Advancing Cross-Cultural Approaches to Forensic Risk Assessment

Linda J. Ashford

This thesis is submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Forensic Behavioural Science
School of Health Sciences
Faculty of Health, Arts & Design
Swinburne University of Technology
Melbourne, Australia

November 2022

Abstract

The cross-cultural fairness of forensic risk assessment instruments has received recent scrutiny due to differences in performance in the literature among cultural majorities and cultural minorities (e.g., African Americans and Indigenous populations of Australia and North America). Definitions of fairness (e.g., error rate balance, calibration, predictive parity, and statistical parity) that can impact a risk assessment instrument's overall utility are less often discussed. The limited literature exploring fairness definitions often notes significant cross-cultural disparities (i.e., differences between cultures in the outcomes of a risk assessment instrument that could adversely impact certain cultural groups). However, a clear way forward for how to reduce such cross-cultural disparities has yet to be established.

To address these gaps in the literature, this research aimed to i) explore the level of fairness of the Level of Service/Risk Need Responsivity (LS/RNR) instrument for male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia (*N* = 380) and ii) increase fairness through statistical learning methods without significantly impacting on the instrument's ability to discriminate individuals who engage in recidivism from individuals who do not. The discrimination of the risk assessment instrument was explored using the area under the curve (AUC) and cross area under the curve (xAUC). Minor differences in discrimination between groups were found when utilising the traditional AUC. However, notable differences were identified when utilising the xAUC. Specifically, the LS/RNR was unable to effectively discriminate Aboriginal and Torres Strait Islander non-recidivists from non-Aboriginal and Torres Strait Islander recidivists. Disparities were also identified for fairness definitions, including error rate balance and statistical parity, with Aboriginal and Torres Strait Islanders consistently having a higher false positive rate and scoring significantly higher on the LS/RNR.

A variety of statistical learning techniques were then used to assess if they could improve the discrimination of the LS/RNR. Transformations were also used, including pre and post-processing modifications, to attempt to increase fairness. A number of statistical learning methods were found to increase the discrimination of the LS/RNR. Pre-processing approaches were found to lead to notable reductions in xAUC, false positive rates, and statistical parity discrepancies between groups. Last, as some of these approaches resulted in algorithms that were not easily interpretable, the importance of predictors was ascertained through Shapley values. Specifically, items relating to criminal history, current drug use, and current unemployment were found to be important predictors of future recidivism.

This research has established that the LS/RNR appears to violate several definitions of fairness across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islander males from Australia. Further, the use of novel statistical learning methods and various processing approaches were shown to demonstrate utility in ameliorating the violated fairness definitions.

Acknowledgements

I acknowledge the Traditional Owners of the land where I work and live, the Wurundjeri people of the Kulin nation and pay my respects to Elders past, present, and emerging. I celebrate the stories, culture, and traditions of Aboriginal and Torres Strait Islander Elders of all communities who also work and live on this land.

I want to profoundly thank my supervisors, Associate Professor Stephane Shepherd and Dr. Ben Spivak. I have learned so much from you both over the past few years and will forever be grateful for the opportunity to have worked with you. Thank you for all of your help in finishing this thesis, for every supervision session, for always easing my countless worries, for every extra opportunity you provided, and for every unrelated and hilarious conversation that occurred when our supervision sessions inevitably unravelled.

I would also like to express my gratitude to Professor Michael Daffern, Professor Denny Meyer, Dr. Stefan Luebbers, and Dr. Caleb Lloyd for their support, encouragement, and direction throughout my candidature. Thank you to the teams at the Department of Justice and Community Safety, Corrections Victoria, and Victoria Police who provided the data that was necessary to complete this thesis. Furthermore, a big thanks to Associate Professor Troy McEwan, Associate Professor Rachael Fullam, and Dr. Janet Ruffles for overseeing ethics, data generation, and responding to my many, many questions.

To my brother John and sister-in-law Kira, thank you for always believing in me and not laughing at me when I told you over ten years ago that I wanted to do a PhD one day. I also appreciate you telling me very early on that you had already purchased a present for me for the day that I finished this thesis. Not knowing what that present was provided me with a profound level of determination to finish and find out. To Nitha and Nick, thank you for always supporting my academic pursuits and for being my voice of reason in moments of doubt.

Last, thank you to my incredibly patient, understanding, and always encouraging partner, Paul, for your unwavering support over the last three years. I do not know how to possibly convey my gratitude for all you have done. During the course of my PhD candidature, we bought our first house, renovated our house, experienced a global pandemic, got married, and finished a PhD thesis. Throughout all of this, you were incredibly selfless and always made sure I was fed, hydrated, sane, and, most importantly, caffeinated so I could keep meeting deadlines and working on my thesis. Here is hoping that the next three years are a little less eventful, but no less rewarding.

Candidate Declaration

In accordance with the Swinburne University of Technology Statement of Practice for the completion of a Higher Degree by Research at Swinburne, the following declarations were made:

I, Linda J. Ashford, declare that the examinable outcome of this thesis titled *“Advancing Cross-Cultural Approaches to Forensic Risk Assessment” c*ontains no material that has been accepted for the award to the candidate of any other diploma or degree, except where due reference has been made in the text of the examinable outcome. I declare that, to the best of my knowledge, this thesis contains no material that has been previously published or written by another individual, except where due reference is made within the text of the examinable outcome. Where the work in this thesis is based on joint research or publications, I have disclosed the relative contributions of the respective workers and authors.

I further warrant that I have obtained, where necessary, permission from the copyright owners to use any third party copyright material reproduced in this thesis (such as artwork, images, and unpublished documents), or to use any of my own published work (such as journal articles) in which the copyright is held by another party (such as publisher, co-author).

Linda J. Ashford

Doctor of Philosophy (PhD) Candidate

Centre for Forensic Behavioural Science

School of Health Sciences

Faculty of Health, Arts and Design

Swinburne University of Technology

Table of Contents

List of Tables

Chapter Two: Literature Review

Chapter Three: General Research Methodology

Chapter Four: Empirical Study One

Chapter Five: Empirical Study Two

Chapter Six: Empirical Study Three

List of Figures

Chapter Four: Empirical Study One

Chapter Six: Empirical Study Three

List of Publications

The following conference presentations, publications, and manuscripts submitted for publication that were produced as part of this thesis include:

Chapter Two: Literature Review

Ashford, L.J., Spivak, B.L., & Shepherd, S.M. (2020, November 4). *Cross-cultural fairness in forensic risk assessment: A review of the literature* [Paper presentation]. Building Bridges: Faculty of Health, Arts and Design Postgraduate Conference, Melbourne, Australia (Online Event).

Ashford, L.J., Spivak, B.L., & Shepherd, S.M. (2021). Racial fairness in violence risk instruments: A review of the literature. *Psychology, Crime & Law, 28*(9), 911-941. https://doi.org/10.1080/1068316X.2021.1972108

Chapter Four: Empirical Study One

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

Chapter Five: Empirical Study Two

Ashford, L.J., Spivak, B.L., Ogloff, J.R.P., & Shepherd, S.M. (2022). Statistical learning methods and cross-cultural fairness: Trade-offs and implications for risk assessment instruments. *Psychological Assessment.* Manuscript under review.

Chapter Six: Empirical Study Three

Ashford, L.J., Spivak, B.L., Ogloff, J.R.P., & Shepherd, S.M. (2022). Increasing the cross-cultural fairness of the LS/RNR and interpretability of statistical learning methods. *Psychiatry, Psychology, and Law.* Manuscript under review.

Further publications and outputs produced throughout my candidature at the Centre for Forensic Behavioural Psychology, Swinburne University of Technology include:

Shepherd, S.M., Spivak, B., **Ashford, L.J.**, Williams, I., Trounson, J., & Paradies, Y. (2020). Closing the (incarceration) gap: Assessing the socio-economic and clinical indicators of Indigenous males by lifetime incarceration status. *BMC Public Health, 20*. https://doi.org/10.1186/s12889-020-08794-3

McEwan, T., **Ashford L.,** Richardson, K. & Fullam, R. (2021). Victorian Fixated Threat Assessment Centre Joint Extracted De-identified database review. Prepared for Victoria Police by the Centre for Forensic Behavioural Science, Swinburne University of Technology, Melbourne Australia.

Anthony, C., Grant, I., **Ashford, L.J.**, Spivak, B., Shepherd, S.M. (2022). Exploring differences in the experiences, perceptions, and reporting of violent incidents in Australia by country of birth. *Journal of Interpersonal Violence, 37*(9). doi:10.1177/0886260520966676

Warner, B., Spivak, B., **Ashford, L.J.**, Fix, R., Ogloff, J., & Shepherd, S.M. (2022). The impact of offender-victim cultural backgrounds on the likelihood of receiving diversion for first-time young offenders. *Criminal Justice Policy Review, 33*(3), 298-316. doi: 10.1177/08874034211046313

Acronyms

AUC = Area Under the Curve

CALD = Culturally and Linguistically Diverse

CANZUS = Canada, Australia, New Zealand, and USA

COMPAS = Correctional Offender Management Profiling for Alternative Sanctions

CRN = Corrections Reference Number

ESB = English Speaking Background

FN = False Negative

FNR = False Negative Rate

FP = False Positive

FPR = False Positive Rate

HCR-20 = Historical Clinical and Risk Management - 20

HR = Hazard Ratio

LEAP = Law Enforcement Assistance Protocol

LIME = Local Interpretable Model-Agnostic Explanations

LS/CMI = Level of Service/Case Management Inventory

LSI-OR = Level of Service - Ontario Revised

LSI-R = Level of Service – Revised

LSI-R: SV = Level of Service Inventory-Revised Screening Version

LS/RNR = Level of Service/Risk Need Responsivity

NPV = Negative Predictive Value

PCL-R = Psychopathy Checklist - Revised

PCL: YV = Psychopathy Checklist - Youth Version

PCRA = Post Conviction Risk Assessment

PPV = Positive Predictive Value

RMGAO = Risk Management Guide for Aboriginal Offenders

ROC = Receiver Operative Characteristic

SAVRY = Structured Assessment of Violence Risk in Youth

SHAP = SHapley Additive exPlanations

SPJ = Structured Professional Judgement

TN = True Negative

TP = True Positive

VRS-SO = Violence Risk Scale - Sexual Offence

xAUC = Cross Area Under the Curve

xROC = Cross Receiver Operating Characteristic

YASI = Youth Assessment and Screening Instrument

YLS/CMI = Youth Level of Service/Case Management Inventory

YLS/CMI: AA = Youth Level of Service/Case Management Inventory: Australian Adaptation

Chapter One: Introduction to the Thesis

1.1 Background

Forensic risk assessment instruments are commonly employed in numerous countries with diverse populations to assess an individual's risk of recidivism (i.e., higher NPV; Yang et al., 2010). Forensic risk assessment originally relied upon clinical judgement (Grove & Meehl, 1996; Singh, 2012). However, they have since been replaced with more structured and reliable assessments that offer a higher level of predictive validity (i.e., accurately labelling someone as high risk who goes on to engage in recidivism; Grove & Meehl, 1996; Grove et al., 2000; Hart, 1998; Monahan, 1981). Risk assessment instruments can be classified as either actuarial risk assessments—those using a formulaic and algorithmic approach to calculate risk from empirical indicators of risk (Doyle & Dolan, 2002; Quinsey et al., 2006); or structured professional judgements (SPJ)—those that utilise a set of guidelines to aid the clinician in arriving at estimates of risk (Douglas et al., 1999; Hart et al., 2017; Webster et al., 1997). Both forms of forensic risk assessment instruments are utilised within criminal justice systems to inform bail and sentencing decisions as well as rehabilitation and interventions (Goel et al., 2018; Gutierrez et al., 2016; Monahan & Skeem, 2016; Schaefer & Hughes, 2019).

The utility of a risk assessment instrument is often assessed through discrimination indices, such as the area under the curve (AUC), that measure an instrument's ability to discriminate individuals who go on to engage in recidivism from those who do not. Risk assessment instruments are often found to have moderate levels of discrimination (Fazel et al., 2012; Singh et al., 2013), which is often comparable across differing cultural groups (e.g., Skeem & Lowenkamp, 2016; Wormith et al., 2015). However, comparable levels of discrimination are not the only way to assess fairness and do not imply that an instrument is cross-culturally fair. The cross-cultural fairness of risk assessment instruments has been an ongoing area of contention in both the academic literature and within the criminal justice system. Specifically, the notion of statistical fairness has become an emerging topic of interest (e.g., Berk & Elzarka, 2020; Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Goel et al., 2018; Mayson, 2019), with several relevant definitions of statistical fairness (e.g., error rate balance, calibration, predictive parity, and statistical parity) having been shown to demonstrate notable cross-cultural differences.

1.1.1 Defining Fairness

Various disciplines, including computer science and statistics, have provided a more complex understanding of what constitutes fairness (Verma & Rubin, 2018). This section details common definitions of fairness in relation to forensic risk assessment. This thesis will take a specific focus on forms of fairness that are related to the prediction of recidivism based on a risk assessment instrument's score.

1.1.1.1 Error Rate Balance. Error rate balance is satisfied when false positive rates and false negative rates are equal across groups (Chouldechova, 2017). The false positive rate is the proportion of non-recidivists who are classified as high risk and/or predicted to engage in recidivism. The false negative rate is the proportion of recidivists who are classified as low risk and/or predicted to not engage in recidivism.

1.1.1.2 Calibration. Calibration among groups is satisfied when the same risk score or classification on a risk assessment instrument reflects the same likelihood of recidivism (Chouldechova, 2017; Corbett-Davies & Goel, 2018; Verma & Rubin, 2018). Calibration can be understood in various ways. For example, it can involve comparing observed recidivism rates against expected recidivism rates or comparing regression equations across groups. Calibration is, therefore, satisfied when groups have similar observed recidivism rates in comparison to expected recidivism rates or regression equations that demonstrate that the relationship between recidivism and a risk assessment instrument's score is comparable across groups.

1.1.1.3 Predictive Parity. Predictive parity is satisfied when positive predictive values and negative predictive values are equal across groups (Berk et al., 2018; Chouldechova, 2017). The positive predictive value is the proportion of those classified as high risk who engage in recidivism. The negative predictive value is the proportion of those classified as low risk who do not engage in recidivism.

1.1.1.4 Statistical Parity. Lastly, statistical parity is concerned with the proportion of risk classifications (e.g., low, medium, and high risk) and risk score distributions being equal among groups (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Huq, 2019).

1.1.2 Research Problem

Cultural minorities from CANZUS nations (i.e., Canada, Australia, New Zealand, and the USA) are already found to experience ongoing disadvantages throughout their dealings with the criminal justice system, including higher arrest rates, a higher chance of being denied parole, and an over-representation in prison (Australian Bureau of Statistics, 2018a; Day, 2003; Dragomir & Tadros, 2020; Hart, 2016; Martel et al., 2011; Morrison, 2009; Shepherd, Adams, et al., 2014). The consequences that can result from unfair risk assessment instruments (i.e., risk assessment instruments not meeting fairness definitions) can further this disadvantage, having both direct legal and medical implications (Hart, 2016; Shepherd, 2018). For example, if bail and sentencing decisions are made on risk classifications that do not reflect an individual's level of risk, personal liberty or public safety can be affected (Hart, 2016). Individuals' access to appropriate treatment and management plans may also be limited (Shepherd, 2018).

This concern has also been raised in a Canadian federal court in which the use of actuarial risk assessment instruments on Indigenous individuals was challenged (*Ewert v. Canada*, 2018; Hart, 2016). The case of *Ewert v. Canada* (2015) heard that particular risk assessment instruments used to assess Indigenous individuals are potentially unreliable due to them being culturally unfair. It was specified that actuarial instruments are vulnerable to test bias that is "built-in" due to the fixed nature of risk factors that are weighted and scored by a formula or algorithm (Hart, 2016). This was also noted in 2014 by Eric Holder, the former attorney-general of the United States of America, where he stated that the use of actuarial or data-driven statistical assessments could lead to the over-criminalisation of already disadvantaged individuals (The United States Department of Justice, 2014). For example, cultural minorities are found to present with a higher number of risk factors in risk assessment instruments (e.g., unemployment, lower levels of income and substance abuse) which can be attributed to the social and economic disadvantage that they already experience (Day et al., 2018; Douglas et al., 2017; Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; Harcourt, 2007; Homel et al., 1999; Jones & Day, 2011; Shepherd, Adams, et al., 2014; Wilson & Gutierrez, 2014).

The introduction of the software-based, statistical risk assessment instrument, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), has also fuelled debate. Although the publishers and other research have shown support for COMPAS and its utility (Dieterich et al., 2016; Flores et al., 2016), it has also been scrutinised due to the limited information surrounding the inner workings of the 137 features used to predict recidivism (Kehl et al., 2017; Wisser, 2019). This was further conveyed by ProPublica, an independent and non-profit newsroom, which published an investigative report and discussed COMPAS as an instrument that was unfair for African American individuals (Angwin et al., 2016). ProPublica specifically noted that African Americans were almost twice as likely to be labelled as high risk and not go on to engage in recidivism, whereas White individuals were almost twice as likely to be labelled as low risk and engage in recidivism (Angwin et al., 2016; Larson et al., 2016). This report was subject to significant criticism due to methodological and conceptual flaws, however; further research into the use of COMPAS still identified disparities between African American and White individuals on false positives and false negatives (Flores et al., 2016).

Furthermore, there is a scarcity of literature reporting on cross-cultural fairness, specifically the varied definitions of statistical fairness, nor a clear consensus on how to increase cross-cultural fairness. Varying proposed solutions have been periodically made to increase fairness, with these seldom trialled or successful. Most often these recommendations have included the development of new culturally appropriate risk assessment instruments (Dawson, 1999; Day et al., 2018; Hart, 2016; Shepherd, Adams, et al., 2014), the alteration of existing assessments (Shepherd, 2018; Shepherd & Anthony, 2018; Shepherd & Lewis-Fernandez, 2016), further clinical training (Hart, 2016; Olver et al., 2014; Shepherd & Lewis-Fernandez, 2016), and a variety of statistical approaches including differential item functioning (Hart, 2016), factorial structures and latent constructs (Hart, 2016; Shepherd & Lewis-Fernandez, 2016), altering approaches to scoring (Skeem & Lowenkamp, 2016; Thompson & McGrath, 2012) and the use of culture as an indicator to increase accuracy (Berk, 2009; Berk et al., 2018; Douglas et al., 2017) and/or fairness (Kleinberg et al., 2018; Skeem & Lowenkamp, 2020).

Recently, disciplines such as data science, statistics, and criminology have been exploring novel statistical learning methods (i.e., machine learning) in an attempt to increase fairness (Berk et al., 2018; Chouldechova & G'Sell, 2017; Chouldechova & Roth, 2018; Corbett-Davies et al., 2017; Wadsworth et al., 2018). This approach involves different techniques that alter the algorithm at varying levels of the algorithm construction process (e.g., Berk et al., 2018; Hajian & Domingo-Ferrer, 2013; Hardt et al., 2016; Zhang et al., 2018) and has demonstrated initial promising results (e.g., Wadsworth et al., 2018), which warrants its use in the discipline of forensic psychology.

1.2 Thesis Aims and Research Questions

The objective of this suite of studies is to contribute further to the debate around cross-cultural fairness in forensic risk assessment instruments. The current literature, although expanding, is still relatively sparse, with limited research exploring the various notions of what constitutes statistical fairness and how these can be applied to risk assessment instruments. In Australia specifically, there is a paucity of research observing the cross-cultural fairness of forensic risk assessment instruments, especially among adult offenders and Aboriginal and Torres Strait Islanders. Further, there is a lack of research discussing and exploring issues that arise when trying to satisfy multiple notions of fairness due to the inherent trade-offs that exist among various types of fairness (Berk, 2019; Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Eckhouse et al., 2018; Huq, 2019; Kleinberg et al., 2016). This also extends to trying to achieve both fairness and optimising accuracy and/or the discrimination of the instrument.

Beyond the lack of exploration around fairness with respect to risk assessment instruments, there is also no clear path forward on how to address any fairness discrepancies. Although there is a body of literature detailing potential causes of these discrepancies and ways to account for them (Day et al., 2018; Douglas et al., 2017; Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; Harcourt, 2007; Homel et al., 1999; Jones & Day, 2011; Schmidt et al., 2020; Shepherd, Adams, et al., 2014; Wilson & Gutierrez, 2014), solutions are rarely, if ever, trailed nor demonstrated to succeed.

Therefore, this thesis has two overarching aims: i) to establish the levels of fairness of the risk assessment instrument, the Level of Service/Risk Needs Responsivity (LS/RNR), between male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia; and ii) to increase the fairness of the LS/RNR by utilising novel statistical approaches while still maintaining the discrimination and utility of the instrument. In achieving these aims, this thesis will explore fairness in greater detail, adding to the limited research that has investigated notions of fairness in risk assessment instruments beyond discrimination indices (e.g., the AUC). It will also demonstrate the usefulness and limitations of certain techniques (e.g., statistical learning approaches for forecasting) in reducing disparity among differing notions of fairness. Furthermore, it will explore the inherent trade-offs between differing fairness types, between fairness and maximising accuracy and/or discrimination, and between the performance of approaches such as statistical learning methods and the reduced transparency of these approaches. This study will also be able to inform interested stakeholders (e.g., Corrections Victoria) of the level of fairness present in the use of the LS/RNR with male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, as well as the potentially usefulness of statistical learning methods in increasing fairness.

The current project addressed the following research questions in order to achieve the research aims.

1.2.1 Research Question One

The first research question explores to what degree Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander males differ on the actuarial risk instrument, the LS/RNR, in terms of discrimination. This thesis addresses this question through Empirical Study One (Chapter Four).

1.2.2 Research Question Two

The second research question is to address the sparsity of research exploring statistical definitions of fairness cross-culturally, especially across Australian adult individuals. Specifically, this research question will assess the cross-cultural fairness of the LS/RNR in terms of error rate balance, calibration, predictive parity, and statistical parity between male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. This thesis addresses this question through Empirical Study One (Chapter Four).

1.2.3 Research Question Three

The third research question will assess if statistical learning methods can improve the discrimination of the LS/RNR overall and for both Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders. This research question is addressed in this thesis through Empirical Study Two (Chapter Five).

1.2.4 Research Question Four

The fourth research question will assess if statistical learning methods using differing processing approaches can increase forms of fairness (error rate balance, calibration, predictive parity, and statistical parity) while still maintaining appropriate levels of discrimination. This research question will also explore the trade-offs that are inherent to statistical forms of fairness, with the aim of achieving the best trade-off across varying forms of fairness and between fairness and discrimination possible. This is addressed in Empirical Study Two (Chapter Five) of this thesis.

1.2.5. Research Questions Five

Statistical learning methods have received criticism due to certain approaches resulting in uninterpretable algorithms. These algorithms may not specify how exactly predictors are being utilised and the impact they have on the predicted outcome and have therefore resulted in scepticism among certain proponents (Brennan & Oliver, 2013; Kehl et al., 2017; Wisser, 2019; Zeng et al., 2017). However, there are statistical approaches that can be employed post hoc that enable an analysis of the algorithm to gain an understanding of the importance of individual predictors to the predicted outcome. Therefore, the final research question will assess if these statistical learning methods can have their interpretability increased so that the impact of individual predictors can be understood. This research question is addressed in the thesis through Empirical Study Three (Chapter Six).

1.3 Overview of Thesis Structure

This thesis in total comprises seven chapters and four main parts: i) a literature review; ii) an extended general methodology; iii) empirical research studies; and iv) an integrated discussion. The following sections detail the four main parts of the thesis and what they incorporate. Within these sections, different cultures (i.e., racial groups) may be referred to in varying ways (e.g., Aboriginal and Torres Strait Islanders, First Nations or Indigenous, Black, or African American, etc.) when discussing existing literature in order to reflect the term that was used by the original authors of each study and/or to maintain consistency within a specific chapter. Terms (i.e., culture vs race) and language (i.e., Australian English vs American English) may also vary within chapters due to being published or submitted to a non-Australian journal.

1.3.1 Literature Review

Chapter Two of this thesis incorporates a wide-ranging and comprehensive literature review that has been published in *Psychology, Crime & Law*. This review encompasses a variety of literature and outlines various definitions of fairness that have previously been delineated in disciplines including computer science, criminology, and statistics. It explores how these fairness definitions can be applied to the forensic risk assessment literature, as well as highlighting a number of considerations regarding cross-cultural fairness, such as how numerous definitions of fairness are unable to be simultaneously achieved. Further, it reviews the forensic risk assessment literature, utilising the statistical definitions of fairness mentioned above, to determine the level of cross-cultural fairness currently evident. This literature review also critiques the commonly proposed suggestion for increasing fairness across groups, with future directions for achieving culturally fair forensic risk assessment instruments discussed.

1.3.2 Extended General Methodology

Chapter Three of this thesis incorporates an extended general methodology to detail the research methodology utilised within the empirical studies. It includes a description of the data, how the data was sourced, and information about the sample used in the present thesis. This chapter also comprises a detailed description of the data analytic approach utilised for all empirical studies. Ethical considerations and ethics approvals for the current thesis are also included.

1.3.3. Empirical Studies

Chapter Four of this thesis is the first empirical study that has been published in *Law and Human Behavior* and focuses on research questions one and two. This chapter focuses on the discrimination and cross-cultural fairness of the LS/RNR within a Victorian sample of adult male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders previously convicted of a serious violent offence. The discrimination of the LS/RNR was compared across cultures. Further, multiple definitions of fairness (error rate balance, calibration, predictive parity, and statistical parity) were computed for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders and compared as a way to assess the cross-cultural fairness of the LS/RNR.

Chapter Five of this thesis is the second empirical study that focuses on research questions three and four. Using statistical learning methods, this empirical study explores the use of this methodological approach in increasing the discrimination of the LS/RNR overall and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders using the LS/RNR items as predictors. Specifically, the use of logistic regression, penalised logistic regression, random forests, stochastic gradient boosting, and support vector machines were employed and compared to assess the impact that statistical learning methods can have on the discrimination of the LS/RNR. Further, this study attempts to increase the fairness (error rate balance, calibration, predictive parity, and statistical parity) between Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander individuals on the LS/RNR and measures the impact of increased fairness on the instruments' discrimination. The algorithms were altered throughout different steps of the construction of the algorithm (pre and post-processing) to attempt to increase fairness across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The trade-offs among different types of fairness and between fairness and discrimination are also explored.

Chapter Six presents the third and final empirical study that addresses the fifth research question. Notable statistical learning algorithms that aid in increasing discrimination and/or fairness are explored post hoc to increase the algorithms' interpretability. As certain algorithms and transformations to algorithms can lead to a lack of interpretability, Shapley values were calculated for the individual variables used as predictors (i.e., LS/RNR items) in the algorithms in order to gain an understanding of their importance in the overall prediction. Specifically, Shapley values were computed for each individual in the study, and then the mean absolute average Shapley value was reported to understand the importance of that predictor for predicting recidivism in this sample and also specifically for Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders.

1.3.4 Integrated Discussion

Chapter Seven incorporates an integrated discussion around the literature review and results from all empirical studies. It discusses the findings in relation to the broad research aims and specific research questions. It further highlights existing limitations within this thesis and specifies the implications of the findings and future directions for this research area.

Chapter Two: Literature Review

2.1 Introduction

This chapter presents a literature review incorporating literature from varying disciplines to provide statistical definitions of fairness that can be utilised to assess the cross-cultural fairness of a forensic risk assessment instrument. As highlighted in Chapter One, there is limited research reporting on statistical definitions of fairness in forensic risk assessment instruments across cultural groups. Although a number of independent studies have reported on certain fairness definitions (primarily statistical parity and discrimination as AUC values) when validating a risk assessment instrument, a consensus has yet to be reached about the prevalence and severity of unfairness among cultural groups. Further, although researchers have periodically suggested a myriad of ways to increase certain statistical definitions of fairness, these suggestions have rarely been tested to demonstrate their utility. This review, therefore, aimed to establish the levels of cross-cultural fairness by drawing upon literature and reporting the disparities between cultural groups on the fairness definitions discussed in Chapter One. Further, it also aimed to review the proposed solutions for increasing fairness, highlighting significant limitations in their efficacy. Last, future directions to aid in the pursuit of a cross-culturally fair forensic risk instrument are discussed.

The literature review in this chapter, titled *"Cross-Cultural Fairness in Forensic Risk Instruments: A Review of the Literature"*, is the authors' original manuscript of an article published by Taylor & Francis Group in *Psychology, Crime & Law* on August 27, 2021, available online: https://doi.org/10.1080/1068316X.2021.1972108. *Psychology, Crime & Law* is a peer-reviewed journal that focuses on the application of psychological methods to crime, criminal behaviour, and the law. Some of the language in the published version of this article has been altered to better reflect the language used internationally. Specifically, "culture" was referred to as “race” within the published version of the article to avoid confusion among international readers. The Author Indication Form that details the contribution of each author to this manuscript is included in Appendix A.

The citation for the published version of this article is as follows:

Ashford, L.J., Spivak, B.L., & Shepherd, S.M. (2021). Racial fairness in violence risk instruments: A review of the literature. *Psychology, Crime & Law*. doi: 10.1080/1068316X.2021.1972108

Cross-Cultural Fairness in Forensic Risk Instruments: A Review of the Literature

Linda J. Ashford, Benjamin L. Spivak & Stephane M. Shepherd

Author Note.

Linda J. Ashford, Centre for Forensic Behavioural Science, Swinburne University of Technology, 1/582 Heidelberg Rd, Alphington, Victoria, lashford@swin.edu.au; Benjamin L. Spivak, Centre for Forensic Behavioural Science, Swinburne University of Technology, 1/582 Heidelberg Rd, Alphington, Victoria, bspivak@swin.edu.au; Stephane M. Shepherd, Centre for Forensic Behavioural Science, Swinburne University of Technology, 1/582 Heidelberg Rd, Alphington, Victoria, sshepherd@swin.edu.au.

This research was funded by Project: DE180100933 from the Australian Research Council.

Corresponding author: Linda J. Ashford, Centre for Forensic Behavioural Science, Swinburne University of Technology, 1/582 Heidelberg Rd, Alphington, Victoria, lashford@swin.edu.au

Abstract

Forensic risk assessment instruments are used in numerous countries to estimate an individual's likelihood of recidivism. The cross-cultural fairness of forensic risk assessment instruments has received increasing attention due to ostensible risk assessment differences among Anglo populations and cultural minorities (e.g., African Americans and Indigenous populations). Fairness, which has numerous definitions (sensitivity fairness, error rate balance, calibration, predictive parity, statistical parity), can affect a risk assessment's utility in varying ways. This literature review explored how notions of fairness are discussed in the risk assessment literature, with a specific focus on cross-cultural fairness. It also examined and critiqued the varying proposed resolutions to increase fairness. Many of these forms of fairness were found to be rarely satisfied in the literature. Further, the complications in achieving multiple forms of fairness simultaneously and the challenges of optimising both fairness and accuracy are discussed. Last, proposed solutions to increase cross-cultural fairness were often found to encompass significant limitations. Future directions for cross-cultural fairness in risk assessment are discussed, with a focus on exploring the trade-offs among varying fairness definitions and between fairness and accuracy.

Keywords: cross-cultural risk assessment; fairness; forensic risk assessment; forensic psychology

Forensic Risk Assessment Instruments

Forensic risk assessment instruments assess an individual's risk of recidivism (Yang et al., 2010). They also inform offender management and treatment plans, as well as decisions around parole and sentencing (Gutierrez et al., 2016; Monahan & Skeem, 2016; Schaefer & Hughes, 2019). Assessing recidivism historically relied upon unstructured clinical judgement (i.e., informal decisions based on clinical experience, theory, and instinct) to assess an individual's level of risk (Grove & Meehl, 1996; Singh, 2012; Westen & Weinberger, 2004). However, unstructured judgements of risk were found to be unreliable, arbitrary, lacking in transparency, and inaccurate. They were also outperformed by assessments that were mechanical, formulaic, and/or relied on statistical algorithms (Grove & Meehl, 1996; Grove et al., 2000; Hart, 1998; Monahan, 1981). Accordingly, the late twentieth century saw an increased focus on developing assessments that were structured, transparent, and evidence-informed (Bonta, 1996; Hurducas et al., 2014; Singh, 2012).

Risk assessment instruments that rely on mechanical and formulaic processes are commonly referred to as "actuarial risk assessment instruments". These instruments are scored by a formula or statistical algorithm that combines the numerical values assigned to factors found to be empirically related to offending (Doyle & Dolan, 2002; Quinsey et al., 2006). The score obtained from an actuarial assessment can be used as a probabilistic estimate of recidivism or further categorised into varying levels (e.g., low, medium, and high) of risk (Doyle & Dolan, 2002; Singh, 2012). However, this type of instrument has been criticised for its rigid structure and nomothetic approach, which may ignore case-specific information (Doyle & Dolan, 2002; Rettenberger et al., 2011). This led to the development of structured professional judgement (SPJ) instruments which allow the clinician to determine an individual's level of risk after considering a set of evidence-based risk factors (Douglas et al., 1999; Hart et al., 2017; Webster et al., 1997). Both actuarial and SPJ assessments have been found to yield similar predictive accuracy estimates (Fazel et al., 2012; Singh et al., 2011). However, they have been criticised for their potential to be unfair towards different minority groups (Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; van Eijk, 2017).

Fairness (i.e., the equal treatment across groups) in forensic risk assessment has been a recent topic of contention in the scientific literature and its consequences for minority individuals, especially cultural minorities (e.g., Day et al., 2018; *Ewert v. Canada*, 2018; Hart, 2016; Shepherd & Lewis-Fernandez, 2016). These concerns have also been raised by those in high office (i.e., Eric Holder, the former Attorney General of the United States; The United States Department of Justice, 2014) and in journalistic investigations (Angwin et al., 2016) where it was proffered that risk assessment instruments (notably actuarial instruments) are disadvantaging already vulnerable populations such as cultural minority groups. These recent occurrences warrant a deeper exploration into the concepts of fairness in risk assessment across cultural groups. Some cultural minority populations are already reported to experience differential contact and treatment within the criminal justice system, including higher arrest rates, a higher chance of being denied parole, and an over-representation in prison (Australian Bureau of Statistics, 2018a; Day, 2003; Dragomir & Tadros, 2020; Hart, 2016; Martel et al., 2011; Shepherd, Adams, et al., 2014). As such, a review of the existing literature to establish the levels of fairness present across cultural groups is necessary to explore how unfairness in risk assessment instruments may be affecting specific cultural groups' interactions within the criminal justice system and how to best rectify these disparities.

This literature review will therefore discuss the concept of fairness in risk assessment in a number of different ways, using examples from the literature. The specific aims of the current literature review are to: (a) review and discuss different types of fairness; (b) examine the risk assessment literature to assess the level of fairness between different cultural groups; (c) ascertain the various explanations in the literature for cross-cultural unfairness; and (d) examine and critique the commonly proposed solutions in the literature aimed at increasing cross-cultural fairness.

Fairness

There are several definitions of what constitutes fairness, many of which have been delineated by researchers in the disciplines of computer science, criminology, and statistics (Verma & Rubin, 2018). Five common notions of fairness are discussed below.

Sensitivity Fairness

Sensitivity is the proportion of those accurately predicted to engage in recidivism from those who were recidivists. Sensitivity fairness refers to sensitivity parity across groups, also known as equality of opportunity (Hardt et al., 2016). If sensitivity fairness is not satisfied, a risk assessment is labelling fewer actual recidivists as high risk for one group compared to another. Sensitivity is commonly examined alongside an instrument's specificity—the proportion of those accurately predicted to not engage in recidivism from those who were not recidivists. Sensitivity and specificity estimates are used to plot the receiver operating characteristic (ROC) curve. The ROC curve plots the sensitivity against 1—specificity (also referred to as the false positive rate) at various thresholds (Cook, 2007; Singh, 2013). The area under the ROC curve is a common discrimination index that assesses how well a risk assessment can distinguish between those who engage in recidivism from those who do not (Cook, 2007). The area under the curve (AUC) can range from 0 to 1, with the midpoint (.50) demonstrating discrimination at chance levels (Cook, 2007; Rice & Harris, 2005). Interpreting an AUC value has been met with some difficulties due to the various benchmarks of what constitutes small, moderate, and large effect sizes (Cohen, 1988; Rice & Harris, 2005; Singh et al., 2013; Swets, 1988). The AUC, however, can be best understood as the probability that a randomly selected individual who engaged in recidivism received a higher risk score than a randomly selected individual who did not engage in recidivism (Singh, 2013; Swets et al., 2000). The AUC value is frequently reported in the risk assessment literature (Helmus & Babchishin, 2017; Singh et al., 2013), and a comparison of AUC values across groups has been used to indicate assessment fairness (e.g., Dieterich et al., 2016). However, AUC values can be approximately equal among groups even with varying sensitivity values as the AUC incorporates sensitivity alongside 1 – specificity at varying thresholds. Nevertheless, if one group is found to have a considerably lower AUC, this indicates that for this group, the instrument is unable to discriminate recidivists from non-recidivists as effectively.

Error Rate Balance

Error rate balance is satisfied when the false positive rate (FPR), or the proportion of non-recidivists classified as high risk, and false negative rates (FNR), or the proportion of recidivists classified as low risk, are equal across groups (Chouldechova, 2017). Error rate balance is similar to sensitivity fairness in that the FNR is the opposite of sensitivity. Sensitivity (and specificity) are classifications, whereas FNR and FPRs are errors in observation. When this form of fairness is unsatisfied, there are differences between groups in the number of errors in observation. A difference in error rate balance across groups could lead to one group being disproportionately disadvantaged. For example, a group with a higher FPR could lead to an impact on the personal liberties of these individuals if they are unnecessarily detained. Conversely, a group with a higher FNR indicates that a higher proportion of recidivists from this group were classified as low risk and were potentially incorrectly released into the public where they engaged in recidivism.

Calibration

Calibration is concerned with the similarity between expected recidivism and observed recidivism across risk scores or classifications. Calibration among groups is satisfied when a risk assessment score or classification reflects the same proportion of recidivists across different groups (Chouldechova, 2017; Corbett-Davies & Goel, 2018; Verma & Rubin, 2018). Occasionally, calibration can involve comparing expected recidivism that is based on the established normative data of a risk assessment instrument (e.g., an assessment predicts that 50% of individuals with a specific risk score will engage in recidivism within 2 years) against observed recidivism of groups (Fazel, 2019; Hanson, 2017; Helmus & Babchishin, 2017). However, only actuarial instruments will have empirical norms, and these are not always provided by instrument developers (Dawes et al., 1989; Gutierrez et al., 2016; Hanson, 2017).

Calibration can be demonstrated through the comparison of regression equations across groups (or against normative data), with similar equations demonstrating comparable recidivism rates across risk assessment scores and a well calibrated instrument (Hanson, 2017). It can also be demonstrated by calculating the E/O index for each group, which is the ratio of the expected (E) number of recidivists divided by the observed (O) number of recidivists (Hanson, 2017; Helmus & Babchishin, 2017; Viallon et al., 2009). When calibration is not satisfied, it indicates that for the same scores/classifications, a risk assessment instrument is predicting recidivism differently for different groups.

Among certain proponents, calibration is discussed as the most logical and important fairness definition (Hébert-Johnson et al., 2018; Royal Statistical Society, 2018) and as another form of predictive validity (Cook, 2007; Fazel, 2019; Singh, 2013). Therefore, when an instrument is not well calibrated, not only does it compromise fairness, but the utility of the instrument for making predictions also comes into question (Hanson, 2017; Lindhiem et al., 2018).

Predictive Parity

Predictive parity is achieved when the positive predictive values (PPV), or the proportion of those who engage in recidivism from those predicted to engage in recidivism (i.e., high risk), are equal across groups (Chouldechova, 2017). Berk et al. (2018) expanded on this in their definition of conditional use accuracy in which equality is also achieved across negative predictive values (NPV), or those who do not go on to engage in recidivism from those predicted to not engage in recidivism (i.e., low risk). Predictive parity differs from calibration in that it is based on a single cut-off score that distinguishes low risk from high risk, instead of across all scores and risk classifications. When this form of fairness is unsatisfied, it signifies that high risk and low risk classifications result in a differing number of recidivists across groups. For example, if one group has a lower NPV, a low risk classification for this group incorporates a higher number of recidivists. If one group has a lower PPV, a high risk classification for this group would incorporate a higher number of non-recidivists. Both of these outcomes lead to predictions that do not reflect the same risk across groups.

Statistical Parity

Statistical parity requires the proportions of classifications (i.e., those predicted to be at low or high risk) to be equal across groups (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Huq, 2019). When this form of fairness is not satisfied, one group is more likely to be labelled as high risk and the other as low risk. Being labelled as high risk more consistently could lead to the perception that one group is at a higher risk of recidivism, or more harmful, than another group.

Limitations of Fairness

It has been established that multiple forms of fairness cannot be achieved simultaneously when group base rates differ, often referred to as an impossibility theorem

(Berk, 2019; Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Eckhouse et al., 2018; Huq, 2019; Kleinberg et al., 2016). Specifically, predictive parity and error rate balance (or sensitivity fairness) cannot be simultaneously achieved when base rates differ as the same values are used in different ways to calculate these fairness metrics. Policymakers will need to determine what form of fairness is prioritised as total fairness (i.e., all forms of fairness being satisfied) is impossible (Berk et al., 2018). They will need to decide, for example, if fairness in the accuracy of predictions (predictive parity) or fairness in the errors in observation (error rate balance) and a risk assessment instrument's ability to discriminate recidivists from non-recidivists (AUC values) is more important. If comparable AUC values were prioritised and a risk assessment was able to discriminate recidivists from non-recidivists equally well across groups, predictive parity would necessarily have inequality across groups such that a high risk classification does not equally predict recidivism.

Further, the statistical definitions of fairness outlined above may not entirely align with what is stipulated by policymakers (Corbett-Davies & Goel, 2018). For example, achieving fairness in statistical parity could involve increasing the number of low risk classifications for one group to be equivalent to the other. This could result in more individuals from that group then being classified as low risk and engaging in recidivism (Berk, 2019; Dwork et al., 2012), an outcome unlikely to be viewed as desirable. This highlights a significant limitation with statistical parity as a form of fairness. Statistical parity is solely concerned with equal rates of acceptance into low and high risk classifications (Chouldechova, 2017), and does not take into account the outcome of recidivism. This can lead to adverse decisions in which groups are classified to achieve parity, potentially contravening other definitions of fairness and leading to lower predictive validity for a specific group or overall. It has therefore been cautioned against as a form of fairness (Berk et al., 2018; Chouldechova, 2017).

In addition to this impossibility theorem, it is also unlikely for fairness definitions to be perfectly satisfied (i.e., zero reported difference between groups). What constitutes an acceptable level of fairness is hard to ascertain. Relying on significance testing poses problems as it assumes the null hypothesis (in this case, no difference between groups on fairness) and makes finding no difference difficult (Amrhein et al., 2017; Cohen, 2011; Fisher, 1925). A non-significant finding does not translate into no effect or a non-meaningful difference between groups (Amrhein et al., 2017; Biau et al., 2010; Cohen, 2011; Lakens et al., 2020). Therefore, it does not signify that fairness has been satisfied. To overcome this, Bayesian analyses or equivalence testing could be used to assess the presence or absence of an effect (Gallistel, 2009; Lakens, 2017; Lakens et al., 2020; Lakens et al., 2018). In short, Bayesian analysis will take into account the null hypothesis and a feasible alternate hypothesis (considering the data/theory), for which the probability favouring either the null or alternative hypothesis is then calculated (Gallistel, 2009; Lakens et al., 2020). Equivalence testing sets an upper and lower bound of equivalence based on the smallest effect size deemed to be meaningful (e.g., 0 ± 0.2). One-sided significant testing is performed against each equivalence bound to assess if a meaningful effect can be rejected, demonstrating a practical equivalence to zero effect (Lakens et al., 2020; Lakens et al., 2018).

However, the idea of having any statistical benchmark to distinguish what constitutes fairness could be considered inappropriate. What is fair is in the eye of the beholder, or in this case, the policymakers, developers, and users of the instruments. Therefore, initially, it might be useful to firstly state if unfairness exists (i.e., are the groups performing perfectly equally or is there a difference?), and then secondly, to what degree is it unfair (i.e., how different are these groups performing from one another?). For example, is it a difference that would negatively affect 1% of individuals or 25% of individuals from one group?

Another consideration is the cut-off score or binning strategy used to create a low and high risk group. Many forms of fairness rely on a single cut-off threshold, and numerous risk assessment instruments have more than two risk classifications (Singh, 2013). Different strategies used to develop two groups include using the median score as the cut-off (Shepherd & Strand, 2016), or the high risk classification being compared against other classifications combined (Larson et al., 2016; Shepherd et al., 2015). Different cut-offs used will inevitably lead to variations in values as the proportions in low and high risk classifications change. As a way to overcome this, two comparisons can be made: the high risk classification is compared against all lower risk classifications combined, and then the low risk classification is compared against all higher risk classifications combined (Singh, 2013; Singh et al., 2011).

Cross-Cultural Fairness in Risk Instruments

The following section will take examples from the cross-cultural risk assessment literature and its consideration of fairness as defined above.

Sensitivity Fairness

Sensitivity is not often reported in isolation in the risk assessment literature as a measure of fairness; however, there is a plethora of studies that have calculated sensitivity alongside 1 – specificity to plot ROC curves and report AUC values. The extensive research comparing the AUC among cultural groups has most often reported that they are generally commensurate among both cultural minorities (e.g., African Americans and Indigenous and First Nations populations of Australia and North America) and Anglo/White populations, with AUC values ranging between .60 and .80 for both cultural groups (e.g., Dieterich et al., 2016; Jones et al., 2016; Lee et al., 2020; Lee et al., 2019; Muir et al., 2020; Olver et al., 2013; Olver et al., 2018; Perrault et al., 2017; Skeem & Lowenkamp, 2016; Thompson & McGrath, 2012; Watkins, 2011; Wormith et al., 2015). These comparable AUC values are found on commonly used instruments, including the Level of Service (LS) measures (Watkins, 2011; Wormith et al., 2015), varying youth instruments (Jones et al., 2016; Muir et al., 2020; Perrault et al., 2017; Thompson & McGrath, 2012), and instruments tailored for sexual recidivists (Lee et al., 2020; Lee et al., 2019; Olver et al., 2018). There are some exceptions (e.g., Helmus et al., 2012; Långström, 2004; Molnar et al., 2020; Shepherd, Luebbers, et al., 2014; Shepherd et al., 2015; Shepherd & Strand, 2016) where wide discrimination differences have been identified between cultural groups, occasionally resulting in instruments being unable to discriminate recidivists from non-recidivists better than chance for some cultural minority groupings (i.e., AUC values, or confidence interval ranges for AUC values, fall below .50). The majority of these studies that identified pronounced differences were smaller sampled studies that predominately focused on Indigenous individuals from Canada and Australia. Despite these findings, scholars have suggested that the generally robust discrimination estimates reported across cultures may be indicative of the continued use of these instruments cross-culturally (e.g., Jones et al., 2016; Lee et al., 2019; Olver et al., 2014; Wormith et al., 2015). However, as highlighted above, comparable AUC values do not indicate that other notions of fairness are satisfied.

Error Rate Balance

The FPR and FNR are scarcely reported in the cross-cultural risk assessment literature. For example, Whiteacre (2006) identified differences among African American, Caucasian, and Hispanic individuals from the United States on the Level of Service Inventory-Revised (LSI-R). The LSI-R was utilised to classify individuals as low or high risk (using cut-off scores of 16 and 25) of a disciplinary incident occurring within a community corrections centre, as well as an unsuccessful program outcome (e.g., rearrest). Among the majority of outcomes with a 25 cut-off, error rates were relatively comparable across all cultures. However, differences were observed for both outcomes when the cut-off was lowered to 16. For both disciplinary incidents and unsuccessful program completion, the African American group had notably higher FPRs (FPRs = .46 and .49, respectively), indicating that they were classified as high risk and did not go on to engage in recidivism more often than Caucasian (FPRs = .26 and .30) and Hispanic (FPRs = .23 and .27) individuals. The reverse was shown for FNRs, in which the African American group had notably lower FNRs for both disciplinary incidents and unsuccessful program completion (FNRs = .36 and .17, respectively) compared to Caucasian (FNRs = .48 and .36) and Hispanic individuals (FNR = .50 for disciplinary incidents). No Hispanic individuals were labelled a false negative for unsuccessful program completion.

Differences in error rate balance have also been demonstrated on the software instrument, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), by combining the medium and high risk classifications to compare against the low risk classification group (Larson et al., 2016). Similar to Whiteacre (2006), the FPRs demonstrated that Black individuals on both general and violent recidivism outcomes (FPRs = .45 and .38 respectively) were almost twice as likely to be classified as high risk and not go on to engage in recidivism when compared to White individuals (FPRs = .23 and .18). The reverse was again demonstrated on the FNRs, in which Larson et al. (2016) reported that White individuals for both general and violent recidivism (FNRs = .48 and .63 respectively) were almost twice as likely to be classified as low risk and not engage in recidivism when compared to Black individuals (FNRs = .28 and .38). Taking these values for general recidivism in a real-world context, out of 1,000 Black non-recidivists, 450 would be labelled as high risk compared to 230 out of 1,000 White non-recidivists. Conversely, out of 1,000 Black recidivists, 280 would be labelled as low risk compared to 480 out of 1,000 White recidivists. This analysis was met with criticism as the binning strategy did not include comparing a combined low and medium risk category against a high risk category (Flores et al., 2016). Flores et al. (2016) reanalyzed the COMPAS data with this binning strategy for general recidivism. The FPR differences between Black (FPR = .14) and White (FPR = .05) individuals were reduced, and the FNR differences between Black (FNR = .62) and White (FNR = .80) individuals remained similar.

The FPRs reported in the above literature indicate that African American individuals are often almost twice as likely to be labelled high risk and not engage in recidivism, potentially resulting in incorrect treatment approaches and being wrongly incarcerated. Simultaneously, the higher FNR found among White people suggests that they were nearly twice as likely to be labelled as low risk and later engage in recidivism However, there is very limited research exploring error rates, especially among other cultural minorities (e.g., Indigenous groups).

Calibration

The majority of the cross-cultural risk assessment literature calculating calibration has compared regression equations or observed recidivism by risk score/classification across cultures. Two studies conducted in the United States compared regression equations obtained from White sexual offenders and cultural minorities, including Black and Hispanic sex offenders, against Static-99R norms at 5 years (Hanson et al., 2014; Lee et al., 2019). The Static-99R expected recidivism rates were significantly higher than the observed rates in all cultures (Hanson et al., 2014; Lee et al., 2019). When comparing regression equations among cultural groups, no meaningful differences were reported between intercept values or the slope. Lee et al. (2019) also reported the E/O index based on 5 year sexual recidivism norms. The Static-99R overpredicted recidivism slightly more for White individuals (E/O = 2.03) than for Black individuals (E/O = 1.63). White and Black individuals have also been shown to have similar regression equations and observed recidivism across risk scores on the software instrument COMPAS (Flores et al., 2016) and the actuarial instrument Post-Conviction Risk Assessment (PCRA; Skeem & Lowenkamp, 2016) for general recidivism. For violent recidivism, however, Skeem and Lowenkamp (2016) reported a significantly lower intercept for White individuals, demonstrating that Black individuals were predicted to engage in violent recidivism more often on the PCRA.

Similarities in recidivism by risk classification have also been found for Aboriginal and non-Aboriginal Canadian youth on the Youth Assessment and Screening Instrument (YASI; Jones et al., 2016), with calibration almost being perfectly satisfied among the low (7.4% and 5.1% respectively) and high risk (50% and 48.4%) classifications. Only in the medium risk classification did this difference increase, with Aboriginal people engaging in recidivism at a higher rate (23.5% versus 16.8%). These similarities were not reflected among young Indigenous Australians on the Youth Level of Service/Case Management Inventory: Australian Adaptation (YLS/CMI: AA; Thompson & McGrath, 2012). Indigenous youth had a higher recidivism rate in the high, medium, and low risk (53.0%, 66.6%, and 75.5%, respectively) classifications. The Australian group (those defined as neither Indigenous nor ethnic) had the second highest (38.1%, 55.6%, and 69.6%) and was followed by the Ethnic group (those defined as having a non-Australian cultural background) with the lowest rates of recidivism per risk classification (31.3%, 52.0%, and 68.3%).

Consistent differences in calibration have been reported among adult Indigenous and Anglo (i.e., White or non-Indigenous) individuals. Lee et al. (2020) found that White Canadians engaged in recidivism at rates significantly lower than both the norms and Indigenous individuals at 5 years on both the Static-99R and Static-2002R, as well as significantly lower than the Static-99R norms at 10 years (Lee et al., 2020). As demonstrated by a larger intercept value, Aboriginal Canadian individuals were also predicted to engage in violent recidivism more often than non-Aboriginal individuals on the actuarial Violence Risk Scale-Sexual Offender version (VRS-SO; Olver et al., 2018). This difference was reduced when observing sexual recidivism as an outcome; however, Aboriginal individuals were still predicted to engage in recidivism more among lower risk scores, with the converse shown among high risk scores (Olver et al., 2018). This finding among lower risk scores has also been reported among Aboriginal and non-Aboriginal groups on the Level of Service Inventory-Ontario Revised (LSI-OR; Wilson & Gutierrez, 2014; Wormith & Hogg, 2012) and the Level of Service/Case Management Inventory (LS/CMI; Wormith et al., 2015), with Aboriginal individuals predicted to engage in recidivism more among lower risk classifications.

The above examples from the literature exploring calibration demonstrate that an instrument's pre-existing norms of expected recidivism rarely align with observed recidivism, regardless of culture. When comparing calibration among cultures, African American, Hispanic, and White American individuals often demonstrated similar regression equations and recidivism rates across risk scores and classifications. However, differences were frequently reported between Indigenous and Anglo individuals. Indigenous individuals were found to recidivate more, especially among lower risk classifications. This demonstrates that these risk assessment instruments are not well calibrated across these cultural groups, such that the same risk classification given to an Indigenous individual does not result in the same chance of recidivism as an Anglo individual, specifically among lower risk classifications.

Predictive Parity

Similar to error rates, few studies in the cross-cultural risk assessment literature have reported PPVs and NPVs. When comparing the PPVs on both general and violent recidivism outcomes on the COMPAS software, Larson et al. (2016) reported that Black (PPVs = .63 and .21 respectively) and White American (PPVs = .59 and .17) individuals had relatively close values. NPVs were also similar and, in this case, White individuals had a higher NPV on both general and violent recidivism outcomes (NPVs = .71 and .93 respectively) compared to Black individuals (NPVs = .65 and .91). Flores et al. (2016) analysed the same dataset, however, with a combined low and medium risk classification compared against high risk. Using this binning strategy, the PPVs for Black and White individuals almost satisfied predictive parity (PPVs = .75 and .73, respectively). However, the NPVs reported by Flores et al. (2016) for Black (NPV = .56) and White (NPV = .65) individuals differed to a greater degree.

Another United States study examined the LSI-R for classification differences among cultures, using two score cut-offs, 25 and 16, to distinguish between low and high risk for disciplinary incidents and unsuccessful program completion (Whiteacre, 2006). For unsuccessful program outcomes, the PPVs and NPVs were relatively similar across all cultures, with more pronounced differences found among disciplinary incidents. With a cut-off score of 25, the Hispanic group had the highest proportion of high risk scores predicting a disciplinary incident (PPV = .71), followed by the African American (PPV = .56) and the Caucasian groups (PPV = .40). Lowering the cut-off score to 16 did demonstrate predictive parity between the African American and Hispanic groups (PPVs = .50), with the Caucasian group still lower (PPV = .38). NPVs also varied across cultures with a cut-off score of 25 and 16, with the highest being identified for the Caucasian group (NPVs = .80 and .84), followed by the Hispanic group (NPVs = .72 and .77), and the African American group (NPVs = .61 and .68).

Two Australian studies observed differences in PPVs and NPVs among Indigenous, culturally and linguistically diverse (CALD) and English speaking background (ESB) young individuals. The proportion of high risk classifications that engaged in recidivism was highest for the Indigenous group for both general and violent recidivism on the Youth Level of Service/Case Management Inventory (YLS/CMI; PPVs = .87 and .74, respectively) and the Psychopathy Checklist: Youth Version (PCL: YV; PPVs = .92 and .83; Shepherd et al., 2015; Shepherd & Strand, 2016). Among ESB and CALD individuals, Shepherd et al. (2015) reported that the PPVs across general recidivism almost reached predictive parity (PPVs = .80 and .81, respectively) and did satisfy predictive parity for violent recidivism (PPVs = .65 for both) on the YLS/CMI. On the PCL: YV, Shepherd and Strand (2016) identified that the high risk ESB group was found to engage in recidivism more (PPVs = .81 and .71) compared to the high risk CALD group (PPVs = .76 and .52). Predictive parity was also unsatisfied across NPVs, with the Indigenous group having the lowest NPVs for both general and violent recidivism on the YLS/CMI (NPVs = .22 and .44, respectively) and the PCL: YV (NPVs = .18 and .41). On the YLS/CMI, for general recidivism, the low risk CALD group (NPV = .41) engaged in recidivism slightly less than the ESB group (NPV = .35), whereas for violent recidivism, the low risk CALD group (NPV = .53) engaged in recidivism slightly more than the ESB group (NPV = .60). On the PCL: YV, the low risk ESB youth engaged in recidivism less for both general and violent outcomes (NPVs = .41 and .63, respectively), compared to the CALD group (NPVs = .35 and .45).

A study with young Canadian individuals also found similar differences on the SPJ instrument the Structured Assessment of Violence Risk in Youth (SAVRY; Muir et al., 2020). Male and female high risk Indigenous individuals were again found to engage in recidivism more both generally (PPVs = .74 and .77, respectively) and violently (PPVs = .51 and .52) compared to male and female Caucasian individuals, who had smaller PPVs on both general (PPVs = .64 and .61, respectively) and violent (PPVs .33 and .28) recidivism outcomes. The converse was again shown for NPVs, with male and female low risk Caucasian individuals engaging in recidivism less both generally (NPVs = .78 and .70, respectively) and violently (NPVs .91 and .92) compared to male and female Indigenous individuals who had smaller NPVs on general (NPVs = .57 and .67, respectively) and violent (NPVs .85 and .87) recidivism outcomes.

The above examples demonstrate a lack of predictive parity in the literature. A high risk classification is less likely to result in recidivism among White individuals, and a low risk classification is more likely to result in recidivism among cultural minorities. Similar to the disparities between cultures identified among calibration, risk assessment classifications are not predicting recidivism the same cross-culturally. When recidivism rates differ by risk classification, cultures may be either advantaged or disadvantaged both legally and medically when using classifications to aid in decision making (Hart, 2016; Shepherd, 2018). However, different to calibration, African American and White individuals were found to differ more among predictive parity metrics when using a single cut-off value compared to calibration that observed recidivism over several risk scores or classifications. This demonstrates that when scores or risk classifications are well calibrated, they can still lead to unfairness in predictive parity at a certain cut-off score (Chouldechova, 2017). It is also worth reiterating the complications of satisfying both predictive parity and error rate balance when base rates differ. If predictive parity were hypothetically satisfied, more disparities would be observed among error rates.

Statistical Parity

Numerous studies in the cross-cultural literature report average risk assessment scores and score distributions, making statistical parity an easy form of fairness to discuss comparisons. A large scale meta-analytic review by Olver et al. (2014) observing a variety of LS measures identified that cultural minorities (i.e., African Americans, Indigenous, Asian, and Hispanic) from numerous countries scored significantly higher than non-minorities (d = 0.24). Higher scores and subsequent higher risk classifications among minorities (i.e., African American and Indigenous individuals) have been consistently found in studies from the United States, Canada and Australia using the actuarial LS measures, including the LSI-R (Chenane et al., 2015; Holsinger et al., 2003, 2006; Hsu et al., 2010; Watkins, 2011; Whiteacre, 2006), the LSI-OR (Wilson & Gutierrez, 2014), and the LS/CMI (Jimenez et al., 2018; Wormith & Hogg, 2012; Wormith et al., 2015). These differences ranged between small (lowest d = 0.21) and large (d = 1.02) in effect size.

On other actuarial risk assessment instruments including the software instrument COMPAS (Angwin et al., 2016; Flores et al., 2016; Larson et al., 2016), the PCRA (Skeem & Lowenkamp, 2016) the Psychopathy Checklist-Revised (PCL-R; Olver et al., 2013), and instruments designed to predict sexual recidivism such as the Static-99 (Hanson et al., 2014; Smallbone & Rallings, 2013), Static-99R (Hanson et al., 2014; Lee et al., 2020; Lee et al., 2019; Olver et al., 2018; Smallbone & Rallings, 2013), Static-2002R (Lee et al., 2020), the VRS-SO (Olver et al., 2018), and the STABLE-2007 (Helmus et al., 2012), cultural minorities are again reported to score higher and be more likely classified as high risk (*d* ranged between 0.27 and 0.50).

For actuarial youth risk assessment instruments, cultural minority individuals scored higher on instruments including the YASI (Jones et al., 2016), the PCL: YV (Schmidt et al., 2006), and among youth LS measures including the YLS/CMI (Shepherd et al., 2015; Thompson & McGrath, 2012) and the YLS/CMI: AA (d ranged between 0.22 and 0.77; Frize et al., 2008; Kenny & Nelson, 2008; McGrath et al., 2018). Further, both Canadian and Australian Indigenous individuals were more likely to be classified as high risk (*d* ranged between 0.22 and 0.77) on the SPJ instrument the SAVRY (Muir et al., 2020; Shepherd, Luebbers, et al., 2014). However, there are occasional examples in which statistical parity is nearly satisfied. Perrault et al. (2017) reported similar scoring and risk classification among young White and Black individuals on the SAVRY and the YLS/CMI. On both assessments, White individuals scored higher, but the differences were trivial in effect size (*d* = 0.06 and 0.02). Similarly, Shepherd and Strand (2016) demonstrated that no specific cultural group was more likely to be labelled low or high risk on the PCL: YV, with Indigenous, ESB, and CALD young Australians scoring comparably. These differences may be due to the cultures under study. The near parity identified by Perrault et al. (2017) differed from other studies observing youth by comparing White and Black individuals instead of an Indigenous cohort. Further, the similarities reported by Shepherd and Strand (2016) among Indigenous, ESB, and CALD youth was the sole study using the PCL: YV for youth in Australia.

For the majority of cross-cultural validation studies, cultural minorities appear to score higher and are classified as having a higher risk of recidivism when compared to predominantly Anglo individuals. Although these differences are often small in magnitude, occasionally they are more pronounced (e.g., Wormith et al., 2015) and demonstrate a lack of statistical parity between cultures on risk assessment instruments. It has been argued that unfairness among statistical parity could lead to negative labelling of specific cultural groups such as cultural minorities, potentially exacerbating and directly contributing to the ongoing inequality already experienced by cultural minorities within the criminal justice system (e.g., higher arrest rates and denial of bail; Hannah-Moffat, 2013; Martel et al., 2011; Shepherd, Adams, et al., 2014). However, as discussed by Skeem and Lowenkamp (2016), differing risk levels may be reflective of actual population differences in risk, and aiming to establish statistical parity can indeed have negative consequences. It could cause or exacerbate already existing differences among other forms of fairness by misclassifying individuals. It could also decrease the predictive validity of the assessments for certain cultural groups or overall, ultimately impeding the utility of the instrument.

Explanations for Unfairness among Cultural Groups

Various explanations have been proposed to unpack the differences found between cultures on risk assessment instruments.

Risk Assessment Development

The majority of risk assessment instruments were originally developed and validated on predominately Anglo samples originating from North America (Hannah-Moffat, 2013; Olver et al., 2014; Shepherd, 2018; Singh et al., 2011; Wilson & Gutierrez, 2014). This has led to concerns using risk assessment instruments on non-Anglo populations or countries outside of where the risk assessment was developed, as the risk factors and items may not be transferable or the most relevant indicators of risk (Day et al., 2018; Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; Schmidt et al., 2020; Shepherd, Adams, et al., 2014). This concern has been demonstrated with studies conducted in Canada producing larger predictive validity effect sizes on a variety of measures and recidivism outcomes (Leistico et al., 2008; Olver et al., 2009) compared to other countries such as the United States (Olver et al., 2014) and Australia (Gutierrez et al., 2013). The same has been identified for different cultures, with larger effect sizes being reported for Anglo populations compared to cultural minorities (e.g., Edens et al., 2007; Gutierrez et al., 2013; Leistico et al., 2008; Singh et al., 2011; Wilson & Gutierrez, 2014; Wormith et al., 2015).

Risk Factor Prevalence

The prevalence of risk factors has also been found to differ. Some cultural minority populations, for instance, often display a higher number of risk factors on risk assessment instruments due to ongoing economic and social disadvantage that leads to higher levels of unemployment, previous criminal histories (potentially due to racial discrimination and policing), substance use, and lower levels of income, which may directly lead to a higher risk classification on risk assessment instruments (Day et al., 2018; Douglas et al., 2017; Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; Harcourt, 2007; Homel et al., 1999; Jones & Day, 2011; Shepherd, Adams, et al., 2014; Wilson & Gutierrez, 2014). This higher prevalence of risk factors and therefore a higher risk classification could directly contribute to the unfairness of statistical parity. As these higher risk scores may not entirely reflect an increased risk among these cultural minorities (Hannah-Moffat, 2013), they may also be contributing to more individuals who do not go on to engage in recidivism being classified as high risk. In other words, a higher FPR and, therefore, disparities among error rates.

Differing Base Rates

As mentioned previously, differing base rates among groups are a direct cause of an impossibility theorem in which multiple forms of fairness are unable to be simultaneously satisfied. To demonstrate the issue of base rates, confusion matrices will be used to show how multiple forms of fairness can potentially be achieved when base rates of recidivism are the same. This, however, cannot be achieved when base rates differ. As shown in Table 1, a confusion matrix represents the predicted outcome against the actual observed outcome.

Table 1

Confusion Matrix of Predicted and Observed Outcomes with Fairness Calculations

	Recidivist	Non-Recidivist	
Predicted to be a recidivist	True Positive (TP)	False Positive (FP)	PPV = TP/(TP+FP)
Not predicted to be a recidivist	False Negative (FN)	True Negative (TN)	NPV = TN/(FN+TN)
	Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	
	FNR = FN/(TP+FN)	FPR = FP/(FP+TN)	

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate; TP true positive; FP false positive; FN false negative; TN true negative.

When an individual is predicted to be a recidivist and does engage in recidivism, this is a True Positive (TP), and when an individual is predicted to not be a recidivist and does not engage in recidivism, this is a True Negative (TN). When an individual is predicted to be a recidivist and does not engage in recidivism, this is a False Positive (FP), and conversely, when an individual is not predicted to be a recidivist but does engage in recidivism, this is a False Negative (FN). These four points of information can be used to calculate relevant indicators of fairness, such as the PPV and NPV for predictive parity, the FPR and FNR for error rate balance, the sensitivity for sensitivity fairness, and sensitivity and 1 – specificity (i.e., FPR) for plotting ROC curves.

As shown in Tables 2 and 3, if there were two groups of individuals (Group A, *N* = 1,000 and Group B, *N* = 1,500) that had the same base rate of recidivism (base rate = .50, or 50% were recidivists), numerous forms of fairness could be simultaneously satisfied among both groups.

Table 2

Group A with 1,000 Individuals and a Base Rate of .50

	Recidivist	Non-Recidivist	
Predicted to be a recidivist	300	100	PPV = .75
Not predicted to be a recidivist	200	400	NPV = .67
	Sensitivity = .60	Specificity = .80	
	FNR = .40	FPR = .20	

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate.

Table 3

Group B with 1,500 Individuals and a Base Rate of .50

	Recidivist	Non-Recidivist	
Predicted to be a recidivist	450	150	PPV = .75
Not predicted to be a recidivist	300	600	NPV = .67
	Sensitivity = .60	Specificity = .80	
	FNR = .40	FPR = .20	

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate.

With the same base rate, Groups A and B have satisfied predictive parity, error rate balance, and sensitivity fairness. There is also equivalence between sensitivity and 1 – specificity, the values used to plot the ROC curve and calculate the AUC. It is also worth noting that in this scenario, statistical parity is also satisfied among groups. However, if the base rate of recidivism for Group A increased (base rate =.70) and Group B remained the same (base rate =.50), predictive parity could not be achieved alongside both error rate balance and sensitivity fairness. As shown in Table 4, error rate balance and sensitivity fairness can be made to remain the same as the previous base rate of .50, although predictive parity values will inevitably differ as a higher proportion of individuals are now found to be recidivists (i.e., higher cell counts in the observed recidivist column).

Table 4

Group A with 1,000 Individuals and a Base Rate of .70

	Recidivist	Non-Recidivist	
Predicted to be a recidivist	420	60	PPV = .88
Not predicted to be a recidivist	280	240	NPV = .46
	Sensitivity = .60	Specificity = .80	
	FNR = .40	FPR = .20	

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate.

Therefore, there is always going to be a trade-off among types of fairness unless there is perfect prediction among all individuals. However, perfect prediction is also extremely improbable, whether due to assessments not measuring every indicator of risk or inherent measurement error such as random error or systematic error (i.e., bias) that prevent accurate prediction in assessments (Cohen et al., 2013; Hair et al., 2019). This again stipulates that when there are unequal base rates or imperfect predictions, even if one form of fairness is satisfied among cultural groups, another form of fairness will always be unsatisfied.

Proposals and Attempts to Identify and Increase Fairness

Regardless of the source of cross-cultural unfairness, suggestions have been made periodically to identify and increase fairness. The following section will explore common approaches to identifying and increasing fairness.

Development of New Instruments

There have been calls to develop new, culturally relevant risk assessment instruments (Dawson, 1999; Day et al., 2018; Hart, 2016; Shepherd, Adams, et al., 2014). Hart (2016) stipulated that the development of risk assessment instruments should incorporate the literature surrounding diverse cultural groups; ensuring relevant risk factors are included. Day et al. (2018) built upon this in their review in which a framework that reconceptualises relevant cultural theories of risk could be developed by involving and consulting with the specific Indigenous communities in which the offenders live. They highlighted concerns with actuarial assessments, specifically that they often fail to incorporate potentially relevant social, contextual, and cultural factors. They discussed placing more effort into the development of SPJ assessments that enable the inclusion of culturally relevant understandings of risk (Day et al., 2018).

A new cultural assessment could offer a solution by ensuring that relevant cultural factors and idiosyncrasies are incorporated. However, the feasibility of this proposed solution is poor regarding it being vague and having no immediate way to ascertain the instruments' predictive utility. However, a new assessment would encompass developing culturally specific predictors of risk that would require extensive data collection before the instrument developers could adequately ascertain the predictive validity. Additionally, the predictive validity of this new measure would need to outperform previous existing measures for that cultural group (Shepherd & Spivak, 2020). This approach also overlooks pre-existing theories (e.g., Andrews and Bonta's central eight) encompassing salient predictors of risk that have demonstrated applicability cross-culturally (Andrews & Bonta, 1994; Gutierrez et al., 2013; Quinsey et al., 2006). The concept of having different cultural assessments further creates complications around legal parity, whereby the law is to treat each individual equally. Different assessments administered to certain individuals for the same legal purposes may be introducing a different type of unfairness. Further, heterogeneity within a culture is often greater than between cultures (Shepherd & Lewis-Fernandez, 2016). The development of a culturally specific instrument would, therefore, need to be made applicable to the broader cultural minority group to attempt to account for the variety of beliefs, values, and traditions expressed (Shepherd, 2015), again leading to feasibility issues in development.

The development of culturally specific instruments in the literature has been scarce. In Australia, Allan and Dawson (2002) developed a 3-Predictor model (poor coping skills, unfeasible release plans, and unrealistic long-term goals) to predict sexual recidivism among Indigenous individuals in Western Australia. Allan et al. (2006) stated that they were able to demonstrate the utility of the 3-Predictor model with it outperforming several other assessments. They reported an AUC value of .84 for the 3-Predictor model, the highest in the overall study (Allan et al., 2006). However, the AUC value was for a mixed group of Indigenous and non-Indigenous individuals, not allowing for a comparison between cultures. Boer et al. (2004) developed the Risk Management Guide for Aboriginal Offenders (RMGAO) with Aboriginal Elders from Canada. The RMGAO comprises a series of culturally relevant questions to consider when assessing and managing Indigenous individuals in custody and in the community. The authors of this instrument state that if an actuarial or SPJ risk assessment has demonstrated utility with Aboriginal individuals, it can be used alongside the RMGAO to assist with the safety of the individual and to ensure they become productive community members through reintegration (Boer et al., 2004). However, beyond face validity, no other testing has been conducted to ensure the predictive utility of the items and assessment overall.

Alteration of Existing Instruments

Other solutions have been proposed in terms of altering or expanding existing instruments to account for cultural differences in items and risk factors (Perley-Robertson et al., 2019; Shepherd, 2016a; Shepherd, 2018; Shepherd & Anthony, 2018; Shepherd & Lewis-Fernandez, 2016; Shepherd & Willis-Esqueda, 2018). One proposed idea is to alter the item content on risk assessment instruments to make them more culturally relevant by amending or translating the wording or removing jargon (Shepherd, 2018; Shepherd & Lewis-Fernandez, 2016; Shepherd & Willis-Esqueda, 2018). Another proposal is to include culturally relevant items and/or remove items that are less predictive when assessing a cultural minority (Ellerby & MacPherson, 2002; Heckbert & Turkington, 2001; LaPrairie, 1995; Mann, 2009; Martel et al., 2011; Perley-Robertson et al., 2019; Shepherd & Willis-Esqueda, 2018; Wilson & Gutierrez, 2014). Last, a stronger consideration of an individual's strengths rather than a disproportionate focus on risk has been suggested for particular cultural groups (Shepherd & Willis-Esqueda, 2018). Shepherd (2016a) and Shepherd and Willis-Esqueda (2018) have also considered the addition of a cultural addendum, which may comprise additional contextual information for each risk item and advice for working effectively inter-culturally in clinical settings.

Similar to creating a new assessment, modifying existing assessments could ensure that culturally relevant content is incorporated into risk assessment instruments, potentially increasing fairness among risk classifications (i.e., statistical parity) and even among the classifications and errors in observation of these instruments (i.e., sensitivity fairness, error rate balance, predictive parity, calibration). However, if certain elements are altered, added, or removed from risk assessment instruments, the predictive validity of the instrument would need to be retested to ensure these changes have not had a detrimental influence (Shepherd & Spivak, 2020). The eradication or modification of items and/or factors that have been periodically shown to predict recidivist behaviours cross-culturally may also impede accuracy (Skeem & Lowenkamp, 2016). Further, as highlighted above, the issue of differing base rates may inhibit this approach from successfully increasing fairness. Varying base rates will ultimately lead to at least one or more forms of fairness being unsatisfied even with the addition or alteration of items on a risk assessment instrument.

Clinician Training

It has been suggested that clinicians should be culturally competent when working with diverse cultures (Shepherd, 2018; Shepherd & Lewis-Fernandez, 2016; Shepherd et al., 2015; Shepherd & Strand, 2016). This includes avoiding or cautiously employing instruments that have lower levels of predictive validity for minority groups (Helmus et al., 2012; Shepherd, 2016b). Professionals have been implored to undertake ongoing and regular education to ensure they can identify potential cultural issues that may result in an unfair assessment (Hart, 2016; Olver et al., 2014; Shepherd & Lewis-Fernandez, 2016). While cultural competence is an important and ongoing educational process, clinician bias will always be present when human error or subjectivity is a factor. Further, there is no evidence to suggest that clinician cultural competence training would increase fairness, with a scarcity of literature examining the impact of clinician training on risk assessment scores (Venner et al., 2021). This suggestion also assumes differences arise due to insensitivity instead of factors such as the higher prevalence of risk factors in certain cultural minorities or differing base rates that cause the inherent trade-offs in risk assessment instruments. One study that specifically reported on the impact of training on risk assessment scores cross-culturally found that although training did lead to some risk factors receiving an increased (or decreased) score post training, there were no differences in risk scores across cultures (Jimenez et al., 2018).

Alternative Statistical Approaches

Numerous supplementary statistical approaches have been proposed in the risk instrument literature to explore and/or resolve unfairness.

Differential Item Functioning, Factorial Structures and Latent Constructs

Differential item functioning is often observed as a way to assess for potential item bias (He & van de Vijver, 2012; van de Vijver & Tanzer, 2004) and has been proposed in the cross-cultural forensic risk assessment literature (Hart, 2016). Specifically, Hart (2016) suggested observing the relationship between the items and the latent trait variable through item response theory to see if it differed cross-culturally. However, this approach is scarcely applied in the cross-cultural risk assessment literature (Schmidt et al., 2020). One example of where it was utilised was when cultural differences in item bias on the PCL: YV were observed between Caucasian, African American, and Hispanic male youth (Tsang et al., 2014). This study found that 15 of the 20 items in the PCL: YV functioned differently across cultural groups, which could contribute to a total PCL: YV score that was approximately 12 points different. Similar to observing differential item functioning on latent traits, observing differences among the factorial structure and latent constructs of a risk assessment instrument across cultures has also been suggested (Hart, 2016; Shepherd & Lewis-Fernandez, 2016).

However, these methods are primarily applicable for assessments that are scored in an additive fashion (Putnick & Bornstein, 2016), making them somewhat impractical for assessments not scored this way (e.g., SPJ instruments). Schmidt et al. (2020) also highlighted the inconsistencies identified among previously conducted factor analyses in risk assessment research, with different factor solutions being reported among the LS measures based on subscale scores (Gordon et al., 2015; Hollin et al., 2003; Loza & Simourd, 1994; Palmer & Hollin, 2007) and item responses (Gordon et al., 2015; Hsu et al., 2011). Some studies have previously demonstrated similar risk structures across cultures on instruments including the VRS-SO (Olver et al., 2020) and the PCL: YV (McCuish et al., 2018). However, even with these similar factorial structures, other studies have still identified predictive parity and statistical parity unfairness on the PCL: YV (Schmidt et al., 2006; Shepherd & Strand, 2016).

Disregarding these limitations, establishing equivalence (or non-equivalence) among item functioning, factorial structures, and latent constructs does little to ultimately increase cross-cultural fairness beyond identifying potential item bias, for which there is no obvious link between item bias and recidivism. None of these approaches take into consideration the outcome (i.e., did the individual go on to engage in recidivism or not?). Therefore, when applying these approaches to the fairness definitions outlined in this paper, the identification of item bias is only directly useful in addressing issues of statistical parity, as statistical parity also does not consider the outcome of recidivism. If these methods are employed to increase statistical parity among cultures, as highlighted previously, achieving statistical parity can have detrimental impacts on other forms of fairness and overall accuracy of the risk assessment instrument.

Alternate Scoring

Other research has suggested alternate ways to score assessments to reduce the statistical parity disparities between cross-cultural groups. Skeem and Lowenkamp (2016) suggested relying less heavily on factors that differed between groups. Specifically, they suggested to focus less on the criminal history of the individuals and instead put a higher weighting on factors that had fewer mean score differences (Skeem & Lowenkamp, 2016). It has also been suggested to choose cut-off scores for different groups (e.g., cultures) that distinguish low from high risk in a way that will help minimise the disparity among risk classifications (Thompson & McGrath, 2012). Although these proposed solutions may increase statistical parity, as discussed previously, this can lead to a variety of detrimental outcomes. For example, lowering the weight of a factor such as criminal history, which has been found to be one of the leading predictors of recidivism (Eisenberg et al., 2019; Wilson & Gutierrez, 2014), will likely result in an overall significant loss in predictive validity and an increase in misclassifications.

Culture as an Indicator

Using culture as an indicator has been suggested as a way to evaluate the presence of unfairness (Flores et al., 2016; Larson et al., 2016; Perrault et al., 2017). This can involve using culture to see if it predicts risk assessment scores on an instrument, in other words, to demonstrate if an individual's culture predicts a higher risk score. Work by Larson et al. (2016) demonstrated that identifying as Black was significantly predictive of higher risk scores for both general and violent recidivism on COMPAS (*p*s < .01). Even though Black individuals had higher recidivism rates, when adjusting for this difference alongside age and gender, they were still 45% more likely to get a higher risk score for general recidivism and 77.3% more likely for violent recidivism.

Alternatively, culture can be used to see if there is a significant interaction between this variable and a risk assessment instrument score in predicting recidivism, demonstrating that risk assessment scores have a different association with recidivism across cultures. Flores et al. (2016) did not find support for an interaction between culture and COMPAS as a predictor of general or violent recidivism. This was mirrored by Perrault et al. (2017) for Black and White youth, in which an interaction between culture and risk assessment scores (on the YLS/CMI and SAVRY) was not statistically significant. This demonstrates that these assessments were not found to predict recidivism differently as a result of culture. However, Jimenez et al. (2018) did report a significant interaction between risk levels on the LS/CMI and culture when comparing a broader minority group (i.e., Black African Americans, Asian Americans, Native Americans, and those of Hispanic descent) to a non-minority group encompassing White European Americans of non-Hispanic descent. However, similar to factorial structures and latent constructs, this approach does little to resolve unfairness beyond being a method to identify its existence and the impact culture has on scores and prediction.

Using culture as a predictor of recidivism has also been discussed as a way to increase accuracy (Berk, 2009; Berk et al., 2018). Although a contentious topic as it can be seen as a moral issue (Berk, 2009), it has been argued that its incorporation may increase predictive validity (Berk, 2009; Berk et al., 2018; Douglas et al., 2017). However, this is another example of a trade-off between fairness and accuracy (Berk, 2019). As cultural minorities are often found to have higher base rates of recidivism (Bonta et al., 1997; Flores et al., 2016; Gutierrez et al., 2013; Jones et al., 2016; Olver, 2016; Shepherd & Strand, 2016; Thompson & McGrath, 2012; Wilson & Gutierrez, 2014; Wormith et al., 2015), the inclusion of culture as a variable may lead to an increased chance of that cultural minatory being classified as high risk, compromising notions of fairness such as statistical parity for an increase in predictive validity. Although recent research has highlighted that algorithms with access to protected variables such as culture can aid in the detection of discrimination as well as increase transparency, predictive validity, and fairness (Kleinberg et al., 2018; Skeem & Lowenkamp, 2020). A recent study trailing numerous algorithms to increase fairness found that providing an algorithm with an individual's culture led to an increase in fairness among error rate balance and maximised the instruments' calibration (Skeem & Lowenkamp, 2020).

Statistical Learning Methods

Methods of statistical learning as an approach to increasing various forms of fairness have been recently discussed in the literature (e.g., Berk et al., 2018; Chouldechova & Roth, 2018). These statistical learning methods aim to increase predictive accuracy (Breiman, 2001b; Spivak & Shepherd, 2020). These approaches can account for large numbers of variables and can establish the relevant predictors and interactions that increase predictive accuracy (Berk & Hyatt, 2015; Breiman, 2001b; Brennan, 2016; Monahan et al., 2005; Spivak & Shepherd, 2020). On the contrary, traditional methods such as linear regression often aim to increase interpretability in which the relationship between predictors and the outcome is transparent

(Breiman, 2001b). These traditional methods also often rely on pre-determining the relevant predictors and interactions to be included in the model (Breiman, 2001b; Duwe & Kim, 2015). Within the forensic risk assessment literature, statistical learning methods have primarily been used with the chief purpose of increasing predictive validity (or discrimination as assessed by the AUC), with research demonstrating its ability to outperform more common methods such as logistic regression in some studies (e.g., Berk & Bleich, 2013; Duwe & Kim, 2015; Ting et al., 2018). Multiple statistical learning methods such as random forests (Breiman, 2001a), stochastic gradient boosting (Friedman, 2002), and support vector machines (Vapnik, 1998; Vapnik, 1999) have demonstrated their ability to be effective classifiers and valuable in criminal justice forecasting (Berk & Hyatt, 2015; Duwe, 2019; Duwe & Kim, 2015; Pflueger et al., 2015; Zaidi et al., 2020; Zeng et al., 2017).

A variety of disciplines, including data science, statistics, and occasionally criminology, have recently expanded the use of these methods and begun exploring them as an approach to increasing fairness (Berk et al., 2018; Corbett-Davies et al., 2017; Wadsworth et al., 2018). This can be achieved by transformations at differing levels of the algorithm construction process (Berk et al., 2018; Hajian & Domingo-Ferrer, 2013). Pre-processing is where the original data is altered in an attempt to remove potential sources contributing to unfairness (e.g., unequal base rates, higher prevalence of predictors for certain racial groups; Calmon et al., 2017; Feldman et al., 2015; Kamiran & Calders, 2012; Zemel et al., 2013). This can involve methods such as using residuals in place of variables found to be predicted by an individual's culture (Berk, 2009; Skeem & Lowenkamp, 2020). Base rates can also be equalled across groups by applying different weights to groups depending on how much they engage in recidivism (Berk et al., 2018) or by altering the recidivism outcome of groups so that base rates are equivalent (Hajian & Domingo-Ferrer, 2013). In-processing is when the algorithms themselves are altered such that the model does not contain any unfair decision rules (Celis et al., 2019; Kamishima et al., 2012; Zhang et al., 2018). This could encompass developing a separate algorithm for each culture (Berk, 2019) or using an adversarial approach that focuses on maximising accuracy whilst simultaneously reducing any indication of culture being found in the prediction (Zhang et al., 2018). Lastly, post-processing is where the resulting data is modified to remove unfairness (Hardt et al., 2016; Kamiran et al., 2012; Pleiss et al., 2017). For example, this could involve the random reassignment of outcome labels (Hardt et al., 2016), where those predicted to engage in recidivism for each culture have their predicted recidivism outcome altered to achieve fairness. Emerging research has shown that statistical learning methods have been able to minimise the fairness disparity among cultures whilst still maintaining, or even improving, overall predictive validity (e.g., Wadsworth et al., 2018).

Interpretability of Statistical Learning Methods. However, as in the previous example of pre-processing, using residuals instead of actual values can result in a model that is somewhat incomprehensible. This has led to scepticism in using statistical learning methods as an approach (Brennan & Oliver, 2013; Kehl et al., 2017; Wisser, 2019; Zeng et al., 2017). Unlike methods such as linear regression, in which the variables' influence on the predicted outcome is explicitly understood through beta weights, certain algorithms (and transformations on the algorithm) can make the relationship between the variables used and the predicted outcome not easily discernible (Breiman, 2001b; Brennan & Oliver, 2013; Duwe & Kim, 2015). However, these approaches are not entirely uninterpretable, with techniques enabling the importance of predictors to be established. Variable importance can be ascertained easily for certain algorithms such as random forests (Berk, 2008; Berk & Bleich, 2013; Breiman, 2001a). Further, approaches such as Shapley values (Shapley, 1953) and local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016) are valuable for numerous algorithms in determining the contribution of variables to the prediction (Lundberg & Lee, 2017).

Conclusion

The current review provided an overview of cross-cultural fairness in forensic risk assessment. It i) articulated and explored numerous definitions of fairness, ii) reviewed the cross-cultural risk assessment literature and critically investigated if different forms of fairness have been satisfied between cultural groups, iii) reviewed explanations for cross-cultural unfairness, and iv) evaluated the proposed solutions to identify and increase fairness. This review highlighted several limitations within the existing literature, including the pursuit of achieving cross-cultural fairness in forensic risk assessment instruments with fairness notions often incompatible with each other and optimising accuracy. First, beyond sensitivity fairness (in the form of AUC) and statistical parity, there has been limited research exploring other notions of fairness. This highlights the need for future research to focus on addressing other varying definitions of fairness to gain an understanding of the fairness or disparities present cross-culturally on risk assessment instruments. This review also established that most forms of fairness are not satisfied cross-culturally. Whether differences ranged from small and inconsequential to substantial, the repeated demonstration of these disparities among differing notions of fairness necessitates continued and ongoing research and efforts to be placed in increasing fairness. The impact of these instruments' not being cross-culturally fair could lead to increased misclassifications for particular cultural groups. Such misclassifications could potentially disadvantage populations already over-represented in the criminal justice system, with treatment and supervision approaches based on misclassifications potentially rendered ineffective.

Second, different forms of fairness themselves are often incompatible with each other and with accuracy, demonstrating a requirement for policymakers to assess which trade-off they find to be the most acceptable and relevant. These decisions are beyond the scope of this review, requiring extensive legal and ethical considerations that policymakers would need to deliberate. Further, these decisions are likely localised decisions that could differ, for example, across countries, jurisdictions, risk assessment instruments, or recidivism types. Policymakers may also have different notions of which form of fairness is most pivotal to satisfy. They may wish to focus on maximizing predictive parity (or calibration) with a focus on public safety, or alternatively, they may wish to place the focus on maximizing error rate balance with the aim of ensuring equal personal liberty across cultures (i.e., racial justice; Skeem & Lowenkamp, 2020). Further, it should be recognized by researchers, policymakers, and users moving forward that if particular forms of fairness are satisfied (e.g., AUC values or error rate balance), yet the base rates of recidivism differ across cultural groups, other forms of fairness are not and cannot be satisfied (e.g., predictive parity). Continued research is therefore encouraged into the best possible trade-offs that can be achieved among varying types of fairness, as well as between fairness and the accuracy of risk assessment instruments.

Finally, varying solutions have been proposed to increase fairness; however, none is without limitations. The creation or alteration of assessments encompasses significant feasibility issues and would need to demonstrate that accuracy had not been considerably impacted. Altering assessments and a focus on clinician training ignores the issue of unequal base rates that inevitably leads to unfairness of some kind. Statistical methods proposed, such as employing culture as an indicator, differential item functioning, and exploring factorial and latent constructs, primarily focus on identifying the nature of the unfairness and do little to resolve it. Although altering test norms may lead to an increase in statistical parity, it will likely result in other fairness definitions and accuracy being compromised. Statistical learning methods, whilst able to overcome issues such as base rates without impeding too heavily on predictive validity, can result in algorithms that require further computations in order to establish how a prediction is being made. Regardless, the continued testing and exploration of these approaches should be encouraged to further the literature. Specifically, the use of novel statistical learning methods should be investigated as they provide a more direct, feasible, statistically rigorous, and time-sensitive solution, with initial research demonstrating promise in increasing cross-cultural fairness on forensic risk assessment instruments (see Wadsworth et al., 2018). Although achieving total fairness is impossible, we should focus our attention on which types of fairness are the most critical to satisfy, how to best achieve this, and the publicly acceptable trade-offs possible among varying types of fairness and accuracy to inform and aid policymakers in decision making moving forward.

References

Allan, A., & Dawson, D. (2002). *Developing a unique risk of violence tool for Australian Indigenous offenders.* Retrieved from Canberra, Australia: http://crg.aic.gov.au/reports/200001-06.pdf

Allan, A., Dawson, D., & Allan, M. M. (2006). Prediction of the risk of male sexual reoffending in Australia. *Australian Psychologist, 41*(1), 60-68. doi:10.1080/00050060500391886

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ, 5*. doi:10.7717/peerj.3544

Andrews, D. A., & Bonta, J. (1994). *The psychology of criminal conduct*. Cincinnati, OH: Anderson.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Australian Bureau of Statistics. (2018). 4517.0 - Prisoners in Australia, 2018. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4517.0~2018~Main%20Features~Aboriginal%20and%20Torres%20Strait%20Islander%20prisoner%20characteristics%20~13

Berk, R. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and Social Problems, 1*(4), 231-242. doi: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. doi: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. doi:10.1111/1745-9133.12047

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1-42. doi:10.1177/0049124118782533

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter, 27*(4), 222-228. doi:10.1525/fsr.2015.27.4.222

Biau, J. D., Jolles, M. B., & Porcher, M. R. (2010). P value and the theory of hypothesis testing: An explanation for new researchers. *Clinical Orthopaedics and Related Research, 468*(3), 885-892. doi:10.1007/s11999-009-1164-4

Boer, D. P., Couture, J., Geddes, C., & Ritchie, A. (2004). *Yókw'tól: Risk Management Guide for Aboriginal Offenders*. British Columbia: Correctional Services of Canada.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply.* (pp. 18-32). Thousand Oaks, CA: Sage Publications, Inc.

Bonta, J., Laprairie, C., & Wallace-Capretta, S. (1997). Risk prediction and re-offending: Aboriginal and non-aboriginal offenders. *Canadian Journal of Criminology, 39*(2), 127-144.

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

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

Brennan, T. (2016). *An alternative scientific paradigm for criminological risk assessment: Closed or open systems, or both?* New York: Taylor & Francis Ltd.

Brennan, T., & Oliver, W. L. (2013). The emergence of machine learning techniques in criminology. *Criminology & Public Policy, 12*(3), 551-562. doi:10.1111/1745-9133.12055

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized pre-processing for discrimination prevention*. Paper presented at the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Celis, L., Huang, L., Keswani, V., & Vishnoi, N. (2019). *Classification with fairness constraints: A meta-algorithm with provable guarantees*. Paper presented at the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA.

Chenane, J. L., Brennan, P. K., Steiner, B., & Ellison, J. M. (2015). Racial and ethnic differences in the predictive validity of the Level of Service Inventory–Revised among prison inmates. *Criminal Justice and Behavior, 42*(3), 286-303. doi:10.1177/0093854814548195

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153-163.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv:1810.08810 [cs:LG].*

Cohen, H. W. (2011). P values: Use and misuse in medical literature. *American Journal of Hypertension, 24*(1), 18-23. doi:10.1038/ajh.2010.205

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). New York: McGraw Hill.

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

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023 [cs:CY]*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). *Algorithmic decision making and the cost of fairness*. Paper presented at the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada. https://doi.org/10.1145/3097983.3098095

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgement. *Science, 243*(4899), 1668-1674. doi:10.1126/science.2648573

Dawson, D. (1999). *Risk of violence assessment: Aboriginal offenders and the assumption of homogeneity*. Paper presented at the Best Practice Interventions in Corrections for Indigenous People Conference, Adelaide, Australia.

Day, A. (2003). Reducing the risk of re-offending in Australian Indigenous offenders: What works for whom? *Journal of Offender Rehabilitation, 37*(2), 1-15. doi:10.1300/J076v37n02_01

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. doi:10.1080/13218719.2018.1467804

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe.

Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence risk assessment: Science and practice. *Legal and Criminological Psychology, 4*(2), 149-184. doi:10.1348/135532599167824

Douglas, T., Pugh, J., Singh, I., Savulescu, J., & Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. *European Psychiatry, 42*, 134-137. doi:https://doi.org/10.1016/j.eurpsy.2016.12.009

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing, 9*(6), 649-657. doi:doi:10.1046/j.1365-2850.2002.00535.x

Dragomir, R. R., & Tadros, E. (2020). Exploring the impacts of racial disparity within the American juvenile justice system. *Juvenile and Family Court Journal, 71*(2), 61-73.

Duwe, G. (2019). Better practices in the development and validation of recidivism risk assessments: The Minnesota Sex Offender Screening Tool–4. *Criminal Justice Policy Review, 30*(4), 538-564. doi:10.1177/0887403417718608

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. doi:10.1177/0887403415604899

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). *Fairness through awareness*. Paper presented at the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, Massachusetts.

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2018). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior, 46*(2), 185-209. doi:10.1177/0093854818811379

Edens, J., Campbell, J., & Weir, J. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior, 31*(1), 53-75. doi:10.1007/s10979-006-9019-y

Eisenberg, M. J., van Horn, J. E., Dekker, J. M., Assink, M., van der Put, C. E., Hendriks, J., & Stams, G. J. J. M. (2019). Static and dynamic predictors of general and violent criminal offense recidivism in the forensic outpatient population: A meta-analysis. *Criminal Justice and Behavior, 46*(5), 732-750. doi:10.1177/0093854819826109

Ellerby, L., & MacPherson, P. (2002). *Exploring the profiles of Aboriginal sexual offenders: Contrasting Aboriginal and non-Aboriginal sexual offenders to determine unique client characteristics and potential implications for sex offender assessment and treatment strategies (research report no. R-122)*. Ottawa, ON: Correctional Service of Canada.

Ewert v. Canada, FC 1093 (2015).

Fazel, S. (2019). The scientific validity of current approaches to violence and criminal risk assessment. In J.W. de Keijser, J.V. Roberts, & J. Ryberg (Eds.), *Predictive sentencing: Normative and empirical perspectives* (1st ed.). Oxford: Hart Publishing.

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 (Clinical Research Ed.), 345*(7868). doi:10.1136/bmj.e4692

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and removing disparate impact*. Paper presented at the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.

Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation, 80*(2), 38-46,66.

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

Frize, M., Kenny, D., & Lennings, C. (2008). The relationship between intellectual disability, Indigenous status and risk of reoffending in juvenile offenders on community orders. *Journal of Intellectual Disability Research, 52*(6), 510-519. doi:10.1111/j.1365-2788.2008.01058.x

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439-453. doi:10.1037/a0015251

Gordon, H., Kelty, S. F., & Julian, R. (2015). Psychometric evaluation of the Level of Service/Case Management Inventory among Australian offenders completing community-based sentences. *Criminal Justice and Behavior, 42*(11), 1089-1109.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323. doi:10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. doi:10.1037/1040-3590.12.1.19

Gutierrez, L., Helmus, L. M., & Hanson, R. K. (2016). What we know and don't know about risk assessment with offenders of Indigenous heritage. *Journal of Threat Assessment and Management, 3*(2), 97-106. doi:10.1037/tam0000064

Gutierrez, L., Wilson, H. A., Rugge, T., & Bonta, J. (2013). The prediction of recidivism with Aboriginal offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology & Criminal Justice, 55*(1), 55-99. doi:10.3138/cjccj.2011.E.51

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Hampshire: Cengage Learning.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1445-1459. doi:10.1109/TKDE.2012.72

Hannah-Moffat, K. (2013). Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly, 30*(2), 270-296. doi:10.1080/07418825.2012.682603

Hannah-Moffat, K., & Maurutto, P. (2010). Re-contextualizing pre-sentence reports: Risk and race. *Punishment & Society, 12*(3), 262-286. doi:10.1177/1462474510369442

Hanson, R. K. (2017). Assessing the calibration of actuarial risk scales: A primer on the E/O index. *Criminal Justice and Behavior, 44*(1), 26-39. doi:10.1177/0093854816683956

Hanson, R. K., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management, 1*(2), 102-117. doi:10.1037/tam0000014

Harcourt, B. E. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age*. Chicago: University of Chicago Press.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. Paper presented at the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain.

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology, 3*(1), 121-137. doi:10.1111/j.2044-8333.1998.tb00354.x

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. doi: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.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2). doi:https://doi.org/10.9707/2307-0919.1111

Hébert-Johnson, U., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). *Calibration for the (computationally-identifiable) masses*. Paper presented at the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, Stockholm, Sweden.

Heckbert, D., & Turkington, D. (2001). *Turning points: A study of the factors related to the successful reintegration of Aboriginal offenders*. Ottawa, Ontario: Correctional Service of Canada.

Helmus, L., Babchishin, K., & Blais, J. (2012). Predictive accuracy of dynamic risk factors for Aboriginal and non-Aboriginal sex offenders: An exploratory comparison using STABLE-2007. *International Journal of Offender Therapy and Comparative Criminology, 56*(6), 856. doi:10.1177/0306624X11414693

Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior, 44*(1), 8-25. doi:10.1177/0093854816678898

Hollin, C. R., Palmer, E. J., & Clark, D. (2003). The Level of Service Inventory-Revised profile of English prisoners: A needs analysis. *Criminal Justice and Behavior, 30*(4), 422-440. doi:https://doi.org/10.1177/0093854803253134

Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2003). Ethnicity, gender, and the Level of Service Inventory-Revised. *Journal of Criminal Justice, 31*(4), 309-320. doi:10.1016/S0047-2352(03)00025-4

Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2006). Exploring the validity of the Level of Service Inventory-Revised with Native American offenders. *Journal of Criminal Justice, 34*(3), 331-337. doi:10.1016/j.jcrimjus.2006.03.009

Homel, R., Lincoln, R., & Herd, B. (1999). Risk and resilience: Crime and violence prevention in Aboriginal communities. *Australian and New Zealand Journal of Criminology, 32*(2), 182-196. doi:10.1177/000486589903200207

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2010). Level of Service Inventory–Revised: Assessing the risk and need characteristics of Australian Indigenous offenders. *Psychiatry, Psychology and Law, 17*(3), 355-367. doi:10.1080/13218710903089261

Hsu, C., Caputi, P., & Byrne, M. K. (2011). The Level of Service Inventory-Revised (LSI-R) and Australian offenders: Factor structure, sensitivity, and specificity. *Criminal Justice and Behavior, 38*(6), 600-618.

Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal, 68*(6), 1043.

Hurducas, C. C., Singh, J. P., De Ruiter, C., & Petrila, J. (2014). Violence risk assessment tools: A systematic review of surveys. *International Journal of Forensic Mental Health, 13*(3), 181-192. doi:10.1080/14999013.2014.942923

Jimenez, A. C., Delgado, R. H., Vardsveen, T. C., & Wiener, R. L. (2018). Validation and application of the LS/CMI in Nebraska probation. *Criminal Justice and Behavior, 45*(6), 863-884. doi:10.1177/0093854818763231

Jones, N. J., Brown, S. L., Robinson, D., & Frey, D. (2016). Validity of the youth assessment and screening instrument: A juvenile justice tool incorporating risks, needs, and strengths. *Law and Human Behavior, 40*(2), 182-194. doi:10.1037/lhb0000170

Jones, R., & Day, A. (2011). Mental health, criminal justice and culture: Some ways forward? *Australasian Psychiatry, 19*(4), 325-330. doi:10.3109/10398562.2011.579613

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1-33. doi:10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012, 10-13 Dec. 2012). *Decision theory for discrimination-aware classification.* Paper presented at the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). *Fairness-aware classifier with prejudice remover regularizer.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg.

Kehl, D., Guo, P., & Kessler, S. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing*. Responsive Communities Initiative, Berkman Klein Center for Internet & Society: Harvard Law School.

Kenny, D. T., & Nelson, P. K. (2008). *Young offenders on community orders: Health, welfare and criminogenic needs*. Sydney, Australia: Sydney University Press.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis, 10*, 113. doi:10.1093/jla/laz001

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807 [cs.LG].*

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B, 75*(1), 45-47. doi:10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Official Journal of the Association for the Treatment of Sexual Abusers (ATSA), 16*(2), 107-120. doi:10.1023/B:SEBU.0000023060.61402.07

LaPrairie, C. (1995). Seen but not heard: Native people in four Canadian inner cities. *The journal of Human Justice, 6*(2), 30-45. doi:10.1007/BF02585441

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lee, S. C., Hanson, R. K., & Blais, J. (2020). Predictive accuracy of the Static-99R and Static-2002R risk tools for identifying Indigenous and white individuals at high risk for sexual recidivism in Canada. *Canadian Psychology/Psychologie canadienne, 61*(1), 42-57. doi:10.1037/cap0000182

Lee, S. C., Hanson, R. K., Calkins, C., & Jeglic, E. (2019). Paraphilia and antisociality: Motivations for sexual offending may differ for American whites and blacks. *Sexual Abuse, 32*(3), 335-365. doi:10.1177/1079063219828779

Leistico, A.-M., Salekin, R., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*(1), 28-45. doi:10.1007/s10979-007-9096-6

Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2018). The importance of calibration in clinical psychology. *Assessment*. doi:10.1177/1073191117752055

Loza, W., & Simourd, D. J. (1994). Psychometric evaluation of the Level of Service Inventory (LSI) among male Canadian federal offenders. *Criminal Justice and Behavior, 21*(4), 468-480.

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Paper presented at the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Mann, M. (2009). *Good intentions, disappointing results: A progress report on federal Aboriginal corrections*. Ottawa, Ontario: Office of the Correctional Investigator.

Martel, J., Brassard, R., & Jaccoud, M. (2011). When two worlds collide: Aboriginal risk management in Canadian corrections. *British Journal of Criminology, 51*(2), 235-255. doi:10.1093/bjc/azr003

McCuish, E. C., Mathesius, J. R., Lussier, P., & Corrado, R. R. (2018). The cross-cultural generalizability of the Psychopathy Checklist: Youth Version for adjudicated

Indigenous youth. *Psychological Assessment, 30*(2), 192-203. doi:10.1037/pas0000468

McGrath, A. J., Thompson, A. P., & Goodman-Delahunty, J. (2018). Differentiating predictive validity and practical utility for the Australian adaptation of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior, 45*(6), 820-839. doi:10.1177/0093854818762468

Molnar, T., Allard, T., McKillop, N., & Rynne, J. (2020). Reliability and predictive validity of the Juvenile Sex Offender Assessment Protocol-II in an Australian context. *International Journal of Offender Therapy and Comparative Criminology*. doi:10.1177/0306624x19900978

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques*: Sage Publications.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*(1), 489-513. doi:10.1146/annurev-clinpsy-021815-092945

Monahan, J., Steadman, H., Appelbaum, P., Banks, S., Grisso, T., Heilbrun, K., . . . Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services, 56*(7), 810-815. doi:10.1176/appi.ps.56.7.810

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and caucasian female and male adolescents on probation. *Psychological Assessment*. doi:10.1037/pas0000816

Olver, M. E. (2016). Some considerations on the use of actuarial and related forensic measures with diverse correctional populations. *Journal of Threat Assessment and Management, 3*(2), 107-121. doi:10.1037/tam0000065

Olver, M. E., Kingston, D. A., & Sowden, J. N. (2020). An examination of latent constructs of dynamic sexual violence risk and need as a function of Indigenous and Nonindigenous ancestry. *Psychological Services*. doi:10.1037/ser0000414

Olver, M. E., Neumann, C. S., Wong, S., & Hare, R. D. (2013). The structural and predictive properties of the Psychopathy Checklist-Revised in Canadian Aboriginal and non-Aboriginal offenders. *Psychological Assessment, 25*(1), 167-179. doi:10.1037/a0029840

Olver, M. E., Sowden, J. N., Kingston, D. A., Nicholaichuk, T. P., Gordon, A., Beggs Christofferson, S. M., & Wong, S. C. P. (2018). Predictive accuracy of Violence Risk Scale–Sexual Offender Version risk and change scores in treated Canadian Aboriginal and non-Aboriginal sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 30*(3), 254-275. doi:10.1177/1079063216649594

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders. *Criminal Justice and Behavior, 36*(4), 329-353. doi:10.1177/0093854809331457

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. doi:10.1037/a0035080

Palmer, E. J., & Hollin, C. R. (2007). The Level of Service Inventory-Revised with English women prisoners: A needs and reconviction analysis. *Criminal Justice and Behavior, 34*(8), 971-984.

Perley-Robertson, B., Helmus, L. M., & Forth, A. (2018). Predictive accuracy of static risk factors for Canadian Indigenous offenders compared to non-Indigenous offenders:

Implications for risk assessment scales. *Psychology, Crime & Law*. doi:10.1080/1068316X.2018.1519827

Perrault, R. T., Vincent, G. M., & Guy, L. S. (2017). Are risk assessments racially biased?: Field study of the SAVRY and YLS/CMI in probation. *Psychological Assessment, 29*(6), 664-678. doi:10.1037/pas0000445

Pflueger, M. O., Franke, I., Graf, M., & Hachtel, H. (2015). Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC Psychiatry, 15*. doi:doi: 10.1186/s12888-015-0447-4

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. (2017). *On fairness and calibration*. Paper presented at the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. doi:10.1016/j.dr.2016.06.004

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC, US: American Psychological Association.

Rettenberger, M., Boer, D. P., & Eher, R. (2011). The predictive accuracy of risk factors in the Sexual Violence Risk–20 (SVR-20). *Criminal Justice and Behavior, 38*(10), 1009-1027. doi:10.1177/0093854811416908

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. 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. doi:10.1007/s10979-005-6832-7

Royal Statistical Society. (2018). *Algorithms in the justice system: Some statistical issues*. Retrieved from https://www.rss.org.uk/Images/PDF/influencing-change/2018/RSS_submission_Algorithms_in_the_justice_system_Nov_2018.pdf

Schaefer, B. P., & Hughes, T. (2019). Examining judicial petrial release decisions: The influence of risk assessments and race. *Criminology, Criminal Justice, Law & Society, 20*(2), 47-58.

Schmidt, F., McKinnon, L., Chattha, H., & Brownlee, K. (2006). Concurrent and predictive validity of the Psychopathy Checklist: Youth Version across gender and ethnicity. *Psychological Assessment, 18*(4), 393-401. doi:10.1037/1040-3590.18.4.393

Schmidt, S., Heffernan, R., & Ward, T. (2020). Why we cannot explain cross-cultural differences in risk assessment. *Aggression and Violent Behavior, 50*. doi:https://doi.org/10.1016/j.avb.2019.101346

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. (2016a). Criminal engagement and Australian culturally and linguistically diverse populations: Challenges and implications for forensic risk assessment. *Psychiatry, Psychology and Law, 23*(2), 256-274. doi:10.1080/13218719.2015.1053164

Shepherd, S. (2016b). Violence risk instruments may be culturally unsafe for use with Indigenous patients. *Australasian Psychiatry, 24*(6), 565-567. doi:10.1177/1039856216665287

Shepherd, S. (2018). Violence risk assessment and Indigenous Australians: A primer. *Alternative Law Journal, 43*(1), 45-47. doi:10.1177/1037969X17748210

Shepherd, S. M. (2015). Finding color in conformity: A commentary on culturally specific risk factors for violence in Australia. *International Journal of Offender Therapy and Comparative Criminology, 59*(12), 1297-1307. doi:10.1177/0306624x14540492

Shepherd, S. M., Adams, Y., McEntyre, E., & Walker, R. (2014a). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law, 20*(3), 281-293. doi:10.1037/law0000017

Shepherd, S. M., & Anthony, T. (2018). Popping the cultural bubble of violence risk assessment tools. *The Journal of Forensic Psychiatry & Psychology, 29*(2), 211-220. doi:10.1080/14789949.2017.1354055

Shepherd, S. M., & Lewis-Fernandez, R. (2016). Forensic risk assessment and cultural diversity: Contemporary challenges and future directions. *Psychology, Public Policy, and Law, 22*(4), 427-438. doi:10.1037/law0000102

Shepherd, S. M., Luebbers, S., Ferguson, M., Ogloff, J., & Dolan, M. (2014b). The utility of the SAVRY across ethnicity in Australian young offenders. *Psychology, Public Policy, and Law, 20*(1), 31-45. doi:10.1037/a0033972

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

Shepherd, S. M., & Spivak, B. L. (2020). Finding color in conformity part II: Reflections on structured professional judgement risk assessment. *International Journal of Offender Therapy and Comparative Criminology*. doi:https://doi.org/10.1177/0306624X20928025

Shepherd, S. M., & Strand, S. (2016). The PCL: YV and re-offending across ethnic groups. *Journal of Criminal Psychology, 6*(2), 51-62. doi:10.1108/JCP-02-2016-0006

Shepherd, S. M., & Willis-Esqueda, C. (2018). Indigenous perspectives on violence risk assessment: A thematic analysis. *Punishment and Society, 20*(5), 599-627. doi:10.1177/1462474517721485

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). New York: Springer.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8-22. doi:doi:10.1002/bsl.2052

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. doi: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. doi:https://doi.org/10.1016/j.cpr.2010.11.009

Skeem, J., & Lowenkamp, C. (in press). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences and the Law*.

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

Smallbone, S., & Rallings, M. (2013). Short-term predictive validity of the Static-99 and Static-99-R for Indigenous and nonindigenous Australian sexual offenders. *Sexual*

Abuse A Journal of Research and Treatment, 25(3), 302-316. doi:10.1177/1079063212472937

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

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293. doi:10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26. doi:10.1111/1529-1006.001

The United States Department of Justice. (2014). Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th annual meeting and 13th State Criminal Justice network conference. Retrieved from https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior, 36*(4), 345-355. doi:10.1037/h0093930

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. doi:10.1177/1468017317743137

Tsang, S., Piquero, A. R., & Cauffman, E. (2014). An examination of the Psychopathy Checklist: Youth Version (PCL:YV) among male adolescent offenders: An item response theory analysis. *Psychological Assessment, 26*(4), 1333-1346. doi:10.1037/a0037500 van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*(2), 119-135. doi:10.1016/j.erap.2003.12.004

van Eijk, G. (2017). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society, 19*(4), 463-481. doi:10.1177/1462474516666282

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

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). Boston, MA: Springer US.

Venner, S., Sivasubramaniam, D., Luebbers, S., & Shepherd, S. M. (2020). Cross-cultural reliability and rater bias in forensic risk assessment: A review of the literature. *Psychology, Crime & Law*, 1-17. doi:10.1080/1068316X.2020.1775829

Verma, S., & Rubin, J. (2018). *Fairness definitions explained*. Paper presented at the International Workshop on Software Fairness, Gothenburg, Sweden. https://doi.org/10.1145/3194770.3194776

Viallon, V., Ragusa, S., Clavel-Chapelon, F., & Bénichou, J. (2009). How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine, 28*(6), 901-916. doi:10.1002/sim.3517

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

Watkins, I. (2011). *The utility of Level of Service Inventory-Revised (LSI-R) assessments within NSW correctional environments. Research bulletin*. NSW, Australia: Corrective Services NSW.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). Assessing risk of violence to others. In C. D. Webster & M. A. Jackson (Eds.), *Impulsivity: Theory, assessment, and treatment* (pp. 251-277). New York: Guilford Press.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*(7), 595-613. doi:10.1037/0003-066X.59.7.595

Whiteacre, K. W. (2006). Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review, 17*(3), 330-342. doi:10.1177/0887403405284766

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. doi:10.1177/0093854813500958

Wisser, L. (2019). Pandora's algorithmic black box: The challenges of using algorithmic risk assessments in sentencing. *American Criminal Law Review, 56*(4), 1811-1832.

Wormith, J., & Hogg, S. (2012). *The predictive validity of Aboriginal offender recidivism with a general risk/needs assessment inventory*. Saskatoon, SK: Program Effectiveness, Statistics, and Applied Research Unit, Ministry of Community Safety and Correctional.

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. doi:10.1177/0093854814552843

Yang, M., Wong, S., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*(5), 740.

Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., & Razali, M. N. (2020). A classification approach for crime prediction. In M. Khalaf, D. Al-Jumeily, & A. Lisitsa (Eds.), *Applied computing to support industry: Innovation and technology* (pp. 68-78). Cham: Springer International Publishing.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). *Learning fair representations*. Paper presented at the 30th International Conference on Machine Learning, Atlanta, GA, USA.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A, 180*(3), 689-722. doi:10.1111/rssa.12227

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating unwanted biases with adversarial learning*. Paper presented at the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA.

Chapter Three: General Research Methodology

3.1 Introduction

This chapter will initially provide a brief overview of the design and sample used for this research. Further information about where the data was sourced, the data linkage, and the data cleaning process is then provided. The coding of specific variables, including recidivism and demographics, is detailed, as is information about the risk assessment instrument that was used for measurement. Information surrounding sample demographic characteristics is then discussed, followed by data analytic approaches and methods utilised in the empirical studies, and research ethics.

3.2 Research Design

The current research was a retrospective study that examined and aimed to increase the discrimination and fairness of a risk assessment instrument for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. All empirical studies involved in this research include quantitative analyses of an individual's score on a risk assessment instrument and recidivism.

3.3 Sample

The sample initially included 561 individuals sentenced to a term of imprisonment for a serious violent offence as detailed in schedule 1 (clause 3) of the *Sentencing Act* 1991 (Vic) from Victoria, Australia, and who were received into prison between January 2015 and December 2017. This included 109 (19.43%) individuals who had not been released from prison by the end of the recruitment period for the current thesis. This resulted in a sample of 452 individuals for which the utility of the Level of Service/Risk Need Responsivity (LS/RNR) could be examined. However, for a clearer sample, the 72 (15.93%) females from the sample were also removed, resulting in a final sample of 380 males.

Within the final sample, 231 (60.79%) individuals were assessed while serving a prison sentence. The mean length of incarceration for those individuals who were assessed while in prison was 607.66 days (*SD* = 385.09, median = 532, range 32 to 3018 days). There were a further 148 (38.95%) individuals who were assessed while completing a community corrections order, and 1 (0.26%) individual who was assessed while on parole. The sample included 180 (47.37%) individuals who identified as Aboriginal and/or Torres Strait Islanders and 200 (52.63%) who were non-Aboriginal and Torres Strait Islanders.

The individuals were assessed by correction officers with the LS/RNR during their respective incarceration, community correction order, or parole periods. As displayed in Table 1, the proportions of Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders who were assessed with the LS/RNR while incarcerated, serving a community corrections order, or on parole were similar.

Table 1

Incarceration, Community Correction Order, and Parole Orders by Group

	Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	n	%	*n*	%
Incarcerated	108	60	123	61.5
Community Correction Order	72	40	76	38
Parole Order	0	0	1	0.5

For individuals to be eligible to be assessed with the LS/RNR, they originally had to be serving a prison sentence of at least six months and receive at least a moderate assessment of risk on an initial assessment of risk, which was established using the Level of Service Inventory-Revised Screening Version (LSI-R: SV; Andrews & Bonta, 1998). An initial LS/RNR was then completed by corrections officers within six weeks of the LSI-R: SV. As there are dynamic items within the LS/RNR, an individual can be assessed with the LS/RNR multiple times. Completions of the LS/RNR, demographics, and dates relating to incarceration, community correction orders, and parole were obtained from Corrections Victoria. The total number of assessors (i.e., corrections officers) who administered the LS/RNR for this specific sample was unknown as this information was not provided. The follow-up recidivism data (i.e., new police charges) was obtained from the Victorian Police Law Enforcement Assistance Program (LEAP) database.

3.4 Data Sources

3.4.1 Corrections Victoria

Corrections Victoria is a unit of the Department of Justice and Community Safety that is responsible for policy, standards, and management of Victoria's correctional facilities. It is also responsible for the development of programs to aid in the management and rehabilitation of prisoners, including the administration of the LS/RNR to prisoners (Corrections Victoria, 2020a). The LS/RNR data for any assessments conducted for those sentenced to a term of imprisonment for a serious violent offence and received into prison between January 2015 and December 2017 was extracted and provided by Corrections Victoria for the current sample. This included assessment information that was obtained for the current sample when they were received into prison, began a community corrections order, or were placed on parole. The assessment data comprised the LS/RNR item-level data, risk factor scores, and overall risk level. Corrections Victoria also provided dates of LS/RNR completion, as well as dates that individuals were received into and released from prison, or the start and end dates of community corrections orders and parole. Numerous demographics for the sample were also provided, including date of birth, ethnicity, marital status, employment, education history, and index offence.

3.4.2 Victoria Police Law Enforcement Assistance Program (LEAP)

LEAP is an online database that was introduced state-wide in Victoria in March 1993 (Victoria Police, 2019). LEAP contains information about Victorian individuals that have had contact with Victoria Police, including information about crimes, missing persons, and family incidents. This database is routinely updated with information and is utilised by Victoria Police to develop criminal statistics and for data analysis purposes. This database incorporates in excess of 5000 individual statutory and common law offences that are grouped into 27 offence categories, further divided into four classes of crime against the person, crime against property, drug offences, and other crimes. Recidivism data was obtained through the LEAP database, with individuals charged by Victoria Police while at risk to the community being labelled as recidivists. The information provided included the report date of any new charge (or charges), the date the offence was committed, the date the offence was charged, and the offence description for the period of January 2015 to December 2019. This data was extracted by Victoria Police staff based on the list of individuals provided by Corrections Victoria.

3.5 Procedure

3.5.1 Data Collection

A study sample was identified by Corrections Victoria for which a master participant list was created. All Aboriginal and Torres Strait Islanders who were received into prison for a serious violent offence within the study recruitment period were eligible to be sampled. Non-

Aboriginal and Torres Strait Islanders were then randomly sampled to have approximately similar sample sizes for both groups. The master participant list for this sample included information such as the individual's full name, date of birth, Corrections Reference Number (CRN), and a unique study identifier. This was used to identify Corrections Victoria offender files in order to extract the LS/RNR and demographic information. This master participant list was also supplied to Victoria Police for the purpose of data extraction from Victoria Police's LEAP database. The information was returned at each step to the researchers with all the identifying information removed. Once all the data was collected, identifying information was also removed from the master participant list. The unique identifiers used by both Corrections Victoria and Victoria Police were retained within the master participant list and information was provided in order to link the datasets.

3.5.2 Data Cleaning and Linkage

The unique study identifier was used to link each of the records provided by Corrections Victoria and Victoria Police, resulting in the final dataset to be analysed. Initially, the first LS/RNR completion for each individual was obtained to enable the longest follow up period. For those individuals whose first assessment was during a period of incarceration, the latest assessment that occurred within that incarceration period prior to release was chosen to account for the LS/RNR items that measured dynamic (i.e., changing) information (e.g., current alcohol and drug use). The unique identifier for these completions was then matched to the unique identifiers within the recidivism data provided by Victoria Police's LEAP database. Any new charges that were recorded to have occurred after the completion of the LS/RNR assessment were retained and regarded as recidivism. The unique identifiers and date of each LS/RNR completion were then matched to either a period of incarceration, community correction orders, or parole. This enabled the identification of individuals who were still incarcerated and those who had previously been released and had the opportunity to engage in recidivism. Last, all demographic information that was provided by Corrections Victoria was matched to the LS/RNR and recidivism data by using the unique identifier for each individual. All linkage of information about the individuals in the current sample was conducted by researchers from the Centre for Forensic Behavioural Science, Swinburne University of Technology.

3.6 Coding of Variables

3.6.1 Demographic Variables

To enable a comparison of cross-cultural fairness and discrimination across groups, the sample was divided into Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. This was based on information provided by Corrections Victoria. Although this was a straightforward way of distinguishing between Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander groups, and it is similar to what has been done in previous research (e.g., Hsu et al., 2010), it does include a number of limitations. First, the non-Aboriginal and Torres Strait Islander group encompasses a significant number of other cultures. Australia is a multi-cultural society (Australian Bureau of Statistics, 2020a, 2020b) and, therefore, the non-Aboriginal and Torres Strait Islander group incorporates both those with an English speaking background and those who are culturally and linguistically diverse (CALD). However, the non-Aboriginal and Torres Strait Islander group was unable to be further classified into more distinct cultural groups for the present study due to inadequate information. Further, information such as country of birth and primary language that can be used to identify those who are CALD, did not have sufficient variation to develop a CALD subgroup. Specifically, the majority (*n* = 166, 83%) of the non-Aboriginal and Torres Strait Islander group identified Australia as their country of birth, and all 200 identified English as their primary language.

Second, Aboriginal and Torres Strait Islanders comprise hundreds of groups, with high levels of diversity existing among different tribes, clans, and language groups (Australian Institute of Health and Welfare, 2020). However, this information was not sufficiently or accurately recorded. Additionally, the limited sample size of the study was not sufficient to divide both the non-Aboriginal and Torres Strait Islander or Aboriginal and Torres Strait Islander groups further into more representative and accurate cultural groups. The Aboriginal and Torres Strait Islander group are further non-representative of the true population as they were oversampled for the present thesis to enable comparisons between groups. Specifically, Aboriginal and Torres Strait Islanders comprised 47.3% of the total sample when they only comprise approximately 9% of the adult prison population in Victoria (Australian Bureau of Statistics, 2018a).

3.6.2 Recidivism Variable

For the current research, recidivism was defined as any future incident within the community that resulted in a police charge. The outcome of recidivism was assessed in a binary format where an individual was classified as a recidivist (1) or a non-recidivist (0). Across the entire sample, the majority were classified as recidivists by the end of the follow up period (*n* = 306, 80.56%), with 154 (85.56%) Aboriginal and Torres Strait Islanders and 152 (76%) non-Aboriginal and Torres Strait Islanders being classified as recidivists. For those individuals who engaged in recidivism, the average time from LS/RNR completion to recidivism was 293.56 days (*SD* = 285.36, median = 220.5, range = 1 to 1533 days). The average was slightly lower for Aboriginal and Torres Strait Islanders (*M* = 279.47, *SD* = 266.63, median = 220, range = 5 to 1362 days) than for non-Aboriginal and Torres Strait Islanders (*M* = 307.84, *SD* = 303.38, median = 220.5, range = 1 to 1533 days).

3.7 Sample Demographics

The average age of the sample at the time of LS/RNR assessment was 32.17 years (*SD* = 8.69, ranging from 19 to 57 years of age). Age at LS/RNR assessment was found to be similar for Aboriginal and Torres Strait Islanders (*M* = 31.29, *SD* = 8.38, range = 18 to 51 years of age) and non-Aboriginal and Torres Strait Islanders (*M* = 32.96, *SD* = 8.92, range = 18 to 57 years of age). Further demographic information that was provided by Corrections Victoria is presented below in Table 2.

The majority of the demographic variables were similar across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Specifically, marital status and highest level of education were comparable, with the non-Aboriginal and Torres Strait Islander group having a slightly higher proportion reaching higher levels of education. The biggest disparities were found in employment status, with Aboriginal and Torres Strait Islanders having a larger proportion of unemployed. Across these demographics, the majority of the sample (including both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders) were never married or single, had completed part of secondary school as their highest level of education, and were unemployed at the time of assessment.

Table 2

Demographic Information of Sample

	Overall		Aboriginal and Torres Strait Islanders		Non-Aboriginal and Torres Strait Islanders	
Variable	n	%	n	%	n	%
Marital Status						
Never Married / Single	248	65.26	120	66.67	128	64.00
De Facto	89	23.42	41	22.78	48	24.00
Married	20	5.26	9	5.00	11	5.50
Separated	11	2.89	6	3.33	5	2.50
Divorced	6	1.58	1	0.56	5	2.50
Widowed	4	1.05	2	1.11	2	1.00
Unknown/Not Stated	2	0.53	1	0.56	1	0.50
Highest Level of Education						
Part of Primary School	3	0.79	2	1.11	1	0.50
Completed Primary School	3	0.79	2	1.11	1	0.50
Part of Secondary School	346	91.05	166	92.22	180	90.00
Secondary School	8	2.11	2	1.11	6	3.00
Trade/Apprenticeship	5	1.32	3	1.67	2	1.00
Tertiary	2	0.53	0	0	2	1.00
Unknown	13	3.42	5	2.78	8	4.00
Prior Employment						
Unemployed	204	53.68	104	57.78	100	50.00
Employee	97	25.53	42	23.33	55	27.5

Variable	Overall		Aboriginal and Torres Strait Islanders		Non-Aboriginal and Torres Strait Islanders	
	n	%	n	%	n	%
Prior Employment (cont.)						
Employer	1	0.26	0	0	1	0.50
Self-Employed	22	5.79	10	5.56	12	6.00
Student	4	1.05	2	1.11	2	1.00
Home Duties	1	0.26	0	0	1	0.50
Pensioner	44	11.58	18	10.00	26	13.00
Other	1	0.26	0	0	1	0.50
Unknown/Not Stated	6	1.58	4	2.22	2	1.00

3.8 Risk Assessment Instrument

3.8.1 The Level of Service/Risk Need Responsivity (LS/RNR; Andrews et al., 2008)

The LS/RNR is an actuarial instrument designed to estimate general recidivism risk and detect criminogenic needs. Further, it provides structure in the management and treatment planning of prisoners. As well as General Risk/Needs section, the LS/RNR also looks at strengths and protective factors, responsivity considerations, specific risk and need factors, and non-criminogenic needs. However, for the purpose of the present thesis, only the General Risk/Needs section was used for analysis.

The General Risk/Needs section comprises eight factors that are scored using 43 items – Criminal History (8 items), Education/Employment (9 items), Family/Marital (4 items), Leisure/Recreation (2 items), Companions (4 items), Alcohol/Drug Problem (8 items), Procriminal Attitude (4 items) and Antisocial Pattern (4 items). Each of these items is ultimately scored as 0 if absent and 1 if present, and the present problematic items are summed up to create respective factor scores and a total General Risk/Needs score. Using the total score, individuals can be categorised into various risk levels that include very low risk (0–4), low risk (5–10), medium risk (11–19), high risk (20–29) and very high risk (30–43).

Where appropriate, an override of the risk level can be considered. Administrators can review and override the General Risk/Needs score using their professional judgement. Corrections Victoria requires that any request for an override be accompanied with supporting comments and formally approved. Further, the risk level of the LS/RNR can only be overridden up or down by one risk category, for example, low risk to medium risk or medium risk to high risk (Corrections Victoria, 2020b). In the case of the present study, no overrides were detected in the final sample. Inter-rater reliability for the present thesis was unable to be determined as the information provided by Corrections Victoria included the item scores and risk factor scores that were completed by a single assessor (i.e., corrections officer) only.

3.9 Data Analytical Approach

All data was analysed through RStudio using R version 4.0.2 (R Core Team, 2021). Numerous packages were utilised, including the *tidyverse* packages (Version 1.3.0; Wickham, 2019) for data cleaning and management, *rms* (Version 6.0-1; Harrell, 2020) for logistic regression, *pROC* (Version 1.16.2; Robin et al., 2020) to generate receiver operating characteristic (ROC) curves and AUC values, *survival* (Version 3.2-7; Therneau, 2020*),* and *survminer* (Version 0.4.8; Kassambara et al., 2020) for survival analysis, *caret* (Version 6.0-88; Kuhn, 2021) for model training and cross-validation, *glmnet* (Version 4.1-2; Friedman et al., 2021) for penalised 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* (Verision 1.7-8; Meyer et al., 2021) for support vector machine algorithms, *cutpointr* (Version 1.1.1; Thiele, 2021) to generate optimal cut-offs, and *iml* (Version 0.10.1; Molnar, 2020) to calculate Shapley Values.

3.9.1 Predictive Validity

Predictive validity refers to whether the total score or the risk category of a risk instrument accurately predicts the probability of recidivism (Singh, 2013). For the current thesis, this refers to the ability of the LS/RNR general risk score to predict recidivism. The predictive validity of the LS/RNR was established through survival analysis.

3.9.1.1 Survival Analysis. Cox regression analyses (Cox, 1972) were used to establish the predictive validity of the LS/RNR risk score while also accounting for individual differences in time at risk to the community. Further, Cox regression analyses were also utilised to observe the impact that Aboriginal and Torres Strait Islander status had on recidivism. A 'time at risk' variable to measure survival time was created that started at the date of release from prison for individuals who were incarcerated or from the date of LS/RNR assessment for those who were not incarcerated (i.e., those who were completing a community corrections order or parole order). The end date was either the date of the first offence for those individuals who went on to receive a charge, or the end of the follow up period (31-12-2010) for those individuals who did not receive a charge. Individuals who were subject to another period of incarceration within these dates had the dates of incarceration summed and removed from their total time at risk. This ensured that the time at risk variable only captured an individual's true time at risk within the community and not periods of incarceration.

Cox regression analyses produce hazard ratios (e^B) that represent the increase in the hazard of recidivism for a 1 unit increase in the predictor variables (LS/RNR total risk score and Aboriginal and Torres Strait Islander status). A hazard ratio greater than 1 represents a predictor that is associated with an increased risk of recidivism, whereas a hazard ratio less than 1 represents a predictor that is associated with a decreased risk of recidivism. To visually observe the survival time differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, Kaplan-Meier survival curves were plotted.

Overall, the whole sample had an average time at risk of 289.95 days (*SD* = 339.72, median = 129, range = 1 to 1589 days). The average time at risk for Aboriginal and Torres Strait Islander people was 232.43 days (*SD* = 270.20, median = 116.5, range = 4 to 1350 days). This was less than the average time at risk for non-Aboriginal and Torres Strait Islanders (*M* = 341.72, *SD* = 385.28, median = 172.50, range = 1 to 1589 days). When observing those individuals who received a charge, the average time at risk before recidivism was 184.75 days (*SD* = 233.80, median = 89.50, range = 1 to 1515 days). Those Aboriginal and Torres Strait Islanders who went on to engage in recidivism were at risk in the community on average for 167.97 days (*SD* = 203.28, median = 83, range = 4 to 1258 days), and non-Aboriginal and Torres Strait Islanders who went on to engage in recidivism were at risk in the community on average for 201.74 days (*SD* = 260.70 days, median = 96.50 days, range = 1 to 1515 days).

3.9.2 Discrimination Indices

Discrimination indices assess how well a risk assessment instrument is able to distinguish between individuals who go on to engage in recidivism from those who do not (Cook, 2007). For the current thesis, the discrimination of the LS/RNR was assessed by two forms of discrimination indices, including the area under the receiver operating characteristic curve and the cross area under the receiver operating characteristic curve.

3.9.2.1 Receiver Operating Characteristic (ROC)/ Area under the Curve (AUC). The area under the receiver operating characteristic (ROC) curve was utilised as a measure of discrimination. The ROC curve considers the sensitivity (the proportion of those accurately predicted to engage in recidivism from all those who were recidivists) and specificity (the proportion of those accurately predicted to not engage in recidivism from all those who were not recidivists). The ROC curve plots the sensitivity against 1 – specificity at various thresholds and is unimpeded by differing base rates across groups (Cook, 2007; Singh, 2013). The area under the curve (AUC) is typically calculated and can range from 0 to 1, with the midpoint (.50) demonstrating discrimination at chance levels (Cook, 2007; Helmus & Babchishin, 2017; Rice & Harris, 2005). The AUC can be understood as the probability that a randomly selected individual who engaged in recidivism received a higher risk score than a randomly selected individual who did not (Helmus & Babchishin, 2017; Singh, 2013; Singh et al., 2013; Swets et al., 2000). There are varying benchmarks of what can constitute a small, medium, and large effect size for the AUC value (Cohen, 1988; Rice & Harris, 2005; Singh et al., 2013; Swets, 1988). A common approach used in previous forensic psychology and criminology research that was adopted for the present thesis is: values between .56-.63 indicate a small effect, .64-.70 a medium effect, and .71 and above as a large effect (Rice & Harris, 2005). AUC was determined for the LS/RNR overall and individually for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

3.9.2.2 Cross Receiver Operating Characteristic (xROC)/ Cross Area under the Curve (xAUC). The cross area under the curve (xAUC; Kallus & Zhou, 2019) is an alteration of the traditional AUC that measures discrimination between groups. The traditional AUC compares within groups (e.g., recidivists and non-recidivists within Aboriginal and Torres Strait Islanders), whereas the xAUC is a better measure for identifying disparities between groups. A cross receiver operating characteristic (xROC) plots the sensitivity against 1 – specificity at various thresholds for two sets of groups for which a xAUC can be calculated. The first set contains the positive outcome (i.e., non-recidivist) from one of the groups (i.e., Aboriginal and Torres Strait Islander) and the negative outcome (i.e., recidivist) from the other group (i.e., non-Aboriginal and Torres Strait Islander). The second set contains the opposite of the first set (i.e., Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists). Therefore, the xAUC measures the probability that a randomly selected individual who engages in recidivism from one of the groups received a higher risk score than a randomly selected individual who did not engage in recidivism from the other group. This way, discrimination is assessed between groups instead of within and can also be understood as a form of fairness. Here, if discrimination is fair between groups, recidivists from one group should receive higher risk scores compared to the non-recidivists from the other group for both sets that the xAUC can be calculated for. In empirical studies one and two, the xAUC is used as a measure of discrimination. However, it is also discussed as a form of fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

3.9.3 Fairness

The fairness of the LS/RNR across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders that was calculated for this research included error rate balance, calibration, predictive parity, and statistical parity. Both predictive parity and error rate balance metrics can be understood with the use of a confusion matrix as represented in Table 3. Using the information contained within a confusion matrix, positive predictive values, negative predictive values, false positive rates, and false negative rates can be calculated.

Table 3

Confusion Matrix of Predicted and Observed Outcomes with Fairness Calculations

	Recidivist	Non-Recidivist	
Predicted to be a recidivist	True Positive (TP)	False Positive (FP)	PPV = TP/(TP+FP)
Not predicted to be a recidivist	False Negative (FN)	True Negative (TN)	NPV = TN/(FN+TN)
	FNR = FN/(TP+FN)	FPR = FP/(FP+TN)	

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate; TP true positive; FP false positive; FN false negative; TN true negative.

For the four fairness metrics (PPV, NPV, FNR, and FPR) that can be calculated using the above confusion matrix, a cut-off point is required in risk scores to distinguish between those predicted to engage in recidivism (i.e., those determined as high risk or above that cut-off value) and those not predicted to engage in recidivism (i.e., those determined as low risk or below the specified cut-off value). However, the LS/RNR has more than two risk classifications and using a single cut-off value does not reflect how the instrument is used in practice. Further, research has also demonstrated that differing cut-off scores result in variations in values as the proportions of low risk and high risk classifications change (Zottola et al., 2021). Therefore, error rate balance and predictive parity were computed for all LS/RNR total risk scores for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders and compared across each LS/RNR risk score (i.e., cut-off value) for Empirical Study One.

For Empirical Study Two, fairness was calculated for numerous statistical learning methods, and due to the number of metrics that needed to be reported, a cut-off value was utilised for this study. There are numerous ways to determine a cut-off value (Kuhn & Johnson, 2013), with none deemed to be the best method. For the present study, the cut-off was determined by the cut-off value that yielded the smallest distance to the point 0, 1 on the ROC space. Although Empirical Study Two aimed to increase fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, achieving this while still maintaining acceptable levels of discrimination was also pivotal. Therefore, this approach was utilised as an instrument that passes through 0, 1 in the ROC space reflects perfect levels of discrimination. A cut-off value was calculated using this method for the LS/RNR total risk score as well as for each statistical learning method.

The following sections will define each fairness definition used in this thesis and discuss how they were assessed.

3.9.3.1 Error Rate Balance. The error rate balance in a risk assessment instrument is satisfied when the false negative rate (FNR) and false positive rate (FPR) is equal across groups (Chouldechova, 2017). The FNR refers to the proportion of individuals who recidivate who are classified as low risk (or were predicted not to recidivate) and was calculated by calculating:

$$\text{False Negative Rate} = \frac{\text{False Negative}}{(\text{True Positive} + \text{False Negative})}$$

This was calculated for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The FPR refers to the proportion of individuals who are classified as high risk (or were predicted to engage in recidivism) who do not engage in recidivism and was calculated by calculating:

$$\text{False Positive Rate} = \frac{\text{False Positive}}{(\text{True Negative} + \text{False Positive})}$$

This was calculated for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The FNR and FPR were then compared across groups for equivalence across all available cut-off thresholds in Empirical Study One and at a specific cut-off threshold in Empirical Study Two.

3.9.3.2 Calibration. Calibration across groups in a risk assessment instrument refers to similarities in the likelihood of recidivism across risk scores or classifications (Chouldechova, 2017; Corbett-Davies & Goel, 2018; Verma & Rubin, 2018). This was observed in the Empirical Study One by initially comparing the proportions of recidivists across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders by risk classification. This was further assessed by utilising an approach that examines if different groups systematically deviate from a similar regression line (i.e., slope and intercept differences) that relates to risk assessment scores and recidivism (see Flores et al., 2016; Monahan et al., 2017). Specifically, four bivariate logistic regression models were conducted and compared to test for differences in regression slopes and intercepts between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

In the first model, only Aboriginal and Torres Strait Islander status was used to predict recidivism. In the second model, the LS/RNR risk score was used to predict recidivism. The third model incorporated both Aboriginal and Torres Strait Islander status and the LS/RNR risk score to predict recidivism. The fourth model included Aboriginal and Torres Strait Islander status, LS/RNR risk score, and an interaction between Aboriginal and Torres Strait Islander status and LS/RNR risk score to predict recidivism.

Differences in intercept were determined by comparing models two and three, with Aboriginal and Torres Strait Islander status adding significant incremental utility to the LS/RNR risk score in predicting recidivism in model three, being indicative of differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders regression intercepts. Differences in slope were determined by comparing models three and four, with Aboriginal and Torres Strait Islander status significantly moderating the utility of the LS/RNR risk score in predicting recidivism, demonstrating differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders regression slopes. If the regression slopes and intercepts did not differ between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, this indicates that a risk score on the LS/RNR relates to the same recidivism rate for both groups. To aid in visualising the regression intercept and slopes for both groups, model four was used to predict the probability of recidivism across all possible LS/RNR risk scores. Predictions by risk score were then grouped for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders and plotted.

For Empirical Study Two, calibration was unable to be assessed in this way due to statistical learning methods producing a predicted probability of recidivism. Therefore, Brier scores (Brier, 1950) were used for Empirical Study Two as a measure of alignment between prediction and outcome. Specifically, Brier scores measure the squared error between the predicted probability of recidivism (which ranged between 0 and 1) and the outcome (specified as 0 if the individual did not engage in recidivism and 1 if the individual did engage in recidivism). A lower Brier score indicates better predictive performance and predictions that are more accurate. Brier scores can range from zero to one, with the worst possible Brier score being one, and the best possible Brier score being zero. Brier scores were calculated for the sample overall, as well as for Aboriginal and Torres Strait Islanders, and non-Aboriginal and Torres Strait Islanders for each statistical learning method.

3.9.3.3 Predictive Parity. Predicted parity in a risk assessment instrument is satisfied when the positive predictive values (PPV) of both groups are equivalent (Chouldechova, 2017).

The PPV refers to the proportion of those individuals who engage in recidivism from all those predicted to engage in recidivism and was determined by calculating:

$$\text{Positive Predictive Value} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

This was calculated for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. An extension of this is that where the negative predictive values (NPV), or those individuals who do not go on to engage in recidivism from all those predicted to not engage in recidivism, is also equivalent across groups (Berk et al., 2018). This was determined by calculating:

$$\text{Negative Predictive Value} = \frac{\text{True Negative}}{(\text{True Negative} + \text{False Negative})}$$

This was calculated for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The PPV and NPV were then compared across groups for equivalence across all available cut-off thresholds in Empirical Study One and at a specific cut-off threshold in Empirical Study Two.

3.9.3.4 Statistical Parity. Statistical parity in a risk assessment instrument is satisfied when the proportion of individuals in risk classifications, or the distribution of risk scores, is equal across groups (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Huq, 2019). For the current research, a comparison of mean scores on the General Risk/Needs section of LS/RNR for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders was conducted for Empirical Study One using an independent samples *t*-test with Cohen's *d* reported as a measure of effect size. This included the overall risk score from the General Risk/Needs section as well as the eight risk factors. For Empirical Study Two, using the cut-off method specified above (closest point to 0,1 on the ROC space), statistical parity was assessed by comparing the proportion of Aboriginal and Torres Strait Islanders to the proportion of non-Aboriginal and Torres Strait Islanders whose predicted probability of recidivism fell above that cut-off. Those whose predicted probability of recidivism was above the cut-off threshold were classified as high risk, and those who fell below the cut-off threshold were classified as low risk. This high risk proportion was then compared across groups for the LS/RNR total risk score and each of the statistical learning methods.

3.9.4 Statistical Learning Methods

The following section details the statistical learning methods (i.e., machine learning algorithms) utilised to increase the discrimination and fairness of the LS/RNR using the items from the LS/RNR. These statistical learning methods were specifically used to predict if an individual would go on to engage in recidivism. This section also details the sampling technique used to assess the performance of each statistical learning method. The best tuning parameters for each statistical learning method were established through cross-validation in the training set by testing a comprehensive range of parameter values and combinations.

3.9.4.1 Logistic Regression. Logistic regression is used to assess the probability of a binary dependent outcome (e.g., an individual who either went on to engage in recidivism or who did not) with specified continuous and/or categorical predictors (Hair et al., 2019). Logistic regression was used in the current research as a baseline comparison to assess the performance of the other algorithms against. This method offers a transparent and interpretable approach to predicting the likelihood of recidivism when compared to some other statically learning methods that were utilised in this thesis, with *B* weights explicitly expressed to enable the individual importance of predictors (Breiman, 2001b). Using logistic regression, therefore, helped to weigh the benefits of a transparent and interpretable model in comparison with a potentially complex and inscrutable model that can offer increased predictive accuracy and discrimination.

3.9.4.2 Penalised Logistic Regression. Penalised logistic regression was also used for the current study as it aims to increase the predictive power of logistic regression by increasing the simplicity of the model and reducing overfitting (i.e., when a model also picks up on the unique noise of the sample and has poorer accuracy when used on a new sample) and the impact of collinearity (Kuhn & Johnson, 2013; Salo et al., 2019; Zou & Hastie, 2005). This is achieved by adding a penalty term as a parameter to predictors and reducing regression coefficients towards zero. Specifically, ridge regression (Hoerl & Kennard, 2000) imposes a penalty term on the squared size of the coefficients and shrinks irrelevant (i.e., the least predictive) predictor coefficients towards zero. All predictors are retained in the final model as no predictors are set to exactly zero. The lasso regression (Tibshirani, 1996; Tibshirani, 2011) imposes a penalty on the absolute value of the coefficients and shrinks irrelevant predictor coefficients completely to zero. This produces a simpler and more interpretable regression model that incorporates only the most important predictors.

Both of these forms of penalised logistic regression can be seen as ordinary logistic regressions when the penalty parameters are set to zero. As the penalty parameters increase, the models approach a null model in which the coefficients are zero. Elastic net regression (Zou & Hastie, 2005) was specifically used for the current study as it combines these two forms of penalised logistic regression. Here, another parameter term is imposed that mixes and signifies the type of penalty used, ranging from 0 to 1, where 0 reflects a pure ridge regression penalty and 1 reflects a pure lasso regression penalty (Friedman et al., 2021; Kuhn & Johnson, 2013). A value between 0 and 1 will shrink some coefficients towards zero (i.e., ridge regression) and some to exactly zero (i.e., lasso regression).

3.9.4.3 Random Forests. Random forests (Breiman, 2001a) is an ensemble based algorithm (i.e., a combination of numerous algorithm predictions) of decision trees. Random forests combine the concepts of bagging with random feature selection. In other words, a number of training datasets are generated through bootstrap sampling of the original training data, which are then used to generate a set of algorithms. 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 minimise the correlation between trees and improve accuracy (Hastie et al., 2009; Kuhn & Johnson, 2013; Lantz, 2015). The predicted probabilities across trees were aggregated and the average predicted probability was used once the ensemble of decision trees (i.e., forest) was generated.

3.9.4.4 Stochastic Gradient Boosting. Stochastic gradient boosting (Friedman, 2002) is a consecutive learning process in which a weak learner (i.e., a learner, often a decision tree, that predicts slightly better than random) is applied repeatedly to the data (Kuhn & Johnson, 2013). It seeks to find an additive algorithm that will minimise the loss function (e.g., squared error). Initially, specified predicted values are utilised (e.g., this can be the average) so that the residual can be established between that predicted value and the observed (Lantz, 2015). 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), and new predicted values being added to the previous.

Similar to 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 final outcome. Instead, each tree is dependent on past trees and is weighted depending on how much of an influence they have over the final outcome (Kuhn & Johnson, 2013; Lantz, 2015). The use of a random subsample helps increase the accuracy, execution speed, and robustness of the algorithm.

3.9.4.5 Support Vector Machines. Support vector machines (Vapnik, 1998; 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., recidivist and non-recidivist) to create the greatest segregation between the two. Therefore, data points that are predicted to fall on either side of this hyperplane can be attributed to an outcome class. The points from each class that are closest to the hyperplane are referred to as support vectors and help distinguish the maximum margin hyperplane between the two classes (Lantz, 2015). As these data points are unlikely to be easily separable in two dimensions, kernels can be used to transform the data into higher dimensions, enabling separation. Support vector machines also enable a cost parameter to be specified, in which misclassifications are penalised (Kuhn & Johnson, 2013). With a large cost parameter, a smaller margined hyperplane would be utilised to avoid any misclassification in the training data, whereas for a small cost parameter, a larger margined hyperplane can be used, which may include some misclassified data points.

3.9.4.6 k-Fold Cross-Validation. *k*-fold cross-validation is a resampling technique that was used to evaluate the performance of the statistical learning methods (Kuhn & Johnson, 2013) utilised in the current study. It is a particularly useful resampling technique on smaller samples and helps avoid overfitting. The sample is randomly portioned into *k* sample sets of equal size. An algorithm is then fit using all the samples except the first sample set (often referred to as the first fold). The held out sample set is then predicted by this algorithm. The first sample set is then returned to the training set and the process is repeated with the second fold now held out and so-on until all folds have been held out. The performance of each algorithm's prediction on the held out sample set is aggregated and summarised to determine the algorithms' average performance. For the current study, a *k* of 10 was used (i.e., an algorithm is trained on 90% of the data and evaluated on 10% of the data 10 times), as this is found to reduce bias and there has been little evidence to suggest that a higher number of folds adds any benefits (Kuhn & Johnson, 2013; Lantz, 2015).

This validation process was also used to trial a number of different parameter options for each statistical learning method. Ultimately, the parameter options that resulted in the best performance (i.e., the highest AUC value) were used. For penalised logistic regression, the parameters included alpha, a parameter that mixes the elastic net between pure ridge and pure lasso regression, and lambda, a parameter that specifies the amount of coefficient shrinkage. For random forest, the parameters included specifying the number of trees grown as well as the number of predictors to be sampled for splitting at each node. For stochastic gradient boosting, the parameters were the number of trees grown, the depth of the tree (i.e., the number of splits on the tree), the shrinkage (i.e., learning rate), and the number of minimum observations within a terminal node. Last, for support vector machines, linear and non-linear (e.g., polynomial) kernels were tested to assess which resulted in the highest level of performance. As the performance of the support vector machines was greatest with a polynomial kernel, the parameters included the degree of the polynomial used to identify the hyperplane that split the data and the cost of a misclassification. Further, as the majority of the sample in the current study engaged in recidivism, upsampling was used to account for the imbalance in the outcome data. Upsampling involves sampling with replacement from the minority class, which in the current study was those who did not go on to engage in recidivism.

3.9.5 Statistical Processing Techniques

The section below details the different alterations applied at different stages of the development of the statistical learning method algorithms to assess the impact on fairness and predictive validity. Pre- and post-processing techniques were utilised for each of the algorithms discussed above. The cross-cultural fairness and discrimination of the LS/RNR were reassessed after applying these alterations to the statistical learning methods to measure the impact these changes had.

3.9.5.1 Pre-Processing. Pre-processing refers to the alteration of the original data used for the algorithm to remove any potential causes of unfairness. For example, although culture is not explicitly used as a predictor in a risk instrument, often there are other predictors (e.g., criminal history) that correlate with cultural minority groupings (Berk, 2009; Skeem & Lowenkamp, 2016). Therefore, the pre-processing technique utilised in this thesis involved using the residual in place of the predictor variables that were found to be predicted by an individual's culture. Specifically, predictor variables were regressed on to culture (i.e., Aboriginal and Torres Strait Islander status) and the residual for each variable was utilised instead of the original predictor value. This approach helps to remove the association between an individual's culture and other predictor variables from the data before the algorithm is constructed (Berk, 2009). This also helps to reduce the influence an individual's culture may have on the predicted outcome.

3.9.5.2 Post-Processing. Post-processing refers to the alteration of the outcome data produced by the algorithm to remove any unfairness in the prediction. The post-processing technique utilised in this thesis involved reassigning the outcome classification (i.e., predicted to be a recidivist, not predicted to be a recidivist) to aid in achieving equality in the predicted outcome across groups. Specifically, this was achieved through a process known as reject option based classification (Kamiran et al., 2012). This process relabels observation outcomes that are deemed to be more uncertain. In other words, where the outcomes are binary (0 = predicted to be a recidivist, 1 = predicted to be a recidivist), an observation with a classification probability of .90 (therefore labelled as 'predicted to be a recidivist') is seen to be specified with a high degree of certainty. Those that fall closer to the cut-off value, for example, at the midpoint of .50 (e.g., a classification probability of .51 or .47) are seen to be classifications that have a higher degree of uncertainty and are influenced by biases (Kamiran et al., 2012). With reject option based classification, a critical region boundary (i.e., margin) is specified around the cut-off threshold, denoted by *θ,* which represents the maximum Euclidian distance to the cut-off that will result in relabelling. The observations that fall within this region, referred to as the critical region, have their labels reassigned specifically to aid in increasing parity among the outcome variables. In regard to the current study, the group that engaged in recidivism more had their observations that fell within the cut-off + θ reassigned to being predicted to not be a recidivist. Conversely, the group that engaged in recidivism less had their observations that fell within the cut-off - θ reassigned to being predicted to be a recidivist. The current study employed numerous θ values in order to identify the one that best balanced discrimination and fairness. The cut-off value was established by the same method previously identified for calculating fairness metrics; the closest point to 0, 1 in the ROC space.

3.9.6 Interpretable Statistical Learning Methods

As a number of the above statistical learning methods (and the transformations made to the statistical learning methods) can result in a lack of transparency in regard to the specific influence predictors have on the predicted outcome, Shapley Values were calculated to aid in increasing interpretability and an understanding of the importance variables had on the predicted outcome.

3.9.6.1 Shapley Values. Shapley values (Shapley, 1953), a concept based on game theory, focuses on the idea that a prediction can be fairly attributed to a group of features (Lundberg & Lee, 2017). In regards to an algorithm, the Shapley value for a feature (i.e., predictor or variable) is the mean marginal contribution of that feature across all possible groups of features (also referred to as coalitions) to the difference between the observed prediction and the average prediction made by the statistical learning method (Molnar, 2019).

One of the benefits of using the Shapley value as an approach for increasing the interpretability of statistical learning methods is that it is considered a fair way of distributing the outcome across the features. This is because a Shapley value satisfies the properties of efficiency, symmetry, null players, and additivity (Molnar, 2019; Peters, 2015). Efficiency is satisfied as the feature contributions (i.e., Shapley values for each feature) equal the difference between a predicted outcome for an observation and the average outcome when summed. Symmetry is satisfied as two features that contribute equally to all possible coalitions will have the same contribution value. A null player is satisfied as a feature that does not alter the predicted value (regardless of the coalition it is added to) has a contribution equal to 0. Last, additivity is satisfied as the approach ensures that for statistical learning methods with multiple predictions, such as random forests, a contribution for a feature can be calculated for each prediction made by a decision tree and then averaged to get the contribution of that feature across the random forest algorithm.

For the current thesis, Shapley values were calculated for the statistical learning methods that aided in increasing discrimination and/or fairness across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Specifically, the Shapley values were calculated for the features used within the statistical learning method to see the influence each feature had on the difference between the observed outcome for a specific observation and the average prediction for that statistical learning method. This was calculated across all observations within the sample, and then the average absolute contribution across observations for each feature was calculated. The absolute average Shapley value was calculated for the overall sample as well as for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. This enabled a comparison between groups to see if the contribution of predictors differed by group. For ease of reporting, the top five highest mean average Shapley values were reported for each statistical learning method.

3.10 Ethical Considerations

Ethics approval for this research was received from the Department of Justice and Community Safety (Victoria) Justice Human Research Ethics Committee (JHREC; Ref CF/18/17759) and Swinburne University Human Research Ethics Committee (SUHREC; SHR Project 2018/293) as part of a broader project aiming to assess the validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample. Support for the project was also received from Corrections Victoria and Victoria Police (see Appendix B for ethics approval and support letters). An ethics amendment was completed and approved with JHREC and SUHREC (see Appendix C), enabling access to the LS/RNR data that was collected as part of the broader research project that was specifically required for this particular thesis. A further ethics amendment was also completed and approved by JHREC (see Appendix D) that enabled empirical studies two and three (Chapters Five and Six) to be completed.

Although the current research thesis was low risk due to the retrospective nature of the data that required no contact or recruitment of participants, several ethical considerations were still considered. Specifically, due to the impracticality of obtaining consent from each participant, this project involved the release of sensitive information about a sample of individuals for which informed individual consent was not obtained. Access to this sensitive information without informed consent was enabled through an exemption that was consistent with Section 2.3.10 of the National Statement on Ethical Conduct in Human Research (National Health and Medical Research Council, 2007).

Further, to ensure that a breach of privacy did not occur and the protection of sensitive and identifying information, strong protections and procedures were employed in order to adhere to ethical guidelines as stipulated under the Information Privacy Principles of the

Privacy and Data Protection Act 2014 Vic) and the National Health and Medical Research Council (NHMRC; National Health and Medical Research Council, 2007). All information that was collected about participants was kept confidential, and identifiable information was stored separately from the final database that was used for the current research project. The information was de-identified prior to gaining access to the database, and each participant was assigned a unique identifier (i.e., a unique numerical identifier which was used instead of personal identifying information). The master participant list that initially contained identifying information about the individuals in the sample and that was utilised in order to collect data and link multiple sources of data was destroyed once the final dataset was collated. Further, the final data set that contained de-identified information was stored in an electronic database that was password protected and stored on password-protected computers. Only researchers named in the ethics approval from the Centre for Forensic Behavioural Science could obtain the passwords necessary to access the electronic database. Additionally, any identifying information was not published or reported on in this thesis. To further protect the confidentiality of the individuals in the sample, only aggregate information was used when discussing results and publishing findings.

Chapter Four: Empirical Study One

4.1 Introduction

Although previous research has established relatively comparable levels of discrimination across cultural groups on forensic risk assessment instruments, there is limited research exploring other statistical definitions of fairness cross-culturally. The limited studies that have explored these other definitions of fairness have often noted disparities between cultural groups that could be further disadvantaging certain cultural minority groupings within the criminal justice system. Yet, in Australia, there remains a scarcity of research examining these fairness definitions and the potential level of unfairness within forensic risk assessment instruments across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

This chapter presents the first empirical study that addresses the first research aim by responding to research questions one and two. Specifically, this paper assesses the discrimination of the LS/RNR with a sample of male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia to see if discrimination was similar or differed by group (research question one). Further, this paper assesses the fairness of the LS/RNR across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders (research question two) by measuring a number of common fairness definitions (i.e., error rate balance, calibration, predictive parity, and statistical parity) that have emerged from a variety of disciplines, including computer science and statistics. Last, this paper also explores the trade-offs that exist among certain fairness definitions and the implications this has for increasing cross-cultural fairness.

Empirical Study One is titled *"The Cross-Cultural Fairness of the LS/RNR: An Australian Analysis"* and is the author's original manuscript of an article that has been published by the American Psychological Association in *Law and Human Behavior. Law and Human Behavior* is a multidisciplinary peer-reviewed journal published by the American Psychological Association that investigates human behaviour and the law, as well as both the criminal justice and legal systems. As *Law and Human Behavior* is a journal based in North America, this paper has been written in American English. The Author Indication Form that details the contribution of each author to this manuscript is included in Appendix A.

The citation for the published version of this article is as follows:

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

The Cross-Cultural Fairness of the LS/RNR: An Australian Analysis

Linda J. Ashford[1], Benjamin L. Spivak[1], James R. P. Ogloff[1] & Stephane M. Shepherd[1]

[1] Centre for Forensic Behavioural Science, Swinburne University of Technology

Author Note.

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

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. We 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.

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

Abstract

Objective: Cross-cultural research into forensic risk assessment instruments has often identified relatively comparable levels of discrimination. However, cross-cultural fairness is rarely addressed. Therefore, this study explored the discrimination and fairness of the Level of Service/Risk Needs Responsivity (LS/RNR) within a sample of Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander males convicted of a violent offense.

Hypotheses: It was hypothesized that Aboriginal and Torres Strait Islanders would have comparable discrimination to non-Aboriginal and Torres Strait Islanders. It was further hypothesized that some fairness definitions would not be satisfied between these two groups.

Method: The present study included 380 males (Aboriginal and Torres Strait Islander *n* = 180; non-Aboriginal and Torres Strait Islander *n* = 200) from Australia. Discrimination was assessed using the area under the curve (AUC) and the cross area under the curve (xAUC). Error rate balance, calibration, predictive parity, and statistical parity were used to determine fairness.

Results: The discrimination of the LS/RNR was relatively commensurate for Aboriginal and Torres Strait Islanders (AUC = .60, 95% CI [.49, .70]) and non-Aboriginal and Torres Strait Islanders (AUC = .63, 95% CI [.55, .72]). The xAUC identified notable disparities with the LS/RNR being unable to discriminate between Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists (xAUC = .46, 95% CI [.35, .57]). Disparities among certain fairness definitions were also identified, with Aboriginal and Torres Strait Islanders scoring higher on the LS/RNR (*d* = 0.52), and non-recidivists being classified as high risk more often.

Conclusions: The findings suggest that the LS/RNR may not be a cross-culturally fair risk assessment instrument for Australian individuals convicted of a violent offense and that comparable levels on more standard discrimination indices do not imply that a risk assessment instrument is cross-culturally fair.

Keywords: cross-cultural fairness, forensic risk assessment, recidivism, LS/RNR

Public Significance Statement

This study identified performance disparities between Aboriginal and Torres Strait Islander peoples and non-Aboriginal and Torres Strait Islander Australians on the LS/RNR. Although the instrument was similarly effective in discriminating recidivists from non-recidivists for both groups, other disparities, such as false positive and false negative rate differences, suggest that the instrument may not assess risk fairly. Future research should aim to reduce these disparities to improve the cross-cultural fairness of the LS/RNR.

Introduction

Forensic risk assessment instruments are currently employed in numerous countries to assess an individual's risk of recidivism. These instruments are used across the criminal justice system to assist in identifying needs for treatment and rehabilitation (Monahan & Skeem, 2016). Risk assessment instruments were developed after approaches relying on unstructured clinical judgements were found to be unreliable (Grove & Meehl, 1996). Two forms of risk assessment instruments have become increasingly prominent. Actuarial risk assessment instruments involve an algorithmic approach to assessing forensic risk. Factors deemed to be empirically related to risk are summed together to arrive at a probabilistic risk estimate (Doyle & Dolan, 2002). Structured professional judgement (SPJ) instruments were developed in response to critiques of actuarial instruments and their nomothetic and rigid approach. SPJ instruments provide clinicians with guidelines surrounding evidence-based risk factors to assist in determining an individual's level of risk (Hart et al., 2017).

The utility of actuarial and SPJ instruments is most often assessed through the area under the curve (AUC), which measures discrimination (Singh, 2013). This refers to a risk assessment instrument's ability to distinguish between those who go on to engage in recidivism and those who do not. AUC values are also frequently compared to see if risk assessment instruments perform similarly across different groups. Research has indicated that these instruments commonly perform well for various cultural groups, with generally equivalent levels of discrimination identified for some cultural minority groups (e.g., African Americans and Indigenous populations of North America and Australia) and predominately White cultural groupings (e.g., Muir et al., 2020; Skeem & Lowenkamp, 2016; Wormith et al., 2015).

However, comparable discrimination (e.g., AUC values) is not the only way to indicate that a risk assessment instrument is equally effective or cross-culturally fair. AUC comparisons among cultural groups are also limited in how they observe cultural groups distinctly and can be understood as a within-group comparison (i.e., is the risk assessment instrument discriminating recidivists from non-recidivists within that cultural group?). This again demonstrates that comparing AUCs obtained from distinct cultural groups does not necessarily imply fairness or unfairness.

As a result, Kallus and Zhou (2019) established the cross AUC (xAUC), which offers a more informed way of demonstrating disparity between cultural groups. The xAUC measures the probability of recidivists from one cultural group receiving a higher risk score than non-recidivists from another cultural group. However, research often reports the AUC in isolation as a measure of discrimination (Singh et al., 2013), neglecting other measures of discrimination and/or measures of calibration that provide a more complete picture of a risk assessment instrument's utility and fairness. Further, the xAUC is yet to be applied within the forensic psychology discipline as a measure to identify discrimination disparities between cultural groups.

Fairness, or the equal treatment across groups by risk assessment instruments, is an ongoing area of contention, with critics arguing that the use of these instruments (primarily actuarial instruments) will lead to the over-criminalization of already disadvantaged cultural minorities (Day et al., 2018; Hart, 2016). Some cultural minorities are already reported to experience inequality within the criminal justice system, including higher chances of being denied parole, higher arrest rates, and an over-representation in prison (Australian Bureau of Statistics, 2018a; Dragomir & Tadros, 2020; Martel et al., 2011; Shepherd, Adams, et al., 2014).

An unfair risk assessment instrument could contribute to this inequality, with risk assessment-based decisions potentially having negative consequences for particular cultural groups. Fairness has also become a developing area of interest across numerous academic disciplines, including computer science, statistics, and criminology (Verma & Rubin, 2018). This has resulted in a more complex and sophisticated understanding of what can constitute fairness, with multiple definitions that can enlighten considerations of cross-cultural fairness within risk instruments beyond comparisons of AUCs.

Fairness

Fairness has numerous definitions, including error rate balance, calibration, predictive parity, and statistical parity. These fairness definitions are outlined in Table 1.

It is also worth noting that an impossibility theorem exists across fairness notions. Specifically, when the base rates of recidivism differ across groups, certain fairness definitions are incompatible. Base rates are found to differ across cultural groups, with cultural minorities often having higher base rates of recidivism (Flores et al., 2016; Shepherd & Strand, 2016; Wilson & Gutierrez, 2014). This disparity has been attributed to the discrimination within the criminal justice system and the social and economic disadvantages that these groups face (Day et al., 2018; Hart, 2016). However, it has been demonstrated that error rate balance and predictive parity cannot be simultaneously achieved with differing base rates (Berk et al., 2018; Chouldechova, 2017).

Table 1

Fairness Definitions

Fairness	Definition	Reference
Error Rate Balance	Equal false negative rates (FNR; the proportion of recidivists who were inaccurately predicted to not be recidivists/labelled low risk) and false positive rates (FPR; the proportion of non-recidivists who were inaccurately predicted to be recidivists/labelled high risk) across groups.	Chouldechova (2017)
Calibration	A risk score or risk classification has the same proportion of recidivists for each group (e.g., 50% of high risk classifications engage in recidivism in each group).	Chouldechova (2017)
Predictive Parity	Positive predictive values (PPV; the proportion of those predicted to be recidivists/labeled high risk who engage in recidivism) and negative predictive values (NPV; the proportion of those not predicted to be recidivists/labeled low risk who do not engage in recidivism) are equal across groups.	Berk et al. (2018)
Statistical Parity	Equal risk score distribution across groups.	Berk et al. (2018)

Regardless, the majority of these definitions of fairness (besides statistical parity) have not been applied in certain disciplines (e.g., forensic psychology), nor are they often discussed in the risk assessment literature. Numerous studies that have explored the statistical parity of risk assessment instruments have identified that cultural minorities score significantly higher on risk scores (e.g., Olver et al., 2014; Wilson & Gutierrez, 2014). However, statistical parity has been critiqued as a form of fairness (Dwork et al., 2012). Statistical parity does not consider the outcome of recidivism and coercing risk classifications to be equivalent could adversely impact other definitions of fairness.

A recent review observing the cross-cultural fairness of risk assessment instruments noted that most other fairness criteria, such as calibration, error rate balance, and predictive parity, are rarely addressed (Ashford et al., 2021). The few studies exploring these other definitions of fairness have also often identified disparities between cultural groups. For example, in studies considering predictive parity, cultural minorities who were classified as high risk were more likely to engage in recidivism (higher PPV) than high risk White individuals, whereas among low risk classifications, White individuals were more likely to not engage in recidivism (Flores et al., 2016; Muir et al., 2020). Further, research by ProPublica demonstrated differences in error rate balance such that African American individuals were almost twice as likely to be classified as high risk and not engage in recidivism (i.e., a higher FPR), whereas White individuals were almost twice as likely to be classified as low risk and later engage in recidivism (i.e., a higher FNR; Angwin et al., 2016). However, Flores et al. (2016) were able to demonstrate with the same sample that calibration was satisfied. For example, calibration differences have been reported between Indigenous (Métis, Inuit, and other First Nation peoples) and non-Indigenous individuals from Canada, with Indigenous groups being predicted to engage in recidivism more (Wilson & Gutierrez, 2014; Wormith et al., 2015).

Cross-Cultural Fairness in Australia

Discrimination

In Australia, the discrimination (i.e., AUCs) of risk assessment instruments has been demonstrated to be lower for Aboriginal and Torres Strait Islander groups when compared to non-Aboriginal and Torres Strait Islander groups (Allan et al., 2006; Thompson & McGrath, 2012; Watkins, 2011). However, these disparities are often not statistically significant (e.g., AUC differences between .01 and .09). Marginally higher AUCs for Aboriginal and Torres Strait Islanders compared to non-Aboriginal and Torres Strait Islanders have also been identified in other research (Shepherd, Luebbers, et al., 2014). However, some Australian studies have identified more substantial disparities, with a notably higher AUC for non-Aboriginal and Torres Strait Islanders and/or risk assessment instruments being unable to effectively discriminate between recidivists and non-recidivists (on certain recidivism outcomes) for Aboriginal and Torres Strait Islanders (Shepherd et al., 2015; Shepherd & Strand, 2016; Smallbone & Rallings, 2013).

Fairness

An unfair risk assessment instrument could exacerbate the disparities already faced by Aboriginal and Torres Strait Islander peoples within the criminal justice system. The Aboriginal and Torres Strait Islander population account for 29% of the total prison population despite comprising only 3.3% of the Australian population (Australian Bureau of Statistics, 2018b, 2020c). Unfair risk assessment instruments could amplify further inequalities, such as incorrect classifications impacting legal decisions as well as inappropriate treatment and management plans that could impact effective rehabilitation. Yet, there is a scarcity of risk assessment studies comparing Aboriginal and Torres Strait Islanders with non-Aboriginal and Torres Strait Islanders across multiple fairness definitions.

The cross-cultural fairness of risk assessment instruments in Australia has most often been assessed by examining statistical parity. Applying this definition of fairness, Aboriginal and Torres Strait Islanders have been found to score higher on a variety of risk assessment instruments when compared to non-Aboriginal and Torres Strait Islander groups (Shepherd, Luebbers, et al., 2014; Shepherd et al., 2015; Thompson & McGrath, 2012). These differences ranged from small to medium in effect size, with adult Aboriginal and Torres Strait Islanders found to score notably higher on risk assessment instruments such as the Level of Service Inventory-Revised (LSI-R; Hsu et al., 2010; Watkins, 2011) and Static-99R (Smallbone & Rallings, 2013). One study examining the Psychopathy Checklist: Youth Version (PCL: YV) reported a slightly lower risk score for Aboriginal and Torres Strait Islander youth compared to those with an English speaking background; however, this difference was inconsequential in effect size, $d = 0.11$ (Shepherd & Strand, 2016).

The remaining definitions of fairness have rarely been explored, both generally and amongst Australians. Two studies have examined the predictive parity of the PCL: YV (Shepherd & Strand, 2016) and the Youth Level of Service/Case Management Inventory (YLS/CMI; Shepherd et al., 2015) for young people. Both studies identified that PPVs were higher for Aboriginal and Torres Strait Islander youth, signifying that a high risk classification had a higher proportion of recidivists from the Aboriginal and Torres Strait Islander group. Conversely, NPVs were higher for non-Aboriginal and Torres Strait Islander youth, indicating that a low risk classification had a higher proportion of non-recidivists from the non-Aboriginal and Torres Strait Islander group. Thompson and McGrath (2012) reported issues in calibration on the YLS/CMI-AA (Australian Adaptation), with Aboriginal and Torres Strait Islanders engaging in recidivism at the highest rate across low, medium, and high risk classifications when compared to non-Aboriginal and Torres Strait Islander groups.

Current Study

In Australia, the limited research exploring the cross-cultural fairness of risk assessment instruments has revealed inequalities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, but this area remains understudied. Few studies have explored multiple statistical definitions of fairness. To address this gap in the literature, the present study examined the discrimination and cross-cultural fairness of the actuarial risk assessment instrument, the Level of Services/Risk Needs Responsivity (LS/RNR). Specifically, the study aimed to assess if there were differences in both discrimination and cross-cultural fairness (based on the statistical definitions outlined earlier) for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in Victoria, Australia. It was hypothesized that: (a) comparable levels of discrimination as assessed by the AUC would be identified across both groups; and (b) numerous fairness definitions would not be satisfied for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

Method

Sample

The sample included 380 males in Victoria, Australia who had been sentenced to a period of incarceration for committing a serious violent offense, as defined in schedule 1 (clause 3) of the *Sentencing Act* 1991 (Vic) and were received into prison between January 2015 and December 2017. They were assessed with the LS/RNR by a corrections officer while serving either a prison sentence (*n* = 231, 60.79%), a community corrections order (*n* = 148, 38.95%), or a parole order (*n* = 1, 0.26 %). The sample included 180 (47.37%) people who identified as Aboriginal and/or Torres Strait Islander people and 200 (52.63%) who were non-Aboriginal and Torres Strait Islanders. Aboriginal and Torres Strait Islander peoples account for approximately 9% of the adult prison population in Victoria (Australian Bureau of Statistics, 2018a). However, for the present study, all eligible Aboriginal and Torres Strait Islanders previously incarcerated for a serious violent offense within the study period were sampled to enable comparisons between groups. The non-Aboriginal and Torres Strait Islanders were then randomly sampled to obtain roughly equivalent group numbers.

The non-Aboriginal and Torres Strait Islander group was unable to be classified into more distinct groups, with the majority being born in Australia (*n* = 166, 83%) and all identifying their primary language as English. Completions of the LS/RNR and dates of incarceration, community correction orders, and parole orders were obtained from Corrections Victoria. Recidivism data (i.e., new police charges) was obtained from the Victorian Police Law Enforcement Assistance Program (LEAP) database for the period of January 2015 to December 2019. Ethical approval for the present study was obtained from the Department of Justice and Community Safety (Victoria) Justice and Human Research Ethics Committee and the Swinburne University Human Research Ethics Committee.

Measures

Levels of Service/Risk Needs Responsivity

The LS/RNR (Andrews et al., 2008) is an actuarial instrument designed to estimate general recidivism risk and identify criminogenic needs. It also provides structure in the management and treatment planning of prisoners. The General Risk/Needs section comprises eight factors that are scored using 43 items – Criminal History (8 items), Education/Employment (9 items), Family/Marital (4 items), Leisure/Recreation (2 items), Companions (4 items), Alcohol/Drug Problem (8 items), Procriminal Attitude (4 items) and Antisocial Pattern (4 items). Each item is scored 0 when absent and 1 when present, and items are summed to create respective factor scores and a total General Risk/Needs score. Using the total score, individuals can be categorized into various risk levels that include very low risk (0–4), low risk (5–10), medium risk (11–19), high risk (20–29) and very high risk (30–43).

Recidivism Outcome

Recidivism was defined as a police charge for any offense while at risk in the community. The average time to first charge was 184.75 days (*SD* = 233.80).

Analytical Approach

All data were analyzed through RStudio using R version 4.0.2 (R Core Team, 2021). A suite of packages was utilized including the *tidyverse* packages (Version 1.3.0; Wickham, 2019) for data cleaning and management, *rms* (Version 6.0-1; Harrell, 2020) for logistic regression, *pROC* (Version 1.16.2; Robin et al., 2020) to generate receiver operating characteristic (ROC) curves and AUC values, and *survival* (Version 3.2-7; Therneau, 2020*)* and *survminer* (Version 0.4.8; Kassambara et al., 2020) for survival analysis.

Survival Analysis

Cox regression analyses (Cox, 1972) were used to estimate the predictive validity of the LS/RNR risk score while accounting for individual differences in time at risk to the community and also observing the impact Aboriginal and Torres Strait Islander status has on recidivism. A "time at risk" (i.e., survival time) variable was created that started at the LS/RNR assessment date for those not incarcerated (those on community correction orders or parole orders) or date of release from prison for those who were incarcerated. The end date was either the date of the first offense for recidivists or the end of follow up data (31-12-2019) for non-recidivists. Any other days of incarceration that fell within the time at risk were excluded from the total number of days at risk of recidivism. The average time at risk to the community (i.e., follow up time) was 280.56 days (*SD* = 329.24). Cox regression also produced hazard ratios (e^B) that represent the increase in the hazard of recidivism for a 1 unit increase in the predictor variables.

Discrimination Indices

Area under the Curve. The receiver operating characteristic (ROC) curve plots the sensitivity against 1 – specificity at various thresholds and is unimpeded by differing base rates (Singh, 2013). The AUC can be understood as the probability that a randomly selected individual who engages in recidivism receives a higher risk score than a randomly selected individual who does not. The AUC value can range from 0 to 1, with the midpoint (.50) demonstrating discrimination at chance levels. There are varying benchmarks of what can constitute a small, medium, or large effect size for the AUC value. For the present study, the following cut-points, which have been commonly adopted in previous forensic psychology and criminology research, were applied: values between .56-.63 indicated a small effect, .64-.70 a medium effect, and .71 and above as a large effect (Rice & Harris, 2005).

Cross Area under the Curve. The xAUC (Kallus & Zhou, 2019) is an alteration of the traditional AUC that measures discrimination between groups to better identify disparities. A cross receiver operating characteristic (xROC) plots the sensitivity against 1 – specificity at various thresholds for two sets of groups for which a xAUC can be calculated. The first set contains a positive outcome (i.e., non-recidivist) from one group (i.e., Aboriginal and Torres Strait Islander) and a negative outcome (i.e., recidivists) from another group (i.e., non-Aboriginal and Torres Strait Islander). The second set is the opposite of the first (i.e., Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists). The xAUC measures the probability that a randomly selected individual who engaged in recidivism from one group received a higher risk score than a randomly selected individual from the other group who did not engage in recidivism.

Fairness

Error rate balance and predictive parity were calculated using the information contained within a confusion matrix. For the fairness definitions that can be calculated using a confusion matrix, a cut-off point is required in risk scores to distinguish between those predicted to engage in recidivism (i.e., those determined as high risk or above that cut-off value) and those not predicted to engage in recidivism (i.e., those determined as low risk or below the specified cut-off value). Previous research has demonstrated that differing cut-off scores result in variations in fairness values as the proportions in low risk and high risk classifications change (see Flores et al., 2016). Therefore, for the present study, predictive parity and error rate balance were computed and reported across varying cut-off thresholds for comparison.

Calibration was observed by initially comparing the proportions of recidivists from each group by risk classification. Calibration was then further assessed by examining if Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders systematically deviated from a similar regression line (i.e., slope and intercept differences) that relates to risk assessment scores and recidivism. Specifically, four bivariate logistic regression models were conducted. In the first model, Aboriginal and Torres Strait Islander status was used to predict recidivism. In the second model, the LS/RNR total risk score was used to predict recidivism. The third model incorporated both Aboriginal and Torres Strait Islander status and the LS/RNR risk score to predict recidivism. The fourth model included Aboriginal and Torres Strait Islander status, LS/RNR risk score, and an interaction between Aboriginal and Torres Strait Islander status and LS/RNR risk score to predict recidivism. Differences in intercept were determined when Aboriginal and Torres Strait Islander status in model three added significant incremental utility to the LS/RNR risk score in predicting recidivism. Differences in slope were determined when Aboriginal and Torres Strait Islander status significantly moderated the utility of the LS/RNR risk score in predicting recidivism in model four. To account for the varying times in follow up, data was constrained to the maximum possible follow up time of six months for the calibration analysis to enable the bivariate logistic regression models to be performed on the full sample. The rest of the results are in relation to the full follow-up period.

Last, the distribution of risk scores was observed and compared to measure statistical parity. A comparison of mean scores on the General Risk/Needs section of LS/RNR for

Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders was also conducted using an independent samples *t*-test with Cohen's *d* reported as a measure of effect size.

Results

Descriptive Statistics

The average age at LS/RNR assessment was 32.17 years (*SD* = 8.69). Moreover, Aboriginal and Torres Strait Islanders were younger (*M* = 31.29, *SD* = 8.38) than non-Aboriginal and Torres Strait Islanders (*M* = 32.96, *SD* = 8.92). However, this difference was not significant and small in effect size, *t* (378) = -1.865, *p* = .063, Cohen's *d* = .19. Overall, 306 (80.53%) individuals engaged in recidivism by the end of the full follow up period. For Aboriginal and Torres Strait Islanders, 154 (85.56%) engaged in recidivism compared to 152 (76%) non-Aboriginal and Torres Strait Islanders. This difference was statistically significant, yet small in effect, χ^2 (1) = 4.924, *p* = .026, Cramer's *V* = .11.

Survival Analysis

A Cox regression analysis was performed to assess the predictive validity of the LS/RNR total risk score and Aboriginal and Torres Strait Islander status to predict recidivism while accounting for varying times at risk to the community. The hazard ratio (HR) for Aboriginal and Torres Strait Islander status suggests that being Aboriginal and Torres Strait Islander was associated with a 14.5% increased risk of recidivism; however, Aboriginal and Torres Strait Islander status was not a statistically significant predictor of recidivism, HR = 1.145, 95% CI [0.903, 1.452], *p* = .263. The LS/RNR risk score was found to be a significant predictor of recidivism, with a one unit increase in risk score being associated with a slightly higher increased risk of recidivism, HR = 1.043, 95% CI [1.027, 1.061], *p* < .001).

For a visual representation of Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander survival curves, please refer to the supplementary materials.

Discrimination Indices

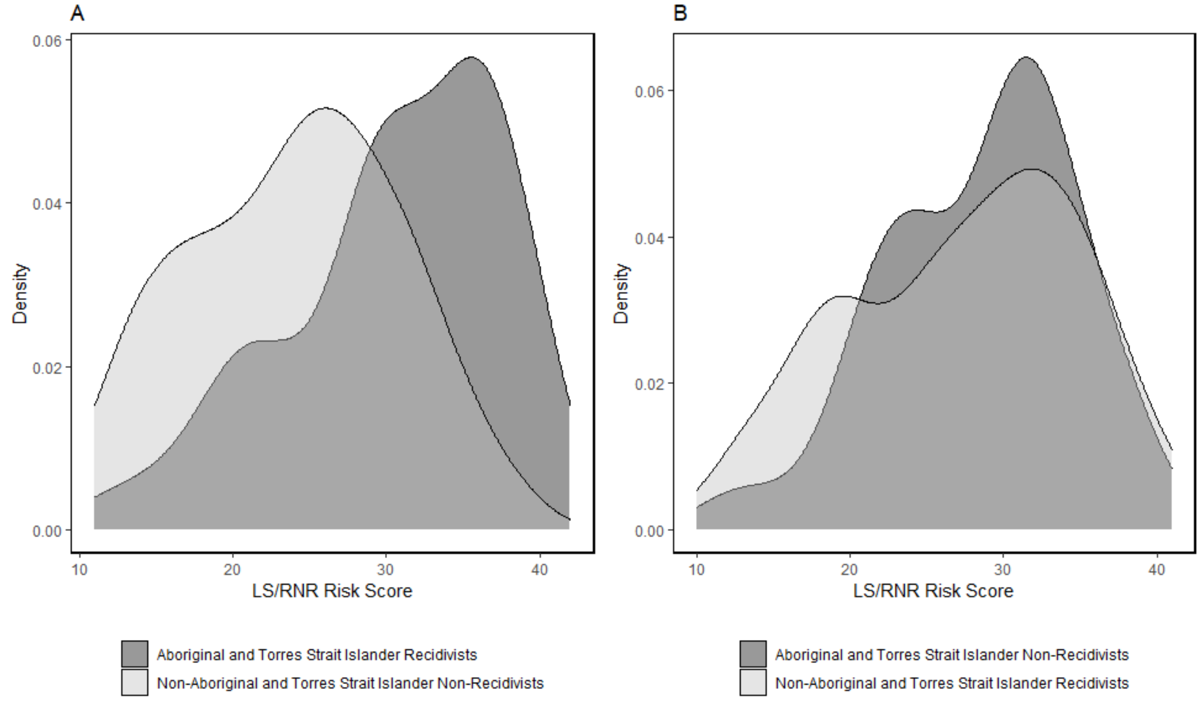
Area under the Curve and Cross Area under the Curve

Overall, the LS/RNR total risk score was found to have moderate discrimination, AUC = .64, 95% CI [.57, .70]. For Aboriginal and Torres Strait Islanders, the AUC was .60, 95% CI [.49, .70], which was slightly lower than the non-Aboriginal and Torres Strait Islanders, whose AUC was .63, 95% CI [.55, .72]. The xAUC identified that the LS/RNR could not effectively discriminate Aboriginal and Torres Strait Islander non-recidivists from non-Aboriginal and Torres Strait Islander recidivists, xAUC = .46, 95% CI [.35, .57]. The LS/RNR was a better discriminator when comparing non-Aboriginal and Torres Strait Islander non-recidivists with Aboriginal and Torres Strait Islander recidivists, with the xAUC being large in effect size, xAUC = .75, 95% CI [.68, .83].

The distribution of risk scores for Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists, as well as for non-Aboriginal and Torres Strait Islander non-recidivists and Aboriginal and Torres Strait Islander recidivists, was plotted using density plots and is presented in Figure 1.

Figure 1

Score Distributions for Between Group Comparisons



Note. Panel A: Density plot of LS/RNR risk scores for Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists. Panel B: Density plot of LS/RNR risk scores for non-Aboriginal and Torres Strait Islander recidivists and Aboriginal and Torres Strait Islander non-recidivists.

Aboriginal and Torres Strait Islander recidivists had higher risk scores compared to non-Aboriginal and Torres Strait Islander non-recidivists. However, there was a significant overlap in risk score distributions between Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists.

Fairness

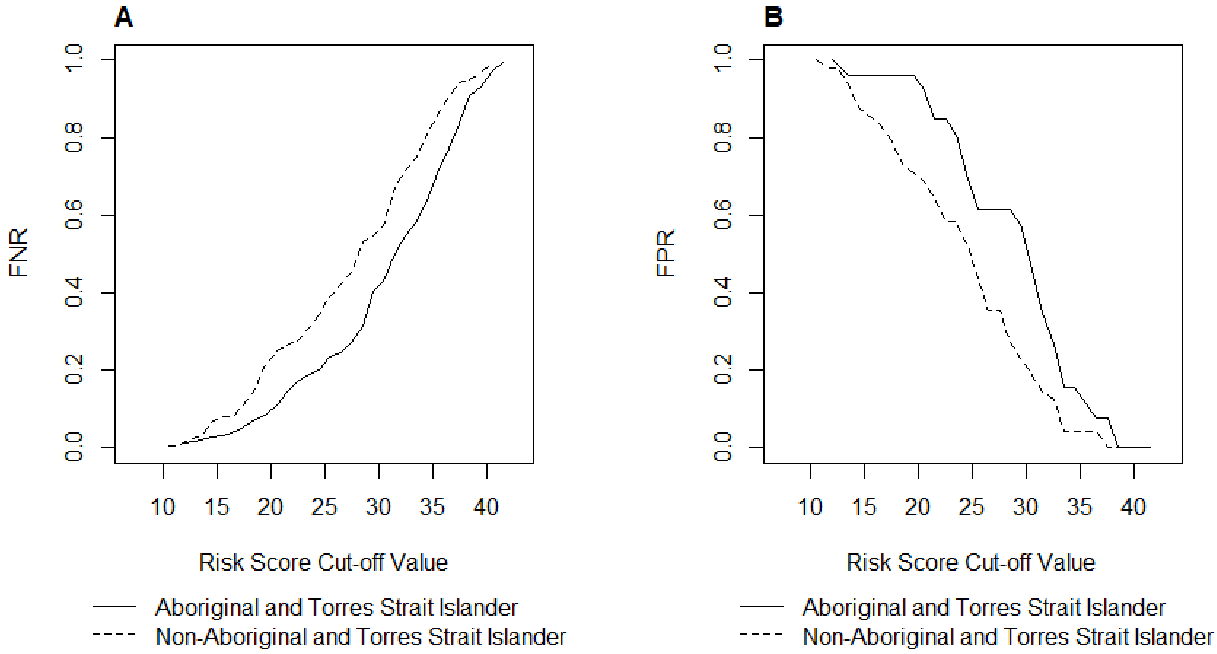
Error Rate Balance

Error rate balance across groups on the LS/RNR was examined by plotting the FNR and FPR for both groups across varying cut-off thresholds. Figure 2 presents the FNRs and FPRs for both groups. The non-Aboriginal and Torres Strait Islander group were found to have a higher FNR across all cut-off thresholds. This indicates that non-Aboriginal and Torres Strait Islanders were more likely to be classified as low risk and later go on to engage in recidivism. These differences were most pronounced among the high risk scores (20-29). At a cut-off value of 20, for example, 25% of non-Aboriginal and Torres Strait Islander groups were classified as low risk and engaged in recidivism, compared to 11.04% of Aboriginal and Torres Strait Islander groups. Across all risk score cut-offs, the non-Aboriginal and Torres Strait Islander group were on average 1.58 times more likely to be classified as low risk and later engage in recidivism.

Across all cut-off thresholds, the Aboriginal and Torres Strait Islander group had higher FPRs, signifying that a higher proportion of Aboriginal and Torres Strait Islanders were classified as high risk and did not engage in recidivism by the end of the follow up period. These disparities were most pronounced among the medium (11-19) and high risk scores (20-29). For example, at a cut-off value of 29, Aboriginal and Torres Strait Islanders were more than twice as likely to be labeled as high risk and not go on to engage in recidivism compared to non-Aboriginal and Torres Strait Islanders (FPRs = .62 and .27 respectively). Across all risk score cut-offs, the Aboriginal and Torres Strait Islander group were on average 1.84 times more likely to be classified as high and not engage in recidivism.

Figure 2

False Negative Rates and False Positive Rates across LS/RNR Risk Scores



Note. Panel A: False negative rates for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders across LS/RNR risk scores. Panel B: False positive rates for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders across LS/RNR risk scores.

Calibration

Calibration was first assessed by examining if the number of individuals from each group that engaged in recidivism differed by their risk classification. No individuals were classified as very low risk. Further, only one non-Aboriginal and Torres Strait Islander was classified as low risk, who subsequently engaged in recidivism. Aboriginal and Torres Strait Islanders engaged in recidivism at a higher rate than non-Aboriginal and Torres Strait Islanders across both medium (92.86% and 68.89%, respectively) and high risk (83.05% and 68.92%, respectively) classifications. Across the very high risk classification, similar proportions of

Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders engaged in recidivism (85.98% and 86.25% respectively).

Calibration was further assessed by ascertaining if Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders had similar regression lines. Four logistic regressions were performed on the full sample using a fixed six month follow up to account for the varying follow up times and are presented in Table 2.

The slope of the relationship between the LS/RNR risk score and recidivism was not significantly different for Aboriginal and Torres Strait Islanders than for non-Aboriginal and Torres Strait Islanders. When models three and four were compared, it was found that adding an interaction term between the LS/RNR risk score and Aboriginal and Torres Strait Islander status did not improve recidivism prediction, χ^2 (1) = 2.015, p = .156, Pseudo-R^2 Δ = .007. This was further reflected in the small odds ratios for the interaction term, which was not statistically significant.

Further, there were also no significant differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in the intercept of the relationship between the LS/RNR risk score and recidivism. Comparisons between models two and three demonstrated that adding Aboriginal and Torres Strait Islander status did not significantly increase incremental utility to the LS/RNR in predicting recidivism, χ^2 (1) = 0.029, p = .864, Pseudo-R^2 Δ = .0001.

Table 2

Odds Ratios from Logistic Regression Models

	Model 1	Model 2	Model 3	Model 4
Aboriginal and Torres Strait Islander Status	1.34	--	1.04	3.58
Risk Score	--	1.08***	1.07***	1.10***
Risk Score X Aboriginal and Torres Strait Islander Status	--	--	--	0.96
Constant	1.02	0.15***	0.15***	0.09***
χ^2	2.02	25.92***	25.95***	27.96***
Pseudo-R^2	.01	.09	.09	.09

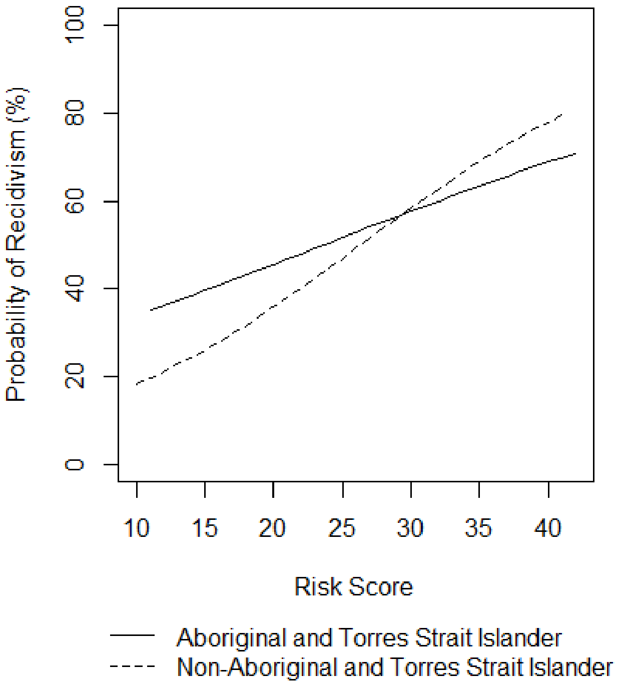
Note. -- = this variable was not included in the model.

**p* < .05. ***p* < .01. ****p* < .001.

The predicted probability of recidivism was estimated using model four, and these probabilities were grouped by risk scores and are presented in Figure 3. This demonstrated that although the LS/RNR was well calibrated, Aboriginal and Torres Strait Islanders had higher predicted probabilities of recidivism and were under-classified across the possible lower risk classifications for this sample (medium and high risk). The reverse was identified across very high risk scores, with non-Aboriginal and Torres Strait Islanders having higher predicted probabilities of recidivism.

Figure 3

Predicted Probabilty of Recidivism across LS/RNR Risk Scores



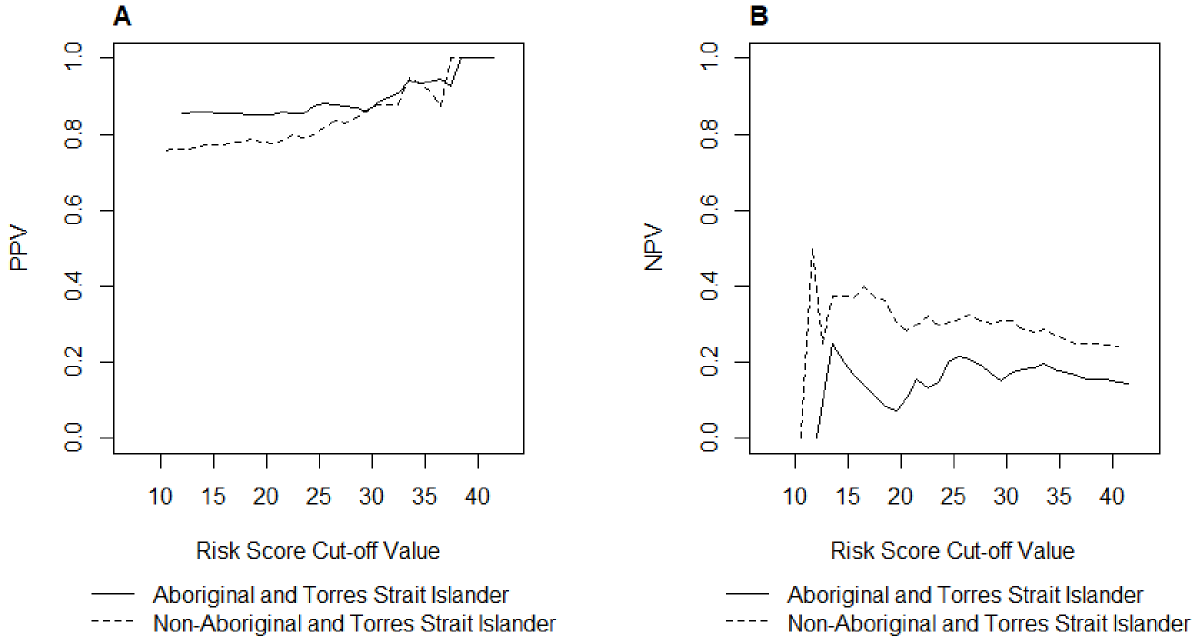
Predictive Parity

Predictive parity was determined by calculating both PPV and NPV for both groups across a range of cut-off thresholds. The PPVs and NPVs for both groups are presented in Figure 4. Aboriginal and Torres Strait Islanders were found to have slightly higher PPVs across the majority of risk scores (for which PPV could be calculated). PPVs were relatively comparable across the very high risk scores, with the majority of the disparity being identified across medium risk scores. PPVs for both groups were also relatively high, indicating that a high risk classification was often associated with an individual who went on to engage in recidivism, regardless of Aboriginal and Torres Strait Islander status.

Across all cut-off thresholds, the non-Aboriginal and Torres Strait Islander group were found to have higher NPVs. The greatest disparities were identified among the higher end of medium risk scores. For example, at a cut-off of 19, 36.11% of non-Aboriginal and Torres Strait Islanders who were classified as low risk did not go on to engage in recidivism compared to 8.33% of Aboriginal and Torres Strait Islanders. Due to the high base rates of recidivism, NPVs were relatively low for both groups, indicating that a low risk classification was often associated with recidivism.

Figure 4

Positive Predictive Values and Negative Predictive Values across LS/RNR Risk Scores



Note. Panel A: Positive predictive values for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders across LS/RNR risk scores. Panel B: Negative predictive values for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders across LS/RNR risk scores.

Statistical Parity

To establish the level of statistical parity, the LS/RNR total risk score and eight risk domains were examined for significant mean differences between groups. The results of independent sample *t*-tests and their respective effect sizes are reported in Table 3. Aboriginal and Torres Strait Islanders were found to score significantly higher on the LS/RNR total risk score as well as six of the eight risk domains – Criminal History, Education/Employment, Family/Marital, Companions, Alcohol/Drug Problems, and Antisocial Pattern. The effect sizes of these differences were generally small. Cohen (1988) notes that when the effect sizes are small, group mean differences are negligible, even if the differences are statistically significant.

The proportion of risk classifications was also compared. No individuals were classified as very low risk. Further, no Aboriginal and Torres Strait Islanders and only one non-Aboriginal and Torres Strait Islander (0.50%) were classified as low risk. The other classifications were found to differ between groups, with a smaller proportion of Aboriginal and Torres Strait Islanders compared to non-Aboriginal and Torres Strait Islanders being classified as medium risk (7.78% and 22.50% respectively) and high risk (32.78% and 37% respectively). Conversely, a higher proportion of Aboriginal and Torres Strait Islanders were classified as very high risk compared to non-Aboriginal and Torres Strait Islanders (59.44% and 40%, respectively). For a visual representation of the distribution of LS/RNR total risk scores, please refer to the supplementary materials.

Table 3

Risk Score Mean Differences

	Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander			
	M	*SD*	*M*	*SD*	*t*	Cohen's *d*
LS/RNR Total Risk Score	30.27	6.97	26.53	7.47	5.03***	0.52
Risk Domains						
1. Criminal History	6.39	1.57	5.88	1.80	2.98**	0.31
2. Education/Employment	6.46	2.24	5.40	2.37	4.45***	0.46
3. Family/Marital	2.45	1.16	2.07	1.32	3.00**	0.31
4. Leisure/Recreation	1.66	0.63	1.58	0.70	1.11	0.11
5. Companions	3.33	1.02	3.10	1.17	2.06*	0.21
6. Alcohol/Drug Problems	5.61	1.92	4.68	2.21	4.33***	0.44
7. Procriminal Attitudes	2.08	1.40	1.97	1.54	0.71	0.07
8. Antisocial Pattern	2.30	1.04	1.86	1.13	3.93***	0.40

*$p < .05$. **$p < .01$. ***$p < .001$.

Discussion

The present study compared the discrimination and assessed the fairness of the LS/RNR across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. Survival analyses demonstrated that the LS/RNR risk score was a significant predictor of recidivism. Aboriginal and Torres Strait Islanders had higher rates of recidivism, a finding consistent with previous research (e.g., Hsu et al., 2010; Shepherd &

Strand, 2016; Thompson & McGrath, 2012), and was associated with an increased risk of recidivism in survival analyses. However, this was not statistically significant.

Discrimination

As anticipated and similar to several studies previously conducted in Australia (e.g., Shepherd et al., 2015; Shepherd & Strand, 2016; Smallbone & Rallings, 2013; Thompson & McGrath, 2012), the discrimination as assessed by the AUC was comparable between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Furthermore, the AUC effect sizes in this study were small to moderate, a finding that, while consistent with previous Australian research that has examined other Level of Service instruments (Shepherd et al., 2015; Thompson & McGrath, 2012), suggests that the LS/RNR is not a highly effective risk assessment instrument in discriminating recidivists from non-recidivists in a sample of individuals previously convicted of a violent offense.

The xAUCs, however, identified notable disparities. Specifically, the probability that a non-Aboriginal and Torres Strait Islander recidivist received a higher risk score than an Aboriginal and Torres Strait Islander non-recidivist was below chance levels. However, Aboriginal and Torres Strait Islander recidivists were more likely to receive a higher risk score compared to non-Aboriginal and Torres Strait Islander non-recidivists. Therefore, when comparing between groups and not within (i.e., the AUC), the ability of the LS/RNR to effectively discriminate between those who do and do not engage in recidivism is not comparable for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Specifically, Aboriginal and Torres Strait Islander peoples were more likely to have received a higher risk score, regardless of whether they engaged in recidivism.

Cross-Cultural Fairness

Error Rate Balance

Disparities were also identified between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on both FPRs and FNRs. Specifically, Aboriginal and Torres Strait Islanders were found to have a consistently higher FPR over the LS/RNR risk score cut-off values. Conversely, non-Aboriginal and Torres Strait Islanders were found to have a consistently higher FNR. In some instances, Aboriginal and Torres Strait Islanders were more than twice as likely to be labeled as high risk and not engage in recidivism at certain LS/RNR risk score cut-off thresholds (e.g., cut-off scores from 29 to 36), while non-Aboriginal and Torres Strait Islanders were more than twice as likely to be labeled as low risk and later engage in recidivism (e.g., cut-off scores of 15, 16, 19, 20, and 21).

Although there is a lack of existing research into error rate balance within Australia, international research has produced comparable findings. Studies have shown similar disparities in the United States on the software risk instrument Correctional Offender Management Profiling for Alternative Sanctions (COMPAS; Angwin et al., 2016; Flores et al., 2016), with African Americans having a higher FPR and White Americans having a higher FNR. Because of the differences found among FPRs and FNRs, individuals who are misclassified may receive the wrong level of management or support, which may impact their rehabilitation and risk of recidivism.

Calibration

Close rates of calibration were identified across the very high risk classification when comparing groups on the proportion of those from each risk classification who engaged in recidivism. However, Aboriginal and Torres Strait Islanders were found to engage in recidivism more across the medium and high risk classifications. This is similar to previous research in Australia in which Aboriginal and Torres Strait Islanders were found to be recidivists at higher rates across low, medium, and high risk classifications (Thompson & McGrath, 2012). When extending this analysis to see whether the form of the relationship between the LS/RNR risk score and recidivism varied by Aboriginal and Torres Strait Islander status, the LS/RNR was found to be calibrated amongst these two groups. However, when plotting the predicted probabilities of recidivism, Aboriginal and Torres Strait Islanders were found to be predicted to be recidivists at a higher rate across the majority of risk scores below a very high risk classification. Conversely, non-Aboriginal and Torres Strait Islanders had a higher predicted probability of recidivism across the very high risk scores.

Predictive Parity

Also supporting previous studies considering predictive parity of risk assessment measures in the Australian population, Aboriginal and Torres Strait Islanders were found to have slightly higher PPVs whereas non-Aboriginal and Torres Strait Islanders were found to have higher NPVs. These differences were also similar in magnitude to previous research involving young Australians (Shepherd et al., 2015; Shepherd & Strand, 2016). The higher PPVs for Aboriginal and Torres Strait Islander peoples indicates that a high risk classification is associated with a larger number of recidivists from this group. However, the PPV differences in the present study were minimal.

The higher NPVs for non-Aboriginal and Torres Strait Islanders, on the other hand, were more pronounced and suggest that a low risk classification is linked to a higher number of Aboriginal and Torres Strait Islanders engaging in recidivism, specifically across the lower risk scores. These findings are in line with the calibration analyses, in which higher rates of predicted recidivism across the lower levels of possible risk scores in this sample were identified for Aboriginal and Torres Strait Islanders. Further, approximately 13.14% more non-

Aboriginal and Torres Strait Islanders who are classified as low risk will not go on to engage in recidivism compared to Aboriginal and Torres Strait Islanders classified as low risk. These discrepancies suggest that the LS/RNR lower risk classifications do not predict recidivism (or non-recidivism) in the same way across these groups. Although these discrepancies were not as pronounced in contrast to what was identified for error rate balance, trying to rectify both predictive parity and error rate balance would be impossible due to the different base rates across groups.

Statistical Parity

As expected, the present study found inequalities between the average risk score on the LS/RNR and the distribution of risk scores between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, echoing previous studies (Hsu et al., 2010; Shepherd, Luebbers, et al., 2014; Shepherd et al., 2015; Smallbone & Rallings, 2013; Thompson & McGrath, 2012; Watkins, 2011). Aboriginal and Torres Strait Islanders were more frequently categorized as high risk, had a higher overall risk score, and scored significantly higher on several risk factors. Previous Australian research looking at variations of the Level of Service risk instrument has identified comparable patterns, with Aboriginal and Torres Strait Islanders scoring significantly higher on similar factors, including previous criminal history, education, employment and substance use issues (Hsu et al., 2010; Shepherd et al., 2015; Thompson & McGrath, 2012; Watkins, 2011).

Aboriginal and Torres Strait Islanders are often found to have higher risk scores for a variety of risk factors, including ongoing social and economic disadvantage, which leads to higher rates of unemployment, criminal involvement, substance abuse, and lower income (Day et al., 2018; Homel et al., 1999; Jones & Day, 2011; Shepherd, Adams, et al., 2014). While these differences in risk scores may reflect real differences in risk, the disproportionate labelling of high risk to Aboriginal and Torres Strait Islander peoples may in practice result in more restrictive interventions and supervision. Its broad use could also unintentionally further marginalize Aboriginal and Torres Strait Islander peoples.

Limitations

The current study was limited primarily due to the sample. The sample consisted of individuals who had been previously sentenced to more serious violent offenses, the majority of whom engaged in recidivism by the end of the follow-up period. As a result, the sample included individuals who received higher risk scores on the LS/RNR and there were insufficient low risk scores to test the usefulness of the LS/RNR for low and very low risk individuals or for those individuals who had not previously been convicted of more serious offenses. Due to this, the sample is also not representative of the general Victorian prison population or those on parole or community correction orders. Consequently, the results may not be broadly generalizable to Victoria and/or to those who have not previously been convicted of a violent offense.

Further, due to the small sample size and limited information, the sample could only be categorized as Aboriginal and Torres Strait Islander or non-Aboriginal and Torres Strait Islander. Australia is a multi-cultural society, with non-Aboriginal and Torres Strait Islanders representing a diverse range of cultures (Australian Bureau of Statistics, 2020a). As a result, when observed across more distinct cultural groups within Australia, discrimination and fairness estimates may vary.

Implications

Although the discrimination of the LS/RNR in the present study was found to be relatively equivalent across groups when utilising traditional measures such as the AUC, the discrimination of the assessment instrument was only small in effect size. This demonstrates that the LS/RNR is acceptable but not overly effective at differentiating recidivists from non-recidivists for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Further, disparities identified across the xAUC and numerous fairness definitions highlight that the LS/RNR is not always fair across these groups.

Policymakers and practitioners who use the LS/RNR should therefore be aware of the potential ways in which this instrument can be unfair towards certain groups and take this into consideration to ensure that pre-existing disadvantages faced by cultural minorities such as Aboriginal and Torres Strait Islander peoples are not exacerbated. Specifically, Aboriginal and Torres Strait Islanders being classified as high risk yet not engaging in recidivism more often may result in not only incorrect interventions and treatment, but also harsher monitoring and surveillance. On the contrary, non-Aboriginal and Torres Strait Islanders who are classified as low risk yet engage in recidivism might not be allocated the appropriate treatment or support that could successfully minimize future risk. Further, if decisions are made based on risk scores or classifications, Aboriginal and Torres Strait Islander non-recidivists are often treated the same way or as higher risk than non-Aboriginal and Torres Strait Islander recidivists.

Further research is needed with a larger and more representative sample to see if the current findings are generalizable and to explore how the disparities identified in the present study could be directly impacting different cultural groups, including through decisions around treatment and rehabilitation. Future research should also aim to reduce these discrepancies between the different fairness definitions. For example, the development of different norms for Aboriginal and Torres Strait Islanders could be useful in helping to account for these disparities. Norms for different groups would involve having different cut-off scores to represent the risk classifications, such that a score for one group would not reflect the same level of risk or risk classification as that same score for another group. This could help reduce any scoring disparities between groups and also calibration, predictive parity, and error rate balance discrepancies at certain cut-off scores. Further, xAUC disparities could also be reduced with the LS/RNR now being able to rank recidivists from one group higher than non-recidivists from another group more efficiently. However, the process of developing new norms would be a time-consuming and expensive exercise.

Other efforts to increase cross-cultural fairness have arisen from different disciplines such as data science, statistics, and criminology. These approaches have concentrated on practical and time-sensitive solutions that use statistical processing methods that require transforming algorithms at different stages (see Berk et al., 2018). This process can, for example, remove bias in data before the data is used to predict an individual's level of risk of recidivism, effectively removing any potential sources contributing to unfairness (Berk, 2009; Hajian & Domingo-Ferrer, 2013). This approach has led to improvements across differing fairness notions, including statistical parity and error rate balance (Wadsworth et al., 2018), and should therefore be tested as a strategy for increasing fairness among Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. However, as identified in the present study, there were disparities in both predictive parity and, more notably, error rate balance. With the differing base rates, satisfying both of these forms of fairness will be unachievable due to the impossibility theorem. Consequently, either fairness in the accuracy of risk classification predictions (i.e., predictive parity) or fairness in the errors in observation (i.e., error rate balance) will need to be prioritized.

Conclusion

This study adds to the growing body of knowledge about the cross-cultural fairness of forensic risk assessment instruments. In particular, it contributes to the limited body of research literature explicitly investigating the fairness of risk assessment instruments in Australia, particularly among Aboriginal and Torres Strait Islander peoples in Victoria. It highlights that comparable AUC values are not sufficient in indicating that a risk assessment instrument is cross-culturally fair. Notable disparities were identified across the xAUC and error rate balance, indicating that the LS/RNR may not be a fair assessment of risk for individuals previously convicted of a serious violent offense in Victoria, Australia.

References

Allan, A., Dawson, D., & Allan, M. M. (2006). Prediction of the risk of male sexual reoffending in Australia. *Australian Psychologist*, *41*(1), 60-68. https://doi.org/10.1080/00050060500391886

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashford, L. J., Spivak, B. L., & Shepherd, S. M. (2021). Racial fairness in violence risk instruments: A review of the literature. *Psychology, Crime & Law*. https://doi.org/10.1080/1068316X.2021.1972108

Australian Bureau of Statistics. (2018a). *4517.0 - Prisoners in Australia, 2018*. http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4517.0~2018~Main%20Features~Aboriginal%20and%20Torres%20Strait%20Islander%20prisoner%20characteristics%20~13

Australian Bureau of Statistics. (2018b). *Estimates of Aboriginal and Torres Strait Islander Australians*. https://www.abs.gov.au/statistics/people/aboriginal-and-torres-strait-islander-peoples/estimates-aboriginal-and-torres-strait-islander-australians/latest-release

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., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1-42. https://doi.org/10.1177/0049124118782533

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-220. http://www.jstor.org/stable/2985181

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

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing*, *9*(6), 649-657. https://doi.org/doi:10.1046/j.1365-2850.2002.00535.x

Dragomir, R. R., & Tadros, E. (2020). Exploring the impacts of racial disparity within the American juvenile justice system. *Juvenile and Family Court Journal*, *71*(2), 61-73. https://doi.org/https://doi.org/10.1111/jfcj.12165

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January 8-10). *Fairness through awareness* [Paper presentation]. 3rd Innovations in Theoretical Computer Science Conference, Cambridge, Massachusetts, United States.

Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation*, *80*(2), 38-46,66.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293-323. https://doi.org/10.1037/1076-8971.2.2.293

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445-1459. https://doi.org/10.1109/TKDE.2012.72

Harrell, F. (2020). *rms: Regression modeling strategies*. (Version 6.0-1) [R package]. https://CRAN.R-project.org/package=rms

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

Homel, R., Lincoln, R., & Herd, B. (1999). Risk and resilience: Crime and violence prevention in Aboriginal communities. *Australian and New Zealand Journal of Criminology*, *32*(2), 182-196. https://doi.org/10.1177/000486589903200207

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2010). Level of Service Inventory–Revised: Assessing the risk and need characteristics of Australian Indigenous offenders. *Psychiatry, Psychology and Law*, *17*(3), 355-367. https://doi.org/10.1080/13218710903089261

Jones, R., & Day, A. (2011). Mental health, criminal justice and culture: Some ways forward? *Australasian Psychiatry*, *19*(4), 325-330. https://doi.org/10.3109/10398562.2011.579613

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

Kassambara, A., Kosinski, M., & Biecek, P. (2020). *survminer: Drawing survival curves using 'ggplot2.'* (Version 0.4.8) [R program]. https://CRAN.R-project.org/package=survminer

Martel, J., Brassard, R., & Jaccoud, M. (2011). When two worlds collide: Aboriginal risk management in Canadian corrections. *British Journal of Criminology*, *51*(2), 235-255. https://doi.org/10.1093/bjc/azr003

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*(1), 489-513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and caucasian female and male adolescents on probation. *Psychological Assessment*. https://doi.org/10.1037/pas0000816

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

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

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

Shepherd, S. M., Adams, Y., McEntyre, E., & Walker, R. (2014). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law*, *20*(3), 281-293. https://doi.org/10.1037/law0000017

Shepherd, S. M., Luebbers, S., Ferguson, M., Ogloff, J., & Dolan, M. (2014). The utility of the SAVRY across ethnicity in Australian young offenders. *Psychology, Public Policy, and Law*, *20*(1), 31-45. https://doi.org/10.1037/a0033972

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

Shepherd, S. M., & Strand, S. (2016). The PCL: YV and re-offending across ethnic groups. *Journal of Criminal Psychology*, *6*(2), 51-62. https://doi.org/10.1108/JCP-02-2016-0006

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, *31*(1), 8-22. https://doi.org/doi:10.1002/bsl.2052

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

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

Smallbone, S., & Rallings, M. (2013). Short-term predictive validity of the Static-99 and Static-99-R for Indigenous and nonindigenous Australian sexual offenders. *Sexual Abuse A Journal of Research and Treatment*, *25*(3), 302-316. https://doi.org/10.1177/1079063212472937

Therneau, T. M. (2020). *survival: Survival analysis*. (Version 3.2-7) [R program]. https://cran.r-project.org/package=survival

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior*, *36*(4), 345-355. https://doi.org/10.1037/h0093930

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]*.

Watkins, I. (2011). *The utility of Level of Service Inventory-Revised (LSI-R) assessments within NSW correctional environments. Research bulletin*. Corrective Services NSW. https://doi.org/https://correctiveservices.dcj.nsw.gov.au/content/dam/dcj/corrective-services-nsw/documents/research-and-statistics/rb29-utility-of-level-of-service-inventory-.pdf

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

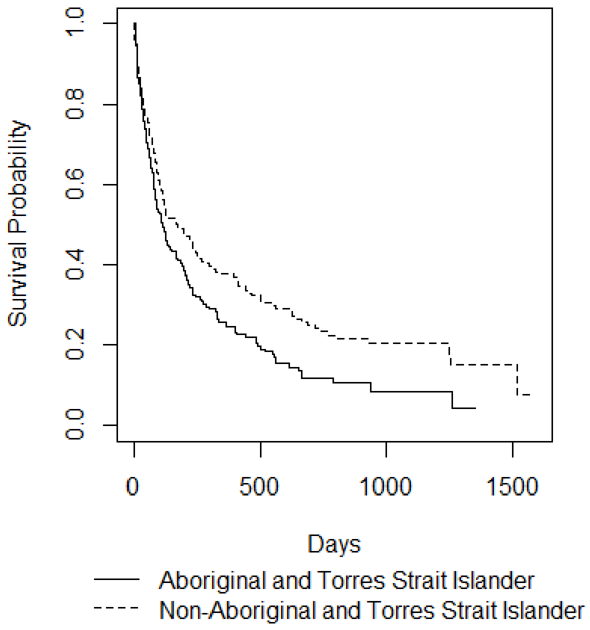
Wormith, J., Hogg, S., & Guzzo, L. (2015). The predictive validity of the LS/CMI with Aboriginal offenders in Canada. *Criminal Justice and Behavior*, *42*(5), 481. https://doi.org/10.1177/0093854814552843

Supplementary Materials

The Cox regression analysis was extended by conducting Kaplan-Meier survival analyses to visually observe the group differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in their trajectories for recidivism. The survival curves are presented in Figure S1. The survival curve demonstrated that the probability of recidivism was greater for Aboriginal and Torres Strait Islanders.

Figure S1

Survival Curves for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders

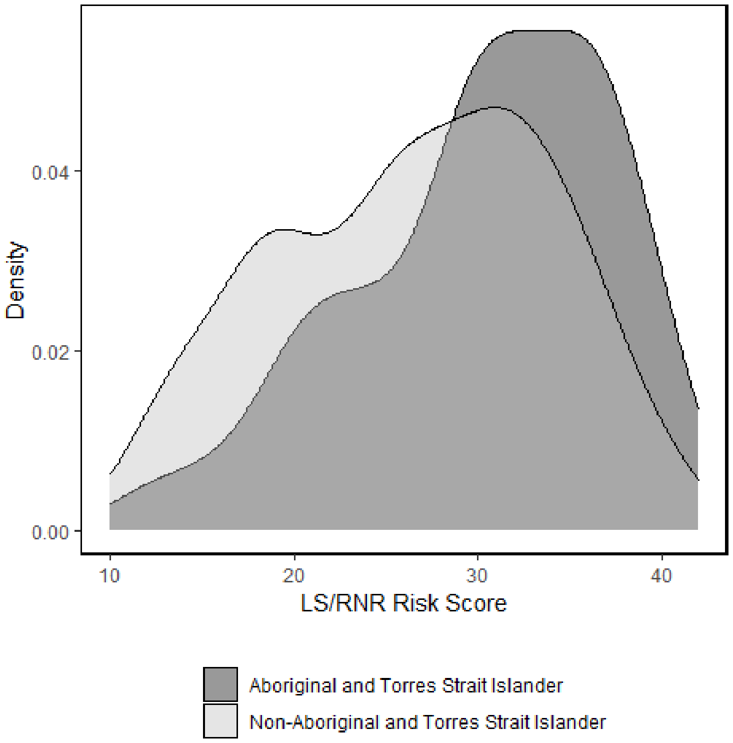


The distribution of the LS/RNR total risk scores for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders is presented in a density plot in Figure S2. Both groups had a large proportion of individuals scoring highly on the LS/RNR, with

Aboriginal and Torres Strait Islanders having a larger proportion of very high risk individuals (scores between 30 and 43). There was a higher representation of medium risk individuals in the non-Aboriginal and Torres Strait Islander group (scores between 11 and 19).

Figure S2

Risk Score Distribution for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders



Chapter Five: Empirical Study Two

5.1 Introduction

Recently, the use of more novel statistical learning approaches (i.e., machine learning) has been used within the forensic risk assessment literature. Primarily, these approaches have been used as a way to increase the predictive accuracy and discrimination of forensic risk assessment instruments. Processing approaches that involve altering statistical learning approaches in varying ways have led to increases in cross-cultural fairness. Yet, even with empirical research demonstrating disparities in numerous fairness definitions, these approaches are yet to be extensively trialled. To date, the majority of research utilising statistical learning approaches has been within computer science disciplines. However, research within the forensic psychology discipline is scarce, especially within Australia.

This chapter presents the second empirical study that addressed the second research aim by responding to research questions three and four. The first empirical study highlighted acceptable levels of discrimination from the LS/RNR total score that were in line with previous Australian studies examining LS instruments; however, they were often small in effect size. This was found for the sample overall, as well as for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Further, it highlighted notable disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on certain fairness definitions (error rate balance and statistical parity) as well as disparities in xAUC.

Therefore, this empirical study assesses the usefulness of statistical learning methods in both increasing the discrimination and fairness of the LS/RNR using the LS/RNR items. Numerous statistical learning methods (logistic regression, penalised logistic regression, random forests, stochastic gradient boosting, and support vector machines) were trialled in this study and compared to the original performance of the LS/RNR total score. Furthermore, pre- and post-processing approaches were applied to the statistical learning methods in order to increase cross-cultural fairness. Similar to Empirical Study One, this paper explores the trade-offs that exist in the pursuit of fairness, as well as other trade-offs that occur due to the use of statistical learning methods.

Empirical Study Two is titled "*Statistical Learning Methods and Cross-Cultural Fairness: Trade-Offs and Implications for Risk Assessment Instruments"* and has been submitted to *Psychological Assessment* for publication. *Psychological Assessment* is an American Psychological Association peer-reviewed journal that publishes empirical research related to assessment instruments within psychology disciplines. The Author Indication Form that details the contribution of each author to this manuscript is included in Appendix A.

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

Linda J. Ashford[1], Benjamin L. Spivak[1], James R. P. Ogloff[1] & Stephane M. Shepherd[1]

[1] Centre for Forensic Behavioural Science, Swinburne University of Technology

Author Note.

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

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. We 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.

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

Abstract

The use of statistical learning methods has recently increased within the risk assessment literature. This has primarily been used to increase predictive accuracy and discrimination. Processing approaches applied to statistical learning methods have also emerged to increase cross-cultural fairness. However, these approaches are rarely trialled in the forensic psychology discipline, nor have they been trialled as an approach to increase discrimination and fairness in Australia. The present study included 380 Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander males that were assessed with the Level of Service/Risk Needs Responsivity (LS/RNR). Discrimination was assessed through the area under the curve (AUC) and fairness was assessed through the cross area under the curve (xAUC), error rate balance, calibration, predictive parity, and statistical parity. Logistic regression, penalised logistic regression, random forest, stochastic gradient boosting, and support vector machine algorithms using the LS/RNR items were used to compare performance against the LS/RNR total risk score. Pre- and post-processing approaches were then performed on each of the algorithms to see if fairness could be increased. Penalised logistic regression (AUC = .73, range [.53, .85]) and stochastic gradient boosting (AUC = .73, range [.59, .88]) were found to have higher levels of discrimination compared to the LS/RNR total risk score (AUC = .64, 95% CI [.57, .70]). Processing approaches (primarily pre-processing) increased a number of fairness definitions between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Specifically, xAUC, error rate balance, and statistical parity were notably improved. 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.

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

Introduction

Risk assessment instruments within the criminal justice system involve estimating the likelihood of an individual engaging in recidivism. 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 judgement (SPJ) instruments. Actuarial risk assessment instruments are scored by a formula or algorithm, combining numerical values 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 risk (Hart et al., 2017). Both of these instrument types have been assessed for their utility, often by observing the instrument's ability to distinguish recidivists from non-recidivists. 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 recidivists from non-recidivists (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 increase predictive accuracy (Spivak & Shepherd, 2020). Statistical learning methods differ from traditional approaches to risk estimation (e.g., linear or logistic regression), which work effectively with careful theorising and pre-specification of predictor interactions and the form of relationship between predictors and the outcome. The form of relationships and predictor interactions does not have to be pre-specified when using statistical learning methods. Instead, certain statistical learning methods can use a large number of predictors and take advantage of non-linear relationships and the predictive capacity of both strong and weak predictions, as well as their interactions. This can aid in maximising predictive accuracy and/or discrimination (Berk & Bleich, 2013; Brennan, 2016; Spivak & Shepherd, 2020).

Within the risk assessment literature, statistical learning methods have primarily been used with the chief purpose of increasing predictive validity and/or discrimination over traditional logistic regression and/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 when using random forests and gradient boosting algorithms compared to logistic regression. Similarly, Liu et al. (2011) reported comparable AUC values across logistic regression, classification and regression trees, and neural network algorithms 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 recidivism. When examining neural networks, random forests, and logistic regression in predicting felonies, drug, violent, and sexual recidivism, Hamilton et al. (2015) noted that logistic regression was a better discriminator, most notably for violent and sexual recidivism.

Comparatively, Ting et al. (2018) utilised 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 recidivism, 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 predictors in decision trees, random forests, and support vector machines and found that the AUC was comparable to the AUC of the original LS/CMI. However, the statistical learning methods were found to be better discriminators along the middle scores of the LS/CMI, with the average AUCs for individual scores improving from .50 to near .60. Breitenbach et al. (2009) explored different algorithms' performance when using all possible predictors and also a subset of predictors. When all possible predictors were used, random forests had the highest AUC for violent recidivism (.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 penalised logistic regression and random forests to logistic regression using different sets of static and dynamic predictors from the Finnish Needs and Risk Assessment Form. Penalised logistic regression and random forest outperformed logistic regression across all sets of items, with logistic regression often found to produce a higher AUC when using a smaller subset of items.

These latter studies that demonstrated more promising results with increases in discrimination often utilised a notably larger number of predictors in their study. This demonstrates one of the benefits of statistical learning methods in that they are able to exploit a large number of predictors to improve predictive accuracy and discrimination, often outperforming traditional approaches such as logistic regression.

Fairness

The use of statistical learning methods has also extended to the complex issue of fairness in risk assessment instruments (Berk et al., 2018). The debate concerning fairness in risk assessment has been increasing over the past decade, with critics of risk assessment instruments arguing that the instruments could disadvantage certain cultural groups (Angwin et al., 2016; 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 risk assessment literature. Statistical parity refers to the distribution of risk scores being equal across different groups (Berk et al., 2018). Error rate balance refers to the false positive rate (FPR), or the proportion of non-recidivists incorrectly predicted to engage in recidivism (or classified as high risk), and the false negative rate (FNR), or the proportion of recidivists incorrectly predicted to not engage in recidivism (or classified as low risk), being the same across groups (Chouldechova, 2017). Calibration refers to a risk instrument's predicted probability (or risk scores/classifications) aligning with actual recidivism (Chouldechova, 2017). Across different groups, predicted probabilities should reflect the same level of recidivism. Last, predictive parity refers to the positive predictive value (PPV), the proportion of those predicted to be recidivists who go on to engage in recidivism, and the negative predictive value (NPV), the proportion of those not predicted to be recidivists who do not go on to engage in recidivism, being the same across groups (Berk et al., 2018).

The utility of these instruments, which are 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. Some cultural minority groupings are often found to score significantly higher on risk assessment instruments (Hsu et al., 2010; Olver et al., 2014), be classified as high risk yet not engage in recidivism more often (i.e., have a higher FPR; Angwin et al., 2016; Flores et al., 2016), and are predicted to be recidivists more often, especially across lower risk scores and risk classifications (Wilson & Gutierrez, 2014). Risk assessment instruments are also often better at predicting recidivism among cultural minority groupings (i.e., a higher PPV), whereas among cultural majority groupings, risk assessment instruments are often better at predicting non-recidivism (i.e., a higher NPV; Muir et al., 2020; Shepherd et al., 2015) and are more likely to classify cultural majorities who engage in recidivism as low risk (i.e., a higher FNR; Angwin et al., 2016; Flores et al., 2016).

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., 2018). Specifically, by altering the algorithm through different stages of its construction and execution, unfairness can be, to a degree, ameliorated. The algorithm can be altered in three stages. Pre-processing 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 algorithm 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. Post-processing involves altering the predictions themselves to aid in improving fairness. For example, predicted outcomes could be randomly reassigned to aid in achieving equivalence between groups.

In the risk assessment literature, a number of pre-processing approaches have been trialled, with often promising results. Berk (2019) reweighted the data to equalise base rates and found that error rate balance and predictive parity were improved among African Americans and Caucasian youth in a gradient boosting algorithm when compared to an unweighted gradient boosting algorithm. Another pre-processing approach trialled by

Johndrow and Lum (2017) involved transforming predictors to achieve independence from an individual's culture. This resulted in a reduction in FPR differences between African Americans, Caucasians, and Hispanics. There was, however, a slightly increased disparity among PPVs and NPVs compared to a random forest with unadjusted predictors. Expanding upon this, Lum and Johndrow (2016) also found that this approach barely reduced 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 with a series of regression based algorithms for which one involved each predictor being regressed onto race and the residuals being used in place of the predictors. Comparing this to an algorithm that included only the Post-Conviction Risk Assessment (PCRA) items, the AUC was found to be slightly reduced. The residual algorithm resulted in an AUC of .71, compared to .72 for the algorithm with PCRA items. Parity among calibration and PPVs was also mildly impeded, with disparities being more pronounced in the residual algorithm. Skeem and Lowenkamp (2020) found that FPR disparities were improved when using residuals in the algorithm, with a difference of 7.21% between African American, and Caucasians being reduced to -3.65%. FNR differences remained similar. However, using residuals in the model resulted in African Americans having a higher FNR (10.92% difference) compared to the comparison model (-9.86% difference).

In-processing and post-processing are less often used. However, these approaches have still demonstrated an ability to increase fairness. For example, Wadsworth et al. (2018) used an in-processing adversarial approach with a neural network algorithm. This led to an increase in discrimination when compared to the original Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) instrument score, with AUC increasing from .66 to .70. 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 recidivism differ (Berk et al., 2018; Chouldechova, 2017).

There is also the issue of an instrument's utility alongside fairness. Altering statistical learning method algorithms for which the main focus has traditionally been maximising accuracy could and will likely lead to detrimental impacts on a risk assessment instrument's overall predictive utility. For example, altering predictors in a pre-processing step that are usually valid and significant predictors of recidivism could impede on the overall accuracy of predictions. This is also relevant to an instrument's ability to discriminate between an individual who is a recidivist and an individual who is not a recidivist.

Last, a common critique regarding statistical learning methods is that the interpretability and transparency of the algorithm is often reduced compared to traditional approaches such as linear or logistic regression (Breiman, 2001b). These 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 predictive performance 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/items that are most predictive of a future offence and therefore are unable to intervene and respond to the relevant needs of the individual.

Current 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 to increasing fairness with cultural minority groups in Australia. Specifically, in Australia, Aboriginal and Torres Strait Islanders already experience inequality within the criminal justice system, such as significant over-incarceration and a decreased likelihood of receiving a diversion (Australian Bureau of Statistics, 2020c; Papalia et al., 2019). Further, as highlighted in the literature above, disparities have been reported between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on a number of fairness definitions.

Therefore, the present study examined a variety of statistical learning methods to explore their ability to increase the discrimination of the Level of Service/Risk Needs Responsivity (LS/RNR; Andrews et al., 2008). Statistical learning methods were also altered at pre-processing and post-processing to see if various forms of fairness could be improved. The study also explored the impact that different processing alterations had on the discrimination of the algorithms 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 violent offence 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 while serving either a prison sentence (*n* = 231, 60.79%), a community corrections order (*n* = 148, 38.95%), or a parole order (*n* = 1, 0.26 %). 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.

LS/RNR completions, demographics, and 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 (LEAP) database for the period of January 2015 through to 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 recidivism and identify their criminogenic needs. The General Risk/Needs section 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. With the total score, individuals can be categorised into risk levels including very low risk (0–4), low risk (5–10), medium risk (11–19), high risk (20–29) and very high risk (30–43).

Recidivism

Recidivism 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 LS/RNR assessment (or release date for those incarcerated) to either the date of first charge for those who engaged in recidivism or the end of the follow up period date for those who did not engage in recidivism (31-12-2019). The average follow up time for the sample was 280.56 days (*SD* = 329.24). The majority of the sample were found to be general recidivists by the end of the follow up period (*n* = 306, 80.53%) for which the average time from LS/RNR assessment to first offence was 184.75 days (*SD* = 233.80). This differed between groups, with more Aboriginal and Torres Strait Islanders engaging in recidivism compared to non-Aboriginal and Torres Strait Islanders (85.56% and 76%, respectively) by the end of the follow up period.

Analytical Approach

All data was analysed 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 penalised 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 cut-offs.

Area under the Curve (AUC)

The AUC is the probability that a randomly selected individual who is a recidivist will receive a higher risk score compared to a randomly selected individual who is not a recidivist.

The AUC is base rate resistant and provides an index of a risk instrument's sensitivity and 1 – specificity across various thresholds (Cook, 2007). The AUC value ranges from 0 to 1, with 0.5 reflecting discrimination at chance levels (Rice & Harris, 2005).

Fairness

Cross Area under the Curve (xAUC). The xAUC (Kallus & Zhou, 2019) is a modification of the AUC that measures discrimination between groups instead of within to better identify disparities. The xAUC is calculated for two sets. The first 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 Aboriginal and Torres Strait Islander recidivists as well as non-Aboriginal and Torres Strait Islander non-recidivists. Set 2 includes non-Aboriginal and Torres Strait Islander recidivists as well as Aboriginal and Torres Strait Islander non-recidivists. The xAUC measures the probability that a random individual who is a recidivist from one group receives a higher risk score than a random individual from the other group who is not a recidivist.

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 zero and the worst possible Brier score being one. These were calculated overall and 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 and computing the difference. In order to distinguish between those who are predicted to engage in recidivist and those who are not to calculate the relevant metrics for predictive parity, a cut-off value is required. For the present study, the optimal cut-off was defined as the cut-off that yielded the smallest distance to the point 0, 1 in the receiver operating characteristic (ROC) space. This approach was utilised as a test that passes through 0, 1 on the ROC space reflects perfect discrimination. Although the present study aims to maximise fairness, doing this while maintaining discrimination of the LS/RNR was also important. Therefore, a cut-off point that prioritised discrimination was utilised and was calculated separately for the LS/RNR total risk score and each algorithm.

Error Rate Balance. Error rate balance was assessed by calculating the FPR and FNR as a percentage for both groups and computing the difference. Like predictive parity, error rate balance metrics require a cut-off value in order to be calculated. The optimal cut-off yielding the smallest distance to 0, 1 in the ROC space was again utilised.

Statistical Parity. Statistical parity was assessed as the difference between the proportions of Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders who were classified as high risk (i.e., the proportion whose predicted probability of recidivism was above the optimal cut-off threshold specified for that algorithm).

Statistical Learning Algorithms

To account for overfitting and a small sample size, *k*-fold cross-validation with 10 folds was used to validate the algorithms. The data was split into 10 equal folds for which each of the folds in turn served as a validation set for the other 90% of the data that is used for training. This validation process also trialled numerous parameter options relevant for each algorithm to choose the parameter values that resulted in the best performance (i.e., the highest AUC value). As the majority of the sample engaged in recidivism, upsampling, or sampling with replacement from the minority class (i.e., those who did not go on to engage in recidivism), was also used when training the algorithms to account for the imbalanced outcome data. The performance of each algorithm's prediction on the held-out fold was aggregated and summarised to determine the algorithms' average performance. The algorithms for the present study include logistic regression, penalised logistic regression, random forest, stochastic gradient boosting, and support vector machine.

Logistic Regression. Logistic regression was used as a baseline comparison to assess the performance of the other algorithms against.

Penalised Logistic Regression. Penalised logistic regression was also used to increase 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 towards zero. The latter imposes a penalty on the absolute value of the coefficients and shrinks irrelevant predictor coefficients completely to zero. For elastic net regression, another parameter term is imposed that mixes and signifies the type of penalty used, ranging from 0 to 1, where 0 reflects a pure ridge regression penalty and 1 reflects a pure lasso regression penalty.

Random Forests. Random forests (Breiman, 2001a) are an ensemble based algorithm (i.e., a combination of numerous algorithm predictions) of decision trees that combine the concepts of bagging with random feature selection. 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 minimise 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, often a decision tree, that predicts slightly better than random) is applied repeatedly to the data. It seeks to find an additive algorithm that will minimise the loss function (e.g., squared error). Initially, specified predicted values are utilised (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), and new predicted values being added to the previous. Similar to 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 final outcome. Instead, each tree is dependent on past trees and is weighted depending on how much of an influence they have over the final 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., recidivist and non-recidivist) to create the greatest segregation between the two. Therefore, data points that are predicted to fall on either side of this hyperplane can be attributed to an outcome class. As these data points are unlikely to be easily separable in two dimensions, kernels are used to transform the data into higher dimensions, enabling separation. For the present study, multiple kernels were tested, and ultimately non-linear polynomial kernels were used as they produced the highest levels of discrimination.

Processing Techniques

Two processing techniques were utilised in the present study. The first was a pre-processing technique that used residuals in place of the predictor variables. Specifically, predictor variables were regressed on to Aboriginal and Torres Strait Islander status and the residual for each variable was utilised instead of the original predictor value. This approach helps to remove the association between an individual's Aboriginal and Torres Strait Islander status, and other predictor variables from the data before the algorithm is constructed.

The second processing technique was a post-processing technique that involved reassigning the outcome classification through a process known as reject option based classification (Kamiran et al., 2012). This process relabels observation outcomes that are deemed to be more uncertain (i.e., close to the cut-off value) and influenced by biases. With reject option based classification, a critical region boundary (i.e., margin) is specified around the cut-off value, denoted by θ. The observations that fall within this region have had their labels reassigned specifically to aid in increasing parity among the outcome variables.

In regard to the present study, the group that engaged in recidivism more had their observations that fell within the cut-off + θ reassigned to being predicted to not be a recidivist. Conversely, the group that engaged in recidivism less had their observations that fell within the cut-off - θ reassigned to being predicted to be a recidivist. Multiple θ values were trailed in the present study (ranging from 0.00625 to 0.01), with the final θ value being the one that led to the biggest increase in fairness and the lowest reduction in discrimination.

Results

Statistical Learning Methods

AUC

The AUC was established for the LS/RNR total risk score, as well as for each statistical learning method. The AUCs are presented in Table 1. For the original measure (i.e., the LS/RNR risk score), the AUC is reported alongside the 95% confidence interval (CI). For the algorithms, the average AUC is reported alongside the range of AUC values from the 10 cross-validation folds. The LS/RNR total risk score yielded the lowest AUC overall, with marginal improvements being found with logistic regression and random forest algorithms. Penalised logistic regression, stochastic gradient boosting, and support vector machine algorithms led to improvements in the LS/RNR's ability to discriminate recidivists from non-recidivists. Improvements were also identified across all algorithms for Aboriginal and Torres Strait Islanders when compared to the original LS/RNR total risk score. Only for non-Aboriginal and Torres Strait Islander people did the performance of algorithms vary. Improvements in discrimination were only notable for the stochastic gradient boosting model.

Table 1

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

	Overall		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	AUC	Range	AUC	Range	AUC	Range
LS/RNR Total Risk Score	.64	.57-.70	.60	.49-.70	.63	.55-.72
Logistic Regression	.65	.45-.85	.70	.43-.92	.59	.31-.85
Penalised Logistic Regression	.73	.53-.85	.80	.56-.94	.66	.49-.82
Random Forest	.67	.50-.78	.66	.21-1	.59	.43-.78
Stochastic Gradient Boosting	.73	.59-.88	.71	.53-.87	.73	.57-.92
Support Vector Machine	.70	.47-.80	.67	.42-.97	.64	.49-.86

Fairness

xAUC. The xAUC was calculated for the LS/RNR risk score and each of the statistical learning methods. The xAUC alongside the 95% CI or range of xAUC values is presented in Table 2. The xAUC was higher for Set 1 (Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists) for the LS/RNR risk score and all algorithms. The biggest disparity between the groups was identified for the LS/RNR total risk score, with the xAUC for Set 2 (non-Aboriginal and Torres Strait Islander recidivists and Aboriginal and Torres Strait Islander non-recidivists) being below chance levels.

Table 2

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

	Set 1		Set 2		
	xAUC	Range	xAUC	Range	xAUC Diff
LS/RNR Total Risk Score	.75	.68-.83	.46	.35-.57	.29
Logistic Regression	.71	.42-1	.60	.25-.77	.11
Penalised Logistic Regression	.79	.70-.90	.64	.42-.85	.15
Random Forest	.77	.68-.87	.57	.17-.95	.20
Stochastic Gradient Boosting	.79	.71-.95	.62	.25-.88	.17
Support Vector Machine	.81	.64-.92	.56	.25-.91	.25

Calibration. Brier scores to assess calibration were calculated for each statistical learning method and are presented in Table 3. 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. The greatest level of calibration overall was identified with the random forest algorithm. Furthermore, greater levels of calibration were also reported for Aboriginal and Torres Strait Islanders for all algorithms. The disparity between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on Brier scores was greatest for the logistic regression algorithm, with the closest Brier scores being found for the stochastic gradient boosting algorithm.

Table 3

Brier Scores for Statistical Learning Methods using LS/RNR Items

	Overall		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	Brier	Range	Brier	Range	Brier	Range
Logistic Regression	.241	.144-.329	.194	.072-.335	.280	.192-.392
Penalised Logistic Regression	.221	.200-.237	.204	.181-.226	.237	.209-.270
Random Forest	.148	.128-.173	.115	.070-.209	.181	.142-.259
Stochastic Gradient Boosting	.211	.194-.225	.199	.176-.228	.222	.196-.240
Support Vector Machine	.228	.177-.296	.196	.118-.285	.255	.206-.318

Predictive Parity, Error Rate Balance, and Statistical Parity. Predictive parity, error rate balance, and statistical parity differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders are reported in Table 4 for the LS/RNR total risk score and algorithms using the LS/RNR items. For each of these fairness definitions, the non-Aboriginal and Torres Strait Islander value was subtracted from the Aboriginal and Torres Strait Islander value. For predictive parity, error rate balance, and statistical parity values by group, please refer to the supplementary materials.

Table 4

Predictive Parity, Error Rate Balance, and Statistical Parity Differences between Groups

	PPV Diff	NPV Diff	FPR Diff	FNR Diff	High Risk Diff
LS/RNR Total Risk Score	4.07	-11.80	26.12	-17.43	20.83
Logistic Regression	8.19	-2.23	-2.27	-12.44	12.70
Penalised Logistic Regression	2.90	-3.59	11.00	-16.04	18.60
Random Forest	5.11	-6.16	22.31	-19.12	21.41
Stochastic Gradient Boosting	1.98	-12.38	14.07	-7.23	11.87
Support Vector Machine	0.73	-8.57	21.84	-14.51	18.82

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 higher FPR, significantly lower FNR, and a larger proportion classified as high risk. All algorithms were found to lower the FPR differences, specifically the regression based models. All the algorithms besides random forests were also found to improve fairness between FNRs and statistical parity. Disparities in PPV were relatively minimal for all algorithms and the LS/RNR risk score, with Aboriginal and Torres Strait Islanders having a higher PPV. Conversely, Aboriginal and Torres Strait Islanders had a lower NPV, which was most notable for the LS/RNR total risk score and the stochastic gradient boosting models.

Processing Techniques

AUC

The AUC was calculated for algorithms for both the pre-processing technique (i.e., residuals) and the post-processing technique (i.e., reject option based classification). The AUC and the range of AUCs from the 10-fold cross validations are presented in Table 5. Overall, the discrimination of the algorithms was minimally impacted after applying pre- or post-processing techniques. For the tree base algorithms (random forest and stochastic gradient boosting), pre-processing resulted in a marginal improvement in AUC, whereas the remaining algorithms resulted in similar or slightly smaller AUCs. For Aboriginal and Torres Strait Islanders, pre-processing resulted in a marginally higher AUC for the random forest and support vector machine algorithms, and unchanged or slightly reduced AUCs for the remaining algorithms. For non-Aboriginal and Torres Strait Islanders, pre-processing led to a mild improvement for the penalised logistic regression, random forest, and stochastic gradient boosting algorithms, and the remaining algorithms had AUCs comparable to the algorithms without processing changes.

Table 5

AUC for Statistical Learning Methods using Pre- and Post-Processing Techniques

		Overall		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	Processing	AUC	Range	AUC	Range	AUC	Range
Logistic Regression	Pre	.63	.46-.88	.70	.43-.92	.59	.31-.90
	Post	.64	.44-.85	.70	.43-.92	.58	.31-.85
Penalised Logistic Regression	Pre	.72	.53-.84	.79	.54-.94	.68	.50-.82
	Post	.73	.53-.85	.80	.56-.94	.66	.49-.83
Random Forest	Pre	.69	.49-.84	.70	.21-1	.61	.39-.85
	Post	.66	.48-.78	.66	.21-1	.59	.41-.78
Stochastic Gradient Boosting	Pre	.74	.56-.90	.71	.52-.93	.75	.56-.98
	Post	.73	.57-.86	.71	.53-.87	.72	.54-.87
Support Vector Machine	Pre	.68	.51-.82	.68	.37-.94	.64	.44-.86
	Post	.69	.46-.79	.66	.42-.97	.63	.49-.84

Fairness

xAUC. The xAUC was calculated for the statistical learning methods using the pre- and post-processing techniques. The xAUC and range of xAUC values from the cross-validation samples are reported in Table 6.

Table 6

xAUC for Statistical Learning Methods using Pre- and Post-Processing Techniques

		Set 1		Set 2		
	Processing	xAUC	Range	xAUC	Range	xAUC Diff
Logistic Regression	Pre	.62	.39-.97	.66	.25-.86	-.04
	Post	.70	.40-1	.61	.25-.80	.09
Penalised Logistic Regression	Pre	.67	.56-.79	.77	.54-.96	-.10
	Post	.78	.70-.90	.64	.42-.90	.14
Random Forest	Pre	.89	.76-.98	.45	.08-.82	.44
	Post	.74	.60-.84	.58	.17-.95	.16
Stochastic Gradient Boosting	Pre	.91	.80-1	.45	.25-.77	.46
	Post	.79	.69-.95	.63	.25-.88	.16
Support Vector Machine	Pre	.67	.52-.84	.68	.33-.94	-.01
	Post	.80	.61-.91	.56	.25-.91	.24

Compared to the original algorithms reported in Table 2, pre and post processing for the majority of algorithms led to a decrease in xAUC for Set 1 (Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists) and an increase in xAUC for Set 2 (non-Aboriginal and Torres Strait Islander recidivists and Aboriginal and Torres Strait Islander non-recidivists). Only the tree-based models with pre-processing had the opposite finding. This led to a decrease in xAUC differences from the original algorithms and from the LS/RNR total risk score for all algorithms besides the pre-processed random forest and stochastic gradient boosting algorithms. In particular, the stochastic gradient boosting algorithm with pre-processing resulted in almost perfect parity between Set 1 and Set 2 xAUCs.

Calibration. Brier scores were established for the statistical learning methods using pre- and post-processing on the LS/RNR items. The Brier scores and the range of scores from the validation samples are reported in Table 7. Compared to the original Brier scores presented in Table 3, pre and post processing for the majority of algorithms led to lower levels of calibration overall and for Aboriginal and Torres Strait Islanders. Only for the random forest model was there mild improvement with pre-processing. However, a number of algorithms with pre-processing led to an increase in calibration for non-Aboriginal and Torres Strait Islanders, namely logistic regression, penalised logistic regression, and support vector machine algorithms. This resulted in lower levels of disparity between Brier scores for these algorithms, with the closest Brier scores being reported for the penalised logistic regression and support vector machine algorithms with pre-processing. Only for the random forest and stochastic gradient boosting models with pre-processing was the disparity between Brier scores exacerbated.

Table 7

Brier Scores for Statistical Learning Methods using Pre- and Post-Processing Technique

		Overall		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	Processing	Brier	Range	Brier	Range	Brier	Range
Logistic Regression	Pre	.244	.152-.328	.219	.097-.357	.265	.168-.385
	Post	.242	.145-.332	.195	.074-.336	.281	.192-.392
Penalised Logistic Regression	Pre	.226	.208-.243	.222	.202-.244	.230	.208-.260
	Post	.222	.200-.237	.205	.181-.226	.238	.207-.270
Random Forest	Pre	.146	.126-.175	.112	.072-.204	.181	.128-.243
	Post	.148	.128-.174	.115	.071-.210	.181	.142-.259
Stochastic Gradient Boosting	Pre	.220	.208-.233	.204	.188-.219	.234	.214-.249
	Post	.211	.194-.225	.199	.176-.228	.222	.196-.241
Support Vector Machine	Pre	.245	.176-.269	.233	.166-.284	.254	.185-.318
	Post	.229	.177-.297	.197	.118-.285	.256	.206-.318

Predictive Parity, Error Rate Balance, and Statistical Parity. Predictive parity, error rate balance, and statistical parity metrics were computed for each algorithm for both the pre- and post-processing algorithms, and the difference between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders is presented in Table 8.

Table 8

Predictive Parity, Error Rate Balance and Statistical Parity Differences between Groups using Pre- and Post-Processing Technique

	Processing	PPV Diff	NPV Diff	FPR Diff	FNR Diff	High Risk Diff
Logistic Regression	Pre	8.93	-7.01	-8.10	-3.32	4.11
	Post	10.44	-2.66	-9.27	-7.34	6.70
Penalised Logistic Regression	Pre	6.36	-9.14	-14.83	0.69	1.51
	Post	8.40	-7.06	-3.96	-8.60	8.50
Random Forest	Pre	-1.46	17.29	38.50	-36.16	39.16
	Post	11.67	-3.65	-6.36	-5.50	4.36
Stochastic Gradient Boosting	Pre	-2.27	5.66	38.57	-29.45	34.39
	Post	4.65	-13.12	3.40	-2.56	5.96
Support Vector Machine	Pre	2.81	-15.29	12.90	-2.31	6.54
	Post	6.11	-7.94	7.15	-8.40	10.21

For each of these fairness definitions, the non-Aboriginal and Torres Strait Islander value was subtracted from the Aboriginal and Torres Strait Islander value. For predictive parity, error rate balance, and statistical parity values by group, please refer to the supplementary materials. Pre- and post-processing had mixed effects on the predictive parity and error rate balance of the LS/RNR items. Specifically, PPV and NPV disparities were mildly increased for the majority of algorithms. However, a number of the NPV differences were smaller than the NPV disparity identified for the LS/RNR total risk score. For error rate balance metrics, FPR and FNR disparities were, for the most part, improved. Only for random forest and stochastic gradient boosting algorithms with pre-processing were the disparities notably increased. Furthermore, for FPR, the disparities increased for logistic regression and penalised logistic regression (pre-processing only). However, these disparities were still smaller when compared to the original LS/RNR total risk score. Last, the statistical parity as assessed by the difference of proportions classified as high risk was improved for all algorithms besides the tree-based models using pre-processing.

Discussion

The present study explored the use of statistical learning methods with LS/RNR items as a way to increase the discrimination and fairness of the LS/RNR instrument. In line with previous findings (Liu et al., 2011), a number of the statistical learning methods did not bring about notable improvements in discrimination over the LS/RNR total score. Some statistical learning methods did demonstrate a significant improvement. Specifically, the biggest improvement was found using penalised logistic regression and stochastic gradient boosting algorithms for the overall sample, with the former performing the best for Aboriginal and Torres Strait Islanders and the latter performing the best for non-Aboriginal and Torres Strait Islanders. Statistical learning methods also often demonstrated an increase in fairness across groups, whether using just LS/RNR items or using pre- or post-processing techniques. Specifically, the stochastic gradient boosting algorithm without any processing techniques was found to decrease discrepancies between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on xAUC, statistical parity, PPV, and error rate balance, whilst improving discrimination overall and for both groups.

The pre-processing approach of using residuals in place of original predictors often led to similar 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 techniques. For example, a support vector machine algorithm that relied on residuals as the predictors suffered a slight drop in AUC overall when compared to the original support vector machine without processing, however, it was improved when compared to the original LS/RNR total risk score. Further, this approach produced almost perfect parity between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on the xAUC. Additionally, closer error rate balance, statistical parity, and calibration estimates (albeit a drop in overall calibration) between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders were also found.

Tree-based models (i.e., random forest and stochastic gradient boosting), however, were found to increase disparities when using residuals. Specifically, xAUC differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders were larger than the tree-based models without processing approaches and the LS/RNR total risk score. Tree-based models employing residuals effectively discriminated Aboriginal and Torres Strait Islander recidivists from non-Aboriginal and Torres Strait Islander non-recidivists. However, they were unable to discriminate between Aboriginal and Torres Strait Islanders who did not engage in recidivism from non-Aboriginal and Torres Strait Islanders who did. Calibration, error rate balance, and statistical parity were also negatively impacted, with wider disparities for these fairness definitions being reported between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The reject option based classification post-processing technique was discovered to be a better approach in reducing disparities for the tree-based algorithms. Specifically, a stochastic gradient boosting algorithm using post-processing led to similar AUC estimates (overall and for each group) to a stochastic gradient boosting algorithm without processing. This, however, reduced the FPR differential between Aboriginal and Torres Straits Islanders and non-Aboriginal and Torres Strait Islanders to only 3.4%.

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 the majority of discrepancies being 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, namely positive predictive values, was often negatively impacted. The increase in disparities among predictive parity was often smaller than the decrease in disparities identified among 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 in FPR discrepancies when using pre-processing 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 prioritised.

The trade-off between discrimination and fairness was not observed in the present study. 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). Instead, calibration, as assessed by Brier scores, which can also be understood as the accuracy of probabilistic predictions, was found to be negatively impeded for the majority of algorithms when using residuals to increase fairness.

Another issue surrounds the loss of interpretability that a number of these statistical learning methods and processing approaches will have resulted in. Unlike the original LS/RNR in which the algorithm to score total risk is known, or approaches like linear regression in which the individual importance of predictors can be understood through beta weights, more complex approaches and/or the processing approaches will have impeded on the transparency of the algorithm.

Limitations

The present study was limited by the sample in a number of ways. First, the sample included those who were originally sentenced to a term of imprisonment for a serious violent offence as outlined by the *Sentencing Act* 1991 (Vic). As a result, the individuals in this study may not be representative of 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 offence. The current sample is therefore not reflective of the wider male prison population nor the serious violent offender population.

Third, the non-Aboriginal and Torres Strait Islander group were unable to be portioned into more distinct groups as the vast majority self-reported that they were born in Australia (n = 166, 83%), and all identified their primary language as English. Australia is a multi-cultural society (Australian Bureau of Statistics, 2020a) and including all of those who do not identify as Aboriginal and/or 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 could also further impede the accuracy and generalisability of the findings. Last, the sample size also posed limitations. The statistical learning methods employed in the present study produce the best estimates for large sample sizes. Although cross-validation approaches were used to mitigate the small sample size, wide estimates were reported across the 10 validation folds. AUC has also been cautioned against with sample sizes of less than 200, as this can result in large inaccuracies (Hanczar et al., 2010).

The present study was further limited by relying on a cut-off in order to easily enable predictive parity, error rate balance, and statistical parity to be calculated. There are numerous ways to determine a cut-off value (Kuhn & Johnson, 2013), with no agreed upon approach to effectively determine which is the best method. Further, using a cut-off to enable the calculation of these fairness metrics is limited as different cut-offs will produce different fairness results (Zottola et al., 2021). Last, although the present study was able to demonstrate the potential usefulness of certain statistical learning methods, the transparency of the algorithms is less certain than traditional approaches. The importance of the predictors (i.e., LS/RNR items) in these algorithms is relatively unknown and therefore poses a real-world issue for corrections officers who make decisions for individuals based on risk assessment factors regarding treatment and rehabilitation in order to mitigate future risk.

Implications

Future research should gather larger, more representative samples to better examine the generalisability of the present findings and how estimates may vary across cultural groups. Further, fairness metrics could be observed across a range of cut-off values to gain a better understanding of the discrepancies that are not limited to a single threshold. Last, the exploration of post hoc analyses such as local interpretable model-agnostic explanations (Ribeiro et al., 2016) and Shapley values (Shapley, 1953) on these statistical learning methods could help counteract the trade-off between the performance of more complex statistical learning methods and interpretability. These approaches could help provide an understanding for corrections officers as to which risk factors and items are the most important in predicting recidivism and therefore which are the most pivotal for treatment plans and rehabilitation. However, post-hoc approaches for increasing the interpretability of statistical learning methods need to have their limitations understood. For example, the interpretations are not always correct, and they may not make sense or provide enough information to understand how the statistical learning method arrived at a prediction (Rudin, 2019).

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, in particular, 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). This discussion requires thoughtful deliberation by policymakers as to what form of fairness is the most relevant to be satisfied for their specific risk assessment, jurisdiction, and cross-cultural fairness needs. For example, if the emphasis is placed on making 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 engage in recidivism, 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 engage in recidivism. Conversely, a higher proportion of non-Aboriginal and Torres Strait Islanders would be classified as low risk and later become recidivists.

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 support vector machine with pre-processing resulted in a slight increase in NPV disparities, error rate balance was noticeably improved such that the absolute mean difference for predictive parity metrics (PPV and NPV) was 9.01% and error rate balance metrics (FPR and FNR) was comparable at 7.61%.

The performance of the statistical learning methods in the present study also helped mitigate the potential loss that processing approaches to increase fairness may have had on the discrimination of the instrument. The ability to discriminate between individuals who went on to engage in recidivism from those who did not for the majority of statistical learning methods with processing approaches still outperformed the original LS/RNR total risk score. The loss in discrimination was mainly noted when compared to the same statistical learning method without processing approaches. This demonstrated that when processing approaches were applied, the loss of performance was not as great as what was initially gained from using these more complex statistical approaches when compared to the original risk assessment instrument.

Conclusion

The present 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 findings in reducing certain fairness disparities (primarily xAUC, error rate balance, and statistical parity) without overly impeding on the discrimination of the instrument and should therefore be continued to be explored moving forward as an approach to mitigate unfairness in risk assessment instruments.

References

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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. (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. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1-42. 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.

Brennan, T. (2016). *An alternative scientific paradigm for criminological risk assessment: Closed or open systems, or both?* Taylor & Francis Ltd.

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. http://dx.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/doi: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 (Clinical Research Ed.)*, *345*(7868). https://doi.org/10.1136/bmj.e4692

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and 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., & GBM Developers. (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.

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.

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2010). Level of Service Inventory–Revised: Assessing the risk and need characteristics of Australian Indigenous offenders. *Psychiatry, Psychology and Law*, *17*(3), 355-367. https://doi.org/10.1080/13218710903089261

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.

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

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and caucasian female and male adolescents on probation. *Psychological Assessment*. https://doi.org/10.1037/pas0000816

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. (2020). *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*, 206–215. https://doi.org/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? *The 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.

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/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*. https://doi.org/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/doi: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*. 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), 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 US.

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. https://doi.org/10.1177/0093854814552843

Zottola, S. A., Desmarais, S. L., Lowder, E. M., & Duhart Clarke, S. E. (2021). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*. https://doi.org/10.1177/00938548211040544

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, *67*, 301-320.

Supplementary Materials

The predictive parity (PPV and NPV), error rate balance (FPR and FNR), and statistical parity (high risk [HR] proportion) for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders for the LS/RNR total risk score and algorithms using the LS/RNR items without processing techniques are reported in Table S1. For Aboriginal and Torres Strait Islanders for the LS/RNR total risk score, all algorithms had a slightly higher PPV and were noticeably higher in the proportion classified as high risk. For the majority of algorithms and the LS/RNR risk score, Aboriginal and Torres Strait Islanders also had a higher FPR. NPVs and FNRs were found to be higher overall for non-Aboriginal and Torres Strait Islanders.

Table S1

Predictive Parity, Error Rate Balance, and Statistical Parity Values by Group

	Aboriginal and Torres Strait Islander					Non-Aboriginal and Torres Strait Islander				
	PPV	NPV	FPR	FNR	HR	PPV	NPV	FPR	FNR	HR
LS/RNR Total Risk Score	87.88	20.83	61.54	24.68	73.33	83.81	32.63	35.42	42.11	52.50
Logistic Regression	93.69	32.97	23.33	33.48	60.31	85.50	35.20	25.60	45.92	47.61
Penalised Logistic Regression	93.30	37.41	33.33	22.00	71.14	90.40	41.00	22.33	38.04	52.54
Random Forest	91.30	32.50	46.67	22.38	72.42	86.19	38.66	24.36	41.50	51.01
Stochastic Gradient Boosting	92.82	31.88	35.00	26.08	67.82	90.84	44.26	20.93	33.31	55.95
Support Vector Machine	91.57	33.69	46.67	19.78	74.68	90.84	42.26	24.83	34.29	55.86

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate; HR high risk.

The predictive parity, error rate balance, and statistical parity for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders for the algorithms with both processing techniques (pre and post) are reported in Table S2. Similar to the original algorithms and the LS/RNR total risk score, Aboriginal and Torres Strait Islanders were still found to have higher PPVs on the majority of algorithms with processing (besides the tree-based models) and a higher proportion classified as high risk for all algorithms. In contrast to the algorithms without processing, logistic regression, penalised logistic regression, and random forest (only with post-processing) instead produced a higher FPR for non-Aboriginal and Torres Strait

Islanders. NPVs and FNRs for the majority of algorithms remained higher for non-Aboriginal and Torres Strait Islanders.

Table S2

Predictive Parity, Error Rate Balance, and Statistical Parity Values by Group using Pre- and Post-Processing Techniques

		Aboriginal and Torres Strait Islander					Non-Aboriginal and Torres Strait Islander				
		PPV	NPV	FPR	FNR	HR	PPV	NPV	FPR	FNR	HR
Logistic Regression	Pre	94.54	30.15	20.00	38.45	55.49	85.61	37.16	28.10	41.77	51.38
	Post	93.44	30.52	23.33	37.92	56.49	83.00	33.18	32.60	45.26	49.79
Penalised Logistic Regression	Pre	95.43	37.86	15.00	28.73	63.34	89.07	47.00	29.83	28.04	61.83
	Post	93.02	30.71	33.33	27.44	66.41	84.62	37.77	37.29	36.04	57.91
Random Forest	Pre	90.80	55.83	52.50	12.83	81.63	92.26	38.54	14.00	48.99	42.47
	Post	92.06	29.81	38.33	31.91	63.06	80.39	33.46	44.69	37.41	58.70
Stochastic Gradient Boosting	Pre	91.15	48.17	50.00	13.69	80.63	93.42	42.51	11.43	43.14	46.24
	Post	92.34	28.90	35.00	29.25	65.31	87.69	42.02	31.60	31.81	59.35
Support Vector Machine	Pre	91.39	26.10	38.33	33.58	61.72	88.58	41.39	25.43	35.89	55.18
	Post	90.98	30.38	46.67	25.26	69.87	84.87	38.32	39.52	33.66	59.66

Note. PPV positive predictive value; NPV negative predictive value; FNR false negative rate; FPR false positive rate; HR high risk.

Chapter Six: Empirical Study Three

6.1 Introduction

The increased use of statistical learning methods within forensic risk assessment has been met with criticism due to the lack of interpretability of this approach. Unlike more traditional approaches to estimating risk, in which the relationship between predictors and the outcome is easily discernible, statistical learning methods can often result in an algorithm that is lacking in transparency. However, there are further analyses that can be conducted on statistical learning methods to help increase the interpretability and gain an understanding of the importance of predictors to the predicted outcome. For example, Shapley values can be calculated, which fairly distribute the difference between the average prediction for the statistical learning method and an individual's prediction across all predictors. This strategy provides a mathematical and fair way to improve the interpretability of statistical learning approaches. Nevertheless, there is limited literature on this approach for forensic risk assessment instruments.

This chapter presents the third empirical study that addresses the second research aim by responding to research question five. The second empirical study identified that a number of statistical learning methods were useful in increasing the discrimination of the LS/RNR total risk score. Further, processing approaches were also beneficial for increasing cross-cultural fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Therefore, this empirical study utilised Shapley values as a way to increase the interpretability of statistical learning methods that increased the discrimination and/or fairness of the LS/RNR total risk score. Specifically, penalised logistic regression, stochastic gradient boosting, and support vector machine algorithms were used alongside pre- and post-processing techniques that were used to further increase fairness. Shapley values were calculated for the entire sample, as well as for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders to compare important predictors (i.e., LS/RNR items) across the two groups. For ease of reporting, this empirical study reports the five highest mean absolute Shapley values for each statistical learning method.

Empirical Study Three is titled "*Increasing the Cross-Cultural Fairness of the LS/RNR and Interpretability of Statistical Learning Methods"* and has been submitted to *Psychiatry, Psychology, and Law* for publication. *Psychiatry, Psychology, and Law* is a peer-reviewed journal that explores research and practice developments in the disciplines of forensic psychiatry, criminology, behavioural science, and law. The Author Indication Form that details the contribution of each author to this manuscript is included in Appendix A.

Increasing the Cross-Cultural Fairness of the LS/RNR and Interpretability of Statistical Learning Methods

Linda J. Ashford[1], Benjamin L. Spivak[1], James R. P. Ogloff[1] & Stephane M. Shepherd[1]

[1] Centre for Forensic Behavioural Science, Swinburne University of Technology

Author Note.

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

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. We 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.

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

Abstract

Statistical learning methods have shown promise in increasing the cross-cultural fairness of risk assessment instruments; however, it has been argued that this comes at the cost of interpretability. This poses problems in practice, especially if risk estimates are used to aid in decision making. Recently, Shapley values have been used as an approach to increase the interpretability of statistical learning methods. The present study calculated Shapley values to improve the interpretability of statistical learning methods that were constructed to optimise discrimination and improve fairness of the Level of Service/Risk Needs Responsivity (LS/RNR), a widely used risk assessment instrument, for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Australia (*N* = 380). Statistical learning methods included penalised logistic regression, stochastic gradient boosting, and support vector machine algorithms using the LS/RNR items, with pre- and post-processing being applied to increase fairness. The area under the curve (AUC) was used to assess discrimination, and the cross area under the curve (xAUC), error rate balance, calibration, predictive parity, and statistical parity were used to assess fairness. Statistical learning methods were found to increase discrimination and fairness (primarily xAUC, error rate balance, and statistical parity) between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders compared to the LS/RNR total risk score. Shapley values revealed that current drug use, current unemployment, and criminal history were the biggest mean marginal contributors to the average prediction of recidivism and individual predictions of recidivism. The findings demonstrate that Shapley values offer a useful approach to increasing the interpretability of more opaque statistical learning methods that have increased the discrimination and/or fairness of risk assessment instruments.

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

Introduction

The past decade has seen the increase of statistical learning methods (i.e., machine learning) within risk assessment (Spivak & Shepherd, 2020). In the criminal justice system, risk assessment instruments are used to estimate the risk of an individual engaging in recidivism, with modern risk instruments often classified as either actuarial risk instruments or structured professional judgement guides (SPJ; Yang et al., 2010). Actuarial instruments use a formula to combine present risk factors (e.g., criminal history, education and employment, and antisocial personality) that are empirically related to recidivism, and SPJ instruments provide guidelines to help clinicians determine an individual's level of risk based on present risk factors (Singh, 2012). The use of statistical learning methods within risk assessment has predominantly been used to increase the predictive accuracy and/or discrimination (i.e., the ability to differentiate an individual who goes on to engage in recidivism from an individual who does not) of risk assessment instruments. This is due to statistical learning methods prioritising predictive accuracy and being an effective approach for incorporating a large number of predictors without the need to pre-specify relationships (e.g., interactions between predictor variables; Breiman, 2001b).

Fairness in Risk Assessment Instruments

When it comes to risk assessment, fairness broadly relates to a risk assessment instrument performing equally across different groups (Verma & Rubin, 2018). More recently, statistical learning methods have also been used as a method to increase the fairness of risk assessment instruments between cultural minorities (e.g., African Americans and Indigenous populations of North America) and cultural majorities (Berk et al., 2018). Processing approaches which alter the statistical learning method algorithm at various stages of construction (i.e., pre, in, or post-processing) can be used to increase fairness. Fairness, which can encompass numerous statistical definitions, is an ongoing problem within risk assessment instruments. Error rate balance, or parity between groups among errors in observation, has not been satisfied in cross-cultural studies, with cultural minorities being classified as high risk and not going on to engage in recidivism more often, and cultural majorities being classified as low risk and engaging in recidivism more frequently (Angwin et al., 2016; Flores et al., 2016; Larson et al., 2016; Whiteacre, 2006).

Predictive parity, or parity between groups among classification predictions, has also been found to show disparities between cultural groups. High risk classifications are found to be more accurate for cultural minorities and low risk classifications are more accurate for cultural majorities (Muir et al., 2020; Shepherd et al., 2015; Whiteacre, 2006). Similarly, calibration disparities, in which risk scores do not reflect the same probability of recidivism, have also been found, primarily among Indigenous and non-Indigenous groups from North America and Australia (Thompson & McGrath, 2012; Wilson & Gutierrez, 2014).

Last, significant scoring differences between cultural minorities and cultural majorities have also been reported, with cultural minorities scoring significantly higher on total risk scores and risk factors that often reflect greater levels of social and economic disparity experienced by cultural minorities (Day et al., 2018; Olver et al., 2014; Shepherd, Adams, et al., 2014; Smallbone & Rallings, 2013). Violating these definitions of fairness could disadvantage certain cultural groupings as risk assessment instruments are sometimes used to inform decision making around sentencing, bail, treatment, and supervision (Monahan & Skeem, 2016). For example, risk scores or classifications that do not align with an observed outcome of recidivism could result in improper treatment and rehabilitation approaches, such as unnecessary restrictions or a lack of intervention that could have helped mitigate future risk.

Transparency of Statistical Learning Methods

Although the use of statistical learning methods is still emerging within this field, initial studies have demonstrated promise in increasing the predictive accuracy, discrimination, and cross-cultural fairness of risk assessment instruments (Lum & Johndrow, 2016; Skeem & Lowenkamp, 2020; Ting et al., 2018; Wadsworth et al., 2018). However, statistical learning methods have been criticised due to a lack of transparency. More complex statistical learning methods, or the use of processing approaches, can result in an algorithm in which the relationship between the predictors and the outcome is relatively unknown.

In the case of risk assessment instruments, this translates to an inability to understand the relationship between risk items and/or factors and the predicted outcome of recidivism. As a result, statistical learning methods with limited transparency have been cautioned for use in high-stakes decision making, such as in the criminal justice system (Rudin et al., 2020). For example, a lack of transparency can result in incomplete or erroneous representations of how the statistical learning method uses predictors to predict an outcome, and errors (e.g., data input errors) can more easily go undiscovered. A lack of transparency also raises a larger ethical concern. As Rudin et al. (2020) state, transparent approaches to risk estimation are more accountable and allow the public to critique the methodology and calculations. Since the inner workings of statistical learning methods can be unknown, scepticism about the potential bias within these processes can arise (Wisser, 2019).

However, Berk (2021) emphasised that the human brain is also not a transparent process, with it being difficult to comprehend how an individual in the criminal justice system (e.g., magistrates, parole boards) makes high-stakes decisions. Further, while risk assessment instruments can influence decision making, they are not the sole reason for a decision and are only one piece of information that is considered (Chouldechova, 2020). There is also evidence that less transparent statistical learning methods can improve a risk assessment instrument's discrimination and/or fairness when compared to transparent approaches (e.g., Johndrow & Lum, 2017; Salo et al., 2019; Ting et al., 2018; Wadsworth et al., 2018). Even if there is a loss of transparency, further research into statistical learning methods is warranted, especially since there are additional analyses that can aid in increasing transparency. For example, researchers have begun to use Shapley values (Shapley, 1953), a concept from cooperative game theory, in which an outcome can be fairly distributed among a coalition of players. Applied to statistical learning methods, this translates to a prediction fairly being distributed among the predictors, therefore enabling some understanding of the predicted value.

Here, Shapley values can be used as a tool to gain an understanding of the importance of predictors to the predicted outcome. This approach has been scarcely applied to risk assessment instruments. Limited studies have utilised Shapley values, or similar approaches to Shapley values such as Shapley additive explanations (SHAP; see Lundberg & Lee, 2017), to explore predictor importance in recidivism predictions, with criminal history and age being important predictors (Bowen & Ungar, 2020; Kaponen, 2020). However, the use of Shapley values is still new within this area, with limited research using this as an approach to explore statistical learning methods that have been useful in increasing the discrimination of a risk assessment instrument and/or ameliorating violations of fairness definitions.

Present Study

In Australia, disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on risk assessment instruments have been documented in the literature. For example, Aboriginal and Torres Strait Islanders are often found to have significantly higher risk scores, and risk assessment instruments are less effective at discriminating recidivists from non-recidivists for this group (Shepherd et al., 2015; Shepherd

& Strand, 2016; Smallbone & Rallings, 2013). Aboriginal and Torres Strait Islanders are also found to experience further disadvantage within the criminal justice system including over-incarceration and a decreased likelihood of receiving diversion (Australian Bureau of Statistics, 2020c; Warner et al., 2021). Therefore, the present study aimed to initially use statistical learning methods to increase the discrimination and fairness of the Level of Service/Risk Need Responsivity (LS/RNR; Andrews et al., 2008) for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in Victoria, Australia. Additionally, the present study also aimed to use Shapley values to increase the interpretability of statistical learning methods so that important LS/RNR items could be identified, thereby overcoming transparency difficulties and making statistical learning methods more usable in practice.

Method

Sample

The sample for the present study included 380 males who had previously been sentenced to a term of imprisonment for a serious violent offence as defined in schedule 1 (clause 3) of the *Sentencing Act 1991* (Vic) and were received into prison between January 2015 and December 2017. These individuals were assessed with the LS/RNR by a corrections officer after receiving at least a medium risk classification on the Level of Service Inventory-Revised: Screening Version (Andrews & Bonta, 1998). The sample included 180 (47.37%) individuals who identified as Aboriginal and/or Torres Strait Islander people, and the remaining 200 (52.63%) individuals identified as non-Aboriginal and Torres Strait Islanders. LS/RNR assessment and demographic information were provided by Corrections Victoria, with any charges post assessment (within the period of January 2015 to December 2019) being obtained from the Victorian Police Law Enforcement Assistance Program (LEAP) database. Ethics approval for the present study was obtained from the Department of Justice and

Community Safety (Victoria) Human Research Ethics Committee and Swinburne University Human Research Ethics Committee.

Measures

Level of Service/Risk Needs Responsivity

The LS/RNR (Andrews et al., 2008) is an actuarial risk assessment instrument developed to ascertain an individual's criminogenic needs and estimate their future risk of general recidivism. The General Risk/Needs section of the instrument consists of the Central Eight risk domains (Criminal History, Education/Employment, Family/Marital, Leisure/Recreation, Companions, Alcohol/Drug Problem, Procriminal Attitude, and Antisocial Pattern), which sum to a total possible score of 43. These items result in a score of either 0 when absent or 1 when present and are summed up to create a total risk score. Individuals can be categorised into risk levels based on their total score – very low risk (0–4), low risk (5–10), medium risk (11–19), high risk (20–29), and very high risk (30–43).

Recidivism

Recidivism in the present study was assessed as a police charge that occurred while at risk in the community (i.e., not during a period of incarceration). The average follow up time (i.e., time at risk in the community) from the LS/RNR assessment (or from prison release date for those who were incarcerated) to either a charge (for those who were recidivists) or the end of the follow up period (31-12-2019) for those who were not recidivists was 280.56 days (*SD* = 329.24). For those individuals who engaged in recidivism, the average time to first offence post LS/RNR assessment was 184.75 days (*SD* = 233.80). Recidivism base rates differed between groups, with more Aboriginal and Torres Strait Islanders engaging in recidivism compared to non-Aboriginal and Torres Strait Islanders (85.56% and 76%, respectively) by the end of the follow up period.

Analytical Approach

The present study is an extension of a previous empirical study which explored the use of statistical learning methods and processing approaches to increase discrimination and/or fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Statistical learning methods which aided in increasing the discrimination and/or fairness of the LS/RNR within Ashford et al. (2022) are reported and utilised within this study to explore the importance of predictors.

All analyses for the present study were conducted through RStudio using R version 4.0.2 (R Core Team, 2021). A suite of statistical packages were utilized, including *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 area under the curve (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 penalised logistic regression, *gbm* (Version 2.1.8; Greenwell et al., 2020) for stochastic gradient boosting, *e1071* (Verision 1.7-8; Meyer et al., 2021) for support vector machine algorithms, *cutpointr* (Version 1.1.1; Thiele, 2021) to generate optimal cut-offs, and *iml* (Version 0.10.1; Molnar, 2020) to calculate Shapley Values.

Discrimination

Discrimination was assessed through the AUC. The AUC provides an index of the sensitivity and 1 – specificity of a risk assessment instrument across numerous thresholds and ranges from 0 to 1, with 0.5 reflecting chance levels of discrimination (Cook, 2007; Rice & Harris, 2005). The AUC is best understood as the probability that a randomly selected individual who engages in recidivism will have received a higher risk score than a randomly selected individual who did not engage in recidivism.

Fairness

Cross Area under the Curve (xAUC). The xAUC is an alteration to the traditional AUC that instead measures discrimination between groups instead of within (Kallus & Zhou, 2019). The traditional AUC compares individuals who are recidivists to individuals who are not recidivists from within one group (e.g., Aboriginal and Torres Strait Islanders), demonstrating the ability of a risk assessment instrument to discriminate recidivists from non-recidivists within that one group. The xAUC is instead calculated across two sets of groups. Set 1 contains a positive outcome from one group and the negative outcome from the other. In the present study, Set 1 refers to Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists. The second set is the opposite of the first. Therefore, Set 2 in the present study is comprised of Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists. Similar to the AUC, the xAUC is best understood as the probability that a randomly selected individual from one group who goes on to be a recidivist receives a higher risk score than a randomly selected individual from the other group who is not a recidivist.

Calibration. Calibration was assessed through the alignment between the predicted probability of recidivism and the recidivism outcome. This was achieved through Brier Scores (Brier, 1950), which measure the squared error between the predicted probability and predicted outcome. Brier scores can range between 0 and 1, with the best possible Brier score being 0. As a Brier score relies on predicted probability, it was unable to be determined for the LS/RNR total risk score. Therefore, the items of the LS/RNR were used in a logistic regression algorithm to develop predicted probabilities, and this was used to report the calibration of the LS/RNR total risk score.

Predictive Parity. Predictive parity was assessed by calculating the positive predictive value (PPV) and negative predictive value (NPV). PPV refers to the proportion of individuals who were classified as high risk who went on to engage in recidivism. NPV refers to the proportion of individuals who were classified as low risk who did not go on to engage in recidivism.

Error Rate Balance. Error rate balance was assessed by calculating the false positive rate (FPR) and the false negative rate (FNR). The FPR refers to the proportion of individuals who were classified as high risk and did not go on to engage in recidivism by the end of the follow up period. The FNR refers to the proportion of individuals who were classified as low risk and went on to engage in recidivism.

Statistical Parity. Statistical parity is a measure of scoring differences between groups. This was assessed by calculating the proportion of individuals who were classified as high risk in each group.

Cut-Offs. In order to calculate predictive parity, error rate balance, and statistical parity, a cut-off was required in order to distinguish high risk from low risk. For the present study, the cut-off was determined by the value that yielded the smallest distance to the point 0, 1 on the receiver operating characteristic (ROC) space. The ROC space is used to calculate the AUC, and an instrument with perfect dissemination passes through 0, 1 in the ROC space. This was therefore chosen as the cut-off as it was an approach that prioritised discrimination. A cut-off value was established for the LS/RNR total risk score and for each statistical learning method.

Statistical Learning Methods

The statistical learning methods used in the present study were penalised logistic regression, stochastic gradient boosting, and support vector machines.

Penalised Logistic Regression. Penalised logistic regression, specifically elastic net penalised logistic regression (Zou & Hastie, 2005), was used as a statistical learning method to increase the simplicity and predictive power of standard logistic regression by reducing overfitting, the number of irrelevant predictors, and the impact of collinearity. Elastic net regression utilises a parameter that mixes the types of penalty (i.e., ridge and lasso penalties). This parameter ranges from 0 to 1, with 0 reflecting pure ridge regression and 1 reflecting pure lasso regression. The penalty for ridge regression is imposed on the squared size of coefficients and shrinks the coefficients of irrelevant predictors closer towards zero. The penalty for lasso regression is imposed on the absolute value of coefficients and shrinks the coefficients of irrelevant predictors completely to zero.

Stochastic Gradient Boosting. Stochastic gradient boosting (Friedman, 2002) is a statistical learning method that involves a weak learner (e.g., a decision tree) being repeatedly applied to the data. This approach is a consecutive learner that aims to find an additive algorithm that minimises the loss function (e.g., the squared error). A decision tree is grown to fit the residuals between the predicted and observed values using a subsample of the data for a specified number of iterations. At each step, the predicted values are updated by adding the newly predicted values to the previous predicted values, with new decision trees grown to fit the residuals of previous learners. The final prediction is based on the ensemble of decision trees; however, each tree's contribution to the outcome is not equal. The decision trees are weighted depending on their performance, which determines their influence over the final prediction.

Support Vector Machine. Support vector machines (Vapnik, 1999) are a statistical learning method that creates a flat boundary known as a hyperplane between data points. For a binary outcome (i.e., a classification with two outcomes), the hyperplane is used to divide and create the greatest separation between the data points from each outcome class. The data points that fall on either side of this hyperplane are then able to be classified as one outcome or the other. Often, data points are unable to be easily separated by a hyperplane in two dimensions, with kernels being used to transform the data into a higher dimension, enabling separation between the classes. Multiple kernels were trialled in the present study, with non-linear polynomial kernels being utilised as they produce the highest performance (i.e., AUC value).

Cross-Validation. To account for the small sample size of the present study and to avoid overfitting (Kuhn & Johnson, 2013), the statistical learning methods were validated using k-fold cross validation with 10 folds. Each of the 10 folds served as a validation set for the remaining 90% of the data that was used to train the statistical learning method. Multiple parameter options for each statistical learning method were also trialled during the validation process, with the final statistical learning method being based on the parameters that produced the highest level of discrimination (i.e., AUC value). The majority of the present sample engaged in general recidivism by the end of the follow up period (i.e., 31-12-2019), resulting in imbalanced outcome data (*n* = 306, 80.53%). Sampling with replacement (i.e., upsampling) was therefore used with the training data for the minority outcome class (i.e., non-recidivists). The performance of each statistical learning method was aggregated across the 10 folds and summarized to determine average performance.

Processing Approaches

Two processing techniques were utilised in the present study as an approach to increase fairness. The first involves altering the data before training the statistical learning method (i.e., a pre-processing approach). To remove the association between predictor variables and an individual's Aboriginal and Torres Strait Islander status, each predictor was regressed onto Aboriginal and Torres Strait Islander status with the residual replacing the original predictor value. The second approach involved reassigning the outcomes based on their proximity to the cut-off value. This is a post-processing approach known as reject option based classification (Kamiran et al., 2012) and involves specifying a margin around the cut-off value. The group that engaged in recidivism more and had their prediction fall within the cut-off value and upper margin were reclassified as non-recidivists. Conversely, the group that engaged in recidivism less and had their prediction fall within the cut-off value and lower margin were reclassified as recidivists. This aids in increasing parity between groups on the predicted outcomes. Multiple margins around the cut-off value were trailed for each statistical learning method, with the margin resulting in an increase in fairness with the lowest level of loss to discrimination being chosen. The cut-off value for each statistical learning method was the same as what was utilised to calculate the fairness definitions.

Shapley Values

Shapley values (Shapley, 1953) is a game theory concept that focuses on the idea that a prediction can be fairly attributed to a group of features (Lundberg & Lee, 2017). For the present study, the Shapley value for a feature (i.e., a LS/RNR item) can be understood as the mean marginal contribution of that feature, across all possible groups of features, to the difference between the observed prediction for that individual and the mean prediction (Molnar, 2019). This provided an understanding of each feature's importance in the deviance of an individual's prediction from the mean prediction. The Shapley value for all LS/RNR items was calculated for all individuals. For global reporting, the absolute mean Shapley value was calculated, and the top five important LS/RNR items for each statistical learning method were reported for the overall sample, as well as for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

Results

Discrimination

Discrimination was established for the overall sample, Aboriginal and Torres Strait Islanders, and non-Aboriginal and Torres Strait Islanders. The discrimination for the LS/RNR total risk score and the statistical learning methods which notably increase discrimination and/or fairness compared to the LS/RNR total risk score are presented in Table 1. Please refer to Ashford et al. (2022) for extended results and discussion.

Table 1

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

	Overall		Aboriginal and Torres Strait Islander		Non-Aboriginal and Torres Strait Islander	
	AUC	Range	AUC	Range	AUC	Range
LS/RNR Total Risk Score	.64	.57-.70	.60	.49-.70	.63	.55-.72
Penalised Logistic Regression	.73	.53-.85	.80	.56-.94	.66	.49-.82
Stochastic Gradient Boosting	.73	.59-.88	.71	.53-.87	.73	.57-.92
Stochastic Gradient Boosting (post-processing)	.73	.57-.86	.71	.53-.87	.72	.54-.87
Support Vector Machine	.70	.47-.80	.67	.42-.97	.64	.49-.86
Support Vector Machine (pre-processing)	.68	.51-.82	.68	.37-.94	.64	.44-.86

For the LS/RNR total risk score, the range refers to the 95% confidence interval. For the remaining statistical learning methods, range refers to the range of discrimination estimates from the 10 folds. Penalised logistic regression, stochastic gradient boosting, and support vector machines using the LS/RNR items all led to an increase in discrimination for the overall sample, and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Further, processing approaches to stochastic gradient boosting and support vector machine algorithms also led to increases in discrimination compared to the LS/RNR total risk score. The support vector machine using pre-processing (i.e., residuals) resulted in only a minor loss in discrimination for the overall sample when compared to the support vector machine without processing. Similarly, the stochastic gradient boosting algorithm using post-processing (i.e., reclassification of outcomes) had comparable levels of discrimination to the stochastic gradient boosting algorithm without processing.

For the penalised logistic regression and support vector machine algorithms, higher levels of discrimination were found for Aboriginal and Torres Strait Islanders. Conversely, the LS/RNR total risk score and stochastic gradient boosting algorithms had higher levels of discrimination for non-Aboriginal and Torres Strait Islanders. However, only the penalised logistic regression disparity between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders was notable.

Fairness

xAUC

xAUC was established for both Set 1 (Aboriginal and Torres Strait Islander recidivists and non-Aboriginal and Torres Strait Islander non-recidivists) and Set 2 (Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists) and are reported in Table 2. The reported range refers to the range of xAUC values obtained from the 10 folds in validation. The xAUC differences highlight that the LS/RNR and the majority of statistical learning methods were more efficient at discriminating Aboriginal and Torres Strait Islander recidivists from non-Aboriginal and Torres Strait Islander non-recidivists (i.e., Set 1).

Table 2

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

	Set 1		Set 2	
	xAUC	Range	xAUC	Range
LS/RNR Total Risk Score	.75	.68-.83	.46	.35-.57
Penalised Logistic Regression	.79	.70-.90	.64	.42-.85
Stochastic Gradient Boosting	.79	.71-.95	.62	.25-.88
Stochastic Gradient Boosting (post-processing)	.79	.69-.95	.63	.25-.88
Support Vector Machine	.81	.64-.92	.56	.25-.91
Support Vector Machine (pre-processing)	.67	.52-.84	.68	.33-.94

The LS/RNR total risk score was unable to effectively discriminate Aboriginal and Torres Strait Islander non-recidivists from non-Aboriginal and Torres Strait Islander recidivists (i.e., Set 2). Discrimination among Set 2 was improved when using statistical learning methods and processing approaches. Further, the gap between the two sets was reduced when using statistical learning methods when compared to the LS/RNR total risk score. When using pre-processing with a support vector machine algorithm, this difference was reduced to almost zero.

Calibration, Predictive Parity, Error Rate Balance, and Statistical Parity

The remaining fairness metrics were established across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders by assessing the calibration, predictive parity (i.e., PPV and NPV), error rate balance (i.e., FPR and FNR), and statistical parity (i.e., proportion classified as high risk). Non-Aboriginal and Torres Strait Islander values on these metrics were subtracted from Aboriginal and Torres Strait Islander values, and the differences between the groups are reported in Table 3. For the values reported by each group for all fairness measures, refer to the supplementary materials.

Table 3

Calibration, Predictive Parity, Error Rate Balance, and Statistical Parity Differences between Groups

	Calibration	PPV	NPV	FPR	FNR	High Risk
LS/RNR Total Risk Score	-.09	4.07	-11.8	26.12	-17.43	20.83
Penalised Logistic Regression	-.04	2.90	-3.59	11.00	-16.04	18.60
Stochastic Gradient Boosting	-.02	1.98	-12.38	14.07	-7.23	11.87
Stochastic Gradient Boosting (post-processing)	-.02	4.65	-13.12	3.40	-2.56	5.96
Support Vector Machine	-.06	0.73	-8.57	21.84	-14.51	18.82
Support Vector Machine (pre-processing)	-.02	2.81	-15.29	12.90	-2.31	6.54

Across the LS/RNR total risk score and statistical learning methods, Aboriginal and Torres Strait Islander predictions were better calibrated to the outcome. Aboriginal and Torres Strait Islanders also had higher PPVs, higher FPRs, and a higher proportion classified as high risk. The largest differences were often found for the LS/RNR total risk score, especially among calibration, FPR, FNR, and high risk proportion estimates. Statistical learning methods reduced the calibration, FPR, FNR, and high risk proportion disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders when compared to the LS/RNR total risk score. Further, the majority of statistical learning methods also reduced PPV disparities, with only penalised logistic regression and support vector machine algorithms reducing NPV disparities compared to the LS/RNR total score. The processing approaches applied to the stochastic gradient boosting and support vector machine algorithms aided in further reducing FPR, FNR, and high risk proportion differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

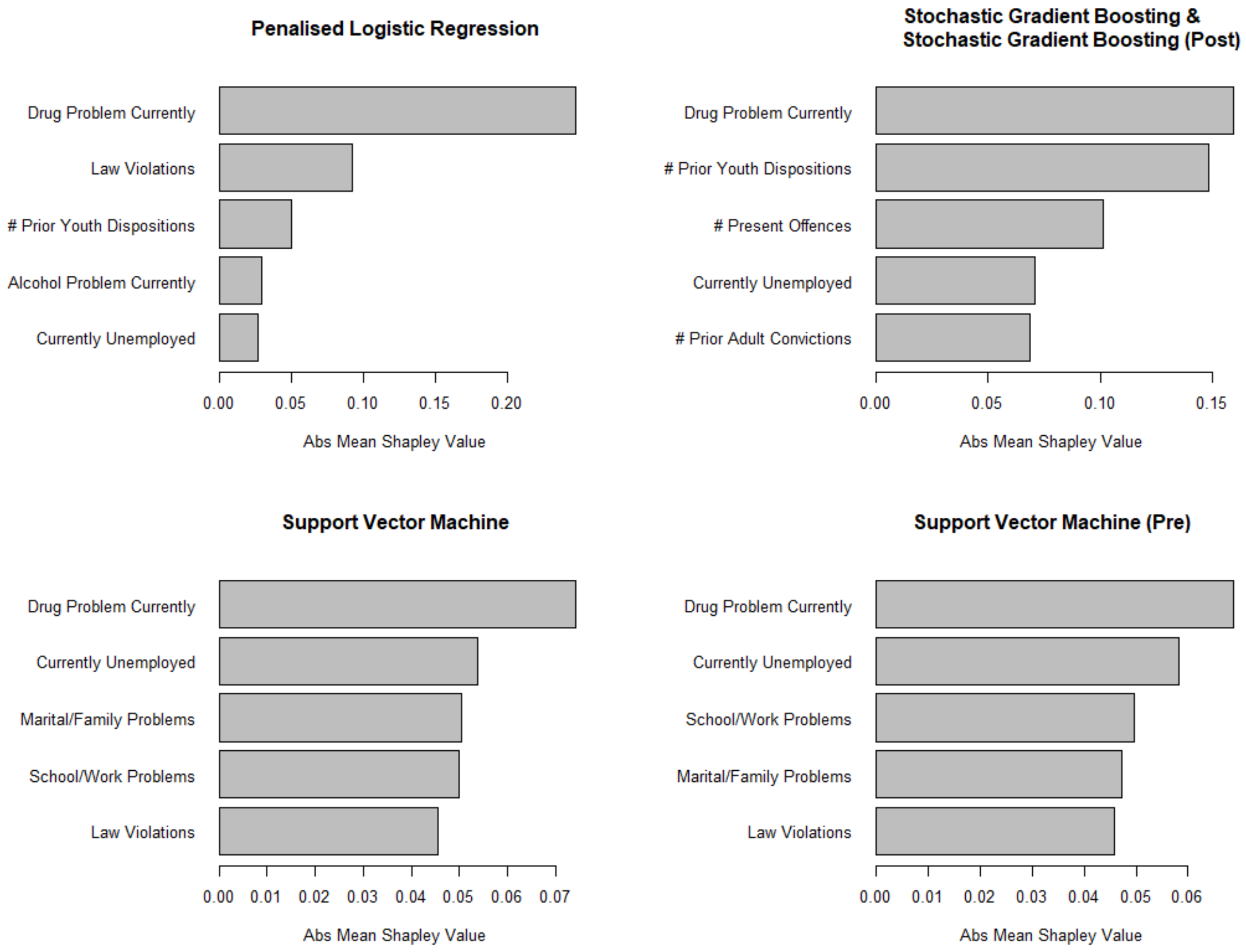
Shapley Values

Overall

Shapley values for the LS/RNR items were calculated for all individuals. This was determined for penalised logistic regression, stochastic gradient boosting, support vector machine, and support vector machine with pre-processing. Shapley values were not calculated for the stochastic gradient boosting algorithm with post-processing as this processing approach made no changes to the Shapley values. The Shapley values for the whole sample were aggregated, and the top five mean absolute Shapley values are presented in Figure 1.

Figure 1

Absolute Mean Shapley Values of the Top Five LS/RNR Items for Overall Sample



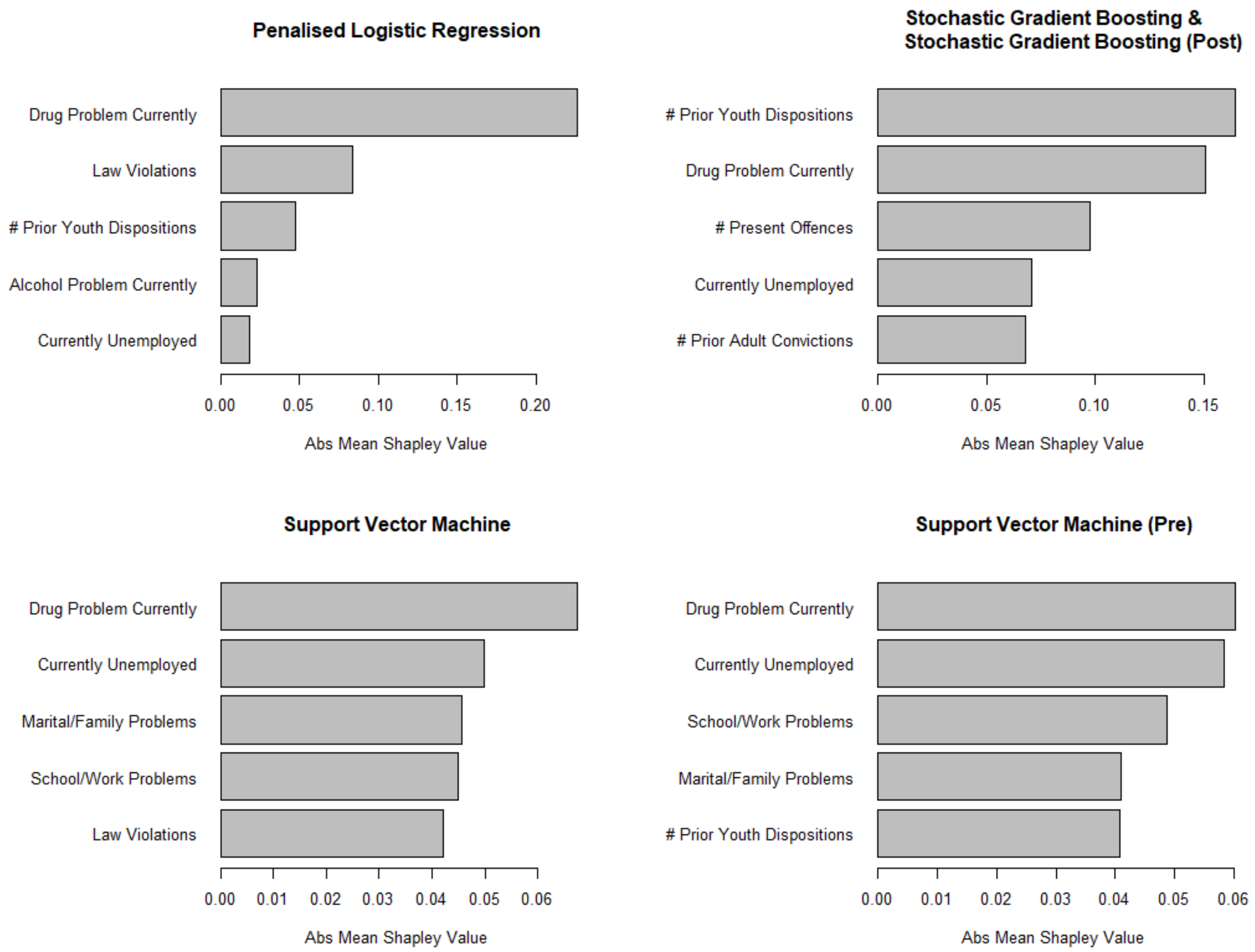
Across all statistical learning methods, current drug use was the most important LS/RNR item in predicting recidivism. Current unemployment was another LS/RNR item that was in the top five important predictors across all statistical learning methods. The other predictors varied across statistical learning methods. The stochastic gradient boosting algorithm mainly had LS/RNR items that related to current and previous criminal activity, whereas the support vector machine algorithms had LS/RNR items that related to current drug and/or alcohol use that resulted in external problems (e.g., family, employment, or law). The remaining items in the penalised logistic regression were items that measured criminal history or current drug and/or alcohol use.

Aboriginal and Torres Strait Islanders

Shapley values were also calculated for Aboriginal and Torres Strait Islanders, and the top five absolute average Shapley values for each statistical learning method are presented in Figure 2.

Figure 2

Absolute Mean Shapley Values of the Top Five LS/RNR Items for Aboriginal and Torres Strait Islanders



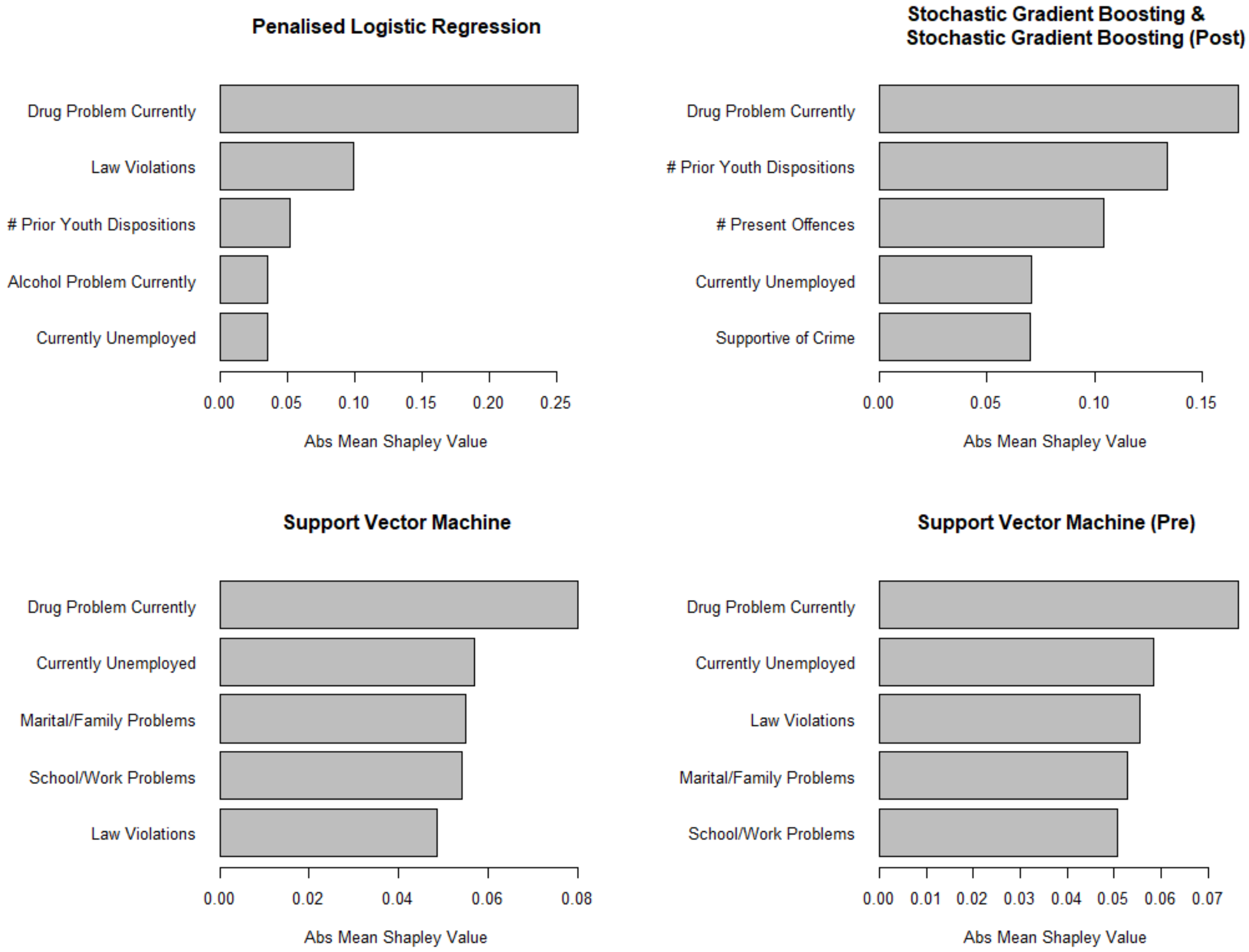
Similar to the overall sample, current drug use and current unemployment were important LS/RNR items in the prediction of recidivism for all statistical learning methods. However, for the stochastic gradient boosting algorithm, the number of prior youth dispositions was a marginally more important predictor than current drug use. The number of prior youth dispositions was also an important predictor for the support vector machine using residuals. The remaining top predictors for the support vector machine algorithms were similar to those for the overall sample, with predictors relating to the problems arising from current drug and/or alcohol use. The predictors and the level of importance of predictors in the penalised logistic regression for Aboriginal and Torres Strait Islanders were the same as those for the overall sample.

Non-Aboriginal and Torres Strait Islanders

Last, the Shapley values were calculated for non-Aboriginal and Torres Strait Islanders, and the top five absolute average Shapley values for each statistical learning method are presented in Figure 3. Consistent with the Aboriginal and Torres Strait Islanders, current drug problems and current unemployment were in the top five important predictors for all statistical learning methods. However, unlike Aboriginal and Torres Strait Islanders, current drug problems were the most important predictor across all algorithms. The stochastic gradient boosting algorithm again had predictors around current and previous criminal activity, as well as a predictor that measures an individual’s level of support for crime. Predictors around problems arising due to current drug and/or alcohol problems were again consistently found for support vector machine algorithms. Further, the predictors and their level of importance in the penalised logistic regression algorithm were the same as those found for Aboriginal and Torres Strait Islanders.

Figure 3

Absolute Mean Shapley Values of the Top Five LS/RNR Items for non-Aboriginal and Torres Strait Islanders



Discussion

The present study aimed to explore the use of Shapley values as an approach to increase the interpretability of statistical learning methods that aided in increasing the discrimination and/or fairness of the LS/RNR. In line with previous research (e.g., Lum & Johndrow, 2016; Wadsworth et al., 2018), the statistical learning methods used in the present study were found to be useful in increasing discrimination and cross-cultural fairness when compared to the LS/RNR total risk score. The increase in discrimination was found for the overall sample and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The statistical learning methods in the present study also demonstrated notable improvements in fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders across the majority of fairness definitions. The use of processing approaches to further increase the fairness of the LS/RNR resulted in only a minor loss in the instruments' ability to discriminate individuals who went on to engage in recidivism from those who did not.

Further, processing approaches were found to increase the fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on the definitions that had the highest levels of disparity. Specifically, Aboriginal and Torres Strait Islanders were more likely to be classified as high risk, classified as high risk and not go on to engage in recidivism (i.e., a higher FPR), and less likely to be classified as low risk and engage in recidivism (i.e., a lower FNR). Processing approaches applied to both stochastic gradient boosting and support vector machine algorithms were found to ameliorate these differences, with the support vector machine with pre-processing also being able to reduce xAUC differences to almost perfect parity.

Shapley Values

The important LS/RNR items in the present study were found to be relatively consistent across all statistical learning methods and for both Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders. Further, nearly all items found to be in the top five important predictors across the algorithms were from three of the Central Eight risk factors in the LS/RNR: Criminal History, Alcohol/Drug Problems, and Education/Employment. One reason why items from these risk factors were the most important predictors could be due to the higher variance that was found among these risk factor scores.

When examining the items specifically, the items related to criminal history and current drug use had higher variance compared to the other LS/RNR items. A higher variance in predictors can result in a more useful predictor, whereas predictors with zero or near-zero variance can often have very little effect on the outcome (Kuhn & Johnson, 2013). The following sections will further discuss why these LS/RNR items and risk factors specifically were useful in predicting the likelihood of recidivism.

Criminal History

In line with previous Shapley values research (Bowen & Ungar, 2020; Kaponen, 2020), criminal history-based items from the Criminal History risk factor in the LS/RNR (e.g., number of prior youth dispositions, number of present offences, number of prior adult convictions) were frequently listed in the penalised logistic regression and stochastic gradient boosting algorithms as important items. Criminal history has consistently been shown to be one of the most important predictors of future recidivism (Eisenberg et al., 2019; Wilson & Gutierrez, 2014). Research from Australia has demonstrated that prior offending in youth and also frequent prior offending in adulthood are linked to an increased risk of later offending (Payne, 2007). Further, the sample for the present study had all previously been sentenced to a term of imprisonment for a serious violent offence, with serious prior offending also being found to be related to increased rates of recidivism (Payne, 2007). The high risk nature of the present sample, therefore, may explain why criminal history-based items were important predictors. The entire sample had previous serious offences and the majority were recidivists by the end of the follow up period (i.e., 31-12-2019).

Drug Problems

Current drug use from the Alcohol/Drug Problems risk factor was found to be the most important predictor across nearly all statistical learning methods. This was found for the overall sample and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Further, issues arising from current alcohol and/or drug use such as law violations, problems within a family and/or marriage, and problems with school and/or work were also consistently found for both groups and primarily for the penalised logistic regression and support vector machine algorithms. Drug use specifically is found to have a close relationship with offending due to the often illegal nature of drug use (Andrews & Bonta, 2010), with the odds of criminal behaviour being approximately 2.79 times greater for those who abuse drugs than for non-drug users (Bennett et al., 2008).

In Australia, illicit drug arrests have increased significantly over the last decade, with research further indicating a relationship between drug use and criminal behaviour among Australians (Australian Institute of Health and Welfare, 2021). For example, in 2018, 65% of prison entrants in Australia reported illicit drug use within the 12 months prior to being incarcerated (Australian Institute of Health and Welfare, 2021). Further, one in three police detainees in 2019 specified that illicit drug use had contributed to their criminal behaviour (Australian Institute of Health and Welfare, 2021), and almost half of police detainees in 2020 reported drug use within the past 30 days (Australian Institute of Criminology, 2021). Current drug use at the time of being assessed with the LS/RNR is therefore understandably an important predictor of future recidivism, with illicit drug use and/or possession being sufficient to be a police charge and therefore labelled as a recidivist in the present study.

Employment

Last, current unemployment from the Education/Employment risk factor was also an important predictor across all algorithms and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. This supports previous research that has demonstrated a notable link between an individual's lifestyle and future offending.

Specifically, those who are unemployed or without stable employment are more likely to engage in recidivism (Payne, 2007). Research in Australia has shown that among previously incarcerated individuals, those who were unemployed were significantly more likely to be re-incarcerated when compared to students or employed individuals (Baldry et al., 2006).

Limitations

The findings of the present study were limited by the small sample size, with statistical learning methods performing the most accurate estimates with considerably larger sample sizes. Although *k*-fold cross-validation was used to account for the small sample size, wide estimates across the fold were reported for discrimination indices. The sample is also not reflective of the general Victorian prison population as it consisted of individuals who had previously been sentenced to a term of imprisonment for a serious violent offence. Additionally, Aboriginal and Torres Strait Islander peoples were oversampled for the present study to enable comparisons between groups. In the present study, they represented 47.37% of the sample. However, in Victoria, Aboriginal and Torres Strait Islanders account for 10% of the male adult prison population (Corrections Victoria, 2022). However, 10% is reflective of the whole adult male prison population in Victoria, and not exclusively those who were incarcerated for a serious violent offence. Therefore, the current sample does not represent either the adult male prison population or the wider serious violent offender population in Victoria.

Also, due to limited demographic information, individuals could only be determined as either Aboriginal and Torres Strait Islanders or non-Aboriginal and Torres Strait Islanders based on self-reported demographic information. Australia is a multi-cultural society that incorporates a large number of culturally and linguistically diverse individuals (Australian Bureau of Statistics, 2020a). The approach to categorising individuals in the present study therefore ignores the cultural diversity and heterogeneity of other cultural groups represented in the non-Aboriginal and Torres Strait Islander group. The self-reported nature of Aboriginal and Torres Strait Islander status may also further limit the accuracy and generalisability of these findings.

Further, the study was limited by the use of a cut-off threshold to determine error rate balance, predictive parity, and statistical parity fairness definitions. Risk assessment instruments often have more than two risk classifications, with the LS/RNR specifically having five risk classifications. Using a single cut-off metric to develop these fairness metrics is therefore not reflective of how the instrument would be used in the real world. Further, different cut-off values result in different values on fairness definitions as the proportion of people who are classified as low risk and high risk changes (Zottola et al., 2021). It may be useful for future research to explore the fairness definitions across all possible cut-off values (e.g., all possible risk scores) to gain an understanding of the fairness between groups at each possible risk score/classification.

Implications

There are several implications for the findings of the present study. First, statistical learning methods can be used as an approach to increase the discrimination and cross-cultural fairness of risk assessment instruments. Second, post hoc analyses such as Shapley values allow you to investigate these statistical learning methods and gain a better understanding of the significance of predictors to the outcome. This helps to overcome one of the biggest limitations of statistical learning methods compared to traditional approaches, with interpretability often being lost. This provides a means for administrators of risk assessment instruments to improve discrimination and fairness while maintaining interpretability for practical usage. If a statistical learning method could be trained and validated on a larger and more representative sample, there is potential for corrections officers to use item scores from a risk assessment instrument to generate a predicted probability of recidivism that is better at discriminating recidivists from non-recidivists and also fairer cross-culturally. For all items, Shapley values could also be calculated to show which items contributed the most to the difference between an individual's prediction and the average prediction from the validated statistical learning algorithm.

However, due to the technological difficulty and accompanying costs of implementing this approach, the feasibility will likely be limited. Further, it needs to be understood that the interpretation of Shapley values is still limited when compared to approaches such as logistic regression. Shapley values do not provide a prediction model and therefore do not provide information on how changes in the predictors correspond to changes in the prediction (Molnar, 2019). Further, Shapley values can be easily misinterpreted and need to be understood by users of risk assessment instruments as the marginal contribution of a predictor, given the current set of predictors, to the difference between the actual and mean prediction. Other approaches such as local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016) can offer a prediction model for individual predictions and could also be explored in future research as an approach to increasing the interpretability of statistical learning methods that have increased the discrimination and/or fairness of risk assessment instruments.

Second, although various statistical learning methods were used in the present study, similar items were found to be important predictors for the sample across all the approaches used in the present study. This demonstrates the validity of these items being important predictors of recidivism and useful for corrections officers to target in treatment and rehabilitation plans. These items also demonstrated their importance cross-culturally, with similar items being important for both Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders. This indicates to corrections officers that these items to be targeted for risk mitigation are comparable across these two groups, with both current drug use and current unemployment specifically being consistently important LS/RNR items in predicting recidivism.

Third, the LS/RNR items belonging to the Alcohol/Drug Problems and Education/Employment risk factors are particularly useful items that were identified as important predictors for the sample as they are dynamic items (i.e., changeable). Specifically, both current drug use and current unemployment at the time of assessment are able to be directly targeted by corrections officers in treatment. However, both of these factors (and treatment/rehabilitation for these factors) are potentially influenced by an individual being in a prison setting and can change over time depending on an individual's length of sentence. Regardless, helping the individual gain access to employment and/or providing them with treatment for their current drug use, when possible, may aid in mitigating their future risk of recidivism.

Conclusion

The present study highlights the potential utility of using statistical learning methods as an approach to increase the discrimination and fairness of the LS/RNR with a sample of Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. It also contributed to the limited research by demonstrating the usefulness of Shapley values for increasing the interpretability of statistical learning. Future research should aim to investigate the use of statistical learning methods and Shapley values on a larger, more representative sample in order to confirm the current findings and to see how these approaches may be utilised in practice.

References

Andrews, D. A., & Bonta, J. (1998). *The Level of Service Inventory–Revised: Screening Version*. Toronto: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct*. Cincinnati: Taylor and Francis. https://doi.org/10.4324/9781315721279

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashford, L. J., Spivak, B. L., & Shepherd, S. M. (2022). *Statistical learning methods and cross-cultural fairness: Trade-offs and implications for risk assessment instruments* [Manuscript submitted for publication]. Centre for Forensic Behavioural Science, Swinburne University of Technology.

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

Australian Institute of Criminology. (2021). *Drug use monitoring in Australia: Drug use among police detainees, 2020*. https://www.aic.gov.au/publications/sr/sr35

Australian Institute of Health and Welfare. (2021). *Alcohol, tobacco & other drugs in Australia*. Australian Institute of Health and Welfare. Australian Government.

https://www.aihw.gov.au/reports/alcohol/alcohol-tobacco-other-drugs-australia/contents/priority-populations/people-in-contact-with-the-criminal-justice-system

Baldry, E., McDonnell, D., Maplestone, P., & Peeters, M. (2006). Ex-prisoners, homelessness and the state in Australia. *Australian & New Zealand journal of criminology*, *39*(1), 20-33. https://doi.org/10.1375/acri.39.1.20

Bennett, T., Holloway, K., & Farrington, D. (2008). The statistical association between drug misuse and crime: A meta-analysis. *Aggression and Violent Behavior*, *13*(2), 107-118. https://doi.org/10.1016/j.avb.2008.02.001

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1-42. https://doi.org/10.1177/0049124118782533

Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. *arXiv:2006.07155 [cs.LG]*.

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

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.CO2

Chouldechova, A. (2020). Transparency and simplicity in criminal risk assessment. *Harvard Data Science Review*, *2*(1). https://doi.org/https://doi.org/10.1162/99608f92.b9343eec

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

Eisenberg, M. J., van Horn, J. E., Dekker, J. M., Assink, M., van der Put, C. E., Hendriks, J., & Stams, G. J. J. M. (2019). Static and dynamic predictors of general and violent criminal offense recidivism in the forensic outpatient population: A meta-analysis. *Criminal Justice and Behavior*, *46*(5), 732-750. https://doi.org/10.1177/0093854819826109

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and 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

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

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.

Kaponen, M. (2020). *Fairness and parameter importance in logistic regression models of criminal sentencing data* [Unpublished master's thesis], Uppsala University. http://uu.diva-portal.org/smash/get/diva2:1459136/FULLTEXT01.pdf

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.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

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

Lundberg, S., & Lee, S. (2017, December 4-9). *A unified approach to interpreting model predictions* [Paper presentation]. 31st International Conference on Neural Information Processing Systems, Long Beach, California, United States.

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

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*.

Molnar, C. (2020). *iml: Interpretable Machine Learning*. (Version 0.10.1) [R package]. https://CRAN.R-project.org/package=iml

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*(1), 489-513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and caucasian female and male adolescents on probation. *Psychological Assessment*. https://doi.org/10.1037/pas0000816

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

Payne, J. (2007). *Recidivism in Australia: Findings and future research*. Australian Institute of Criminology. https://www.aic.gov.au/sites/default/files/2020-05/rpp080.pdf

R Core Team. (2020). *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., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1). https://doi.org/https://doi.org/10.1162/99608f92.6ed64b30

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., Adams, Y., McEntyre, E., & Walker, R. (2014). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law*, *20*(3), 281-293. https://doi.org/10.1037/law0000017

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

Shepherd, S. M., & Strand, S. (2016). The PCL: YV and re-offending across ethnic groups. *Journal of Criminal Psychology*, *6*(2), 51-62. https://doi.org/10.1108/JCP-02-2016-0006

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.

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

Smallbone, S., & Rallings, M. (2013). Short-term predictive validity of the Static-99 and Static-99-R for Indigenous and nonindigenous Australian sexual offenders. *Sexual Abuse A Journal of Research and Treatment*, *25*(3), 302-316. https://doi.org/10.1177/1079063212472937

Spivak, B. L., & Shepherd, S. M. (2020). Machine learning and forensic risk assessment: New frontiers. *Journal of Forensic Psychiatry & Psychology*. 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

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior*, *36*(4), 345-355. https://doi.org/10.1037/h0093930

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

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 US.

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]*.

Warner, B., Spivak, B., Ashford, L., Fix, R., Ogloff, J., & Shepherd, S. (2021). The impact of offender–victim cultural backgrounds on the likelihood of receiving diversion. *Criminal Justice Policy Review*, 08874034211046313. https://doi.org/10.1177/08874034211046313

Whiteacre, K. W. (2006). Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review*, *17*(3), 330-342. https://doi.org/10.1177/0887403405284766

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

Wisser, L. (2019). Pandora's algorithmic black box: The challenges of using algorithmic risk assessments in sentencing. *American Criminal Law Review*, *56*(4), 1811-1832.

Yang, M., Wong, S., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*(5), 740. https://doi.org/https://doi.org/10.1037/a0020473

Zottola, S. A., Desmarais, S. L., Lowder, E. M., & Duhart Clarke, S. E. (2021). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*, 00938548211040544. https://doi.org/10.1177/00938548211040544

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, *67*, 301-320.

Supplementary Materials

Calibration as assessed by Brier scores, predictive parity (i.e., PPV and NPV), error rate balance (i.e., FPR and FNR), and statistical parity (i.e., high risk proportion) for Aboriginal and Torres Strait Islander people are presented in Table S1. Calibration estimates were comparable for the LS/RNR total risk score and the majority of statistical learning methods, with the support vector machine with pre-processing having the lowest levels of calibration. Aboriginal and Torres Strait Islanders were found to have higher PPVs, which were higher among the statistical learning methods when compared to the LS/RNR total risk score. However, NPVs were low, indicating that the majority of Aboriginal and Torres Strait Islanders classified as low risk engaged in recidivism.

The FNR for the LS/RNR total score was higher than the statistical learning methods, indicating that for the LS/RNR total risk score, more than half of Aboriginal and Torres Strait Islanders were classified as high risk and did not go on to engage in recidivism by the end of the follow up period. In contrast, the FNR was lower across the LS/RNR total risk score and all statistical learning methods, indicating that Aboriginal and Torres Strait Islanders were less likely to be classified as low risk and later go on to engage in recidivism. For the LS/RNR total risk score and all statistical learning methods, the majority of Aboriginal and Torres Strait Islanders were found to be classified as high risk. The lowest levels of high risk classification were found for the support vector machine with pre-processing.

Table S1

Calibration, Predictive Parity, Error Rate Balance, and Statistical Parity Values for Aboriginal and Torres Strait Islanders

	Calibration	PPV	NPV	FPR	FNR	High Risk
LS/RNR Total Risk Score	.19	87.88	20.83	61.54	24.68	73.33
Penalised Logistic Regression	.20	93.30	37.41	33.33	22.00	71.14
Stochastic Gradient Boosting	.20	92.82	31.88	35.00	26.08	67.82
Stochastic Gradient Boosting (post-processing)	.20	92.34	28.90	35.00	29.25	65.31
Support Vector Machine	.20	91.57	33.69	46.67	19.78	74.68
Support Vector Machine (pre-processing)	.23	91.39	26.10	38.33	33.58	61.72

Calibration, predictive parity, error rate balance, and statistical parity for non-Aboriginal and Torres Strait Islanders are presented in Table S2. Compared to Aboriginal and Torres Strait Islanders, non-Aboriginal and Torres Strait Islanders were found to have lower levels of calibration. Furthermore, in contrast to Aboriginal and Torres Strait Islanders, the LS/RNR total risk score had the lowest levels of calibration. PPV was also high for non-Aboriginal and Torres Strait Islanders, indicating that individuals who were classified as high risk often went on to engage in recidivism. Across the LS/RNR total risk score and statistical learning methods, the PPV was slightly lower for non-Aboriginal and Torres Strait Islanders. However, NPVs were found to be consistently higher for non-Aboriginal and Torres Strait Islanders, demonstrating that more non-Aboriginal and Torres Strait Islanders who were classified as low risk did not go on to engage in recidivism by the end of the follow up period. However, the majority of those classified as low risk still engaged in recidivism.

Non-Aboriginal and Torres Strait Islanders also consistently had a lower FPR and a higher FNR when compared to Aboriginal and Torres Strait Islanders. This demonstrates that non-Aboriginal and Torres Strait Islanders were less likely to be classified as high risk and not go on to engage in recidivism. However, they were more likely to be classified as low risk and later engage in recidivism than Aboriginal and Torres Strait Islanders. Only for the stochastic gradient boosting algorithm with post-processing were the FPR and FNR estimates found to be essentially equivalent for non-Aboriginal and Torres Strait Islanders. Similar to Aboriginal and Torres Strait Islanders, the majority of non-Aboriginal and Torres Strait Islanders were found to be classified as high risk for the LS/RNR total risk score and across all statistical learning methods. However, the proportion of non-Aboriginal and Torres Strait Islanders who were classified as high risk was notably lower than that of Aboriginal and Torres Strait Islanders.

Table S2

Calibration, Predictive Parity, Error Rate Balance, and Statistical Parity Values for non-Aboriginal and Torres Strait Islanders

	Calibration	PPV	NPV	FPR	FNR	High Risk
LS/RNR Total Risk Score	.28	83.81	32.63	35.42	42.11	52.50
Penalised Logistic Regression	.24	90.40	41.00	22.33	38.04	52.54
Stochastic Gradient Boosting	.22	90.84	44.26	20.93	33.31	55.95
Stochastic Gradient Boosting (post-processing)	.22	87.69	42.02	31.60	31.81	59.35
Support Vector Machine	.26	90.84	42.26	24.83	34.29	55.86
Support Vector Machine (pre-processing)	.25	88.58	41.39	25.43	35.89	55.18

Chapter Seven: Integrated Discussion

7.1 Introduction

This final chapter integrates the findings from the literature review (Chapter Two) and three empirical studies (Chapters Four to Six). The research aims and questions outlined in Chapter One are discussed with reference to the findings from the empirical studies. These findings are also examined in relation to their implications for both theory and practice. Last, the methodological limitations of the thesis are discussed, and considerations for future research are proposed.

7.2 Research Overview

The overall goal of this thesis was to examine and improve the cross-cultural fairness of forensic risk assessment. In order to explore this, the thesis had two primary aims. After defining fairness and gaining an understanding of how this can be conceptualised within forensic risk assessment, the thesis initially aimed to establish the levels of fairness of the widely used risk assessment instrument, the LS/RNR, between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders from Victoria, Australia. To address this aim, two research questions were posed. The first research question asked to what degree Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders differed on the actuarial risk instrument, the LS/RNR, in terms of discrimination (i.e., AUC). This question was asked as the majority of the literature has reported on the AUC as the primary measure of a risk assessment instrument's utility (Singh, 2013), and has also used the AUC as a metric to compare across cultures to demonstrate comparable performance (e.g., Jones et al., 2016; Wormith et al., 2015). The second research question explored the level of cross-cultural fairness of the LS/RNR with consideration to error rate balance, calibration, predictive parity, and statistical parity between Aboriginal and Torres Strait Islanders and non-Aboriginal and

Torres Strait Islanders. Both of these research questions and the first research aim were explored within Empirical Study One (see Chapter Four).

The second research aim was to attempt to increase the fairness of the LS/RNR by utilising novel statistical approaches while still maintaining an acceptable level of discrimination and therefore the utility of the instrument. This was achieved through addressing three research questions. The first research question asked if statistical learning methods could improve the discrimination of the LS/RNR overall and for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The second research question asked if statistical learning methods using differing processing approaches could increase forms of fairness (i.e., error rate balance, calibration, predictive parity, and statistical parity) across the two groups while still maintaining appropriate levels of discrimination. These two questions were responded to in Empirical Study Two (see Chapter Five). The final research question asked if statistical learning methods can have their interpretability increased so that the impact of individual predictors can be understood. This was addressed in Empirical Study Three (see Chapter Six) by using Shapley values to unpack the importance of predictors (i.e., LS/RNR items) for predicting recidivism overall, and for Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders.

7.3 Overview of Literature Review

In Chapter Two, the literature review highlighted that there are numerous ways to define fairness, including error rate balance, calibration, predictive parity, and statistical parity (Verma & Rubin, 2018). Most commonly, risk assessment instruments were evaluated to see if they were performing the same cross-culturally by assessing their predictive accuracy or the discrimination of the instrument through the AUC. The findings demonstrate that there are often comparable levels of discrimination between cultural majority and cultural minority groupings (e.g., Dieterich et al., 2016; Lee et al., 2020; Muir et al., 2020; Perrault et al., 2017; Skeem & Lowenkamp, 2016), with certain cultural minorities (e.g., Indigenous populations of North America and Australia) occasionally reporting lower levels of discrimination (e.g., Helmus et al., 2012; Molnar et al., 2020; Shepherd, Luebbers, et al., 2014). However, when observing these other fairness definitions in the literature, consistent disparities were identified within the limited research that had been conducted. Specifically, error rate balance highlighted discrepancies between groups. Cultural minorities were classified as high risk and did not go on to engage in recidivism at a higher rate when compared to cultural majorities. Moreover, cultural majorities were classified as low risk and went on to engage in recidivism at higher rates when compared to cultural minorities (Flores et al., 2016). Differences among predictive parity metrics were less pronounced; however, cultural minorities still consistently had a higher proportion of individuals who were recidivists both in the low risk and high risk categories (Muir et al., 2020; Whiteacre, 2006). Similarly, Indigenous populations from North America were found to have higher predicted rates of recidivism across lower risk scores when compared to non-Indigenous populations, demonstrating issues with calibration (Wilson & Gutierrez, 2014; Wormith & Hogg, 2012; Wormith et al., 2015). Last, statistical parity disparities were identified with specific cultural minority groupings consistently found to score significantly higher on risk assessment instruments (e.g., Hsu et al., 2010; Lee et al., 2020; Lee et al., 2019; Olver et al., 2018; Olver et al., 2014; Shepherd et al., 2015).

This literature review also highlighted that certain fairness definitions (e.g., error rate balance and predictive parity) are incompatible with each other when groups report different base rates of recidivism (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017). This demonstrates a trade-off that is often unavoidable when pursuing fairness. Last, this review discussed the approaches often used to identify and/or address unfairness among cultural groups. The review identified that all approaches had significant limitations that needed to be considered. Nor did they solve the potential causes of unfairness. Statistical learning methods and processing approaches, on the other hand, provided a time-sensitive and feasible approach to increasing fairness, with limited research demonstrating promising results (Skeem & Lowenkamp, 2020; Wadsworth et al., 2018).

7.4 Overview of Empirical Findings

The findings of the literature review were used to shape the research aims of this thesis. This following section will address the key findings of the empirical studies in relation to the research aims.

7.4.1 Research Aim One

Due to the disparities identified in the literature review across fairness definitions and the limited research that has examined cross-cultural fairness, research aim one explored the cross-cultural fairness of a commonly used risk assessment instrument, the LS/RNR, among Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. As discrimination had been a primary measure of a risk assessment instrument's utility and a point of comparison between cultural groups, Empirical Study One (Chapter Four) initially analysed the AUC across groups. Similar to a number of previous findings, comparable discrimination was found across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, with a slightly lower AUC reported for Aboriginal and Torres Strait Islanders (Shepherd & Strand, 2016; Smallbone & Rallings, 2013; Thompson & McGrath, 2012). The AUC values found in Empirical Study One also signified that the LS/RNR in general was relatively poor at discriminating individuals who went on to engage in recidivism from those who did not. To explore discrimination further, this study also incorporated the xAUC (Kallus & Zhou, 2019) to compare the discrimination between groups (instead of within). To the author's knowledge, this is the first application of the xAUC for the LS/RNR and within the forensic psychology discipline. The xAUC in the present thesis provided useful information about the instruments' ability to differentiate recidivists from non-recidivists alongside the traditional AUC. This study found comparable AUC values across groups, but significant disparities in xAUC within groups, such that the LS/RNR was unable to distinguish between Aboriginal and Torres Strait Islander non-recidivists and non-Aboriginal and Torres Strait Islander recidivists.

Even with comparable levels of discrimination as assessed by the standard AUC, disparities among fairness definitions were found in this empirical study that were in line with findings from the literature review. Predictive parity and calibration differences between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders were minor, with Aboriginal and Torres Strait Islanders having higher rates of recidivism primarily across lower risk scores and classifications. However, error rate balance and statistical parity disparities were more notable between groups. Specifically, Aboriginal and Torres Strait Islanders were on average 1.84 times more likely to be classified as high risk and not go on to become a recidivist, while non-Aboriginal and Torres Strait Islanders were on average 1.58 more likely to be classified as low risk and later become a recidivist. Aboriginal and Torres Strait Islanders also scored significantly higher on the total LS/RNR risk score and numerous risk factors.

Overall, the LS/RNR was able to discriminate individuals who went on to engage in recidivism from those who did not to a similar degree within Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander groups. However, this was found alongside disparities in discrimination between the groups (i.e., xAUC) and among fairness definitions, primarily error rate balance and statistical parity. A focus solely on comparable AUC values, or within group discrimination, does not effectively demonstrate a risk assessment instrument that is performing equitably across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

7.4.2 Research Aim Two

The literature review explored numerous approaches to increasing the cross-cultural fairness of forensic risk assessment instruments. Statistical learning methods and processing approaches have emerged as a time-sensitive and feasible solution that can be used to increase fairness, with the limited previous research finding utility in this methodology for improving both the AUC and various fairness definitions such as calibration and error rate balance (Johndrow & Lum, 2017; Skeem & Lowenkamp, 2020; Wadsworth et al., 2018). Increases in AUC highlight another advantage of statistical learning methods, which are designed to optimise predictive accuracy and/or discrimination (Spivak & Shepherd, 2020). This was particularly useful for the current thesis, as discrimination (specifically the AUC) is often the main metric reported to demonstrate a risk assessment instrument's utility (Singh, 2013; Singh et al., 2013), and the AUC of the LS/RNR from Empirical Study One (Chapter Four) was relatively poor. Introducing statistical learning methods to increase fairness may therefore also aid in increasing the AUC.

Research aim two, therefore, aimed to increase the cross-cultural fairness and discrimination of the LS/RNR using statistical learning methods and processing approaches. Empirical Study Two (Chapter Five) found that multiple statistical learning methods using the LS/RNR items were able to increase the discrimination of the LS/RNR when compared to the original LS/RNR total risk score. This was in line with previous research that had used a large number of predictors with statistical learning methods (Breitenbach et al., 2009; Ting et al., 2018). A few of these statistical learning methods only led to minor increases in discrimination (e.g., logistic regression and random forest algorithms). However penalised logistic regression, stochastic gradient boosting, and support vector machine algorithms resulted in a notable increase. These increases were found for the overall sample, and also often for both Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders.

Statistical learning methods were also used with processing approaches including residuals (i.e., pre-processing) and reject option based classifications (i.e., post-processing) to increase fairness in Empirical Study Two (Chapter Five). This led to mixed results. Overall, the pre-processing approach of using residuals often led to an increase in fairness (including fewer disparities between xAUC values) compared to the LS/RNR total score and compared to statistical learning methods with no processing approaches. These findings were similar to previous research that had explored pre-processing approaches in an attempt to increase fairness definitions across cultural groups (Johndrow & Lum, 2017; Skeem & Lowenkamp, 2020). Only minor losses in AUC were found compared to statistical learning methods with no processing, with an increase in AUC when compared to the LS/RNR total risk score. However, tree-based models (i.e., random forest and stochastic gradient boosting algorithms) using residuals produced increased disparities among fairness definitions and the xAUC. The fairness definitions that had notable increases in disparities were those that were already reported to have significant disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders in Empirical Study One (Chapter Four), namely error rate balance and statistical parity. These tree-based statistical learning methods had better results when using reject option based classification as a post-processing approach, with reductions in disparities among fairness definitions.

Certain approaches, such as support vector machines with pre-processing and stochastic gradient boosting with post-processing demonstrated notable increases in fairness, primarily across xAUC, error rate balance, and statistical parity, which had the largest disparities with the LS/RNR total score and from Empirical Study One (Chapter Four). These algorithms specifically also suffered only a minor loss in their ability to discriminate recidivists from non-recidivists. However, predictive parity metrics and calibration suffered a minor increase in disparity between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. This highlights the trade-off between fairness definitions, specifically error rate balance and predictive parity. Increases in fairness among error rate balance was often accompanied with an increase in disparity for predictive parity. Comparatively, tree-based models with pre-processing that led to increases in disparity among error rate balance metrics often had an increase in fairness among one or both of the predictive parity metrics (i.e., positive predictive value or negative predictive value). The implications of this finding are discussed in more detail within the implications section of this chapter.

A further trade-off that arose due to the use of statistical learning methods was that these approaches are less transparent than traditional causal approaches to forensic risk assessment, in which the relationship between the outcome and the predictors is known and able to be explicitly stated (Breiman, 2001b). Therefore, in Empirical Study Three (Chapter Six), Shapley values were used to help address this limitation and explore the important predictors of the statistical learning methods that were useful in increasing the discrimination and/or fairness of the LS/RNR in Empirical Study Two (Chapter Five). Specifically, penalised logistic regression, stochastic gradient boosting, stochastic gradient boosting with reject options based classification (i.e., post-processing), support vector machine, and support vector machine with residual (i.e., pre-processing) algorithms were examined using Shapley values.

Although Shapley values are unable to provide a transparent understanding of the link between predictors and the outcome, they do aid in providing useful information about which predictors were the most important to the outcome. Empirical Study Three (Chapter Six) found similarities in the top five important predictors (i.e., the LS/RNR items with the highest absolute average Shapley value) across all statistical learning methods examined. Specifically, both current drug use and current unemployment were notably important predictors of recidivism across all statistical learning methods, including those with processing approaches. The other predictors that were consistently within the top five highest absolute average Shapley values focused on either criminal history items (e.g., number of prior youth dispositions, number of present offences, and number of prior adult convictions) or issues resulting from current alcohol and/or drug use (e.g., law violations, problems with family/marital, problems with school/work). The criminal history items were more notable within the stochastic gradient boosting algorithms and the problems arising from current alcohol and/or drug use were more notable within the support vector machine algorithms. These findings were also similar across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. The biggest disparity was a slightly higher importance being placed on criminal history items for Aboriginal and Torres Strait Islanders; however, this difference was minor.

Overall, empirical studies two and three (Chapters Five and Six) demonstrated that certain statistical learning methods and processing approaches using the LS/RNR items were able to increase the discrimination and ameliorate violations of certain fairness definitions between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Furthermore, Shapley values offered a potential solution to the transparency issue that statistical learning methods can present. Specifically, Empirical Study Three (Chapter Six) found that the highest absolute average Shapley values were relatively consistent across all statistical learning methods, with current drug use and current unemployment being important LS/RNR items that contributed to the prediction of recidivism.

7.5 Implications

The following section will discuss the implications of the present findings and how they relate to theory, policy, and practice.

7.5.1 Utility of the LS/RNR

The present thesis highlights that the LS/RNR was an acceptable but relatively poor discriminator. In line with Rice and Harris (2005), the AUC values found for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders can be considered weak, and the AUC for the overall sample was just on the threshold to be considered moderate in strength. In other words, the LS/RNR total risk score was poor at differentiating an individual who went on to engage in recidivism from an individual who did not, for both Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders.

Therefore, if this instrument is being used in practice to estimate an individual's level of risk of recidivism, it needs to be understood that an individual with a higher risk score is not more likely to go on and engage in recidivism when compared to an individual with a lower risk score. Instead, approximately one out of every three individuals, a non- recidivist will have a higher LS/RNR risk score than a recidivist. Further, for Aboriginal and Torres Strait Islanders in the present thesis, there is only a 60% chance that an individual who engaged in recidivism received a higher risk score than an individual who did not engage in recidivism. This raises concerns about decisions that may be made on the basis of a risk assessment score. High stakes decisions made on the basis of LS/RNR risk scores, and the expectations of recidivism that accompany those scores, may be inappropriate or misguided for serious violent offenders. The weak AUC values for this sample suggest that the LS/RNR total risk score has a large number of false positives among the higher risk scores and classifications (Cook, 2007). When used for high stakes decision making, which may result in punitive measures, these decisions will likely have negative implications for those who are false positives (e.g., restrictions that impede on personal liberty).

Further, the ability of the LS/RNR total risk score to discriminate between recidivists and non-recidivists was comparable across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. This comparability was not maintained between the two groups across different metrics of fairness, demonstrating that equivalent discrimination is not always indicative of a risk assessment instrument that is fair or performing equally across groups. Specifically, comparable AUC between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders was also found alongside disparate xAUC and violations of error rate balance and statistical parity.

It is worth noting another limitation in reporting the AUC in isolation as this is a measure that assesses discrimination within a group (e.g., discriminates Aboriginal and Torres Strait Islander recidivists from Aboriginal and Torres Strait Islander non-recidivists). The xAUC (Kallus & Zhou, 2019), which is only a minor alteration of the traditional AUC, can measure discrimination between groups (e.g., discriminates non-Aboriginal and Torres Strait Islander recidivists from Aboriginal and Torres Strait Islander non-recidivists). This discrimination measure therefore better assesses the fairness of a risk assessment instrument across groups and should be utilised as a metric in the future to be reported alongside the traditional AUC. In the present thesis specifically, the xAUC identified that if decisions were being made on the basis of the LS/RNR total risk score, Aboriginal and Torres Strait Islander non-recidivists are essentially being treated the same as non-Aboriginal and Torres Strait Islander recidivists.

Further, the LS/RNR was found to consistently violate error rate balance across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. Aboriginal and Torres Strait Islander non-recidivists were more likely to be classified as high risk of recidivism, while non-Aboriginal and Torres Strait Islander recidivists were more likely to be classified as low risk of recidivism. This finding again calls into question the validity of decisions made for this sample on the basis of the LS/RNR total risk score, and also how decisions may be disadvantaging certain groups. In this case, non-recidivist Aboriginal and Torres Strait Islanders being classified as high risk more often may result in harsher monitoring or surveillance that impedes on their personal liberty. Conversely, non-Aboriginal and Torres Strait Islander recidivists being classified as low risk of recidivism more often could result in lower levels of intervention and/or rehabilitation that may have mitigated their risk of recidivism and also increased public safety. In summary, the LS/RNR for the present sample resulted in low levels of discrimination and also violated a number of fairness definitions that may have direct implications in practice if the instrument's total risk score is being used to influence decision making.

It is worth noting that the current thesis focused on fairness definitions that were in the context of prediction only. This does not necessarily mean that an unfair risk assessment instrument will directly result in unfair decision making or treatment of a specific cultural group. For example, a risk assessment instrument that was found to be fair (i.e., performed the same) for different cultural groups could still result in unfair outcomes for one of those cultural groups due to other forms of unfairness or bias within the criminal justice system. On the contrary, a risk assessment instrument that was found to perform disparately for different cultural groups could still result in fair treatment across groups. The current thesis does, however, highlight the ways in which a risk assessment instrument can perform differently among cultural groups, which in turn has the potential to translate into unfair outcomes.

7.5.2 Statistical Learning Methods as an Approach to Increase Fairness

Statistical learning methods were found to be useful in the present thesis for increasing the ability of the LS/RNR to discriminate between individuals who went on to engage in recidivism from individuals who did not. Improvements were found for a number of the different statistical learning methods trialled for the overall sample and for both Aboriginal and Torres Strait Islanders as well as non-Aboriginal and Torres Strait Islanders. Further, processing approaches, namely post-processing (i.e., reject option based classification) for tree-based statistical learning methods (i.e., random forests and stochastic gradient boosting), and pre-processing (i.e., residuals) for the remaining statistical learning methods, aided in ameliorating the violations of fairness definitions that were the most disparate between Aboriginal and Torres Strait Islander and non-Aboriginal and Torres Strait Islander individuals. Specifically, xAUC, error rate balance, and statistical parity disparities were often decreased between the two groups when using these approaches. These findings demonstrate the potential usefulness of statistical learning methods and processing approaches in forensic risk assessment to increase the utility and fairness of a risk assessment instrument. With limited research to date utilising these methodologies, these approaches should continue to be explored and potentially trialled in practice.

However, often, when error rate balance disparities were reduced between the two groups, predictive parity disparities were increased. Furthermore, although statistical learning methods and processing approaches can aid in increasing cross-cultural fairness, they do not solve the original cause of the unfairness. For example, even though processing approaches could be used to manufacture equal base rates of recidivism and/or equal prevalence of risk factors across Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, the base rates and prevalence of risk factors still differ between the groups in reality. Therefore, statistical learning methods that ameliorate violations of fairness definitions may also mask the underlying causes of unfairness, such as social and economic disadvantage, which lead to a higher prevalence of a number of risk factors, such as lower education levels, employment, and income (Day et al., 2018; Douglas et al., 2017; Hannah-Moffat, 2013; Hannah-Moffat & Maurutto, 2010; Harcourt, 2007). Furthermore, as mentioned previously, the complexity of statistical learning methods is also often associated with a loss of interpretability (Spivak & Shepherd, 2020), which further poses problems with this approach in practice.

Nevertheless, statistical learning methods do demonstrate promise as a way to create a fairer assessment of risk (e.g., Berk, 2019; Skeem & Lowenkamp, 2020; Wadsworth et al., 2018). Although there is significant and warranted scepticism and concern around statistical learning methods with regards to entrenching existing discrimination with data and/or the transparency of this approach that poses practical limitations (Kehl et al., 2017; Rudin et al., 2020), statistical learning methods can be developed with the improvement of cross-cultural fairness as a primary aim. As in the case of the present thesis, this can result in a fairer risk assessment instrument with a greater capability to differentiate recidivists from non-recidivists than what is currently being used.

7.5.3 Shapley Values for Increasing Interpretability

As mentioned previously, a common concern and critique surrounding statistical learning methods is the lack of transparency of these approaches. Relationships between predictors and between predictors and the outcome can be obscured (Breiman, 2001b). The present thesis demonstrated that Shapley values can offer a way to increase the interpretability of statistical learning methods, thereby increasing the usefulness of statistical learning methods in practice. Although Shapley values do not explain the statistical learning method entirely (Molnar, 2019), they do provide useful information to the administrators and those being assessed of what predictors (in the case of the present thesis, the LS/RNR items) had the largest contribution to the difference between their prediction and the average prediction. Beyond just using this approach for information for an individual, Shapley values can also be aggregated, and the predictor's importance can be observed across a larger sample of individuals. There is also the potential for Shapley values to be generated at the time of assessment, providing necessary information to a clinician around which predictors were the most important for this individual's prediction of recidivism. This could be used as another piece of information to help with treatment and/or intervention decisions that could reduce the likelihood of offending in the future.

However, the limitations of Shapley values need to be understood by those both developing statistical learning methods and also users of the risk assessment instrument. Shapley values are easily misinterpreted and can be seen as quite complex. Therefore, additional training may be required if Shapley values are to be used in practice so that they are not misconstrued, and the information is used adequately. Furthermore, Shapley values do not provide total transparency of a statistical learning method and do not produce a predictive model of the statistical learning method (Molnar, 2019). For example, if a predictor variable were to increase in value, Shapley values would not be able to provide information on the respective increase or decrease in the predicted probability of recidivism. Therefore, this approach can only aid in increasing interpretability, but does not provide complete transparency. Nevertheless, the ability of Shapley values to aid in creating a more open and interpretable, albeit not completely transparent, approach to decision making is also both useful and a necessity, especially if decisions made on the basis of a prediction from a statistical learning method may impact an individual's life.

7.5.4 Importance of Dynamic LS/RNR Items for Clinical Intervention

The use of Shapley values in the present thesis identified similar LS/RNR items as the biggest mean marginal contributors to the prediction of recidivism. Specifically, current drug use and current unemployment were consistently in the top five highest mean Shapley values for all statistical learning methods examined. Previous research has noted the association between both drug use and unemployment with future offending behaviours (Andrews & Bonta, 2010; Baldry et al., 2006; Bennett et al., 2008). Of note, current drug use on average in the present thesis was the single biggest contributor to the difference between the average prediction and an individual prediction of recidivism across the vast majority of statistical learning methods (besides the stochastic gradient boosting for Aboriginal and Torres Strait Islanders). Current drug use and current unemployment are potentially useful LS/RNR items to have identified as having high Shapley values as they are dynamic items and are therefore changeable. Changeable items at the time of assessment may give clinicians an opportunity to intervene. For example, prioritising treatment for current drug use and/or helping an individual gain employment could help in reducing the future risk of recidivism for this sample. Further, there is evidence that addressing dynamic risk factors contributes to a reduction of recidivism risk (Bonta, 2002), with research demonstrating that a reduction in drug use and assisting individuals in obtaining stable employment specifically can reduce recidivism (Belenko et al., 2013; Ramakers et al., 2017).

7.5.5 Trade-Offs are Inherent and Unavoidable

The present thesis explores numerous trade-offs in the pursuit of fairness that all have a number of implications to be considered. The following section will outline the three main trade-offs identified in this thesis.

7.5.5.1 Error Rate Balance vs Predictive Parity. As previously discussed, total fairness is unachievable in forensic risk assessment when base rates of recidivism differ (Berk et al., 2018). As equal base rates between groups are highly improbable, with particular cultural minorities often being found to engage in recidivism at a higher rate (Bonta et al., 1997; Olver, 2016; Shepherd & Strand, 2016; Wilson & Gutierrez, 2014), a trade-off exists among various fairness definitions. Namely, error rate balance and predictive parity have been previously shown to be unable to be achieved simultaneously (see Chouldechova, 2017). This was also demonstrated in the present thesis. Predictive parity disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders were minimal; however, error rate balance discrepancies were more pronounced. Further, there were base rate differences between these two groups, with 85.56% of Aboriginal and Torres Strait Islanders engaging in recidivism and 76% of non-Aboriginal and Torres Strait Islanders engaging in recidivism. When attempts were made to rectify violations of fairness definitions through the use of statistical learning methods and processing approaches, error rate balance discrepancies between the two groups were lessened; however, predictive parity discrepancies often increased.

The trade-off between these fairness definitions therefore requires thoughtful deliberation around which fairness definition may be the most pivotal to satisfy between groups. In the case of the findings from this thesis, a high risk classification (or high predicted probability of recidivism) for Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders will have a similar proportion of individuals who go on to engage in recidivism. However, when looking at the outcomes of recidivism, non-recidivists are more likely to be classified as high risk (or have a higher predicted probability of recidivism) if they are Aboriginal and Torres Strait Islanders. On the other hand, recidivists are more likely to be classified as low risk (or have a lower predicted probability of recidivism) if they are non-Aboriginal and Torres Strait Islanders. If this disparity was improved, then high risk classifications (or higher predicted probabilities) would have more disparate proportions of Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders going on to engage in recidivism.

Therefore, is it more essential to have equal predictive accuracy across groups (i.e., predictive parity) and maximise public safety, or to have equal errors in observation across groups (i.e., error rate balance) and maximise personal liberty? Careful deliberation is required to make this decision, which may also differ depending on numerous circumstances such as country, jurisdiction, the use of the risk assessment instrument, and recidivism type (e.g., any, violent, sexual). Furthermore, policymakers, clinicians, and researchers all need to be aware of this trade-off and its potential implications moving forward. If fairness is able to be demonstrated on one fairness definition, it is likely that another fairness definition has been violated (Chouldechova, 2017; Huq, 2019; Skeem & Lowenkamp, 2020).

As an alternative, this trade-off could be explored in order to find an acceptable level of fairness for both predictive parity and error rate balance. In Empirical Study Two (Chapter Five), pre-processing with a support vector machine algorithm reduced large error rate balance disparities to have on average a 7.61% difference between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders on FPR and FNR. Simultaneously, predictive parity discrepancies were mildly increased, resulting in an average difference of 9.01% between groups on PPV and NPV. This could, for example, be seen as an acceptable trade-off between these two fairness definitions, with relatively mild disparities between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders being similar for both error rate balance and predictive parity metrics.

7.5.5.2 Performance vs Fairness of Risk Assessment Instruments. Another trade-off that can often arise in the pursuit of fairness is between the performance (i.e., accuracy and/or discrimination) and the fairness of a risk assessment instrument (Berk, 2019; Berk et al., 2018). For example, pre-processing that alters the data before being used in a statistical learning method can result in a loss of useful information and therefore impede on the performance of the risk assessment instrument. In the present thesis, although the ability of the LS/RNR to distinguish recidivists from non-recidivists (i.e., AUC) increased when using statistical learning methods compared to the LS/RNR total risk score, processing approaches occasionally led to a minor reduction in discrimination. Further, the accuracy (as assessed by Brier scores) was also often negatively impacted once processing approaches were applied. Again, careful deliberation is required to consider the trade-off between a loss in performance and a gain in cross-cultural fairness. Is a risk assessment instrument with the highest level of accuracy more important than a risk assessment that is cross-culturally fair? However, similar to the trade-off between fairness definitions, there may be the potential for an acceptable level of trade-off between performance and fairness.

As an example, in the present thesis, although processing approaches occasionally resulted in a minor reduction in performance, the fairness definitions that were the most violated between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders were significantly ameliorated. Further, the gain in the AUC value when using statistical learning methods compared to the LS/RNR total risk score meant that after processing approaches were applied and the AUC was mildly lowered, the AUC of the statistical learning method with processing was still greater than the total risk score of the LS/RNR (e.g., the overall AUC for the support vector machine with pre-processing was .68 and the overall AUC of the LS/RNR total risk score was .64).

75.5.3 Performance vs Interpretability of Statistical Learning Methods. The use of statistical learning methods and processing approaches, like a number of those utilised in the present thesis, introduces another trade-off that needs to be considered. This increased performance (i.e., discrimination) and cross-cultural fairness found between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders using statistical learning methods often resulted in reduced interpretability of the algorithm. Unlike the LS/RNR, in which the contribution of each item to risk factor scores and the overall risk score is explicitly understood, when using the LS/RNR items in a statistical learning method, the interpretability was often reduced. This has been a common criticism of statistical learning methods, with critics highlighting that an uninterpretable risk assessment instrument can have negative consequences (Kehl et al., 2017; Rudin et al., 2020). For example, if risk assessment instruments are used to aid decision making, an uninterpretable instrument cannot easily be held accountable or understood and therefore scrutinised by the public (Rudin et al., 2020).

In an attempt to increase the interpretability of statistical learning methods, Shapley values were used in the present thesis. Although this approach was useful in identifying the largest mean marginal contributors to a recidivism prediction, it did not create a completely transparent algorithm, and many other approaches to increasing the interpretability of statistical learning methods often come with significant limitations (Rudin, 2019). Therefore, another consideration to be deliberated is whether a risk assessment instrument with an increased capacity to discriminate between recidivists and non-recidivists and increased cross-cultural fairness is more important than a risk assessment instrument that is lower in these performance metrics but is more interpretable, accountable, and therefore can be considered more ethical. In other words, is it more important to have a better performing risk assessment instrument or one in which it is easily understood how the prediction was made?

7.5.5.4 Publicly Acceptable Trade-Offs? In summary, there are a number of trade-offs both inherent and unavoidable in the pursuit of cross-cultural fairness in forensic risk assessment. Each of these trade-offs needs to be understood moving forward and factored into the development of risk assessment instruments, especially when fairness and/or the use of statistical learning methods is involved. Further, policymakers and stakeholders need to carefully deliberate each side of these respective trade-offs and decide which is the most pivotal to prioritise, or alternatively, what a publicly acceptable trade-off might look like.

