Evaluating Actuarial Risk Assessment Instruments and Their Ethical Implications

Dylan Doby

December 12, 2024

Introduction

Imagine being judged by a single number—a value that determines your freedom, your future, and your place in society. In today's world, actuarial risk assessment instruments (ARAIs) make this scenario a reality. Originally developed for the insurance industry, these statistical models predict the likelihood of future events based on historical data, but they now wield significant influence in criminal justice, mental health, and other high-stakes domains.[1] With decisions about parole, sentencing, and treatment often hinging on these tools, their impact on individuals and communities cannot be overstated.[2] By 2024, ARAIs are projected to influence decisions for over 2 million people in the United States alone.[3] This growing reliance on ARAIs promises efficiency and objectivity but raises profound ethical concerns about fairness, reliability, and the reduction of human behavior to a simple probability. When errors in prediction can drastically affect someone's life—determining their freedom or access to care—precision is not just statistical; it's a moral imperative.

As ARAIs play a major role in shaping legal, insurance, and mental health outcomes, a pressing question emerges: Can a tool with inherent imprecision be ethically justifiable when used to make life-altering decisions? Moreover, is it fair—or even morally acceptable—to assign a numerical value to the complexities of human life and behavior?

In their paper, *Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, Stephen D. Hart and David J. Cooke tackle these questions by examining the reliability of ARAIs in making predictions. Their work highlights fundamental concerns about the wide margins of error in these tools, particularly when applied to individual cases. This paper critiques their methods and, through the lens of utilitarianism and deontology, concurs that while ARAIs may offer practical efficiency at the group level, their application to individual decision-making raises profound moral concerns that challenge their legitimacy.

Analysis of Methods

The methodology of Hart and Cooke's study, *Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, is built on a rigorous statistical analysis of data from Karla Jackson's ongoing research into sexually violent recidivism. The dataset consisted of 90 adult male sex offenders who had completed a community-based treatment program between 2002 and 2004. These participants were referred to an outpatient forensic mental health clinic as part of their probation or parole conditions under the Criminal Code of Canada. The study captured a diverse demographic sample, with

[1] https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment
[2] https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments
[3] *Ibid*

participants ranging in age from 19 to 77, and included detailed treatment records encompassing criminal history, police reports, mental health assessments, and treatment notes.

To assess the likelihood of recidivism, Hart and Cooke employed the Sexual Violence Risk-20 (SVR-20), an actuarial model (specifically a structured professional judgment (SPJ) tool) that evaluates 20 risk factors across four domains: psychological adjustment, social adjustment, history of sexual offenses, and future plans. Each factor was rated on a three-point scale (0 = absent, 1 = possible or partial presence, 2 = definite presence), and evaluators reached consensus ratings to enhance reliability. The aggregated domain scores served as predictor variables in the statistical analysis, with inter-rater reliability coefficients demonstrating strong agreement across domains (e.g., ICC = 0.92 for psychological adjustment).

For their predictive analysis, Hart and Cooke utilized logistic regression, a versatile model that estimates the probability of a binary outcome based on predictor variables. The choice of logistic regression for this context is appropriate given its ability to handle binary outcomes and interpret predictor impacts via odds ratios. However, logistic regression models require certain assumptions to hold, including linearity of the log-odds, independence of observations, and sufficiently large sample sizes to avoid overestimation of odds ratios. Hart and Cooke accounted for some of these challenges by using domain scores from the SVR-20 as predictors, aggregating them into meaningful variables with robust inter-rater reliability. Yet, their relatively small sample size (n = 90) presents significant limitations. Smaller datasets can lead to overestimation of odds ratios, high variance in parameter estimates, and reduced generalizability to broader populations.

Moreover, logistic regression alone does not provide class labels; it predicts probabilities. To transition from probabilities to actionable decisions, a decision rule is necessary. Logistic regression, as employed by Hart and Cooke, models recidivism as a binary outcome (e.g., failure or no failure), capturing the influence of multiple domains (psychological adjustment, social adjustment, history of sexual offenses, and future plans). Each domain's score is treated as a predictor variable, and the model outputs a probability that a given individual will reoffend. Hart and Cooke subsequently used these probabilities to classify individuals into high-risk and low-risk categories. This decision-making process implicitly relies on a threshold, though the paper does not explicitly state the value. A common approach in binary classification is to use a probability threshold of $P(Y \geq 0.5)$, where probabilities greater than or equal to 0.5 are classified as high-risk and probabilities below this threshold are classified as low-risk.[4] Alternatively, the threshold could have been adjusted based on specific considerations, such as minimizing false negatives (failing to identify high-risk individuals) or false positives (misclassifying low-risk individuals as high-risk).

