Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments

Dylan Doby

10/25/2024

Have you ever been judged by a number? Actuarial risk assessment instruments (ARAIs), statistical models designed to predict the likelihood of future events based on historical data, do exactly that. Originally developed for the insurance industry, these tools are also now widely used in legal and mental health settings to estimate the probability of future criminal behavior or recidivism (What is risk assessment, n.d.). With their growing influence, ARAIs have become central to decisions regarding parole, sentencing, and even involuntary commitment (Lewis, n.d.). However, despite their widespread application, a critical debate persists about their accuracy and fairness—especially at the individual level. In forensic psychology, where even a slight miscalculation can profoundly impact a person's freedom or treatment, precision is not just statistical; it's a moral imperative.

In "Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments," Stephen D. Hart and David J. Cooke examine the reliability of ARAIs in assessing individual risk. This paper explores their research methodologies and results, as well as the ethical implications of using these tools. As ARAIs play a major role in shaping legal, insurance, and mental health outcomes, a pressing question emerges: how can a tool that potentially lacks precision be ethically justifiable when used to make life-altering decisions? Moreover, is it even morally acceptable to assign a quantifiable value to human life?

The methodology of Hart and Cooke's study is built on the analysis of previously collected data from Karla Jackson's ongoing research into sexually violent recidivism. The dataset consisted of 90 adult male (aged 19 to 77 with a median age of 40) sex offenders (under the Criminal Code of Canada) who had completed a community-based treatment program between 2002 and 2004. These participants were all referred to an outpatient forensic mental health clinic for assessment and treatment, having received sentences that included probation or parole. The sample's demographic breakdown consisted of 66% European descent, 18% Asian descent, and 17% Aboriginal descent. Nearly half (47%) were cohabiting with a partner, and 60% were employed during the study period. Prior to the study's start, 76% of participants had documented histories of sexual violence, and 59% had prior convictions for sexual offenses.

To create the ARAI for this study, the researchers utilized the Sexual Violence Risk-20 (SVR-20), a structured professional judgment (SPJ) instrument that assesses 20 risk factors across four domains: psychological adjustment, social adjustment, history of sexual offenses, and future plans. The SVR-20 was chosen for its well-documented reliability in past research and its comprehensive scope in capturing risk factors for sexual violence. The evaluators, who were doctoral-level researchers with clinical-forensic psychology training, independently coded the risk factors for each subject based on extensive review of treatment records. These records included criminal history, police reports, mental health assessments, and treatment notes.

The coding of the SVR-20 risk factors followed a three-point scale, with 0 indicating the absence of a factor, 1 indicating possible or partial presence, and 2 indicating definite presence. After completing independent ratings, the evaluators reached consensus ratings to enhance reliability. The consensus ratings were then aggregated into domain scores by summing the unit-weighted ratings for each domain. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC), which were found to be strong across all four domains: 0.92 for psychological and social adjustment, 0.89 for history of sexual offenses, and 0.77 for future plans. The final domain scores ranged as follows: psychological adjustment (0–12), social adjustment (1–10), history of sexual offenses (0–12), and future plans (0–4).

For the statistical analysis, the authors employed logistic regression (LR) to predict recidivism, utilizing the SVR-20 domain scores as predictor variables. Logistic regression was selected because it is capable and well-suited for estimating the probability of binary outcomes; in this case, the probability of reoffending. In this study, recidivism was defined as any police investigation, charge, or conviction for a sexual offense within the follow-up period, which averaged 4.21 years. The LR model generated regression coefficients (B-values) that quantified the weight of each domain in predicting failure. The coefficients varied across domains, with psychological adjustment and history of sexual offenses showing stronger associations than social adjustment and future plans, though none of the domains reached statistical significance independently. The ARAI scores produced by the model ranged from -3.64 to 0.77, with a mean score of -1.74 and a standard deviation of 0.83. This score allowed for an initial categorization of individuals into varying risk levels.

The authors first employed receiver operating characteristic (ROC) analysis, which is a common method for assessing model accuracy in binary classification tasks, to evaluate the predictive validity of the ARAI model. The area under the curve (AUC) for the ROC was calculated to be 0.72, suggesting that the ARAI was moderately capable of correctly identifying whether a subject would reoffend, with a 72% chance of distinguishing between recidivists and non-recidivists. To further validate the ARAI's predictive performance, the researchers then calculated confidence intervals (CIs) for group-level estimates and prediction intervals (PIs) for individual-level estimates.

