Psychological Assessment

Statistical Learning Methods and Cross-Cultural Fairness: Trade-Offs and Implications for Risk Assessment Instruments

Linda J. Ashford, Benjamin L. Spivak, James R. P. Ogloff, and Stephane M. Shepherd

Online First Publication, March 2, 2023. https://dx.doi.org/10.1037/pas0001228

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

https://doi.org/10.1037/pas0001228

Statistical Learning Methods and Cross-Cultural Fairness: Trade-Offs and Implications for Risk Assessment Instruments

Linda J. Ashford, Benjamin L. Spivak, James R. P. Ogloff, and Stephane M. Shepherd
Centre for Forensic Behavioural Science, Swinburne University of Technology

The use of statistical learning methods has recently increased within the risk assessment literature. They have primarily been used to increase accuracy and the area under the curve (AUC, i.e., discrimination). Processing approaches applied to statistical learning methods have also emerged to increase cross-cultural fairness. However, these approaches are rarely trialed in the forensic psychology discipline nor have they been trialed as an approach to increase fairness in Australia. The study included 380 Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander males assessed with the Level of Service/Risk Needs Responsivity (LS/RNR). Discrimination was assessed through the AUC, and fairness was assessed through the cross area under the curve (xAUC), error rate balance, calibration, predictive parity, and statistical parity. Logistic regression, penalized logistic regression, random forest, stochastic gradient boosting, and support vector machine algorithms using the LS/RNR risk factors were used to compare performance against the LS/RNR total risk score. The algorithms were then subjected to pre- and postprocessing approaches to see if fairness could be improved. Statistical learning methods were found to produce comparable or marginally improved AUC values. Processing approaches increased several fairness definitions (namely xAUC, error rate balance, and statistical parity) between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The findings demonstrate that statistical learning methods may be a useful approach to increasing the discrimination and cross-cultural fairness of risk assessment instruments. However, both fairness and the use of statistical learning methods encompass significant trade-offs that need to be considered.

Public Significance Statement

The purpose of this study is to investigate the use of algorithms in increasing fairness between cultural groups by utilizing a commonly used instrument that assesses the risk of reoffending. Although this preliminary study demonstrated positive findings in increasing cross-cultural fairness, various trade-offs emerged, including among different forms of fairness, that require thoughtful deliberation.

Keywords: fairness, risk assessment, statistical learning methods, cross-cultural

Supplemental materials: https://doi.org/10.1037/pas0001228.supp

Risk assessment instruments within the criminal justice system involve estimating the likelihood of reoffending. These instruments are used to inform offender management decisions, including parole, bail, supervision, and treatment (Heilbrun et al., 2010). Risk was previously assessed intuitively by clinicians and has since advanced into numerous structured instruments that aid in estimating future risk (Monahan & Skeem, 2014). Current risk assessment instruments primarily include actuarial and structured professional judgment (SPJ) instruments. Actuarial risk assessment instruments are scored by a formula or algorithm, combining numerical values

Linda J. Ashford https://orcid.org/0000-0003-2617-5645
Benjamin L. Spivak https://orcid.org/0000-0002-9051-3349
James R. P. Ogloff https://orcid.org/0000-0002-3137-5556
Stephane M. Shepherd https://orcid.org/0000-0002-3078-9407

The authors thank the members of the Catalyst Consortium, notably Rachael Fullam and Janet Ruffles for their contribution to the generation of the data used in this article.

This research was supported by funding from the Australian Research Council (Project: DE180100933), the Victorian Corrections Minister, and the Victorian Institute of Forensic Mental Health.

Linda J. Ashford played a lead role in formal analysis, writing of original draft, and writing of review and editing and an equal role in conceptualization and methodology. Benjamin L. Spivak played a supporting role in writing of review and editing and an equal role in conceptualization, methodology, and supervision. James R. P. Ogloff played a supporting role in writing of review and editing and an equal role in funding acquisition. Stephane M. Shepherd played a supporting role in writing of review and editing and an equal role in conceptualization, funding acquisition, and supervision.

This study was not preregistered; data and study materials are not publicly available.

Study analysis code is available within the Supplemental Materials.

Correspondence concerning this article should be addressed to Linda J. Ashford, Centre for Forensic Behavioural Science, Swinburne University of Technology, 1/582 Heidelberg Road, Alphington, VIC 3078, Australia. Email: lashford@swin.edu.au

assigned to evidence-based risk factors (Singh, 2012). SPJ instruments assist clinicians in determining an individual's level of risk by providing guidelines for factors empirically related to offending (Hart et al., 2017). Both instrument types have been assessed for their utility, often by observing the instrument's ability to distinguish individuals who reoffend from those who do not. This is referred to as discrimination and has been frequently assessed by the area under the curve (AUC) in the risk assessment literature (Singh et al., 2013). Meta-analytic and systematic reviews have often highlighted that risk assessment instruments are moderately effective in their ability to discriminate between individuals who reoffend from those who do not (Fazel et al., 2012; Singh et al., 2011).

Statistical Learning Methods in Risk Assessment

Recently, the use of statistical learning methods (i.e., machine learning algorithms) has increased in the area of risk assessment as an approach to increasing predictive accuracy and discrimination (Spivak & Shepherd, 2020). The aim of statistical learning methods differs from traditional approaches to risk estimation. Traditional approaches entail meticulous theorizing and prespecifying of predictor interactions, as well as comprehending the nature of the relationship between predictors and the outcome. The form of relationship and predictor interactions are not required to be prespecified with statistical learning methods. Instead, more complex statistical learning methods can use a large number of predictors and take advantage of nonlinear relationships and the predictive capacity of both strong and weak predictors, as well as their interactions. This can aid in achieving the aim of statistical learning methods, which is ultimately to maximize predictive accuracy and discrimination (Berk & Bleich, 2013; Spivak & Shepherd, 2020).

Within the risk assessment literature, more complex statistical learning methods (e.g., random forests, gradient boosting, and neural networks) have primarily been used to achieve this aim of increasing predictive validity and discrimination over more traditional and straightforward statistical learning methods (e.g., logistic regression) or an existing risk assessment instrument. This has resulted in mixed findings. Tollenaar and van der Heijden (2019) found that logistic regression performed as well as more complex statistical learning methods, with no improvement in AUC. Similarly, Liu et al. (2011) reported comparable AUC values across logistic regression and more complex statistical learning methods when using the Historical, Clinical, and Risk Management–20 (HCR-20) items. These approaches also did not lead to an improvement in the discrimination of the original HCR-20 risk score for violent reoffending. When examining a variety of complex statistical learning methods and logistic regression in predicting felonies, drug, violent, and sexual reoffending, Hamilton et al. (2015) noted that logistic regression was a better discriminator, most notably for violent and sexual reoffending.

Comparatively, Ting et al. (2018) used the Youth Level of Service/Case Management Inventory (YLS/CMI) items in a random forest algorithm to yield an AUC of .69 with Singapore youth, a marginal improvement on previous research from Singapore that produced an AUC of .64 when using the YLS/CMI risk score (Chu et al., 2015). Duwe and Kim (2015) also reported positive findings when using a random forest model to predict reoffending, with random forests having the highest AUC across 12 different algorithms, including logistic regression. However, this difference was again minimal. Ghasemi et al. (2020) used the Level of Service/Case Management Inventory (LS/CMI) as predictor in complex statistical learning methods and found that the AUC was comparable to the AUC of the original LS/CMI. The statistical learning methods, on the other hand, were found to be better discriminators along the LS/CMI's middle scores, with average AUCs for individual scores improving from .50 to near .60. Breitenbach et al. (2009) explored the performance of different statistical learning methods when using all possible predictors and a subset of predictors. When all possible predictors were used, random forests had the highest AUC for violent reoffending (.70) compared to logistic regression (.63). However, when using a subset of predictors, logistic regression was found to outperform random forests. Salo et al. (2019) compared the performance of more complex statistical learning methods to logistic regression using different sets of static and dynamic predictors from the Finnish Needs and Risk Assessment Form. The complex algorithms outperformed logistic regression across all sets of items, with logistic regression often producing a higher AUC when using a smaller subset of items. These latter studies, which showed more promising results with increased discrimination, frequently used a significantly larger number of predictors. This demonstrates one of the benefits of more complex statistical learning methods in that they can exploit a large number of predictors to improve predictive accuracy and discrimination.