Hart and Cooke's classification process highlights an inherent trade-off in threshold selection: a lower threshold may capture more high-risk individuals but at the cost of increasing false positives, while a higher threshold does the opposite. The study does not clarify whether the threshold was data-driven (e.g., determined by optimizing model performance metrics) or predefined based on clinical judgment.

Hart and Cooke validated their model using receiver operating characteristic (ROC) analysis, with the area under the curve (AUC) calculated at 0.72. This metric reflects the model's moderate ability to distinguish between recidivists and non-recidivists. While an AUC of 0.72 suggests some predictive validity at the group level, the authors extended their evaluation to include confidence intervals (CIs) for group-level estimates and prediction intervals (PIs) for individual-level estimates. At the group level, their model effectively differentiated between high-risk and low-risk categories. For example, the failure rate in the low-risk group was 10%, compared to 33% in the high-risk group, with statistically significant differences (p = 0.006). However, confidence intervals for these estimates were relatively wide, indicating uncertainty in the aggregate predictions and suggesting that larger sample sizes would be necessary for greater precision. At the individual level, the model's limitations became more pronounced. Prediction intervals for individual estimates were significantly wider than the confidence intervals for group-level estimates, often spanning 14 to 64 percentage points. This wide range undermines the model's ability to provide actionable insights for individual decision-making, as even individuals classified as high-risk and low-risk exhibited overlapping prediction intervals. These findings highlight a central critique of actuarial risk assessment instruments: their tendency to perform well in aggregate but falter when applied to individuals.

[4] https://www.yourdatateacher.com/2021/06/14/are-you-still-using-0-5-as-a-threshold/

Building on the methodological foundation established by Hart and Cooke, I will apply their approach to the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset, a widely studied resource in the criminal justice domain.[5] This dataset contains extensive information on criminal recidivism, including demographics, prior offenses, and COMPAS risk scores, making it ideal for testing the reliability of actuarial risk assessment instruments in a different context. By utilizing logistic regression to model recidivism probabilities, similar to Hart and Cooke's approach with the SVR-20, I aim to evaluate the precision of group-level predictions compared to individual-level predictions. This analysis will explore whether the challenges identified in Hart and Cooke's study—such as the limitations of prediction intervals at the individual level—persist in a larger and more diverse dataset. Furthermore, this application allows for a critical examination of the ethical and practical implications of using ARAIs in high-stakes decisions, such as parole and sentencing, while assessing the generalizability of their method to broader contexts. The results will serve to validate, critique, and potentially expand upon Hart and Cooke's findings.

Novel Analysis

To validate the authors' described methodology, I followed a three-stage process: model generation, selection, and validation, using the COMPAS dataset. This dataset, provided a large and diverse sample of 7,214 observations, capturing demographic, criminal history, and risk score information. My analysis, conducted using the R programming language, closely mirrored Hart and Cooke's methodology but extended it to a broader context to test the robustness of their findings.

Stage 1: Model Generation

I began by generating two logistic regression models: a simple model without interaction terms and a more complex interaction model that included second-order interaction terms between predictors (e.g., age, sex, race, priors count, and COMPAS decile scores). The simple model serves as a baseline, while the interaction model evaluates whether higher-order relationships improve predictive accuracy.

Simple Model

$$\text{logit}(P(\text{is\_recid} = 1)) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sexMale} + \beta_3 \cdot \text{race} + \beta_4 \cdot \text{priors\_count} + \beta_5 \cdot \text{decile\_score} + \varepsilon$$

Interaction Model

$$\begin{aligned}\text{logit}(P(\text{is\_recid} = 1)) = \beta_0 &+ \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sexMale} + \beta_3 \cdot \text{race} \\ &+ \beta_4 \cdot \text{priors\_count} + \beta_5 \cdot \text{decile\_score} \\ &+ \beta_6 \cdot (\text{age} \cdot \text{race}) + \beta_7 \cdot (\text{age} \cdot \text{priors\_count}) \\ &+ \beta_8 \cdot (\text{priors\_count} \cdot \text{decile\_score}) + \varepsilon\end{aligned}$$

Table 1: Summary of Model Performance for Stage 1.