Confidence intervals were used to assess the degree of uncertainty in the aggregate estimates of recidivism for subjects within similar score categories (high-risk or low-risk based on their ARAI scores). In the low-risk group, the failure rate was 10%, while the high-risk group experienced a 33% failure rate. These differences were statistically significant ($p = 0.006$), with the odds of failure for high-risk individuals being 4.5 times greater than those for low-risk individuals. However, the CIs for these group estimates were relatively wide, ranging from 16 to 34 percentage points, reflecting the limitations posed by the small sample size. The authors noted that achieving a narrower CI of about ±3% would require a sample size of approximately 500 subjects per category.

When assessing the ARAI's performance at the individual level, the results were more concerning. Prediction intervals for individual estimates were significantly wider than the CI's for group-level estimates, often spanning 14 to 64 percentage points. The mean individual risk estimate was 18%, but the prediction intervals for individual scores demonstrated considerable overlap, even between subjects categorized as high-risk and low-risk. The average individual risk estimate in the low-risk category was 11%, compared to 32% in the high-risk category, yet the PI's overlapped almost entirely. In fact, only one subject had an individual prediction interval that distinctly diverged from the sample's base failure rate of 18%, emphasizing the model's challenges in achieving accurate individual predictions.

The ethical considerations surrounding the use of ARAIs extend beyond statistical validity to fundamental questions about justice, fairness, and human dignity. A core normative concern involves the reduction of human behavior to numerical values—a concept that resonates with critiques of utilitarianism, which often assigns a "common currency of value" to diverse aspects of human life. Just as the aftermath of 9/11 saw lawyers and actuaries grappling with the complex task of assigning a monetary value to human lives lost in the attacks, ARAIs similarly attempt to quantify the potential for human behavior into a calculable risk. By assigning a numeric risk score to individuals, ARAIs imply that all factors (e.g. psychological adjustment, social adjustment, history of sexual offenses, and future plans) can be equated and summed to a single predictive measure. While efficient, this procedure risks oversimplifying the complex, complicated nature of human behavior, thereby reducing individuals to mere data points.

While these instruments may offer an efficient and objective way to estimate risk at the group level, their precision diminishes significantly at the individual level. This variability in individual risk predictions can lead to morally questionable outcomes, such as unjustly denying a person's freedom based on an inaccurately inflated risk score. From a deontological perspective, this practice could be seen as a violation of the Categorical Imperative, particularly the principle of treating individuals as ends in themselves rather than merely as means to a broader social goal, like reducing overall recidivism rates. When the margin of error is so wide that it blurs meaningful distinctions between high- and low-risk individuals, the fairness and legitimacy of using ARAIs in such consequential decisions are called into question.

Additionally, the ethical issues associated with ARAIs reflect broader concerns about structural biases and inequality embedded in the criminal justice system. The data used to train these models often mirror historical biases related to race, socioeconomic status, and other demographic factors. Consequently, ARAIs may unintentionally perpetuate these biases, leading to outcomes that disproportionately disadvantage marginalized communities. This problem aligns with another one of the downsides of utilitarianism—its potential to erode personal liberties in the pursuit of the "greater good." Just as utilitarian frameworks can justify harmful decisions in the name of overall utility, ARAIs may justify ethically problematic decisions by prioritizing aggregate risk reduction over individual rights. The ethical risk here is not only that ARAIs might misrepresent the likelihood of recidivism for certain individuals but that they could also reinforce systemic inequities, thus undermining the ethical principle of impartiality.

ARAIs are rapidly reshaping sectors like criminal justice, mental health, and insurance, following a legacy of other data-driven technologies that have become central to decision-making in modern society. The work by Hart and Cooke sheds light on both the strengths and the limitations of ARAIs, offering important insights for improving predictive models. Their study showcases that while these tools can enhance decision-making efficiency, they also carry significant ethical risks, such as the potential to commodify human behavior and impose a "common currency of value." As ARAIs gain further adoption, navigating these moral concerns becomes increasingly important. Balancing innovation with ethical responsibility is difficult, but it is imperative to ensure that these models serve as just tools in human-centric decisions rather than as blunt instruments of statistical convenience. Ultimately, achieving this balance is necessary if ARAIs are to contribute positively to the fields they aim to transform.

References

Hart, S.D. and Cooke, D.J. (2013), Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments. Behav. Sci. Law, 31: 81-102. https://doi.org/10.1002/bsl.2049

What is risk assessment: PSRAC. Bureau of Justice Assistance. (n.d.). https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

Lewis, R. A. (n.d.). Best practices for improving the use of criminal justice risk assessments: Insights from NIJ's Recidivism Forecasting Challenge Winners Symposium. National Institute of Justice. https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments