the low-risk juveniles reoffended during the follow-up period compared with 30.0% of the high-risk juveniles (see Table 15). The results for the two medium-risk categories are similar—28.6% and 32.4%, respectively. Nonetheless, the fact that the medium-high-risk group reoffended at a higher rate than the high-risk group suggests the classification is not discriminating well on the basis of risk.

TABLE 15
WISCONSIN RISK CLASSIFICATION BY REOFFENDING RATE

Risk	No	Reoffend
Low	75.6%	24.4%
Low-Medium	71.4%	28.6%
Medium-High	67.6%	32.4%
High	70.0%	30.0%
Total	**70.5%**	**29.5%**

Separating the results by gender highlights the poor overall discrimination of the Wisconsin Tool, and the higher rate of reoffending for medium-high-risk group compared with the high-risk group is replicated for both males and females (see Table 16). Indeed, for females the rate of reoffending is also lower for the low-medium than the low-risk group (10.4% to 13.0%, respectively).

TABLE 16
WISCONSIN RISK CLASSIFICATION BY GENDER AND REOFFENDING RATE

Risk	Males		Females	
	No	Reoffend	No	Reoffend
Low	73.5%	26.5%	87.0%	13.0%
Low-Medium	68.8%	31.2%	89.6%	10.4%
Medium-High	65.4%	34.6%	84.7%	15.3%
High	67.7%	32.3%	87.9%	12.1%
Total	**68.1%**	**31.9%**	**87.3%**	**12.7%**

Repeating the analysis for race indicates poor overall discrimination for both whites and non-whites. For white juveniles, the low-medium-risk group had the highest actual rate of reoffending (28.1%), though all groups differed only slightly from the overall average of 24.9% (see Table 17). For non-whites, the separation of reoffending rates is a little more marked, but the highest rate of reoffending remains with the medium-high-risk group (33.7% compared with 31.4% for the high-risk group).

The results for the configural assessment represent an improvement, with 17.4% of low-risk juveniles reoffending compared with 31.2% of medium-risk juveniles and 43.3% of the high-risk juveniles (see Table 18). The configural risk tool achieves considerably better discrimination among the risk categories in terms of reoffending. Moreover, the model places a similar proportion of juveniles into the high-risk category (18.5% compared with 21.3% for the Wisconsin Risk Assessment Tool) and almost twice as many juveniles into the

low-risk category (28.5% compared with 14.7%). From a decision-making perspective, the configural model is superior—it allows for the identification of a larger group of low-risk juveniles whose actual reoffending rate is lower than that for low-risk juveniles identified by the Wisconsin Risk Assessment Tool (17.4% compared with 24.4%), and it is correct more often for the high-risk juveniles. If the configural assessment was used as part of dispositional decisions, a larger proportion of juveniles would be considered for less intrusive interventions, *and* a larger proportion of those individuals would remain delinquency free for the period of study. Among the high-risk juveniles, there would be a lower rate of false positives.

TABLE 17
WISCONSIN RISK CLASSIFICATION BY RACE AND REOFFENDING RATE

Risk	White		Non-White	
	No	Reoffend	No	Reoffend
Low	78.8%	21.3%	75.3%	24.7%
Low-Medium	71.9%	28.1%	71.3%	28.7%
Medium-High	75.5%	24.5%	66.3%	33.7%
High	76.2%	23.8%	68.6%	31.4%
Total	**75.1%**	**24.9%**	**69.8%**	**30.2%**

TABLE 18
CONFIGURAL RISK CLASSIFICATION BY REOFFENDING RATE

Risk	No	Reoffend
Low	82.6%	17.4%
Medium	68.8%	31.2%
High	56.7%	43.4%
Total	**70.5%**	**29.5%**

Separate analysis by gender shows that the improved discrimination of reoffending is evident both for males and females (see Table 19). The actual reoffending rate for high-risk males is twice that for low-risk males (43.8% to 21.0%, respectively) and for females the differences are even greater (25.6% to 11.9%). The most striking contrast between the two models is found in classification of females. The Wisconsin Risk Assessment Tool places 193 females in the low-risk category and, for this group, the reoffending rate is 13.0%. If one combines the low- and low-medium-risk categories the figures

become 519 and 11.4%, respectively. The configural analysis places 929 females in the low-risk category and their reoffending rate is 11.9%.