Model	Degrees.of.Freedom	AIC
Simple Model	10	8716.57
Interaction Model	36	8640.54

[5] https://github.com/propublica/compas-analysis

The results from the simple model indicated that certain predictors, such as prior convictions ($p < 2e-16$) and decile scores ($p < 2e-16$), were highly significant, aligning with expectations in recidivism prediction. Conversely, predictors such as race and sex showed varying levels of significance, with only specific subcategories (e.g., Hispanic race, Male sex) approaching significance.

Adding interaction terms in the complex model revealed nuanced relationships between variables. For example, the interaction between age and Caucasian race ($p = 0.009$) and priors_count and decile_score ($p < 2e-6$) demonstrated significant contributions to the model. The inclusion of interactions, however, came at the cost of increased model complexity, reflected by a higher number of coefficients and a slight reduction in the AIC value (from 8716.57 in the simple model to 8640.54 in the interaction model).

Stage 2: Model Selection

To simplify the interaction model, I applied backward elimination based on Akaike Information Criterion (AIC), which iteratively removed non-significant terms to improve model parsimony. The final model retained only significant predictors and interactions, reducing complexity while maintaining predictive power. This process ensured that the final model preserved key predictors and interactions without overfitting, enhancing its practical utility in high-stakes applications like criminal justice.

Stage 3: Cross-Validation and Classification

To evaluate the predictive performance of the models, I performed 10-fold cross-validation, splitting the dataset into training (80%) and testing (20%) subsets. The simple model achieved an accuracy of 67.8%, with a sensitivity of 74.5% and specificity of 60.5%. These metrics reflect the model's ability to identify true recidivists and non-recidivists, respectively.

The interaction model marginally improved accuracy to 68.1%, with a sensitivity of 75.1% and specificity of 61.0%. Receiver Operating Characteristic (ROC) analysis further supported these findings, with the area under the curve (AUC) increasing from 0.731 in the simple model to 0.746 in the interaction model. The ROC curve for both models highlights a moderate ability to distinguish between recidivists and non-recidivists. However, the interaction model's higher AUC showcases its slight improvement in predictive accuracy at the cost of complexity.

Table 2: Confusion Matrix for Simple Model

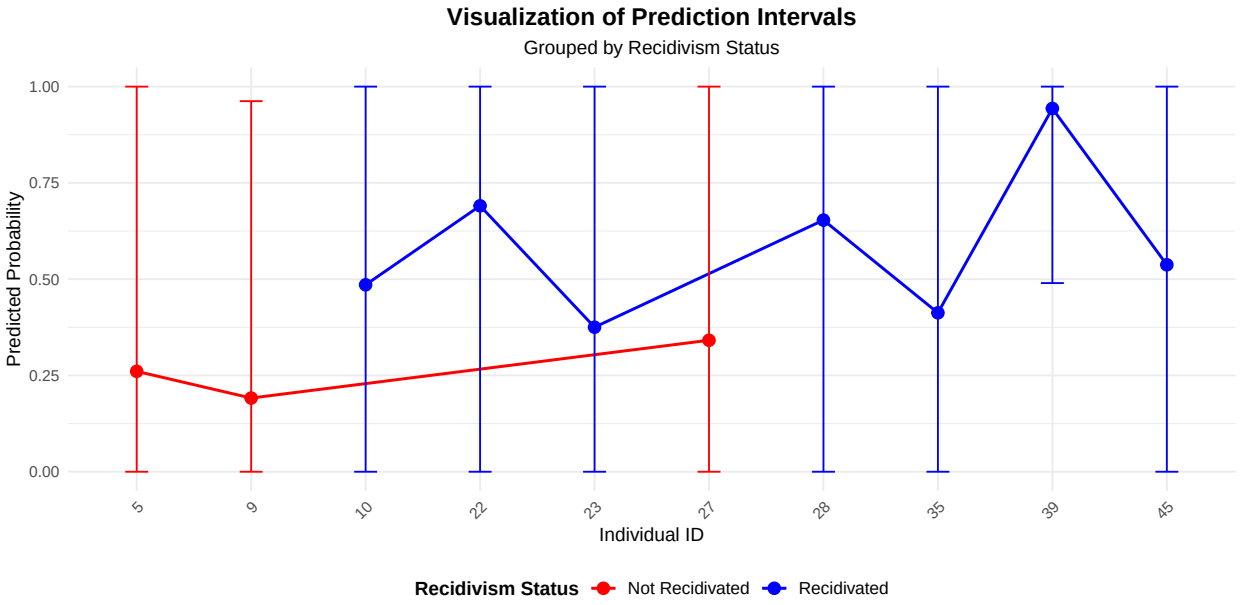
	Predicted: No	Predicted: Yes
Actual: No	557	191
Actual: Yes	274	420