Fairness

The use of statistical learning methods has also extended to the complex issue of fairness in risk assessment instruments (Berk et al., 2021). The debate concerning fairness in risk assessment has been increasing over the past decade, with critics of risk assessment arguing that the instruments could disadvantage certain cultural groups (Day et al., 2018; Hart, 2016). Disciplines such as computer science and statistics have established a more nuanced understanding of what constitutes statistical fairness (Verma & Rubin, 2018). Specifically, definitions including statistical parity, error rate balance, calibration, and predictive parity have been receiving increasing attention within the literature. Statistical parity is achieved when the distribution of risk scores is equal across different groups (Berk et al., 2021). Error rate balance refers to the false positive rate (FPR), or the proportion of individuals who do not reoffend being incorrectly predicted to reoffend (or classified as high risk), and the false negative rate (FNR), or the proportion of individuals who do reoffend being incorrectly predicted to not reoffend (or classified as low risk), being the same across groups (Chouldechova, 2017). Calibration refers to a risk instrument's predicted probability (or risk scores or classifications) aligning with actual reoffending (Chouldechova, 2017). Across different groups, predicted probabilities should reflect the same level of reoffending. Last, predictive parity refers to the positive predictive value (PPV), the proportion of those predicted to reoffend who go on to reoffend, and the negative predictive value (NPV), the proportion of those not predicted to reoffend who do not go on to reoffend, being the same across groups (Berk et al., 2021).

The utility of these instruments, which is frequently assessed by discrimination indices such as the AUC, has often demonstrated

comparable findings between minority and majority cultural groups (Skeem & Lowenkamp, 2016; Wormith et al., 2015). However, these other definitions of fairness have often highlighted unfairness among particular cultural minority groupings (e.g., African Americans and Indigenous populations of North America and Australia) and cultural majority groupings on risk assessment instruments such as the Level of Service instruments (e.g., Level of Service Inventory, Level of Service/Risk–Needs–Responsivity [LS/RNR]). Some cultural minority groupings are often found to score significantly higher (Ashford et al., 2022; Olver et al., 2014), be classified as high risk yet not reoffend more often (i.e., have a higher FPR; Ashford et al., 2022), and are predicted to reoffend more, especially across lower risk scores and risk classifications (Wilson & Gutierrez, 2014). Risk assessment instruments are also often better at predicting reoffending among cultural minority groupings (i.e., a higher PPV), whereas among cultural majority groupings, risk assessment instruments are often better at predicting nonreoffending (i.e., a higher NPV; Shepherd et al., 2015) and are more likely to classify cultural majorities who reoffend as low risk (i.e., a higher FNR; Ashford et al., 2022).

Increasing Fairness

The disparities identified across various fairness definitions have led to the use and exploration of statistical learning methods as a way of increasing fairness (Berk et al., 2021). Specifically, by altering the statistical learning method through different stages of its construction and execution, unfairness can be, to a degree, ameliorated. The statistical learning method can be altered in three stages. Preprocessing involves altering the original data to remove or reduce any potential causes of unfairness. For example, the protected variable (e.g., culture) can be used to predict each of the predictor variables, and the residuals are then used in place of the original predictor variable. In-processing involves altering the statistical learning method itself so that it contains no unfair decision rules that may impact a specific group, such as having separate statistical learning method algorithms for each group. Postprocessing involves altering the predictions made by the statistical learning method to improve fairness. To help achieve group equivalence, predicted outcomes might be randomly reassigned, for instance, so that the group with a higher base rate of reoffending would have its "predicted to reoffend" outcomes that fall closer to the cutoff value (i.e., the value that separates those who are predicted to reoffend from those who are not) altered to "predicted to not reoffend," and vice versa.

In the risk assessment literature, a number of preprocessing approaches have been trialed with often promising results. Berk (2019) equalized base rates of offending across cultural groups when using statistical learning methods and found that error rate balance and predictive parity were improved among African Americans and Caucasian young offenders. Another preprocessing approach trialed by Johndrow and Lum (2017) involved transforming predictors to achieve independence from an individual's culture. This resulted in a reduction in FPR differences among African Americans, Caucasians, and Hispanics but also a slightly increased disparity among PPVs and NPVs. Lum and Johndrow (2016) also found that this approach barely impeded the AUC, with the unadjusted data yielding an AUC of .71 and the adjusted data increasing it slightly to .72.

Skeem and Lowenkamp (2020) used a similar approach that involved each predictor being regressed onto culture and the residuals being used in place of the original predictors. When the statistical learning method algorithm with residuals was compared to an algorithm that only included the Post Conviction Risk Assessment items, the AUC was reduced from .72 to .71. Parity among calibration and PPVs was also mildly impeded, with disparities being more pronounced when using residuals. However, FPR disparities were improved when using residuals, with a difference of 7.21% between African Americans and Caucasians being reduced to −3.65%.

In-processing and postprocessing are less often utilized. However, these approaches have still demonstrated an ability to increase fairness. For example, Wadsworth et al. (2018) used an in-processing approach that led to an increase in discrimination, with the AUC increasing from .66 to .70 when compared to the original Correctional Offender Management Profiling for Alternative Sanctions instrument score. Further, Wadsworth et al. (2018) found that both error rate balance and statistical parity were improved, with FPR differences between African American and Caucasian individuals being reduced from 17% to 1%, FNR differences being reduced from 22% to 2%, and statistical parity differences being reduced from 18% to 2%.

Trade-Offs

There are inherent trade-offs that exist when attempting to achieve multiple forms of fairness or predictive accuracy and fairness simultaneously. It has been established that total fairness (i.e., achieving all forms of fairness) is impossible and that an impossibility theorem exists among different types of fairness (e.g., error rate balance and predictive parity) being achieved concurrently when the base rates of reoffending differ (Berk et al., 2021; Chouldechova, 2017). There is also the issue of an instrument's utility alongside its fairness. Changing statistical learning method algorithms, which have traditionally been optimized for accuracy, could have a negative impact on a risk assessment instrument's overall predictive utility. For example, altering predictors in a preprocessing step (e.g., using residuals) that are usually valid and significant predictors of reoffending could impede the overall accuracy of predictions. This is also relevant for an instrument's ability to discriminate between an individual who reoffends and an individual who does not reoffend. However, previous research using preprocessing approaches such as residuals has reported no notable changes to the discrimination of the instrument (Skeem & Lowenkamp, 2020).

Last, a common critique regarding the more complex and flexible statistical learning methods is that the interpretability and transparency of certain algorithms are often reduced compared to more traditional statistical learning methods such as linear or logistic regression (Breiman, 2001b). These more complex approaches and the use of processing can often result in an algorithm in which the direct relationship between the predictors and the outcome is unclear. Therefore, another trade-off exists between the interpretability and the flexibility of an algorithm. This trade-off results in a practical issue for clinicians assessing risk, as they may be unable to ascertain the specific risk factors or items that are most predictive of a future offense and therefore are unable to intervene and respond to the relevant needs of the individual.

The Present Study

Although these trade-offs are unavoidable in the pursuit of fairness in risk assessment instruments, finding publicly acceptable trade-offs is an avenue worth exploring. However, the application of statistical learning methods has been scarcely applied within the forensic psychology discipline as an approach to increasing cross-cultural fairness (Spivak & Shepherd, 2020) nor has it been explored as an approach for increasing fairness with cultural minority groups in Australia. In Australia, Aboriginal and Torres Strait Islanders already experience inequality within the criminal justice system, such as significant overincarceration and a decreased likelihood of receiving a diversion (Australian Bureau of Statistics, 2020b; Papalia et al., 2019). Further, a recent study found that Aboriginal and Torres Strait Islanders scored notably higher, and nonreoffenders were more likely to be classified as high risk when compared to non-Aboriginal and Torres Strait Islanders on the LS/RNR (Ashford et al., 2022).

Therefore, the present study acted as an exploratory and preliminary study that aimed to examine a variety of statistical learning methods to explore their ability to increase the discrimination (i.e., AUC) of the LS/RNR. This study also sought to improve various forms of fairness between Aboriginal and Torres Strait Islander peoples and non-Aboriginal and Torres Strait Islander peoples by employing statistical learning methods and modifying them during pre- and postprocessing. Last, the study aimed to explore the impact that different processing alterations had on the discrimination of the statistical learning methods and the trade-offs across different types of fairness.

Method

Sample

The sample comprised 380 male individuals who were previously sentenced to a term of imprisonment for a serious violence offense as defined in Schedule 1 (Clause 3) of the *Sentencing Act* 1991 (Vic) from Victoria, Australia, and received into prison during the period of January 2015 to December 2017. These individuals were assessed with the LS/RNR by corrections officers while serving either a prison sentence ($n = 231$, 60.79%), a community corrections order (i.e., a flexible sentencing order that is served in the community, $n = 148$, 38.95%), or a parole order ($n = 1$, 0.26%). Those who were assessed during a prison sentence had since been released to accurately reflect those who were at risk to the community. The sample included 180 (47.37%) individuals who identified as Aboriginal and Torres Strait Islander peoples and 200 (52.63%) who identified as non-Aboriginal and Torres Strait Islanders.

Aboriginal and Torres Strait Islanders were oversampled in the present study to enable comparisons between groups. Specifically, all Aboriginal and Torres Strait Islanders who were received into prison for a serious violent offense within the study recruitment period were eligible to be sampled. Non-Aboriginal and Torres Strait Islanders were then randomly sampled to have equal sample sizes across both groups. Due to the small sample size and lack of demographic information, the non-Aboriginal and Torres Strait Islander group was unable to be portioned into more distinct cultural groups. The vast majority of the non-Aboriginal and Torres Strait Islander group self-reported that they were born in Australia ($n = 166$, 83%), and all identified their primary language as English.

LS/RNR completions, demographics, dates of incarceration, community correction orders, and parole orders were provided by Corrections Victoria. Information regarding charges post-LS/RNR assessment was obtained from the Victorian Police Law Enforcement Assistance Program database for the period of January 2015 through December 2019. The Department of Justice and Community Safety (Victoria) Human Research Ethics Committee and the Swinburne University Human Research Ethics Committee provided ethical approval for the present study.

Measures

Level of Service/Risk Needs Responsivity

The LS/RNR (Andrews et al., 2008) is an actuarial instrument that was developed to estimate an individual's risk of general reoffending and identify their criminogenic needs. The General Risk/Needs section includes 43 risk items and has eight factors: criminal history, education/employment, family/marital, leisure/recreation, companions, alcohol/drug problem, procriminal attitude, and antisocial pattern. All items within these factors produce a score of either 0 when absent or 1 when present. They are summed to create factor scores and a total risk score. Individuals can be classified into risk levels based on their total score, which includes very low risk (0–4), low risk (5–10), medium risk (11–19), high risk (20–29), and very high risk (30–43). The eight risk factor scores from the General Risk/Needs section were used as predictors in the algorithms for the present study and compared to findings from the LS/RNR total risk score. The Level of Service instruments have been well validated, with total risk scores often demonstrating an acceptable level of predictive validity (i.e., moderate-to-large in effect size) in predicting reoffending and discriminating individuals who reoffend from individuals who do not (Olver et al., 2014; Wilson & Gutierrez, 2014; Wormith et al., 2015).

Reoffending

Reoffending was defined as any police charge while at risk in the community (i.e., not during a period of incarceration). The follow-up for the present study was from the LS/RNR assessment (or release date for those incarcerated) to either the date of the first charge for those who reoffended or the end of the follow-up period for those who did not reoffend (December 31, 2019). The average follow-up time for the sample was 280.56 days ($SD = 329.24$). Most of the sample had generally reoffended by the end of the follow-up period ($n = 306$, 80.53%), for which the average time from LS/RNR assessment to the first offense was 184.75 days ($SD = 233.80$). A larger proportion of Aboriginal and Torres Strait Islanders reoffended ($n = 154$, 85.56%) compared to non-Aboriginal and Torres Strait Islanders ($n = 152$, 76%). This difference, although statistically significant, was small in effect size, $\chi^2(1) = 4.92$, $p = .03$, and Cramer's $V = .11$.

Analytical Approach

All data were analyzed through RStudio using R Version 4.0.2 (R Core Team, 2021). Numerous packages were used including the *tidyverse* packages (Version 1.3.0; Wickham, 2019) for data cleaning and management, *pROC* (Version 1.16.2; Robin et al., 2020) to

generate receiver operating characteristic (ROC) curves and AUC values, *caret* (Version 6.0-88; Kuhn, 2021) for model training and cross-validation, *glmnet* (Version 4.1-2; Friedman et al., 2021) for penalized logistic regression, *randomForest* (Version 4.6-14; Liaw & Wiener, 2018) for random forest algorithms, *gbm* (Version 2.1.8; Greenwell et al., 2020) for stochastic gradient boosting, *e1071* (Version 1.7-8; Meyer et al., 2021) for support vector machine algorithms, and *cutpointr* (Version 1.1.1; Thiele, 2021) to generate optimal cutoffs. The R script for the analyses can be found within the Supplemental Materials.

The following sections detail the statistical learning methods used in the present study and the process approaches applied. The performance of the LS/RNR total risk score and statistical learning methods were determined by examining the AUC and various fairness definitions (cross area under the curve [xAUC], calibration, predictive parity, error rate balance, and statistical parity), all of which are outlined in the following sections.

Statistical Learning Methods

To train the statistical learning methods, the data were randomly split into a training set ($n = 229$, 60%) and a testing set ($n = 151$, 40%). To account for overfitting (i.e., when an algorithm also picks up on the unique noise of the sample and has poorer accuracy when used on a new sample) and a small sample size, bootstrapping with 1,000 resamples was used with the training data set. For each algorithm (other than logistic regression), a variety of parameters were tested to determine which produced the best performance (i.e., the highest AUC). As most of the sample reoffended, upsampling, or sampling with replacement from the minority class (i.e., those who did not go on to reoffend) was also used when training the algorithms to account for the imbalanced outcome data. The parameters that produced the highest AUC were retained as the final model, and this was then used to predict reoffending in the held-out testing sample.

The statistical learning methods for the present study included logistic regression, penalized logistic regression, random forest, stochastic gradient boosting, and support vector machines, which are detailed in turn below. The predictors used for each statistical learning method included the eight risk scores from each of the LS/RNR General Risk/Needs factors: criminal history, education/employment, family/marital, leisure/recreation, companions, alcohol/drug problem, procriminal attitude, and antisocial pattern. For descriptive information about the eight risk factors and preliminary information on the importance of the predictor variables in each of the algorithms, please refer to the Supplemental Materials.

Logistic Regression. Logistic regression was used as a baseline comparison against which the performance of other, more complex, and flexible statistical learning methods was evaluated.

Penalized Logistic Regression. Penalized logistic regression was used to improve the predictive power of logistic regression by increasing the simplicity of the model and reducing overfitting and the impact of collinearity (Zou & Hastie, 2005). Specifically, elastic net regression (Zou & Hastie, 2005) was used as it combines both ridge regression and lasso regression. The former imposes a penalty term on the squared size of the coefficients and shrinks irrelevant predictor coefficients toward zero. The latter imposes a penalty on the absolute value of the coefficients and shrinks irrelevant predictor coefficients completely to zero.

Random Forests. A random forest (Breiman, 2001a) is an ensemble-based algorithm (i.e., a combination of numerous algorithm predictions) of decision trees. Each tree is grown on a new training set in which only a random subset of features is tried through each split in the tree. This introduces randomness to the tree construction process and helps to minimize the correlation between trees and improve accuracy. Once the ensemble of decision trees (i.e., forest) has been generated, the predictions are aggregated and result in an overall predicted probability.