This contrast in models reflects a common problem in risk-prediction modeling, where male-dominated samples are analyzed with statistical techniques that prioritize predictors with explanatory power for the entire sample. Such an approach fails to recognize that the predictors of reoffending for females may have a different etiology to those for males. This problem is compounded when the resulting risk model—dominated as it is by male predictors—is applied to females. The model incorrectly places some females in medium- and even high-risk categories and simultaneously fails to identify females with the highest risk of reoffending. The Wisconsin Risk Assessment Tool shows a 12.1% reoffending rate for its high-risk females whereas the configural model identifies a smaller but more at-risk group, with an actual reoffending rate of 25.6%, or twice the female base rate.

TABLE 19
CONFIGURAL RISK CLASSIFICATION BY GENDER AND REOFFENDING RATE

Risk	Males		Females	
	No	Reoffend	No	Reoffend
Low	79.0%	21.0%	88.1%	11.9%
Medium	68.6%	31.4%	83.9%	16.1%
High	56.2%	43.8%	74.4%	25.6%
Total	**68.1%**	**31.9%**	**87.3%**	**12.7%**

The results by race reveal superior statistical discrimination of reoffending rates by risk category for both whites and non-whites (see Table 20). For white juveniles, the low-risk group in the configural assessment is larger in size (26.5% of juveniles) than the Wisconsin low-risk category (11.3%), yet has a lower rate of reoffending (14.0% compared to 21.3%). In contrast, the configural assessment high-risk white group is half the size of the Wisconsin high-risk group (15.8% to 30.0%, respectively) and has a higher rate of reoffending (36.5% to 23.8%, respectively).

For non-whites, the results are slightly different. The low-risk non-white group in the configural assessment is larger in size (28.8% of juveniles) than the Wisconsin low-risk category (15.3%) and has a lower rate of reoffending (17.9% compared to 24.7%). The configural assessment non-white high-risk group is about the same size as the Wisconsin high-risk group (18.9% to 20.0%, respectively) and has a higher actual rate of reoffending (44.2% to 31.4%, respectively).

Though race is not a predictor in either model, it is likely that many of the predictors identified are themselves closely correlated with race and therefore act as proxy measures for race in the assessment models. The results presented

here suggest that the Wisconsin Risk Assessment Tool is not recognizing those variables that affect reoffending differentially for whites and non-whites or for females compared with males. As a result, the Wisconsin Risk Assessment Tool reflects very little discrimination of reoffending rates by risk group for either gender or racial groups. The configural model has superior performance for gender—which is included in the model—and for race, even though race is not directly included in the model.

TABLE 20
CONFIGURAL RISK CLASSIFICATION BY RACE AND REOFFENDING RATE

Risk	White		Non-White	
	No	Reoffend	No	Reoffend
Low	86.0%	14.0%	82.1%	17.9%
Medium	73.3%	26.7%	68.0%	32.0%
High	63.5%	36.5%	55.8%	44.2%
Total	**75.1%**	**24.9%**	**69.8%**	**30.2%**

The results for the neural networks assessment represent a significant improvement over both the Wisconsin and the configural models (see Table 21). Creating only two categories, it identifies a group of low-risk juveniles for whom the reoffending rate is three percent and a high-risk group for which it is almost eighty-one percent. The predictive validity of this model far exceeds the estimated prediction capabilities forecast by Stephen Gottfredson[115] and Peter Schmidt and Ann Witte[116] less than twenty years ago. The predictive validities vary slightly by gender (see Table 22) and race (see Table 23), but it is evident that the neural networks model is far superior to either of the two other classifications.

TABLE 21
NEURAL NETWORK RISK CLASSIFICATION BY REOFFENDING RATE

Risk	No	Reoffend
Low	97.0%	3.0%
High	19.4%	80.6%
Total	**70.5%**	**29.5%**

115. *See* Gottfredson, *supra* note 5, at 24 (suggesting that "no clear-cut empirical advantages of selecting one prediction method over another" exist).

116. *See* SCHMIDT & WITTE, *supra* note 39, at 15 (finding it encouraging that research studies kept false positives and false negatives below fifty percent).

TABLE 22
NEURAL NETWORK RISK CLASSIFICATION BY GENDER AND REOFFENDING RATE

Risk	Males		Females	
	No	Reoffend	No	Reoffend
Low	96.9%	3.1%	97.5%	2.5%
High	19.5%	80.5%	18.2%	81.8%
Total	**68.1%**	**31.9%**	**87.3%**	**12.7%**

TABLE 23
NEURAL NETWORK RISK CLASSIFICATION BY RACE AND REOFFENDING RATE

Risk	White		Non-White	
	No	Reoffend	No	Reoffend
Low	98.6%	1.4%	96.7%	3.3%
High	12.3%	87.7%	20.2%	79.8%
Total	**75.1%**	**24.9%**	**69.8%**	**30.2%**

IV. DISCUSSION

During the past several decades, there has been a large body of research that has established the superior predictive validity of actuarial compared with clinical risk models.[117] This research has also established many features of best practice in the field of risk prediction. Samples need to be large enough to sustain multivariate analyses and division into construction and validation subsamples.[118] Appropriate multivariate techniques need to be employed and the researcher must be guided either by theory or by the data to identify and include complex interactions among predictor variables—gender being a primary example. The quality of the data used is vital to the reliability of the final results—analyses that are seriously compromised by copious missing data are simply not to be trusted. And finally, the samples being used to develop a risk instrument will inevitably place some constraints on the range of applications for which the resulting risk assessment tools can be used—a model developed on a

117. *See, e.g.*, EILEEN GAMBRILL, CRITICAL THINKING IN CLINICAL PRACTICE: IMPROVING THE QUALITY OF JUDGMENTS & DECISIONS 442-43 (2d ed. 2005) (stating that actuarial model is superior because it can better integrate substantial amounts of diverse data).

118. *See id.* at 234-35 (suggesting that relying on small sample sizes may cause inaccurate estimates).

Midwestern, primarily white urban or rural sample cannot be transferred to an urban, inner city, predominantly minority setting with any sense of confidence.

Despite this growing body of knowledge, the fact remains that many of the risk assessments being utilized in juvenile justice and child welfare settings leave a great deal to be desired. Risk assessments are borrowed, adapted to fit local data conditions (by adding or excluding carefully identified predictors or by changing the risk scores associated with established predictors), and then utilized without validation to produce risk classifications to support potentially life-changing decisions. There is a general but misplaced sense that actuarial risk assessment is so robust, so superior to clinical decision making, that it can sustain these modifications without being seriously compromised.

The data presented here examines the classifications and the predictive validity of three risk assessments—all of which could easily be employed in the field. The data indicates that the tools do not produce the same classifications—very few juveniles would be placed in the same risk category if assessed by all three tools. The data further indicate that the Wisconsin Risk Assessment Tool performs poorly, displaying little predictive validity overall, and particularly poor validity for females. The configural assessment represents an improvement, and the neural network assessment sets a new and higher standard.

The implications of the disparity among these three assessment models are not trivial. Actuarial risk instruments—even the very poor ones—have an aura of objectivity and general superiority to clinical decision making. They appear "scientific." The results of this study demonstrate that such assumptions need to be justified. The predictive validity of actuarial tools must be examined, especially when they are used in settings other than those where they were developed. For example, if a low-risk classification was a factor in securing a less intrusive response from the juvenile justice system, then the Wisconsin Risk Assessment Tool would classify 1,215 of the current sample as low risk, and it would be correct in about seventy-five percent of cases (i.e., they would not reoffend during the study period). The configural model would identify 2,349 as low risk and get eighty-three percent of them correct. The neural networks model would identify 5,424 as low risk and get ninety-seven percent correct.

The appropriateness and predictive validity of risk assessment tools is clearly an issue that will continue to generate considerable research activity. The degree to which these instruments shape the rights and freedoms of juveniles is less well-known. This becomes particularly pertinent when it involves predictions of high risk that result in more intrusive interventions requiring juveniles to be removed from their home and school communities. Also pertinent is the risk instruments contributing directly or indirectly to discriminatory decisions based on such factors as gender and race. The Canadian Human Rights Commission has provided evidence that actuarial tools discriminate against women in general, and Aboriginal women and women with disabilities in particular.[119] Such

119. CANADIAN HUMAN RIGHTS COMM'N, PROTECTING THEIR RIGHTS: A SYSTEMIC REVIEW OF HUMAN RIGHTS IN CORRECTIONAL SERVICES FOR FEDERALLY SENTENCED WOMEN 24-25 (2003).

criticism is found increasingly in Canadian research,[120] but is not as evident in the United States. Professor Ivan Zinger cites critics who argue that actuarial risk assessment not only reflects, but also contributes to, the overrepresentation of minorities in the juvenile and criminal justice systems[121] and states that "the enthusiasm demonstrated by its promoters is 'cult-like.'"[122]

Actuarial risk assessment has not yet become the target of legalistic scrutiny and challenge, and there is little existing case law that addresses the problems identified in the present study. Judge David P. Cole and Former Director of Special Projects for the Corrections Branch of the Solicitor General Glenn Angus identify a few Canadian cases and one American case that directly questioned the scientific validity of actuarial risk assessment.[123] Ivan Zinger reports a review of Canadian case law as revealing no cases arguing that actuarial risk assessment violated individual rights.[124]

A review of federal and state case law in the United States does show that actuarial risk assessment is facing some legal challenge. A recent Illinois Supreme Court decision addressed a case where the respondent initially argued that actuarial risk assessment is "a novel scientific methodology that has yet to gain general acceptance in the psychological and psychiatric communities."[125] The respondent further argued that any expert testimony based on actuarial risk assessment must be excluded under *Frye v. United States*.[126] "[T]he State argued that (1) actuarial principles are not the least bit novel and therefore are not subject to *Frye*; and (2) even if the particular actuarial instruments at issue are novel, they have gained general acceptance in the relevant psychological and psychiatric communities."[127] After conflicting opinions by the circuit and appellate courts, the Illinois Supreme Court ruled that actuarial risk assessment of a sexually violent offender meets the *Frye* general acceptance test and is admissible evidence in an Illinois court of law.[128]

As Ivan Zinger notes, most criticism of actuarial risk assessment fails to offer any reasonable alternative, and a return to clinical assessment takes us in the direction of increased and less explicit and recognizable bias.[129] The answer

120. *See* Zinger, *supra* note 25, at 610 (citing series of criticisms of actuarial tools for discrimination against women).

121. *Id.* at 613.

122. *Id.* at 613-14 (quoting Kelly Hannah-Moffat & Margaret Shaw, *Situation Risquee: Le Risque et les Services Correctionnels au Canada*, 34 CRIMINOLOGIE 47 (2001)).

123. David P. Cole & Glenn Angus, *Using Pre-Sentence Reports to Evaluate and Respond to Risk*, 47 CRIM. L.Q. 302, 313-14 (2003) (citing, among other cases, Bains v. Canada (Nat'l Parole Bd.), [1989] 3 F.C. 450; Moore v. Valdez, unreported op. (Fla. C.I. Aug. 21, 2000)).

124. Zinger, *supra* note 25, at 615.

125. *In re* Commitment of Simons, 821 N.E.2d 1184, 1186 (Ill. 2004).

126. *Simons*, 821 N.E.2d at 1186. *Frye* provides that scientific evidence is an admissible methodology if it has gained mainstream, general acceptance. Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

127. *Simons*, 821 N.E.2d at 1186-87.

128. *Id.* at 1196.

129. Zinger, *supra* note 25, at 615-16.

is not to throw the baby out with the bathwater—it is to recognize the gap that exists between theory and practice in the field of actuarial risk assessment in the United States and to move quickly to establish minimum quality control standards. The present study has established that actuarial risk assessments can vary enormously in terms of where they place juveniles and how accurate their predictions can be. For change to occur, there needs to be explicit recognition that the potential benefits of actuarial risk assessment can be achieved only by carefully maintaining the integrity of the risk prediction process. The methodological "high ground" needs to be earned and not simply assumed.