In line with Hart and Cooke's methodology, I used a probability threshold of 0.5 to classify individuals as high-risk (probability 0.5) or low-risk (probability < 0.5). This decision was driven by its widespread use in binary classification tasks. However, the choice of threshold is inherently context-dependent, balancing false positives (incorrectly classifying low-risk individuals as high-risk) against false negatives (failing to identify high-risk individuals).

To address these trade-offs, I computed group-level confidence intervals and individual prediction intervals for the decile scores. High-risk individuals exhibited an average decile score of 5.56, compared to 3.57 for low-risk individuals. While group-level predictions showed statistically significant differences, individual prediction intervals were wide, often overlapping, undermining their utility for case-specific decisions.

Table 3: Group-Level Confidence Intervals for Decile Scores

Group	Mean_Decile_Score	CI_Lower	CI_Upper
Low Risk	3.57	3.47	3.66
High Risk	5.56	5.46	5.67



Critique of Methods

A significant methodological issue in Hart and Cooke's study lies in the lack of alignment between the predictors they claim to have used and the data they cite. For instance, while their analysis ostensibly includes 20 factors from the SVR-20, the aggregation of these into domain scores obscures how individual predictors contribute to the model. This aggregation choice simplifies the analysis but forfeits transparency and granularity, which are important for understanding the dynamics at play. The omission of specifics about individual predictors raises questions about the integrity of the model's construction, particularly given the importance of accurately assessing risk factors in high-stakes decision-making.[6]

Another troubling aspect is the unclear treatment of interactions and second-order terms in their logistic regression model. While Hart and Cooke indicate that their model incorporates various domain scores, they fail to adequately describe whether interaction effects between these domains were tested or included. This oversight is significant, as interactions could capture how different risk factors combine to predict recidivism. The absence of such details leaves open the possibility that their model oversimplifies complex relationships, thereby reducing its explanatory power and predictive accuracy.

Hart and Cooke's reliance on a relatively small sample size (n = 90) compounds these concerns. Small datasets are susceptible to overfitting, where the model captures noise rather than meaningful patterns. While the authors address inter-rater reliability to enhance the robustness of their domain scores, they do not discuss the potential for overfitting in their logistic regression model. This omission is significant given that overfitting can inflate performance metrics during model validation, leading to overly optimistic conclusions about the model's utility.[7]

The authors validate their model using ROC analysis and report an AUC of 0.72, which suggests moderate discriminative ability. However, this performance metric primarily evaluates group-level predictions and does not address the wide prediction intervals observed for individual cases. Hart and Cooke acknowledge these wide intervals, which often span 14 to 64 percentage points, but they fail to explore how this imprecision undermines the model's practical applicability in individual decision-making contexts. Without clearer

[6] https://doi.org/10.1002/bsl.2050

[7] https://journals.lww.com/psychosomaticmedicine/Abstract/2004/05000/What_You_See_May_Not_Be_What_You_Get___A_Brief,.6.aspx

threshold selection criteria or methods for narrowing prediction intervals, their model risks misclassification and unjust outcomes.

Finally, Hart and Cooke's methodology misses an opportunity to employ alternative modeling techniques that could address the identified shortcomings. For instance, ensemble methods like random forests or gradient boosting could improve predictive performance while offering richer insights into variable importance. Bayesian models could provide more robust uncertainty estimates, helping to contextualize the imprecision of individual-level predictions. The authors' exclusive reliance on traditional logistic regression—without testing these alternatives—limits the scope of their findings and leaves critical methodological gaps unaddressed.[8]

Despite these shortcomings, Hart and Cooke's study highlights important challenges in using actuarial risk assessment tools. Their work sheds light on the difficulties of making accurate predictions for both groups and individuals, raising questions about fairness and reliability in high-stakes decisions. While there are clear flaws in their methods, the study serves as a meaningful starting point for improving these tools.