Stochastic Gradient Boosting. Stochastic gradient boosting (Friedman, 2002) is a consecutive learning process in which a weak learner (i.e., a learner that predicts slightly better than random) is applied repeatedly to the data. It seeks to find an additive algorithm that will minimize the loss function (e.g., squared error). Initially, specified predicted values are utilized (e.g., this can be the average), so that the residual can be established between that predicted value and the observed value. Then, using a random subsample of the training data, a weak learner (e.g., a decision tree) is grown to fit the residuals, and the algorithm is then used to predict that subsample. The predicted values are then updated by adding the newly predicted values to the previously predicted values. This continues for a specified number of iterations, with new decision trees being grown to fit the residuals of previous trees (i.e., the difference between the most recent predicted value and the observed value), and new predicted values being added to the previous. Like random forests, the final prediction is based on an ensemble of trees; however, with gradient boosting, the trees are not created independently nor are they equal in their contribution to the outcome. Instead, each tree is dependent on past trees and is weighted depending on their performance and how much influence they have over the outcome. The use of a random subsample helps increase the accuracy, execution speed, and robustness of the algorithm.

Support Vector Machines. Last, support vector machines (Vapnik, 1999) aim to create a hyperplane (i.e., a flat boundary) between data points. In a classification example with two outcome classes, the hyperplane divides the space between the outcome classes (e.g., those who reoffended and those who did not reoffend) to create the greatest segregation between the two. Therefore, data points that are predicted to fall on either side of this hyperplane can then be attributed to an outcome class. As these data points are unlikely to be easily separable in two dimensions, kernels (i.e., a set of mathematical functions) are used to transform the data into a different form that better enables separation. For the present study, multiple kernels were trialed, and ultimately, nonlinear polynomial kernels were used as they produced the highest levels of discrimination.

Processing Techniques

For the present study, two processing techniques were used. The first was a preprocessing technique that used residuals in place of the predictor variables (Berk, 2009). For the training sample, the eight risk factor scores were regressed onto Aboriginal and Torres Strait Islander status, and the residual (i.e., the difference between the actual risk factor score and the predicted risk factor score) was used instead of the original variable as the new variable for the statistical learning methods to be trained on. This approach helps to remove the association between an individual's Aboriginal and

Torres Strait Islander status and the risk factor scores from the data before the algorithm is constructed. The same regression models used on the training data were then used to estimate the residuals for the testing data set.

The second processing technique was a postprocessing technique that involved reassigning the outcome classification through a process known as reject option–based classification (see Kamiran et al., 2012). This process relabels observation outcomes that are deemed to be more uncertain (i.e., close to the cutoff value that distinguishes a prediction of reoffended from a prediction of did not reoffend) and therefore influenced by biases and more likely to be incorrectly classified. With reject option–based classification, a critical region boundary (i.e., margin) is specified around the cutoff value (which for the present study was determined using the training data), denoted by θ. The observations that fall within this region have their labels reassigned specifically to reduce bias by aiding in increasing parity among the outcome variable. Although reject option–based classification could result in overall lower predictive accuracy due to the reclassification of outcome classes (Berk, 2019), this processing approach does provide decision makers with a better level of control for the level of outcome class disparity between groups as well as control over the trade-off between fairness and accuracy (Kamiran et al., 2012).

Regarding the present study, the group that reoffended more (Aboriginal and Torres Strait Islanders) had their observations that fell within the cutoff + θ reassigned to being predicted to not reoffend. Conversely, the group that reoffended less (non-Aboriginal and Torres Strait Islanders) had their observations that fell within the cutoff − θ reassigned to being predicted to reoffend. Multiple θ values were trailed in the present study (ranging from .00625 to .01) on the training data, with the final θ value being the one that led to the biggest increase in fairness and the lowest reduction in discrimination to ensure that this approach was not resulting in a notable loss in AUC. The same θ value and cutoff value used on the training data were then used with the testing data to calculate the AUC and fairness outcomes.

Area Under the Curve

The AUC is the probability that a randomly selected individual who reoffended received a higher risk score compared to a randomly selected individual who did not reoffend. The AUC is base rate resistant and provides an index of a risk instrument's sensitivity and 1 – specificity across all observed values (Cook, 2007). The AUC value ranges from 0 to 1, with a value of .50 reflecting discrimination at chance levels and a higher AUC indicating stronger discriminative power (Rice & Harris, 2005).

Fairness

Cross Area Under the Curve. The xAUC (Kallus & Zhou, 2019) is a modification of the AUC that measures discrimination between groups instead of within to better identify disparities. Specifically, the standard AUC evaluates an instrument's capability to differentiate individuals who reoffend from individuals who do not within a single group. The xAUC evaluates an instrument's ability to differentiate individuals who reoffend in one group (e.g., Aboriginal and Torres Strait Islanders) from individuals who do not in another (e.g., non-Aboriginal and Torres Strait Islanders). A cross-receiver operating characteristic (xROC) plots the sensitivity against 1 – specificity at various thresholds for the two sets of groups for which an xAUC can be calculated. The first set contains a positive outcome from one group and a negative outcome from the other group. The second set is the opposite of the first. For the present study, Set 1 includes all Aboriginal and Torres Strait Islander individuals who reoffended and all non-Aboriginal and Torres Strait Islander individuals who did not. Set 2 includes all non-Aboriginal and Torres Strait Islander individuals who reoffended and all Aboriginal and Torres Strait Islander individuals who did not. The xAUC measures the probability that a randomly selected individual who reoffended from one group received a higher risk score than a randomly selected individual from the other group who did not reoffend.

Calibration. The calibration of the statistical learning methods was assessed by Brier scores (Brier, 1950), which measure the squared error between a predicted probability (ranging between 0 and 1) and the outcome (coded as 0 if the outcome did not occur and 1 if the outcome did occur). Lower Brier scores indicate better performance and more accurate forecasts, with the best possible Brier score being 0 and the worst possible Brier score being 1. These were calculated for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

Predictive Parity. Predictive parity was assessed by calculating the PPV and NPV as a percentage for both groups. To distinguish between those who are predicted to reoffend and those who are not, in order to calculate the relevant metrics for predictive parity, a cutoff value is required. For the present study, the optimal cutoff was defined as the cutoff that yielded the smallest distance to the Point 0, 1 on the ROC space. This approach was utilized as a test that passes through 0, 1 on the ROC space reflects perfect discrimination. Although the present study aims to maximize fairness, doing so while maintaining discrimination of the LS/RNR was also important. Therefore, a cutoff point that prioritized discrimination was utilized and was calculated separately for the LS/RNR total risk score and each algorithm. The cutoff was determined using the training sample only, and the same cutoff was then utilized for the testing data.

Error Rate Balance. Error rate balance was assessed by calculating the FPR and FNR as a percentage for both groups. Like predictive parity, error rate balance metrics require a cutoff value to be calculated. The optimal cutoff yielding the smallest distance to 0, 1 on the ROC space for the training data was again utilized.

Statistical Parity. Statistical parity was assessed as the proportion of Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders whose predicted probability of reoffending was above the optimal cutoff threshold established for the training data.

This study was not preregistered. The data and study materials are not publicly available; however, the study analysis code is available within the Supplemental Materials.

Results

AUC

The AUC and 95% confidence intervals (CI) were established for the LS/RNR total risk score as well as for each statistical learning

method using the held-out testing sample and are presented in Table 1. For the whole sample, similar or slightly higher AUC values were reported for all statistical learning methods, including those with pre- and postprocessing approaches, when compared to the LS/RNR. The stochastic gradient boosting and penalized logistic regression algorithms demonstrated the highest levels of discrimination. For Aboriginal and Torres Strait Islanders, the LS/RNR total risk score had the lowest levels of discrimination, with statistical learning methods often notably improving the AUC. However, statistical learning methods also frequently led to a comparable or reduced level of discrimination for non-Aboriginal and Torres Strait Islanders when compared to the LS/RNR, increasing the disparity between the two groups on overall AUC values. Only stochastic gradient boosting and penalized logistic regression produced higher AUC values for non-Aboriginal and Torres Strait Islanders; however, these differences were negligible. Most often, for the whole sample and both groups, processing approaches resulted in similar or slightly reduced AUCs when compared to the original statistical learning method. However, due to the small sample, wide CIs were reported, especially when examining Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in separate groups, as this further reduced the sample size. This frequently resulted in CIs falling below the .50 threshold, primarily for non-Aboriginal and Torres Strait Islander people, reflecting random levels of discrimination and emphasizing the uncertainty surrounding the obtained values.

Fairness

xAUC

The xAUC was calculated for the LS/RNR total risk score and for each of the statistical learning methods using the testing sample and are presented in Table 2. The xAUC was higher for Set 1 (Aboriginal and Torres Strait Islanders who reoffended and non-Aboriginal and Torres Strait Islanders who did not reoffend) for the LS/RNR risk score and for all statistical learning methods. Based on the LS/RNR risk score or predicted probability of reoffending, Aboriginal and Torres Strait Islander offenders are often distinguishable from non-Aboriginal and Torres Strait Islander offenders. However, the same cannot be said for Set 2, with Aboriginal and Torres Strait Islanders who did not reoffend having similar risk scores (or predicted probabilities of reoffending) and therefore being less distinguishable from non-Aboriginal and Torres Strait Islanders who did reoffend, resulting in these two groups potentially being treated equally if decisions were made on the basis of risk scores or predicted probabilities. Only the support vector machine had a greater disparity between the two groups than the LS/RNR total risk score, with the xAUC for Set 2 being below chance levels for both the LS/RNR and support vector machine algorithms. This indicates that the LS/RNR and support vector machine were unable to discriminate between Aboriginal and Torres Strait Islander who did not reoffend from non-Aboriginal and Torres Strait Islanders who did.

All the remaining statistical learning methods (including those with pre- or postprocessing techniques) led to an increase in xAUC for Set 2 and a decrease in the disparity between Set 1 and Set 2 xAUC values. Preprocessing led to a further decrease in the disparity between Set 1 and Set 2 for all statistical learning methods. Specifically, logistic regression and support vector machine algorithms with preprocessing resulted in near parity between the two sets. This indicates that these statistical learning methods were discriminating between individuals who reoffended from individuals who did not equally well between groups. Postprocessing techniques resulted in no (or marginal) levels of improvement in xAUC disparities across all statistical learning methods. Wide CIs were also reported for xAUC values, with all xAUC values

Table 1
AUC for the LS/RNR Risk Score and Statistical Learning Methods

		All		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
Algorithm	Processing	AUC	95% CI	AUC	95% CI	AUC	95% CI
LS/RNR total risk score	None	.64	[.57, .70]	.60	[.49, .70]	.63	[.55, .72]
Logistic regression	None	.66	[.54, .77]	.77	[.63, .91]	.57	[.43, .72]
	Pre	.64	[.52, .75]	.78	[.64, .92]	.55	[.41, .70]
	Post	.66	[.55, .77]	.77	[.63, .91]	.57	[.43, .72]
Penalized logistic regression	None	.73	[.62, .83]	.76	[.59, .94]	.64	[.51, .78]
	Pre	.72	[.61, .82]	.76	[.59, .94]	.64	[.51, .78]
	Post	.73	[.63, .83]	.81	[.63, .98]	.64	[.50, .77]
Random forest	None	.68	[.57, .79]	.70	[.48, .93]	.63	[.50, .77]
	Pre	.63	[.53, .73]	.64	[.48, .80]	.61	[.48, .74]
	Post	.68	[.57, .79]	.70	[.48, .93]	.63	[.50, .77]
Stochastic gradient boosting	None	.73	[.64, .83]	.79	[.61, .98]	.65	[.52, .77]
	Pre	.71	[.60, .81]	.79	[.59, .99]	.62	[.48, .76]
	Post	.73	[.64, .83]	.79	[.61, .98]	.65	[.52, .77]
Support vector machine	None	.71	[.61, .81]	.76	[.59, .93]	.63	[.50, .77]
	Pre	.68	[.58, .79]	.81	[.64, .97]	.61	[.48, .75]
	Post	.71	[.61, .81]	.76	[.59, .93]	.63	[.50, .77]

Note. AUC = area under the curve; CI = confidence interval; LS/RNR = Level of Service/Risk Needs Responsivity.

Table 2
xAUC for the LS/RNR Risk Score and Statistical Learning Methods

Algorithm	Processing	Set 1 xAUC	Set 1 95% CI	Set 2 xAUC	Set 2 95% CI	xAUC difference
LS/RNR total risk score	None	.75	[.68, .83]	.46	[.35, .57]	.29
Logistic regression	None	.72	[.58, .87]	.59	[.40, .78]	.13
	Pre	.68	[.53, .83]	.62	[.44, .80]	.06
	Post	.72	[.58, .87]	.62	[.43, .80]	.10
Penalized logistic regression	None	.85	[.74, .97]	.60	[.38, .81]	.25
	Pre	.79	[.66, .92]	.68	[.48, .88]	.11
	Post	.84	[.73, .96]	.63	[.43, .83]	.21
Random forest	None	.75	[.63, .87]	.61	[.37, .85]	.14
	Pre	.68	[.56, .80]	.56	[.37, .75]	.12
	Post	.75	[.63, .87]	.61	[.37, .85]	.14
Stochastic gradient boosting	None	.84	[.74, .94]	.64	[.41, .86]	.20
	Pre	.81	[.70, .91]	.62	[.38, .87]	.19
	Post	.84	[.74, .94]	.64	[.41, .86]	.20
Support vector machine	None	.83	[.72, .94]	.44	[.25, .64]	.39
	Pre	.73	[.59, .87]	.65	[.49, .82]	.08
	Post	.83	[.72, .94]	.56	[.37, .76]	.27

Note. xAUC = cross area under the curve; CI = confidence interval; LS/RNR = Level of Service/Risk Needs Responsivity. Set 1 refers to Aboriginal and Torres Strait Islander individuals who reoffended and non-Aboriginal and Torres Strait Islander individuals who did not reoffend. Set 2 refers to non-Aboriginal and Torres Strait Islander individuals who reoffended and Aboriginal and Torres Strait Islander individuals who did not reoffend.

for Set 2 crossing the .50 threshold, indicating uncertainty in the estimates.

Calibration

Brier scores to assess calibration were calculated for each statistical learning method using the testing sample and are presented in Table 3. For the differences between the two groups, please refer to the Supplemental Materials. A Brier score was unable to be calculated for the original LS/RNR total risk score as predicted probabilities are required as part of the calculation. The logistic regression algorithm was used here as the baseline comparison model. Overall, greater levels of calibration were reported for Aboriginal and Torres Strait Islanders. The greatest levels of calibration for both groups were often found for the random forest algorithms; however, random forest algorithms also demonstrated one of the greatest disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The disparities between the two groups were often similar or reduced for most statistical learning methods when using pre- or postprocessing approaches. Compared to logistic regression, only penalized logistic regression and stochastic gradient boosting algorithms produced closer levels of calibration between the two groups. Stochastic gradient boosting specifically produced nearly equal Brier scores for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, demonstrating

Table 3
Calibration, Predictive Parity, Error Rate Balance, and Statistical Parity Values for Both Groups

Algorithm	Processing	Aboriginal and Torres Strait Islander PPV	NPV	FPR	FNR	SP	Brier	Non-Aboriginal and Torres Strait Islander PPV	NPV	FPR	FNR	SP	Brier
LS/RNR total risk score	None	87.88	20.83	61.54	24.68	73.33	—	83.81	32.63	35.42	42.11	52.50	—
Logistic regression	None	93.33	22.73	37.50	28.81	67.16	.203	83.78	31.91	28.57	50.79	44.05	.276
	Pre	91.23	30.00	62.50	11.86	85.07	.137	74.60	23.81	76.19	25.40	75.00	.218
	Post	95.12	23.08	25.00	33.90	61.19	.204	85.37	34.88	28.57	44.44	48.81	.275
Penalized logistic regression	None	92.16	25.00	50.00	20.34	76.12	.225	88.89	35.42	19.05	49.21	42.86	.248
	Pre	93.44	66.67	50.00	3.39	91.04	.202	78.26	40.00	71.43	14.29	82.14	.225
	Post	94.87	21.43	25.00	37.29	58.21	.226	87.80	37.21	23.81	42.86	48.81	.248
Random forest	None	92.11	17.24	37.50	40.68	56.72	.130	85.71	32.65	23.81	52.38	41.67	.223
	Pre	93.02	20.83	37.50	32.20	64.18	.124	83.72	34.15	33.33	42.86	51.19	.199
	Post	92.11	17.24	37.50	40.68	56.72	.130	86.11	33.33	23.81	50.79	42.86	.223
Stochastic gradient boosting	None	95.24	15.22	12.50	66.10	31.34	.242	92.31	28.17	4.76	80.95	15.48	.248
	Pre	91.94	60.00	62.50	3.39	92.54	.133	79.66	36.00	57.14	25.40	70.24	.212
	Post	95.24	15.22	12.50	66.10	31.34	.242	92.31	28.17	4.76	80.95	15.48	.248
Support vector machine	None	91.49	20.00	50.00	27.12	70.15	.160	86.11	33.33	23.81	50.79	42.86	.266
	Pre	91.94	60.00	62.50	3.39	92.54	.102	77.05	30.43	66.67	25.40	72.62	.194
	Post	94.74	20.69	25.00	38.98	56.72	.162	86.11	33.33	23.81	50.79	42.86	.266

Note. PPV = positive predictive value; NPV = negative predictive value; FPR = false positive rate; FNR = false negative rate; SP = statistical parity; LS/RNR = Level of Service/Risk Needs Responsivity.

similar levels of alignment between the prediction and outcome for both cultural groups.

Predictive Parity, Error Rate Balance, and Statistical Parity

Predictive parity, error rate balance, and statistical parity were calculated for the LS/RNR total risk score and for each statistical learning method using the testing sample and are also presented in Table 3. For the differences between the two groups, please refer to the Supplemental Materials. The LS/RNR total risk score had notable differences in error rate balance and statistical parity, with Aboriginal and Torres Strait Islanders having a significantly larger proportion being predicted to reoffend, a higher FPR (or higher proportion of those who did not reoffend being predicted to reoffend), and a significantly lower FNR (or lower proportion of those who reoffended being predicted to not reoffend). All statistical learning methods (besides penalized logistic regression and support vector machine algorithms with no processing) were found to lower the FPR differences, with pre- and postprocessing often decreasing the differences further. Similarly, when compared to the LS/RNR total risk score, the majority of statistical learning methods resulted in a reduced FNR discrepancy between groups. Pre- and postprocessing again often further reduced these disparities to lower levels than the LS/RNR total risk score. For statistical parity, almost all statistical learning methods (including those with pre- and postprocessing) produced lower discrepancies between groups in the proportion who were predicted to reoffend when compared to the LS/RNR total risk score.

Predictive parity differences between groups were often lower overall compared to the disparities found for error rate balance. For almost all of the statistical learning methods and the LS/RNR, Aboriginal and Torres Strait Islanders had a higher PPV and lower NPV. PPV differences were lower for the LS/RNR total risk score when compared to statical learning methods besides penalized logistic regression and stochastic gradient boosting. Processing approaches often resulted in either a similar or increased disparity between groups. Similarly, for NPV, almost all statistical learning methods produced a greater disparity between groups compared to the LS/RNR total risk score expected for logistic regression and penalized logistic regression.

Discussion

The present study explored the use of statistical learning methods with LS/RNR risk factors as predictors to increase the discrimination and fairness of the LS/RNR instrument. In line with previous findings that used a small number of predictors (Liu et al., 2011; Tollenaar & van der Heijden, 2019), many of the statistical learning methods did not bring about notable improvements in discrimination over the LS/RNR total score. Some statistical learning methods (e.g., penalized logistic regression and stochastic gradient boosting) did demonstrate a significant improvement in discrimination, with most approaches performing well for Aboriginal and Torres Strait Islanders. Statistical learning methods also often demonstrated an increase in fairness across groups, whether using just the statistical learning method with no processing applied or using pre- or postprocessing techniques. For example, stochastic gradient boosting with no processing was found to decrease discrepancies between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on xAUC, calibration, PPVs, statistical parity, and error rate balance, while improving discrimination for the whole sample, Aboriginal and Torres Strait Islanders, and non-Aboriginal and Torres Strait Islanders. However, this approach did also lead to a minor increase in the disparity among the NPVs.

The preprocessing approach of using residuals in place of original predictors and the postprocessing approach of reject option–based classification often led to improvements in increasing fairness between Aboriginal and Torres Straits Islanders and non-Aboriginal and Torres Strait Islanders, not only when compared to the original LS/RNR total risk score but also to the algorithms with no processing. This was primarily found for error rate balance, statistical parity, and xAUC.

Overall, statistical learning methods that used LS/RNR risk factors as predictors often aided in ameliorating disparities among various fairness definitions between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders when compared to the traditional LS/RNR total risk score. This was achieved while producing a comparable ability to discriminate individuals who reoffended from individuals who did not when compared to the LS/RNR total risk score (i.e., similar AUC values). However, there was no single statistical learning method that performed best, with more traditional statistical learning methods (i.e., logistic regression) often producing similar findings to more complex approaches (i.e., support vector machines). The biggest difference between the statistical learning methods was found with AUC and xAUCs, with more complex and flexible approaches often producing higher values.

Trade-Offs

For the original LS/RNR total risk score, predictive parity was relatively comparable for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, with most discrepancies identified among error rate balance and statistical parity. These latter two definitions of fairness were the most improved when using statistical learning methods and processing approaches. However, predictive parity was often negatively impacted. The increase in disparities among predictive parity was often smaller than the decrease in disparities identified in error rate balance. This was in line with previous findings by Skeem and Lowenkamp (2020) and Johndrow and Lum (2017), who also reported a slight increase in PPV disparity but improvement among FPR discrepancies when using preprocessing approaches. This also highlights the trade-off that exists between these two forms of fairness. With a risk assessment instrument that does not have perfect accuracy or equal base rates, predictive parity and error rate balance cannot be satisfied simultaneously, and one will need to be prioritized.

The trade-off between discrimination and fairness was not notably observed in the present study when altering the statistical learning methods. In line with previous research, AUC values were scarcely impacted negatively, with no notable losses in discrimination when using processing approaches (Lum & Johndrow, 2016; Skeem & Lowenkamp, 2020; Wadsworth et al., 2018). Another issue is the loss of interpretability caused by some of these more complex statistical learning methods and the use of processing

approaches. In contrast to the original LS/RNR, where the procedure used to calculate the total risk score is known, or simple statistical learning methods like linear regression, where the individual importance of predictors can be explicitly understood through β weights, more flexible statistical learning methods (e.g., stochastic gradient boosting, support vector machines, random forests) will impair algorithm transparency.

Limitations

The present study was limited by the sample in several ways. First, the sample included those who were originally sentenced to a term of imprisonment for a serious violent offense as outlined by the *Sentencing Act* 1991 (Vic). As a result, the individuals in this study are a specific subsample of the Victorian prison population and may not be representative of or generalizable to the general prison population for which the LS/RNR was developed. Second, Aboriginal and Torres Strait Islander peoples were oversampled in order to enable comparisons between groups. In Victoria, Aboriginal and Torres Strait Islanders account for 10% of the adult male custodial population (Corrections Victoria, 2022); however, this is reflective of all adult prisoners in Victoria and not those specifically incarcerated for a serious violent offense. As a result, the current sample does not represent the larger male prison population or the serious violent offender population. Third, the non-Aboriginal and Torres Strait Islander group was unable to be portioned into more accurate cultural groups. Australia is a multicultural society (Australian Bureau of Statistics, 2020a), and including all of those who did not identify as Aboriginal and Torres Strait Islander into one group is not reflective of the cultural diversity that exists within Australia and ignores the heterogeneity of this group. The self-reported nature of this demographic information and the quality and accuracy of the collected data could also further impede the accuracy and generalizability of the findings.

The sample size also posed notable constraints. The statistical learning methods employed in the present study produce the best estimates with large sample sizes. Although bootstrapping with 1,000 resamples was used to mitigate the small training sample size, wide 95% CIs were often reported across the testing sample. These broad CIs reflect substantial uncertainty in the obtained AUC and xAUC values, and they should therefore be interpreted with significant caution. Although certain statistical learning methods may have resulted in closer AUC or xAUC values between groups, these findings may not be particularly robust or meaningful due to the small sample and wide CIs. Furthermore, the AUC has also been cautioned against sample sizes of less than 200, as this can result in large inaccuracies (Hanczar et al., 2010).

Relying on a cutoff to easily enable predictive parity, error rate balance, and statistical parity to be calculated further limited this study. There are numerous ways to determine a cutoff value (Kuhn & Johnson, 2013), but there is no agreed-upon approach to effectively determining which is the best method. Further, using a cutoff to enable the calculation of these fairness metrics is problematic, as different cutoffs will produce different fairness results (Zottola et al., 2022). Last, this study has treated predictive parity and error rate balance as equally important definitions of fairness to satisfy, which may not be reflective of forensic practice.

Implications

The findings from the present study demonstrate a potential approach that could be employed to increase fairness between cultural groups. Although no single approach performed best overall, penalized logistic regression and stochastic gradient boosting algorithms appeared to offer an advantage in improving overall statistical discrimination, with higher AUC values reported overall and for both groups, as well as higher xAUC values for both sets. A random forest algorithm, on the other hand, was better suited for achieving higher levels of statistical parity and error rate balance between groups, which were further improved by using residuals instead of original predictors. Using the LS/RNR risk factor scores, these statistical learning methods could be applied to provide risk estimates to corrections officers and clinicians that have higher statistical discrimination and are fairer than what is currently being used in practice. However, as a first step, future research should gather larger, more representative samples to better examine the generalizability of the present findings and how estimates may vary across cultural groups. Larger samples will also allow more predictors (e.g., LS/RNR items) to be used, which may aid in increasing the discrimination of the instrument, especially with the more flexible statistical learning methods. Further, fairness metrics could be observed across a range of cutoff values to gain a better understanding of the discrepancies that are not limited to a single threshold. For calibration specifically, overall calibration metrics like Brier scores in the present study can mask where the differences in calibration between groups arise (e.g., low or high predicted probabilities of reoffending). Future research should therefore explore calibration further by examining the differences across predicted probabilities (e.g., plotting calibration curves for each group).

Although a basic and preliminary examination of variable importance was conducted, which demonstrated that reoffending predictions were most often driven by the alcohol and drug problems risk factor (please refer to the Supplemental Materials), the exploration of more informative post hoc analyses such as local interpretable model-agnostic explanations (Ribeiro et al., 2016) and Shapley values (Shapley, 1953) on these statistical learning methods should be explored to help counteract the trade-off between the flexibility of more complex statistical learning methods and interpretability. These approaches could help provide a more in-depth understanding for corrections officers and clinicians of which risk factors and items are the most important in predicting reoffending and therefore the most pivotal for treatment plans and rehabilitation. However, post hoc approaches for increasing the interpretability of complex statistical learning methods need to have their limitations understood. For example, they can be easily misunderstood and may not provide enough information to understand explicitly how the statistical learning method arrived at a prediction (Rudin, 2019). Furthermore, if more simple statistical learning methods such as logistic regression are able to produce similar findings to more flexible approaches, the use of logistic regression may be worth prioritizing to ensure complete transparency.

The present study also highlighted the issues surrounding the trade-offs that exist across certain fairness definitions. The trade-off between error rate balance and predictive parity raises the question of whether it is more necessary to have equality in predictive

accuracy (i.e., predictive parity) or equality in the number of errors (i.e., error rate balance). Although they are treated as equally important in this study, this is unlikely to reflect what is relevant in practice. This discussion requires thoughtful deliberation by policymakers as to what form of fairness is most relevant to satisfy their specific risk assessment, jurisdictional, and cross-cultural fairness needs. For example, if the emphasis is placed on ensuring cross-culturally fair predictions (i.e., predictive parity), so that similar proportions of Aboriginal and Torres Strait Islander people and non-Aboriginal and Torres Strait Islander people classified as high-risk reoffend, the number of errors in observation (i.e., error rate balance) will inevitably differ. In the case of the present findings, this would result in a higher proportion of Aboriginal and Torres Strait Islanders being classified as high risk on the LS/RNR and not going on to reoffend. Conversely, a higher proportion of non-Aboriginal and Torres Strait Islanders would be classified as low risk and later reoffend.

However, in the case of the present study, certain approaches may potentially lead to a publicly acceptable trade-off between these two forms of fairness. For example, even though the random forest with postprocessing resulted in a slight increase in PPV and NPV disparities, error rate balance was notably improved such that the absolute mean difference for predictive parity metrics (PPV and NPV) was 11.05% and error rate balance metrics (FPR and FNR) were comparable at 11.90%. This was achieved alongside a reduction in xAUC differences and comparable or improved AUC values. Therefore, if both forms of fairness want to be prioritized, the approaches in the present study may provide a way to achieve an acceptable trade-off. However, what represents an acceptable level of fairness (or unfairness) between groups also requires careful consideration. There is currently no agreed-upon classification system for differences among fairness definitions, and what constitutes a meaningful difference between groups may differ between policy-makers and users of risk assessment instruments.

Conclusion

The present preliminary study explored the impact of using statistical learning methods and processing approaches on the discrimination and fairness of the LS/RNR for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. These approaches demonstrated positive results in reducing certain fairness disparities (primarily xAUC, error rate balance, and statistical parity) without impairing the instrument's discrimination and should thus be investigated further as a method to mitigate unfairness in risk assessment instruments in the future.

References

Andrews, D. A., Bonta, J., & Wormith, J. (2008). *The Level of Service/Risk Need Responsivity Inventory (LS/RNR): Scoring guide*. Multi-Health Systems.

Ashford, L. J., Spivak, B. L., Ogloff, J. R. P., & Shepherd, S. M. (2022). The cross-cultural fairness of the LS/RNR: An Australian analysis. *Law and Human Behavior*, *46*(3), 214–226. https://doi.org/10.1037/lhb0000486

Australian Bureau of Statistics. (2020a). *Australia's population: Over 7.5 million born overseas*. https://www.abs.gov.au/articles/australias-population-over-75-million-born-overseas

Australian Bureau of Statistics. (2020b). *Prisoners in Australia*. https://www.abs.gov.au/statistics/people/crime-and-justice/prisoners-australia/latest-release

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and Social Problems*, *1*(4), 231–242. https://doi.org/10.1007/s12552-009-9017-z

Berk, R. (2019). Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies*, *16*(1), 175–194. https://doi.org/10.1111/jels.12206

Berk, R., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*(3), 513–544. https://doi.org/10.1111/1745-9133.12047

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Breitenbach, M., Dieterich, W., Brennan, T., & Fan, A. (2009). Creating risk-scores in very imbalanced datasets: Predicting extremely low violent crime among criminal offenders following release from prison. In Y. S. Koh & N. Rountree (Eds.), *Rare association rule mining and knowledge discovery: Technologies for infrequent and critical event detection* (pp. 231–254). Information Science Reference.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Chu, C. M., Lee, Y., Zeng, G., Yim, G., Tan, C. Y., Ang, Y., Chin, S., & Ruby, K. (2015). Assessing youth offenders in a non-Western context: The predictive validity of the YLS/CMI ratings. *Psychological Assessment*, *27*(3), 1013–1021. https://doi.org/10.1037/a0038670

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*(7), 928–935. https://doi.org/10.1161/circulationha.106.672402

Corrections Victoria. (2022). *Annual prisoner statistical profile 2009–10 to 2019–20*. https://www.corrections.vic.gov.au/annual-prisoner-statistical-profile-2009-10-to-2019-20

Day, A., Tamatea, A. J., Casey, S., & Geia, L. (2018). Assessing violence risk with Aboriginal and Torres Strait Islander offenders: Considerations for forensic practice. *Psychiatry, Psychology and Law*, *25*(3), 452–464. https://doi.org/10.1080/13218719.2018.1467804

Duwe, G., & Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review*, *28*(6), 570–600. https://doi.org/10.1177/0887403415604899

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ*, *345*(7868), Article e4692. https://doi.org/10.1136/bmj.e4692

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J. H., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., & Simon, N. (2021). *glmnet: Lasso and elastic-net regularized generalized linear models* (Version 4.1-2) [R program]. https://CRAN.R-project.org/package=glmnet

Ghasemi, M., Anvari, D., Atapour, M., Stephen Wormith, J., Stockdale, K. C., & Spiteri, R. J. (2020). The application of machine learning to a general risk–need assessment instrument in the prediction of criminal

recidivism. *Criminal Justice and Behavior*, *48*(4), 518–538. https://doi.org/10.1177/0093854820969753

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. B. M. (2020). *gbm: Generalized boosted regression models* (Version 2.1.8) [R program]. https://CRAN.R-project.org/package=gbm

Hamilton, Z., Neuilly, M.-A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*, *11*(2), 299–318. https://doi.org/10.1007/s11292-014-9221-8

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*(6), 822–830. https://doi.org/10.1093/bioinformatics/btq037

Hart, S. D. (2016). Culture and violence risk assessment: The case of Ewert v. Canada. *Journal of Threat Assessment and Management*, *3*(2), 76–96. https://doi.org/10.1037/tam0000068

Hart, S. D., Douglas, K. S., & Guy, L. (2017). The structured professional judgment approach to violence risk assessment: Origins. nature. and advances. In D. P. Boer, A. R. Beech, T. Ward, L. A. Craig, M. Rettenberger, L. E. Marshall, & W. L. Marshall (Eds.), *The Wiley handbook on the theories, assessment, and treatment of sexual offending* (pp. 643–666). Wiley-Blackwell. https://doi.org/10.1002/9781118574003.wattso030

Heilbrun, K., Yasuhara, K., & Shah, S. (2010). Violence risk assessment tools: Overview and critical analysis. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 1–17). Routledge/Taylor & Francis Group.

Johndrow, J., & Lum, K. (2017). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. arXiv:1703.04957 [stat.AP].

Kallus, N., & Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. arXiv:1902.05826 [cs.LG].

Kamiran, F., Karim, A., & Zhang, X. (2012, December 10–13). *Decision theory for discrimination-aware classification* [Paper presentation]. 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium.

Kuhn, M. (2021). *caret: Classification and regression training* (Version 6.0-88) [R package]. https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Liaw, A., & Wiener, M. (2018). *randomForest: Breiman and Cutler's random forests for classification and regression* (Version 4.6-14) [R program]. https://CRAN.R-project.org/package=randomForest

Liu, Y., Yang, M., Ramsay, M., Li, X., & Coid, J. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, *27*(4), 547–573. https://doi.org/10.1007/s10940-011-9137-7

Lum, K., & Johndrow, J. (2016). A statistical framework for fair predictive algorithms. arXiv:1610.08077 [stat.ML].

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc functions of the Department of Statistics, probability theory group (formerly: E1071), TU Wien* (Version 1.7-8) [R package]. https://CRAN.R-project.org/package=e1071

Monahan, J., & Skeem, J. L. (2014). The evolution of violence risk assessment. *CNS Spectrums*, *19*(5), 419–424. https://doi.org/10.1017/S1092852914000145

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*(1), 156–176. https://doi.org/10.1037/a0035080

Papalia, N., Shepherd, S. M., Spivak, B., Luebbers, S., Shea, D. E., & Fullam, R. (2019). Disparities in criminal justice system responses to first-time juvenile offenders according to Indigenous status. *Criminal Justice and Behavior*, *46*(8), 1067–1087. https://doi.org/10.1177/0093854819851830

R Core Team. (2021). *R: A language and environment for statistical computing*. https://www.R-project.org/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 13–17). *"Why should I trust you?": Explaining the predictions of any classifier* [Paper presentation]. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States. https://doi.org/10.1145/2939672.2939778

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2020). *pROC: Display and analyze ROC curves* (Version 1.16.2) [R program]. https://CRAN.R-project.org/package=pROC

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salo, B., Laaksonen, T., & Santtila, P. (2019). Predictive power of dynamic (vs. static) risk factors in the Finnish Risk and Needs Assessment Form. *Criminal Justice and Behavior*, *46*(7), 939–960. https://doi.org/10.1177/0093854819848793

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games* (pp. 307–317). Princeton University Press.

Shepherd, S. M., Singh, J. P., & Fullam, R. (2015). Does the Youth Level of Service/Case Management Inventory generalize across ethnicity? *International Journal of Forensic Mental Health*, *14*(3), 193–204. https://doi.org/10.1080/14999013.2015.1086450

Singh, J. P. (2012). The history. development. and testing of forensic risk assessment tools. In E. Grigorenko (Ed.), *Handbook of juvenile forensic psychology and psychiatry* (pp. 215–225). Springer. https://doi.org/10.1007/978-1-4614-0905-2_14

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law*, *31*(1), 55–73. https://doi.org/10.1002/bsl.2053

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*(3), 499–513. https://doi.org/10.1016/j.cpr.2010.11.009

Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences and the Law*, *38*(3), 259–278. https://doi.org/10.1002/bsl.2465

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*(4), 680–712. https://doi.org/10.1111/1745-9125.12123

Spivak, B. L., & Shepherd, S. M. (2020). Machine learning and forensic risk assessment: New frontiers. *Journal of Forensic Psychiatry & Psychology*, *31*(4), 571–581. https://doi.org/10.1080/14789949.2020.1779783

Thiele, C. (2021). *cutpointr: Determine and evaluate optimal cutpoints in binary classification tasks* (Version 1.1.1) [R package]. https://CRAN.R-project.org/package=cutpointr

Ting, M. H., Chu, C. M., Zeng, G., Li, D., & Chng, G. S. (2018). Predicting recidivism among youth offenders: Augmenting professional judgement with machine learning algorithms. *Journal of Social Work*, *18*(6), 631–649. https://doi.org/10.1177/1468017317743137

Tollenaar, N., & van der Heijden, P. G. M. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLOS ONE*, *14*(3), Article e0213245. https://doi.org/10.1371/journal.pone.0213245

Vapnik, V. (1999). Support vector method for function estimation. In J. A. K. Suykens & J. Vandewalle (Eds.), *Nonlinear modeling: Advanced black-box techniques* (pp. 55–85). Springer.

Verma, S., & Rubin, J. (2018, May 29). *Fairness definitions explained* [Paper presentation]. International Workshop on Software Fairness, Gothenburg, Sweden. https://doi.org/10.1145/3194770.3194776

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: An application to recidivism prediction. arXiv:1807.00199 [cs.LG].

Wickham, H. (2019). *tidyverse: Easily install and load the 'tidyverse'* (Version 1.3.0) [R program]. https://CRAN.R-project.org/package=tidyverse

Wilson, H. A., & Gutierrez, L. (2014). Does one size fit all? A meta-analysis examining the predictive ability of the Level of Service Inventory (LSI) with Aboriginal offenders. *Criminal Justice and Behavior*, *41*(2), 196–219. https://doi.org/10.1177/0093854813500958

Wormith, J., Hogg, S., & Guzzo, L. (2015). The predictive validity of the LS/CMI with Aboriginal offenders in Canada. *Criminal Justice and Behavior*, *42*(5), 481–508. https://doi.org/10.1177/0093854814552843

Zottola, S. A., Desmarais, S. L., Lowder, E. M., & Duhart Clarke, S. E. (2022). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*, *49*(3), 389–410. https://doi.org/10.1177/00938548211040544

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Received April 7, 2022
Revision received January 12, 2023
Accepted January 26, 2023 ■