The analyses presented in this Article show the variation that exists between a reputable, widely recognized and commonly used tool such as the Wisconsin Risk Assessment Tool and alternative risk models developed for a specific sample.[130] For Philadelphia's delinquent population, the Wisconsin Risk Assessment Tool performed only slightly better than chance (the relative improvement over chance ("RIOC") statistic is seven percent). The configural model provided modest improvement (RIOC = twenty percent), and the neural networks model a very significant improvement (RIOC = ninety percent).

Neural networks represent a level of analysis that is not presently available on a widespread basis. Immediate improvements to the field of risk assessment are possible, however, with the application of several quality control measures. The benefits of a "scientific" label should adhere only if actuarial risk assessments meet or exceed the following criteria:

> 1. The risk assessment should have been tested for reliability and validated at the time of development. Both indicators should be reassessed at least every three years.
>
> 2. If a risk assessment developed for a specific population is being used in a different setting—whether this be defined by geography, race, gender, age, etc.—then it must be validated prior to implementation where data allow, or within two years of implementation if required historic data are not available.
>
> 3. If a risk assessment of a specific outcome is being used for a different purpose, its predictive validity needs to be established—e.g., an assessment of general reoffending likely has little predictive validity for outcomes such as dangerousness, violent reoffending, or sexual reoffending.
>
> 4. The tests of reliability and validity need to be conducted by researchers who do not have a vested interest in the promotion of the assessment tools.[131]
>
> 5. The content or format of a risk assessment should match the original risk assessment—it must be recognized that any unauthorized

130. See *supra* Part III.E for a comparison of the three models discussed.

131. As Ivan Zinger notes, risk assessment has become big business and independent evaluations are required to support those of the individuals or companies that developed the tools. Zinger, *supra* note 25, at 608.

deletions or additions may not only render the assessment tool invalid but introduce inappropriate, and not always recognizable, bias.

6. Risk assessments should be required to identify any possible gender and racial bias—i.e., tests that the predictive validity is comparable for appropriate subgroups of the population.

7. Risk assessment tools should have a clearly defined and reasonable articulation of the number of risk levels it produces and the cutoff scores for those categories (i.e., the risk score that moves a juvenile from medium to high risk).

8. All risk assessments are dependent on the quality and availability of the risk measures. Data quality should be explicitly reviewed and reported—e.g., adequate reliability and psychometric validity of the risk variables employed and, for each risk item, the proportion of cases for which data are missing.

This Article clearly demonstrates the importance and value of using more sophisticated research protocols in juvenile justice and child welfare. The results presented also have implications for the professional development of many of the criminal justice and juvenile justice researchers in academic and practice settings. Configural analysis is rarely used and neural networks are something about which many researchers are quite unfamiliar. The results also demonstrate the value of enlisting support and participation of professionals from disciplines outside the traditional social and behavioral science fields. We must reach out to well-trained experts in computer and information sciences, engineering, and bioinformatics—experts trained and familiar with the advanced research protocols that are needed to advance the quality of work currently being completed in the field.

There is a desperate need to inform and educate practitioners in the field of juvenile justice and child welfare—particularly judges and administrators—about new advances that can significantly improve the quality of services and decision making. If they remain uninformed about new developments—particularly ones that involve advanced research protocols—they are unlikely to embrace promising practices anytime soon. This may explain why there continues to be widespread use and abuse of actuarial models of risk assessment despite their limited utility, and the fact that, while vintage for their time, they are based on obsolete technology.

Maintaining the integrity of the risk assessment process by identifying and requiring adherence to minimal standards of quality is imperative. After more than two decades of relative neglect, it is unlikely that sufficient impetus for improvement will come solely from the researchers and statisticians who develop the models, or even from decision makers who use the models. There is an opportunity for lawyers and child advocates to inject increased rigor into the use of actuarial risk assessment practices of the field. Given the current poor standards of risk assessment applications and the significant potential for improvement, there is no time for delay.