Analysis of Normative Considerations

The widespread adoption of actuarial risk assessment instruments (ARAIs), particularly in fields such as criminal justice and mental health, has sparked significant ethical debates. Through the lens of utilitarian philosophy, which argues that actions should maximize overall well-being, ARAIs can be defended as a means of improving efficiency and fairness in decision-making. However, their limitations and potential harms—particularly when applied to individual cases—challenge this justification. In this section, I examine how ARAIs align with utilitarian and deontological ethical frameworks, ultimately arguing that their current use in high-stakes decision-making often fails to uphold these principles.

ARAIs promise to enhance social welfare by offering consistent, data-driven evaluations of risk. This objective aligns with utilitarian goals, as tools like the Sexual Violence Risk-20 (SVR-20) aim to reduce recidivism rates and promote public safety. Studies suggest that ARAIs can outperform human judgment in some contexts, providing standardized risk scores that minimize bias and improve decision accuracy. For example, data-driven tools have been shown to reduce subjectivity in parole decisions, potentially leading to more equitable outcomes and greater societal trust in justice systems.[9] When used at the group level, these tools can help allocate resources, such as mental health services or rehabilitation programs, to those who need them most. From a utilitarian perspective, such benefits may justify their implementation—provided that the collective gains outweigh the harms to individuals.

However, applying ARAIs to individual cases raises profound ethical concerns. The imprecision of these instruments, as highlighted by Hart and Cooke, often results in wide prediction intervals that undermine their reliability in case-specific contexts.[10] For instance, an individual classified as "high risk" might have a predicted recidivism probability ranging from 14% to 64%. Such variability not only questions the accuracy of predictions but also risks unjust outcomes, such as unwarranted denial of parole or access to critical resources. These consequences diverge from utilitarian ideals by inflicting disproportionate harm on individuals without clear evidence of societal benefit.

Deontological ethics, which emphasize the intrinsic rights and dignity of individuals, further complicate the moral justification of ARAIs. Assigning numerical risk scores to individuals inherently reduces them to probabilities, ignoring the complexities of human behavior and individual circumstances. This approach violates Kantian principles of treating people as ends in themselves, rather than as means to an institutional goal. For example, a person labeled as "high risk" might face significant stigma or discrimination, regardless of whether the prediction is accurate.[11] Such outcomes conflict with the deontological imperative to respect individual autonomy and fairness.

[8] https://web.stanford.edu/~hastie/ElemStatLearn/

[9] https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments

[10] https://doi.org/10.1002/bsl.2055

[11] https://journals.sagepub.com/doi/full/10.1177/1526443721997057

The reliance on ARAIs also introduces systemic ethical challenges, particularly in reinforcing existing biases. Many ARAIs rely on historical data that may reflect structural inequalities, such as over-policing in minority communities or disparities in mental health diagnoses. By incorporating these biases into their algorithms, ARAIs risk perpetuating discriminatory practices under the guise of objectivity.[12] For instance, studies of the COMPAS algorithm have revealed significant racial disparities in its predictions, with higher false-positive rates for Black defendants compared to White defendants.[13] These findings raise serious questions about whether ARAIs can be ethically justified in systems that already struggle with equity and fairness.

Despite these criticisms, ARAIs hold significant potential if implemented responsibly. For instance, integrating complementary tools, such as structured professional judgment (SPJ) methods, could address some of their limitations by incorporating contextual factors and human oversight. Additionally, adopting ethical guidelines and rigorous validation protocols could mitigate issues related to bias and imprecision. Transparency in how risk scores are calculated and used is crucial for fostering trust and ensuring accountability in high-stakes decisions.[14]

Ultimately, the ethical application of ARAIs hinges on their ability to balance societal benefits with individual rights. While these tools can contribute to public safety and resource allocation, their limitations necessitate cautious and context-sensitive use. By addressing concerns related to fairness, transparency, and precision, ARAIs have the potential to align more closely with both utilitarian and deontological ideals, promoting justice while minimizing harm. Until such safeguards are widely implemented, however, their use in individual decision-making remains ethically fraught.

Conclusion

Impact of Paper

Stephen D. Hart and David J. Cooke's paper, *Another Look at the (Im-)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, makes a significant contribution to the study of actuarial risk assessment tools by critically examining their precision and ethical implications. Their analysis addresses an essential gap in the literature by highlighting the limitations of these instruments, particularly their wide prediction intervals and the challenges of applying group-level models to individual cases.

The impact of their work extends beyond academic discourse, as it calls for greater scrutiny and refinement of the tools used in high-stakes decision-making contexts like criminal justice and mental health. By emphasizing the trade-offs between efficiency and fairness, Hart and Cooke's research highlights the need for methodologies that balance societal benefits with individual rights. Their findings lay a foundation for developing more transparent, reliable, and ethically sound risk assessment tools, potentially influencing both policy and practice.

Moreover, this paper sets a benchmark for future research by proposing a strong framework for evaluating the precision of risk estimates. This contribution is vital for fostering innovation in predictive modeling, as it encourages the adoption of alternative techniques that address biases and enhance predictive accuracy. Essentially, Hart and Cooke's work provides an important starting point for rethinking how actuarial tools can be responsibly integrated into systems that impact human lives, promoting justice and equity in their application.

Future Work & Wraap-Up

Building on the insights from Hart and Cooke's study, future research into actuarial risk assessment instruments (ARAIs) should focus on addressing the methodological and ethical challenges identified. A key

[12] https://doi.org/10.1037/lhb0000145
[13] https://github.com/propublica/compas-analysis
[14] https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

priority is testing the precision of ARAIs across larger and more diverse datasets, which would help ensure their reliability and applicability across various populations. Such research could involve developing models that account for intersectional factors—such as race, gender, and socioeconomic background—to better understand how these variables interact in predicting outcomes.

Another promising avenue for exploration involves integrating alternative predictive techniques, such as ensemble methods or Bayesian approaches, to enhance the accuracy and transparency of risk predictions. These methods could mitigate issues like overfitting and provide richer uncertainty estimates, which are particularly critical in high-stakes decision-making contexts. Moreover, future studies could examine the potential for real-time adaptation of ARAIs based on evolving individual or situational factors, offering a more dynamic and nuanced approach to risk assessment.

Ethical considerations should remain central to these advancements. Research that incorporates input from ethicists, practitioners, and impacted communities could help shape policies and practices that prioritize fairness and justice. Investigations into how ARAIs are perceived and experienced by those subject to their predictions could also inform guidelines for their implementation, ensuring that these tools are used responsibly and equitably.

Hart and Cooke's work provides a supremely important foundation for understanding both the potential and the limitations of ARAIs. Their study highlights the urgent need for refinement in the design, application, and ethical oversight of these tools. By addressing methodological flaws, embracing innovative modeling techniques, and centering ethical considerations, future research can help ensure that ARAIs serve as instruments of fairness and precision rather than perpetuators of systemic bias. Ultimately, this will enable the responsible integration of these tools into decision-making processes, aligning their use with societal values and individual rights.

References

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine, 66(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bureau of Justice Assistance. (n.d.). What is risk assessment? Retrieved [Month Day, Year], from https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

Hart, S. D., & Cooke, D. J. (2022). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. Behavioral Sciences & the Law. https://doi.org/10.1002/bsl.2055

Harris, G. T., & Rice, M. E. (2007). Adjusting actuarial violence risk assessments based on aging or the passage of time. Behavioral Sciences & the Law, 25(6), 831–845. https://doi.org/10.1002/bsl.2050

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. Retrieved from https://web.stanford.edu/~hastie/ElemStatLearn/

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). ProPublica COMPAS Analysis. GitHub. https://github.com/propublica/compas-analysis

National Institute of Justice. (n.d.). Best practices for improving the use of criminal justice risk assessments. Retrieved [Month Day, Year], from https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, & recidivism: Predictive bias and disparate impact. Law and Human Behavior, 41(3), 258–271. https://doi.org/10.1037/lhb0000145

Your Data Teacher. (2021, June 14). Are you still using 0.5 as a threshold? Retrieved from https://www.yourdatateacher.com/2021/06/14/are-you-still-using-0-5-as-a-threshold/