For example, and as demonstrated in the present thesis, for the trade-off between fairness definitions, perhaps similar minor violations of both error rate balance and predictive parity can be acceptable. Further, the trade-off between performance and fairness could be accepted if only a minor loss in performance was met with a notable increase in fairness between groups. Or, alternatively, an increase in fairness and certain performance metrics (e.g., discrimination indices) when compared to the original risk assessment instrument. Finally, if all items within the instrument are known (e.g., using the LS/RNR items) and post-hoc approaches like Shapley values provide a degree of insight into item importance, an opaque risk assessment instrument with better discrimination and cross-cultural fairness may be acceptable when compared to a transparent risk assessment instrument with lower cross-cultural fairness.

7.6 Limitations

The current section will discuss a number of the limitations of the present thesis, including both sample and methodological limitations. A number of these limitations have been previously raised within the empirical studies. However, this section will provide an overview of the limitations that may have influenced the findings of the present thesis.

7.6.1 Study Sample

The study sample in the present thesis posed a number of limitations. First, there was not enough information provided to effectively construct distinctive or representative cultural groups. Therefore, only two groups were able to be established. Those who identified as either Aboriginal and/or Torres Strait Islanders and those who did not (i.e., non-Aboriginal and Torres Strait Islanders). Australia is a multi-cultural society that has a large number of individuals born overseas (Australian Bureau of Statistics, 2020a). Previous research in Australia has therefore often divided cross-cultural research into three groups; Aboriginal and Torres Strait Islanders, those from an English speaking background (ESB); and those who are culturally and linguistically diverse (CALD; e.g., Shepherd et al., 2015; Shepherd & Strand, 2016). Different pieces of information can be used to establish CALD groups, including the individual's country of birth and primary language spoken. However, in the non-Aboriginal and Torres Strait Islander group, the majority of the individuals identified Australia as their country of birth (83%) and all stated that English was their primary language. Previous research that has been able to distinguish between an ESB and a CALD group has often noted differences in risk assessment performance between the ESB and CALD groups (Shepherd, Luebbers, et al., 2014; Shepherd et al., 2015; Shepherd & Strand, 2016; Thompson & McGrath, 2012), something which was unable to be determined in the present thesis. The non-Aboriginal and Torres Strait Islander group therefore do not reflect the cultural heterogeneity that exists within Australia.

Second, the Aboriginal and Torres Strait Islander group were also not representative. Aboriginal and Torres Strait Islanders do not constitute a homogenous group, with Aboriginal and Torres Strait Islanders actually comprising hundreds of various language groups, clans, and tribes (Australian Institute of Health and Welfare, 2020). However, sufficient information was not available in order to establish more representative groups. It is also worth noting that Aboriginal and Torres Strait Islanders were also oversampled in the present thesis to enable comparisons between groups, comprising 47.37% of the total sample when they only comprise approximately 9% of the adult Victorian prison population (Australian Bureau of Statistics, 2018a).

Third, the sample was potentially high risk in nature, with all individuals having been previously incarcerated for a serious violent offence as outlined in schedule 1 (clause 3) of the *Sentencing Act* 1991 (Vic). This was also demonstrated by a large proportion of the sample classified as high or very high risk (84.21%) and also the majority of the sample engaged in recidivism (80.56%) by the end of the follow up period. This further highlights that the sample in the present thesis is not reflective of the general prison population, nor is it reflective of the general population that is on a community corrections order or parole order.

A potential reason for the high risk nature of the sample is the assessment protocols undertaken by Corrections Victoria. As mentioned within the methodology, the assessment protocol to assess the general risk of recidivism has two stages. The first involves an initial classification process using LSI-R: SV that occurs upon reception into prison. Individuals who are assessed as having a low risk of general recidivism do not receive any further assessment. The second stage involves those who have a sentence of six months or more and who were assessed with either a medium or high risk of general recidivism on the LSI-R: SV. These individuals go on to receive a more comprehensive assessment of risk by being assessed with the full LS/RNR. This is initially completed within six weeks of being assessed with the LSI-R: SV, and individuals can go on to be reassessed with the LS/RNR throughout their incarceration or once released into the community. As only those who initially received an assessment of at least medium risk of recidivism on the LSR-R: SV are assessed with the LS/RNR, this aids in explaining the high proportion of the current sample that was classified as high risk.

Fourth, the sample size in the present thesis was small, especially for a number of the analyses conducted. Specifically, statistical learning methods perform best when there is a significantly large sample size. Although *k*-fold classification was used as a way to account for the small sample size, empirical studies two and three (Chapters Five and Six) demonstrated large variations in the reported metrics (e.g., AUC, Brier scores) across the 10 folds. More accurate estimates would have been achieved with a larger sample size. Further, AUC with small samples can result in large inaccuracies and has therefore been cautioned against with sample sizes less than 200 (Hanczar et al., 2010). The size of the sample within the folds in empirical studies two and three (Chapters Five and Six) was often significantly smaller than this.

Fifth, there were not enough females in the sample to examine the cross-cultural fairness of the LS/RNR for this group. The original sample had only 72 (15.93%) females who were ultimately removed to have a clearer sample for analyses. Overall, these issues reflect that the present sample is not representative of the general prison population and the results may not be generalizable beyond male Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders who were previously incarcerated for a serious violent offence.

7.6.2 Time of Assessment

Another potentially limiting aspect of the study is that the individuals in the sample were assessed with the LS/RNR at different times. The majority of the sample was assessed while incarcerated (60.79%), while the remaining sample was assessed while at risk in the community, either on a community corrections order or parole order (39.21%). To overcome these differences, the present thesis looked at time at risk to the community (i.e., removing days of incarceration) for survival analysis. However, this then ignored those who engaged in recidivism during a period of incarceration. There was a small portion of the sample who were found to engage in recidivism while incarcerated (n = 16, 4.21%), of whom the majority also engaged in recidivism while at risk to the community. Only two individuals (0.53%) who were deemed recidivists while incarcerated did not also engage in recidivism while in the community. These individuals were labelled as non-recidivists for the present thesis.

Further, for those who were assessed while incarcerated, there was variation in the time between being assessed by the LS/RNR and their release from prison. Although the most recent LS/RNR completion before being released was chosen for those who were assessed while incarcerated, the days from LS/RNR completion to release from prison ranged from 1 day to 903 days (M = 194.45 days, SD = 179.49, median = 153 days). This potentially poses a problem as the LS/RNR contains a number of dynamic items that may have changed by the time the individual was deemed at risk to the community (i.e., released from prison). However, when examining to see if there was a relationship between the days from LS/RNR completion to release from prison and recidivism, no statistically significant relationship was found ($p > .05$).

7.6.3 Cut-off Thresholds

Cut-off thresholds are required in order to calculate a number of the fairness definitions examined in the present thesis. Specifically, error rate balance and predictive parity require a threshold to establish those who are high risk (or predicted to engage in recidivism) and those who are low risk (or not predicted to engage in recidivism) in order to be calculated. Limitations around the use of cut-off thresholds were raised within the literature review and have also been demonstrated in recent research (Zottola et al., 2021). Differing cut-off thresholds lead to different results due to the proportions being classified as high and low risk changing. To overcome this, Empirical Study One (Chapter Four) reported on error rate balance and predictive parity metrics over all possible cut-off thresholds (i.e., all possible LS/RNR total risk scores). However, Empirical Studies Two and Three (Chapters Five and Six) relied on a single cut-off threshold. Due to the number of statistical learning methods trialled, a single cut-off threshold was used for ease of reporting.

As highlighted previously, if a different cut-off threshold had been used within these empirical studies, different fairness metrics would have been reported. Further, there is no agreed upon way to determine the best cut-off threshold and the method chosen for this thesis, the cut-off that was closest to 0, 1 in the ROC space, might not reflect the best cut-off threshold for another study examining fairness. This also does not reflect how a risk assessment instrument would be used in practice, with two or more cut-offs being used to develop multiple risk classifications. In the case of the LS/RNR, this instrument has five different risk classifications dependent on multiple risk score thresholds. The results of this thesis, specifically those from empirical studies two and three, are therefore limited by the use of a single cut-off threshold and would differ if a variation of cut-offs were utilised.

7.6.4 Charges as Recidivism

Recidivism was defined in the current thesis as future contact with the police that results in a formal police charge. Although police records are frequently used as a measure of recidivism, they have the potential to overestimate recidivism when compared to other common measures of offending behaviour, such as convictions, as not all police charges will result in a conviction. (Payne, 2007). Ultimately, there can be no perfect or precise measurement of recidivism. Despite the use of police charges in the current thesis, this outcome measure only includes criminal behaviours that were either reported to or identified by police, making it impossible to assess the actual recidivism that has occurred.

Another possible issue with the use of police charges as recidivism is the potential for biased policing practices to result in an outcome that is already biased towards either Aboriginal and Torres Strait Islanders or non-Aboriginal and Torres Strait Islanders. However, this is an issue that is likely to be faced regardless of which measurement is used to assess recidivism, with bias towards a specific cultural group being able to occur at any stage of the criminal justice process.

7.7 Future Research

Overall, it would be beneficial for future research to continue exploring the cross-cultural fairness of forensic risk assessment instruments and aim to explore approaches that may increase fairness across groups. Initially, research should aim to replicate the approach of the present thesis on larger and more representative samples with distinct cultural groups and females. A larger sample size will help produce more accurate estimates when using statistical learning methods and processing approaches to increase the discrimination and cross-cultural fairness of risk assessment instruments. Furthermore, a more representative sample will help with clarifying the generalisability of the findings from the present thesis.

When exploring the cross-cultural fairness and/or utility of risk assessment instruments, research moving forward should ensure that the AUC is not the sole measure being used to indicate fairness or equivalent performance between groups. The present thesis was able to demonstrate comparable AUC between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders. However, notable violations of the xAUC and fairness definitions were also found simultaneously. Future research should therefore continue to explore various fairness definitions and also the xAUC alongside the traditional AUC to get a better understanding of the performance and fairness across groups and where differences exist. Further, when examining fairness definitions that require a cut-off to be computed, specifically error rate balance and predictive parity, these should be examined across all possible cut-offs to ensure a better understanding of the level of fairness across all possible thresholds.

More work is also needed to explore the inherent trade-offs that exist within the pursuit of cross-culturally fair forensic risk assessment instruments. Exploring the trade-offs between fairness definitions, fairness and instrument performance, and performance and transparency will all be useful in gaining an understanding of the best possible acceptable trade-offs that can be achieved. Further, having these trade-offs and their consequences explicitly demonstrated will aid policymakers in determining which side of each trade-off may be the most applicable for them to satisfy.

Future research should examine other forms of processing approaches that may be beneficial in increasing fairness. The present thesis examined one pre-processing approach and one post-processing approach; however, there are numerous other processing approaches that have been developed within the computer science discipline that could be utilised. For example, utilising a pre-processing approach that could equal out rates of recidivism between groups (see Berk, 2019) could aid in overcoming the issue of base rates causing multiple forms of fairness to be unable to be satisfied simultaneously. However, research that uses statistical learning methods and processing approaches should also aim to increase the interpretability of these algorithms. Explicit information should be provided around the items used within the algorithm, as well as any information around which items were the biggest contributors to the predicted outcome.

This could involve calculating Shapley values or local interpretable model-agnostic explanations (LIME) in order to ensure that there is some level of understanding of the relationship between predictors and the predicted outcome or of the important predictors in the algorithm. Research should also explore if clinical intervention to improve important dynamic items (e.g., current drug use and current unemployment) is beneficial in reducing recidivism.

Transparent approaches (e.g., logistic regression) should also always be explored alongside more complex statistical learning methods, as often they can produce useful algorithms that have higher levels of discrimination and fairness between groups. If a transparent statistical learning method with similar levels of performance to an opaque statistical learning method can be developed, this should always be prioritised to ensure accountability and the capacity for public scrutiny. Having transparency enables another form of fairness to be satisfied, with the public (e.g., defendants) being given the opportunity to understand and scrutinise the data that was used and how that data was used to reach a risk estimate (Rudin et al., 2020). If not, significant effort should be made to increase the interpretability of more opaque statistical learning methods.

7.8 Conclusion

The empirical studies within this thesis contributed to the literature in a number of ways. First, this thesis synthesised research from various disciplines, including computer science and statistics and the forensic psychology discipline, to establish an understanding of what constitutes fairness in relation to forensic risk assessment and how this can be explored and potentially increased across groups. Second, it contributed to the limited literature that explores the cross-cultural fairness of forensic risk assessment instruments, specifically beyond AUC comparisons between groups. Third, it contributed to the small number of studies that have explored statistical learning methods as an approach to increasing the cross-cultural fairness of risk assessment instruments. To the author's knowledge, this is the first significant empirical investigation into the cross-cultural fairness of forensic risk assessment with an adult Australian sample, and one of the few that exist internationally that has explored numerous fairness definitions and attempted to mitigate violations of fairness definitions.

Overall, the present thesis aided in advancing approaches to cross-culturally fair forensic risk assessment. It provided an understanding of what fairness is, how fairness can be assessed, and how fairness can potentially be increased between groups. Although this thesis demonstrated positive findings in regard to increasing cross-cultural fairness among Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders, it did not solve the causes of unfairness. However, it did demonstrate that using existing risk assessment instrument items within statistical learning methods can result in greater fairness between Aboriginal and Torres Strait Islanders and non-Aboriginal and Torres Strait Islanders than what is currently in practice.

References

Allan, A., & Dawson, D. (2002). *Developing a unique risk of violence tool for Australian Indigenous offenders.* http://crg.aic.gov.au/reports/200001-06.pdf

Allan, A., Dawson, D., & Allan, M. M. (2006). Prediction of the risk of male sexual reoffending in Australia. *Australian Psychologist*, *41*(1), 60-68. https://doi.org/10.1080/00050060500391886

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ*, *5*. https://doi.org/10.7717/peerj.3544

Andrews, D. A., & Bonta, J. (1994). *The psychology of criminal conduct*. Anderson Publishing Co.

Andrews, D. A., & Bonta, J. (1998). *The Level of Service Inventory–Revised: Screening Version*. Toronto: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct*. Cincinnati: Taylor and Francis. https://doi.org/10.4324/9781315721279

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashford, L. J., Spivak, B. L., & Shepherd, S. M. (2021). Racial fairness in violence risk instruments: A review of the literature. *Psychology, Crime & Law*. https://doi.org/10.1080/1068316X.2021.1972108

Ashford, L. J., Spivak, B. L., & Shepherd, S. M. (2022). *Statistical learning methods and cross-cultural fairness: Trade-offs and implications for risk assessment instruments* [Manuscript submitted for publication]. Centre for Forensic Behavioural Science, Swinburne University of Technology.

Australian Bureau of Statistics. (2018a). *4517.0 - Prisoners in Australia, 2018*. http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4517.0~2018~Main%20Features~Aboriginal%20and%20Torres%20Strait%20Islander%20prisoner%20characteristics%20~13

Australian Bureau of Statistics. (2018b). *Estimates of Aboriginal and Torres Strait Islander Australians*. https://www.abs.gov.au/statistics/people/aboriginal-and-torres-strait-islander-peoples/estimates-aboriginal-and-torres-strait-islander-australians/latest-release

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). *Migration, Australia*. https://www.abs.gov.au/statistics/people/population/migration-australia/2018-19

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

Australian Institute of Criminology. (2021). *Drug use monitoring in Australia: Drug use among police detainees, 2020*. https://www.aic.gov.au/publications/sr/sr35

Australian Institute of Health and Welfare. (2020). *Indigenous Australians*. https://www.aihw.gov.au/reports-data/population-groups/indigenous-australians/overview

Australian Institute of Health and Welfare. (2021). *Alcohol, tobacco & other drugs in Australia*. Australian Institute of Health and Welfare. Australian Government. https://www.aihw.gov.au/reports/alcohol/alcohol-tobacco-other-drugs-australia/contents/priority-populations/people-in-contact-with-the-criminal-justice-system

Baldry, E., McDonnell, D., Maplestone, P., & Peeters, M. (2006). Ex-prisoners, homelessness and the state in Australia. *Australian & New Zealand journal of criminology*, *39*(1), 20-33. https://doi.org/10.1375/acri.39.1.20

Belenko, S., Hiller, M., & Hamilton, L. (2013). Treating substance use disorders in the criminal justice system. *Current psychiatry reports*, *15*(11), 414-414. https://doi.org/10.1007/s11920-013-0414-z

Bennett, T., Holloway, K., & Farrington, D. (2008). The statistical association between drug misuse and crime: A meta-analysis. *Aggression and Violent Behavior*, *13*(2), 107-118. https://doi.org/10.1016/j.avb.2008.02.001

Berk, R. (2008). *Statistical learning from a regression perspective*. Springer.

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., & Elzarka, A. A. (2020). Almost politically acceptable criminal justice risk assessment. *Criminology & Public Policy*. https://doi.org/https://doi.org/10.1111/1745-9133.12500

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1-42. https://doi.org/10.1177/0049124118782533

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, *27*(4), 222-228. https://doi.org/10.1525/fsr.2015.27.4.222

Biau, J. D., Jolles, M. B., & Porcher, M. R. (2010). P value and the theory of hypothesis testing: An explanation for new researchers. *Clinical Orthopaedics and Related Research*, *468*(3), 885-892. https://doi.org/10.1007/s11999-009-1164-4

Boer, D. P., Couture, J., Geddes, C., & Ritchie, A. (2004). *Yókw'tól: Risk Management Guide for Aboriginal Offenders*. Correctional Services of Canada.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply.* (pp. 18-32). Sage Publications, Inc.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, *29*(4), 355-379. https://doi.org/10.1177/0093854802029004002

Bonta, J., Laprairie, C., & Wallace-Capretta, S. (1997). Risk prediction and re-offending: Aboriginal and non-aboriginal offenders. *Canadian Journal of Criminology*, *39*(2), 127-144.

Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. *arXiv:2006.07155 [cs.LG]*.

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.

Brennan, T. (2016). *An alternative scientific paradigm for criminological risk assessment: Closed or open systems, or both?* Taylor & Francis Ltd.

Brennan, T., & Oliver, W. L. (2013). The emergence of machine learning techniques in criminology. *Criminology & Public Policy*, *12*(3), 551-562. https://doi.org/10.1111/1745-9133.12055

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

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017, December 4-9). *Optimized pre-processing for discrimination prevention* [Paper presentation]. 31st International Conference on Neural Information Processing Systems, Long Beach, California, United States.

Celis, L., Huang, L., Keswani, V., & Vishnoi, N. (2019, January 29-31). *Classification with fairness constraints: A meta-algorithm with provable guarantees* [Paper presentation]. Conference on Fairness, Accountability, and Transparency, Atlanta, GA, United States.

Chenane, J. L., Brennan, P. K., Steiner, B., & Ellison, J. M. (2015). Racial and ethnic differences in the predictive validity of the Level of Service Inventory–Revised among prison inmates. *Criminal Justice and Behavior*, *42*(3), 286-303. https://doi.org/10.1177/0093854814548195

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

Chouldechova, A. (2020). Transparency and simplicity in criminal risk assessment. *Harvard Data Science Review*, *2*(1). https://doi.org/https://doi.org/10.1162/99608f92.b9343eec

Chouldechova, A., & G'Sell, M. (2017). Fairer and more accurate, but for whom?

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv:1810.08810 [cs:LG]*.

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

Cohen, H. W. (2011). P values: Use and misuse in medical literature. *American Journal of Hypertension*, *24*(1), 18-23. https://doi.org/10.1038/ajh.2010.205

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). McGraw Hill.

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

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023 [cs:CY]*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August 13-17). *Algorithmic decision making and the cost of fairness* [Paper presentation]. 23rd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada. https://doi.org/10.1145/3097983.3098095

Corrections Victoria. (2020a). *Corrections Victoria*. https://www.corrections.vic.gov.au/corrections-victoria

Corrections Victoria. (2020b). *Sentence management manual - Part 2*. https://www.corrections.vic.gov.au/sentence-management-manual-part-2

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-220. http://www.jstor.org/stable/2985181

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgement. *Science*, *243*(4899), 1668-1674. https://doi.org/10.1126/science.2648573

Dawson, D. (1999, October 13-15). *Risk of violence assessment: Aboriginal offenders and the assumption of homogeneity* [Paper presentation]. Best Practice Interventions in Corrections for Indigenous People Conference, Adelaide, Australia.

Day, A. (2003). Reducing the risk of re-offending in Australian Indigenous offenders: What works for whom? *Journal of Offender Rehabilitation*, *37*(2), 1-15. https://doi.org/10.1300/J076v37n02_01

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

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe.

https://doi.org/https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence risk assessment: Science and practice. *Legal and Criminological Psychology*, *4*(2), 149-184. https://doi.org/10.1348/135532599167824

Douglas, T., Pugh, J., Singh, I., Savulescu, J., & Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. *European Psychiatry*, *42*, 134-137. https://doi.org/https://doi.org/10.1016/j.eurpsy.2016.12.009

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing*, *9*(6), 649-657. https://doi.org/doi:10.1046/j.1365-2850.2002.00535.x

Dragomir, R. R., & Tadros, E. (2020). Exploring the impacts of racial disparity within the American juvenile justice system. *Juvenile and Family Court Journal*, *71*(2), 61-73. https://doi.org/https://doi.org/10.1111/jfcj.12165

Duwe, G. (2019). Better practices in the development and validation of recidivism risk assessments: The Minnesota Sex Offender Screening Tool–4. *Criminal Justice Policy Review*, *30*(4), 538-564. https://doi.org/10.1177/0887403417718608

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January 8-10). *Fairness through awareness* [Paper presentation]. 3rd Innovations in Theoretical Computer Science Conference, Cambridge, Massachusetts, United States.

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2018). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, *46*(2), 185-209. https://doi.org/10.1177/0093854818811379

Edens, J., Campbell, J., & Weir, J. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior*, *31*(1), 53-75. https://doi.org/10.1007/s10979-006-9019-y

Eisenberg, M. J., van Horn, J. E., Dekker, J. M., Assink, M., van der Put, C. E., Hendriks, J., & Stams, G. J. J. M. (2019). Static and dynamic predictors of general and violent criminal offense recidivism in the forensic outpatient population: A meta-analysis. *Criminal Justice and Behavior*, *46*(5), 732-750. https://doi.org/10.1177/0093854819826109

Ellerby, L., & MacPherson, P. (2002). *Exploring the profiles of Aboriginal sexual offenders: Contrasting Aboriginal and non-Aboriginal sexual offenders to determine unique client characteristics and potential implications for sex offender assessment and treatment strategies (research report no. R-122)*. Correctional Service of Canada. https://doi.org/https://www.publicsafety.gc.ca/lbrr/archives/e%2078.c2%20e45e%202002-eng.pdf

Ewert v. Canada, SCC 30 (2018).

Fazel, S. (2019). The scientific validity of current approaches to violence and criminal risk assessment. In J.W. de Keijser, J.V. Roberts, & J. Ryberg (Eds.), *Predictive sentencing: Normative and empirical perspectives* (1st ed.). Hart Publishing. https://doi.org/10.5040/9781509921447.ch-011

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 (Clinical Research Ed.)*, *345*(7868). https://doi.org/10.1136/bmj.e4692

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August 10-13). *Certifying and removing disparate impact* [Paper presentation]. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.". *Federal Probation*, *80*(2), 38-46.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and 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*. In (Version 4.1-2) [R program]. https://CRAN.R-project.org/package=glmnet

Frize, M., Kenny, D., & Lennings, C. (2008). The relationship between intellectual disability, Indigenous status and risk of reoffending in juvenile offenders on community orders. *Journal of Intellectual Disability Research*, *52*(6), 510-519. https://doi.org/10.1111/j.1365-2788.2008.01058.x

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439-453. https://doi.org/10.1037/a0015251

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

Goel, S., Shroff, R., Skeem, J., & Slobogin, C. (2018). *The accuracy, equity, and jurisprudence of criminal risk assessment*. https://ssrn.com/abstract=3306723

Gordon, H., Kelty, S. F., & Julian, R. (2015). Psychometric evaluation of the Level of Service/Case Management Inventory among Australian offenders completing community-based sentences. *Criminal Justice and Behavior*, *42*(11), 1089-1109.

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

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293-323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19-30. https://doi.org/10.1037/1040-3590.12.1.19

Gutierrez, L., Helmus, L. M., & Hanson, R. K. (2016). What we know and don't know about risk assessment with offenders of Indigenous heritage. *Journal of Threat Assessment and Management*, *3*(2), 97-106. https://doi.org/10.1037/tam0000064

Gutierrez, L., Wilson, H. A., Rugge, T., & Bonta, J. (2013). The prediction of recidivism with Aboriginal offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology & Criminal Justice*, *55*(1), 55-99. https://doi.org/10.3138/cjccj.2011.E.51

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445-1459. https://doi.org/10.1109/TKDE.2012.72

Hamilton, Z., Neuilly, M.-A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*, *11*(2), 299.

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

Hannah-Moffat, K. (2013). Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, *30*(2), 270-296. https://doi.org/10.1080/07418825.2012.682603

Hannah-Moffat, K., & Maurutto, P. (2010). Re-contextualizing pre-sentence reports: Risk and race. *Punishment & Society*, *12*(3), 262-286. https://doi.org/10.1177/1462474510369442

Hanson, R. K. (2017). Assessing the calibration of actuarial risk scales: A primer on the E/O index. *Criminal Justice and Behavior*, *44*(1), 26-39. https://doi.org/10.1177/0093854816683956

Hanson, R. K., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management*, *1*(2), 102-117. https://doi.org/10.1037/tam0000014

Harcourt, B. E. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

Hardt, M., Price, E., & Srebro, N. (2016, December 5-10). *Equality of opportunity in supervised learning* [Paper presentation]. 30th International Conference on Neural Information Processing Systems, Barcelona, Spain.

Harrell, F. (2020). *rms: Regression modeling strategies*. In (Version 6.0-1) [R package]. https://CRAN.R-project.org/package=rms

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, *3*(1), 121-137. https://doi.org/10.1111/j.2044-8333.1998.tb00354.x

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd edition. ed.). Springer.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2). https://doi.org/https://doi.org/10.9707/2307-0919.1111

Hébert-Johnson, U., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018, July 10-15). *Calibration for the (computationally-identifiable) masses* [Paper presentation]. 35th International Conference on Machine Learning, Stockholm, Sweden.

Heckbert, D., & Turkington, D. (2001). *Turning points: A study of the factors related to the successful reintegration of Aboriginal offenders*. Correctional Service of Canada.

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.

Helmus, L., Babchishin, K., & Blais, J. (2012). Predictive accuracy of dynamic risk factors for Aboriginal and non-Aboriginal sex offenders: An exploratory comparison using STABLE-2007. *International Journal of Offender Therapy and Comparative Criminology*, *56*(6), 856. https://doi.org/10.1177/0306624X11414693

Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior*, *44*(1), 8-25. https://doi.org/10.1177/0093854816678898

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80-86. https://doi.org/10.1080/00401706.2000.10485983

Hollin, C. R., Palmer, E. J., & Clark, D. (2003). The Level of Service Inventory-Revised profile of English prisoners: A needs analysis. *Criminal Justice and Behavior*, *30*(4), 422-440. https://doi.org/https://doi.org/10.1177/0093854803253134

Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2003). Ethnicity, gender, and the Level of Service Inventory-Revised. *Journal of Criminal Justice*, *31*(4), 309-320. https://doi.org/10.1016/S0047-2352(03)00025-4

Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2006). Exploring the validity of the Level of Service Inventory-Revised with Native American offenders. *Journal of Criminal Justice*, *34*(3), 331-337. https://doi.org/10.1016/j.jcrimjus.2006.03.009

Homel, R., Lincoln, R., & Herd, B. (1999). Risk and resilience: Crime and violence prevention in Aboriginal communities. *Australian and New Zealand Journal of Criminology*, *32*(2), 182-196. https://doi.org/10.1177/000486589903200207

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2010). Level of Service Inventory–Revised: Assessing the risk and need characteristics of Australian Indigenous offenders. *Psychiatry, Psychology and Law*, *17*(3), 355-367. https://doi.org/10.1080/13218710903089261

Hsu, C., Caputi, P., & Byrne, M. K. (2011). The Level of Service Inventory-Revised (LSI-R) and Australian offenders: Factor structure, sensitivity, and specificity. *Criminal Justice and Behavior*, *38*(6), 600-618.

Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal*, *68*(6), 1043.

Hurducas, C. C., Singh, J. P., De Ruiter, C., & Petrila, J. (2014). Violence risk assessment tools: A systematic review of surveys. *International Journal of Forensic Mental Health*, *13*(3), 181-192. https://doi.org/10.1080/14999013.2014.942923

Jimenez, A. C., Delgado, R. H., Vardsveen, T. C., & Wiener, R. L. (2018). Validation and application of the LS/CMI in Nebraska probation. *Criminal Justice and Behavior*, *45*(6), 863-884. https://doi.org/10.1177/0093854818763231

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

Jones, N. J., Brown, S. L., Robinson, D., & Frey, D. (2016). Validity of the youth assessment and screening instrument: A juvenile justice tool incorporating risks, needs, and strengths. *Law and Human Behavior*, *40*(2), 182-194. https://doi.org/10.1037/lhb0000170

Jones, R., & Day, A. (2011). Mental health, criminal justice and culture: Some ways forward? *Australasian Psychiatry*, *19*(4), 325-330. https://doi.org/10.3109/10398562.2011.579613

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., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33. https://doi.org/10.1007/s10115-011-0463-8

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.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September 24-28). *Fairness-aware classifier with prejudice remover regularizer* [Paper presentation]. The European Conference on Machine Learning and Knowledge Discovery in Databases, Bristol, United Kingdom.

Kaponen, M. (2020). *Fairness and parameter importance in logistic regression models of criminal sentencing data* [Unpublished master's thesis], Uppsala University. http://uu.diva-portal.org/smash/get/diva2:1459136/FULLTEXT01.pdf

Kassambara, A., Kosinski, M., & Biecek, P. (2020). *survminer: Drawing survival curves using 'ggplot2'*. In (Version 0.4.8) [R program]. https://CRAN.R-project.org/package=survminer

Kehl, D., Guo, P., & Kessler, S. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing*. Harvard Law School. https://doi.org/http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041

Kenny, D. T., & Nelson, P. K. (2008). *Young offenders on community orders: Health, welfare and criminogenic needs*. Sydney University Press. https://www.justicehealth.nsw.gov.au/publications/ch1-3.pdfhttps://books.google.com.au/books?id=g_alyfiXyjQC

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113. https://doi.org/10.1093/jla/laz001

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807 [cs.LG]*.

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

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45-47. https://doi.org/10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Official Journal of the Association for the Treatment of Sexual Abusers (ATSA)*, *16*(2), 107-120. https://doi.org/10.1023/B:SEBU.0000023060.61402.07

Lantz, B. (2015). *Machine learning with R* (2nd ed.). Packt Publishing.

LaPrairie, C. (1995). Seen but not heard: Native people in four Canadian inner cities. *The journal of Human Justice*, *6*(2), 30-45. https://doi.org/10.1007/BF02585441

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lee, S. C., Hanson, R. K., & Blais, J. (2020). Predictive accuracy of the Static-99R and Static-2002R risk tools for identifying Indigenous and white individuals at high risk for sexual recidivism in Canada. *Canadian Psychology/Psychologie canadienne*, *61*(1), 42-57. https://doi.org/10.1037/cap0000182

Lee, S. C., Hanson, R. K., Calkins, C., & Jeglic, E. (2019). Paraphilia and antisociality: Motivations for sexual offending may differ for American whites and blacks. *Sexual Abuse*, *32*(3), 335-365. https://doi.org/10.1177/1079063219828779

Leistico, A.-M., Salekin, R., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior*, *32*(1), 28-45. https://doi.org/10.1007/s10979-007-9096-6

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

Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2018). The importance of calibration in clinical psychology. *Assessment*, *27*(4). https://doi.org/10.1177/1073191117752055

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

Loza, W., & Simourd, D. J. (1994). Psychometric evaluation of the Level of Service Inventory (LSI) among male Canadian federal offenders. *Criminal Justice and Behavior*, *21*(4), 468-480.

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

Lundberg, S., & Lee, S. (2017, December 4-9). *A unified approach to interpreting model predictions* [Paper presentation]. 31st International Conference on Neural Information Processing Systems, Long Beach, California, United States.

Mann, M. (2009). *Good intentions, disappointing results: A progress report on federal Aboriginal corrections*. Office of the Correctional Investigator. https://www.oci-bec.gc.ca/cnt/rpt/pdf/oth-aut/oth-aut20091113-eng.pdf

Martel, J., Brassard, R., & Jaccoud, M. (2011). When two worlds collide: Aboriginal risk management in Canadian corrections. *British Journal of Criminology*, *51*(2), 235-255. https://doi.org/10.1093/bjc/azr003

Mayson, S. G. (2019). Bias in, bias out. *Yale Law Journal*, *128*(8), 2218.

McCuish, E. C., Mathesius, J. R., Lussier, P., & Corrado, R. R. (2018). The cross-cultural generalizability of the Psychopathy Checklist: Youth Version for adjudicated Indigenous youth. *Psychological Assessment*, *30*(2), 192-203. https://doi.org/10.1037/pas0000468

McGrath, A. J., Thompson, A. P., & Goodman-Delahunty, J. (2018). Differentiating predictive validity and practical utility for the Australian adaptation of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior*, *45*(6), 820-839. https://doi.org/10.1177/0093854818762468

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*. In (Version 1.7-8) [R package]. https://CRAN.R-project.org/package=e1071

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*.

Molnar, C. (2020). *iml: Interpretable Machine Learning*. In (Version 0.10.1) [R package]. https://CRAN.R-project.org/package=iml

Molnar, T., Allard, T., McKillop, N., & Rynne, J. (2020). Reliability and predictive validity of the Juvenile Sex Offender Assessment Protocol-II in an Australian context. *International Journal of Offender Therapy and Comparative Criminology*. https://doi.org/10.1177/0306624x19900978

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques*. Sage Publications.

Monahan, J., Skeem, J., & Lowenkamp, C. (2017). Age, risk assessment, and sanctioning: Overestimating the old, underestimating the young. *Law and Human Behavior*, *41*(2), 191-201. https://doi.org/10.1037/lhb0000233

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

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*(1), 489-513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Monahan, J., Steadman, H., Appelbaum, P., Banks, S., Grisso, T., Heilbrun, K., Mulvey, E., Roth, L., & Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services*, *56*(7), 810-815. https://doi.org/10.1176/appi.ps.56.7.810

Morrison, B. (2009). *Identifying and responding to bias in the criminal justice system: A review of international and New Zealand research*. https://www.justice.govt.nz/assets/Documents/Publications/Identifying-and-responding-to-bias-in-the-criminal-justice-system.pdf

Muir, N. M., Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., & Rogers, B. J. (2020). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with Indigenous and caucasian female and male adolescents on probation. *Psychological Assessment*, *32*(6), 594-607. https://doi.org/10.1037/pas0000816

National Health and Medical Research Council. (2007). *National statement on ethical conduct in human research*. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018#block-views-block-file-attachments-content-block-1

Olver, M. E. (2016). Some considerations on the use of actuarial and related forensic measures with diverse correctional populations. *Journal of Threat Assessment and Management*, *3*(2), 107-121. https://doi.org/10.1037/tam0000065

Olver, M. E., Kingston, D. A., & Sowden, J. N. (2020). An examination of latent constructs of dynamic sexual violence risk and need as a function of Indigenous and Nonindigenous ancestry. *Psychological Services*. https://doi.org/10.1037/ser0000414

Olver, M. E., Neumann, C. S., Wong, S., & Hare, R. D. (2013). The structural and predictive properties of the Psychopathy Checklist-Revised in Canadian Aboriginal and non-Aboriginal offenders. *Psychological Assessment*, *25*(1), 167-179. https://doi.org/10.1037/a0029840

Olver, M. E., Sowden, J. N., Kingston, D. A., Nicholaichuk, T. P., Gordon, A., Beggs Christofferson, S. M., & Wong, S. C. P. (2018). Predictive accuracy of Violence Risk Scale–Sexual Offender Version risk and change scores in treated Canadian Aboriginal and non-Aboriginal sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *30*(3), 254-275. https://doi.org/10.1177/1079063216649594

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders. *Criminal Justice and Behavior*, *36*(4), 329-353. https://doi.org/10.1177/0093854809331457

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

Palmer, E. J., & Hollin, C. R. (2007). The Level of Service Inventory-Revised with English women prisoners: A needs and reconviction analysis. *Criminal Justice and Behavior*, *34*(8), 971-984.

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

Payne, J. (2007). *Recidivism in Australia: Findings and future research*. Australian Institute of Criminology. https://www.aic.gov.au/sites/default/files/2020-05/rpp080.pdf

Perley-Robertson, B., Helmus, L. M., & Forth, A. (2019). Predictive accuracy of static risk factors for Canadian Indigenous offenders compared to non-Indigenous offenders: Implications for risk assessment scales. *Psychology, Crime & Law*, *25*(3). https://doi.org/10.1080/1068316X.2018.1519827

Perrault, R. T., Vincent, G. M., & Guy, L. S. (2017). Are risk assessments racially biased?: Field study of the SAVRY and YLS/CMI in probation. *Psychological Assessment*, *29*(6), 664-678. https://doi.org/10.1037/pas0000445

Peters, H. (2015). *Game theory: A multi-leveled approach* (2nd ed.). Springer.

Pflueger, M. O., Franke, I., Graf, M., & Hachtel, H. (2015). Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC Psychiatry*, *15*. https://doi.org/doi: 10.1186/s12888-015-0447-4

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. (2017, December 4-9). *On fairness and calibration* [Paper presentation]. 31st International Conference on Neural Information Processing Systems, Long Beach, California, United States.

Privacy and Data Protection Act 2014. (Vic). Retrieved from https://content.legislation.vic.gov.au/sites/default/files/2020-08/14-60aa026%20authorised.pdf

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). American Psychological Association. https://doi.org/10.1037/11367-000

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

Ramakers, A., Nieuwbeerta, P., Van Wilsem, J., & Dirkzwager, A. (2017). Not just any job will do: A study on employment characteristics and recidivism risks after release. *International Journal of Offender Therapy and Comparative Criminology*, *61*(16), 1795-1818. https://doi.org/10.1177/0306624X16636141

Rettenberger, M., Boer, D. P., & Eher, R. (2011). The predictive accuracy of risk factors in the Sexual Violence Risk–20 (SVR-20). *Criminal Justice and Behavior*, *38*(10), 1009-1027. https://doi.org/10.1177/0093854811416908

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*. In (Version 1.16.2) [R program]. https://CRAN.R-project.org/package=pROC

Royal Statistical Society. (2018). *Algorithms in the justice system: Some statistical issues*. https://www.rss.org.uk/Images/PDF/influencing-change/2018/RSS_submission_Algorithms_in_the_justice_system_Nov_2018.pdf

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

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1). https://doi.org/https://doi.org/10.1162/99608f92.6ed64b30

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

Schaefer, B. P., & Hughes, T. (2019). Examining judicial petrial release decisions: The influence of risk assessments and race. *Criminology, Criminal Justice, Law & Society*, *20*(2), 47-58.

Schmidt, F., McKinnon, L., Chattha, H., & Brownlee, K. (2006). Concurrent and predictive validity of the Psychopathy Checklist: Youth Version across gender and ethnicity. *Psychological Assessment*, *18*(4), 393-401. https://doi.org/10.1037/1040-3590.18.4.393

Schmidt, S., Heffernan, R., & Ward, T. (2020). Why we cannot explain cross-cultural differences in risk assessment. *Aggression and Violent Behavior*, *50*. https://doi.org/https://doi.org/10.1016/j.avb.2019.101346

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. (2016a). Criminal engagement and Australian culturally and linguistically diverse populations: Challenges and implications for forensic risk assessment. *Psychiatry, Psychology and Law*, *23*(2), 256-274. https://doi.org/10.1080/13218719.2015.1053164

Shepherd, S. (2016b). Violence risk instruments may be culturally unsafe for use with Indigenous patients. *Australasian Psychiatry*, *24*(6), 565-567. https://doi.org/10.1177/1039856216665287

Shepherd, S. (2018). Violence risk assessment and Indigenous Australians: A primer. *Alternative Law Journal*, *43*(1), 45-47. https://doi.org/10.1177/1037969X17748210

Shepherd, S. M. (2015). Finding color in conformity: A commentary on culturally specific risk factors for violence in Australia. *International Journal of Offender Therapy and Comparative Criminology*, *59*(12), 1297-1307. https://doi.org/10.1177/0306624x14540492

Shepherd, S. M., Adams, Y., McEntyre, E., & Walker, R. (2014). Violence risk assessment in Australian Aboriginal offender populations: A review of the literature. *Psychology, Public Policy, and Law*, *20*(3), 281-293. https://doi.org/10.1037/law0000017

Shepherd, S. M., & Anthony, T. (2018). Popping the cultural bubble of violence risk assessment tools. *The Journal of Forensic Psychiatry & Psychology*, *29*(2), 211-220. https://doi.org/10.1080/14789949.2017.1354055

Shepherd, S. M., & Lewis-Fernandez, R. (2016). Forensic risk assessment and cultural diversity: Contemporary challenges and future directions. *Psychology, Public Policy, and Law*, *22*(4), 427-438. https://doi.org/10.1037/law0000102

Shepherd, S. M., Luebbers, S., Ferguson, M., Ogloff, J., & Dolan, M. (2014). The utility of the SAVRY across ethnicity in Australian young offenders. *Psychology, Public Policy, and Law*, *20*(1), 31-45. https://doi.org/10.1037/a0033972

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

Shepherd, S. M., & Spivak, B. L. (2020). Finding color in conformity part II: Reflections on structured professional judgement risk assessment. *International Journal of Offender Therapy and Comparative Criminology*, *65*(1). https://doi.org/https://doi.org/10.1177/0306624X20928025

Shepherd, S. M., & Strand, S. (2016). The PCL: YV and re-offending across ethnic groups. *Journal of Criminal Psychology*, *6*(2), 51-62. https://doi.org/10.1108/JCP-02-2016-0006

Shepherd, S. M., & Willis-Esqueda, C. (2018). Indigenous perspectives on violence risk assessment: A thematic analysis. *Punishment and Society*, *20*(5), 599-627. https://doi.org/10.1177/1462474517721485

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.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, *31*(1), 8-22. https://doi.org/doi:10.1002/bsl.2052

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/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*. https://doi.org/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/doi:10.1111/1745-9125.12123

Smallbone, S., & Rallings, M. (2013). Short-term predictive validity of the Static-99 and Static-99-R for Indigenous and nonindigenous Australian sexual offenders. *Sexual*

Abuse A Journal of Research and Treatment, *25*(3), 302-316. https://doi.org/10.1177/1079063212472937

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

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293. https://doi.org/10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1-26. https://doi.org/10.1111/1529-1006.001

The United States Department of Justice. (2014). *Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th annual meeting and 13th State Criminal Justice network conference*. https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Therneau, T. M. (2020). *survival: Survival analysis*. In (Version 3.2-7) [R program]. https://cran.r-project.org/package=survival

Thiele, C. (2021). *cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks*. In (Version 1.1.1) [R package]. https://CRAN.R-project.org/package=cutpointr

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior*, *36*(4), 345-355. https://doi.org/10.1037/h0093930

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 273-282.

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), e0213245. https://doi.org/10.1371/journal.pone.0213245

Tsang, S., Piquero, A. R., & Cauffman, E. (2014). An examination of the Psychopathy Checklist: Youth Version (PCL:YV) among male adolescent offenders: An item response theory analysis. *Psychological Assessment*, *26*(4), 1333-1346. https://doi.org/10.1037/a0037500

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119-135. https://doi.org/10.1016/j.erap.2003.12.004

van Eijk, G. (2017). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, *19*(4), 463-481. https://doi.org/10.1177/1462474516666282

Vapnik, V. (1998). *Statistical learning theory*. Wiley.

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 US.

Venner, S., Sivasubramaniam, D., Luebbers, S., & Shepherd, S. M. (2021). Cross-cultural reliability and rater bias in forensic risk assessment: A review of the literature. *Psychology, Crime & Law*, *27*(2), 1-17. https://doi.org/10.1080/1068316X.2020.1775829

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

Viallon, V., Ragusa, S., Clavel-Chapelon, F., & Bénichou, J. (2009). How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine*, *28*(6), 901-916. https://doi.org/10.1002/sim.3517

Victoria Police. (2019). *Crime statistics*. https://www.police.vic.gov.au/crime-statistics

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

Warner, B., Spivak, B., Ashford, L., Fix, R., Ogloff, J., & Shepherd, S. (2021). The impact of offender–victim cultural backgrounds on the likelihood of receiving diversion. *Criminal Justice Policy Review*. https://doi.org/10.1177/08874034211046313

Watkins, I. (2011). *The utility of Level of Service Inventory-Revised (LSI-R) assessments within NSW correctional environments. Research bulletin*. Corrective Services NSW. https://doi.org/https://correctiveservices.dcj.nsw.gov.au/content/dam/dcj/corrective-services-nsw/documents/research-and-statistics/rb29-utility-of-level-of-service-inventory-.pdf

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). Assessing risk of violence to others. In C. D. Webster & M. A. Jackson (Eds.), *Impulsivity: Theory, assessment, and treatment* (pp. 251-277). Guilford Press.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, *59*(7), 595-613. https://doi.org/10.1037/0003-066X.59.7.595

Whiteacre, K. W. (2006). Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review*, *17*(3), 330-342. https://doi.org/10.1177/0887403405284766

Wickham, H. (2019). *tidyverse: Easily install and load the 'tidyverse.'*. In (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

Wisser, L. (2019). Pandora's algorithmic black box: The challenges of using algorithmic risk assessments in sentencing. *American Criminal Law Review*, *56*(4), 1811-1832.

Wormith, J., & Hogg, S. (2012). *The predictive validity of Aboriginal offender recidivism with a general risk/needs assessment inventory*. Program Effectiveness, Statistics, and Applied Research Unit, Ministry of Community Safety and Correctional. https://doi.org/https://cfbsjs.usask.ca/documents/research/research_papers/LSI-OR%20Aboriginal%20Paper%20w%20Abstract.pdf

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. https://doi.org/10.1177/0093854814552843

Yang, M., Wong, S., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*(5), 740. https://doi.org/https://doi.org/10.1037/a0020473

Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., & Razali, M. N. (2020). A classification approach for crime prediction. In M. Khalaf, D. Al-Jumeily, & A. Lisitsa (Eds.), *Applied computing to support industry: Innovation and technology* (pp. 68-78). Springer International Publishing.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, June 16-21). *Learning fair representations* [Paper presentation]. 30th International Conference on Machine Learning, Atlanta, GA, USA.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A*, *180*(3), 689-722. https://doi.org/10.1111/rssa.12227

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, February 2-3). *Mitigating unwanted biases with adversarial learning* [Paper presentation]. 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, United States.

Zottola, S. A., Desmarais, S. L., Lowder, E. M., & Duhart Clarke, S. E. (2021). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*. https://doi.org/10.1177/00938548211040544

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, *67*, 301-320.

Appendix A

Authorship Indication Forms

Swinburne Research

Authorship Indication Form

For HDR students

NOTE

This Authorship Indication form is a statement detailing the percentage of the contribution of each author in each submitted/published 'paper.' This form must be signed by each co-author and the Principal Supervisor. This form must be added to the publication of your final thesis as an appendix. Please fill outa separate form for each published paper to be included in your thesis.

DECLARATION

We hereby declare our contribution to the publication of the 'paper' entitled:

Cross-Cultural Fairness in Forensic Risk Instruments: A Review of the Literature

First Author

Name: Linda J. Ashford Signature:

Percentage of contribution: 80% Date: 17 / 01 / 2022

Brief description of contribution to the 'paper' and your central responsibilities/role on project:

Reviewed and analysed the literature to be incorporated in this paper. Wrote a complete manuscript that was revised in line with suggestions from my supervisors.

Second Author

Name: Benjamin L. Spivak Signature:

Percentage of contribution: 10% Date: 17 / 01 / 2022

Brief description of your contribution to the 'paper':

Involved in the conceptualisation, review, and editing of the manuscript.

Third Author

Name: Stephane M. Shepherd Signature:

Percentage of contribution: 10% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in the conceptualisation, review, and editing of the manuscript.

Principal Supervisor:

Name: Stephane M. Shepherd Signature:

Date: 17 / 01 / 2022

In the case of more than four authors please attach another sheet with the names, signatures, and contribution of the authors.

SWINBURNE UNIVERSITY OF TECHNOLOGY

Swinburne Research

Authorship Indication Form

For HDR students

NOTE

This Authorship Indication form is a statement detailing the percentage of the contribution of each author in each submitted/published 'paper.' This form must be signed by each co-author and the Principal Supervisor. This form must be added to the publication of your final thesis as an appendix. Please fill outa separate form for each published paper to be included in your thesis.

DECLARATION

We hereby declare our contribution to the publication of the 'paper' entitled:

The Cross-Cultural Fairness of the LS/RNR: An Australian Analysis

First Author

Name: Linda J. Ashford Signature:

Percentage of contribution: 80% Date: 17 / 01 / 2022

Brief description of contribution to the 'paper' and your central responsibilities/role on project:

Completed data cleaning, linkage, and analysis. Reviewed appropriate literature and wrote complete manuscript. Revised manuscript and made changes in line with feedback from supervisors and co-authors.

Second Author

Name: Benjamin L. Spivak Signature:

Percentage of contribution: 10% Date: 17 / 01 / 2022

Brief description of your contribution to the 'paper':

Involved in the conceptualisation and review of the manuscript. Assisted with data analysis questions.

Third Author

Name: James R. P. Ogloff Signature:

Percentage of contribution: 3% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in reviewing the manuscript.

Fourth Author

Name: Stephane M. Shepherd Signature:

Percentage of contribution: 7% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in the conceptualizing and review of manuscript.

Principal Supervisor:

Name: Stephane M. Shepherd Signature:

Date: 17 / 01 / 2022

In the case of more than four authors please attach another sheet with the names, signatures, and contribution of the authors.

SWINBURNE UNIVERSITY OF TECHNOLOGY

Swinburne Research

Authorship Indication Form

For HDR students

NOTE

This Authorship Indication form is a statement detailing the percentage of the contribution of each author in each submitted/published 'paper.' This form must be signed by each co-author and the Principal Supervisor. This form must be added to the publication of your final thesis as an appendix. Please fill outa separate form for each published paper to be included in your thesis.

DECLARATION

We hereby declare our contribution to the publication of the 'paper' entitled:

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

First Author

Name: Linda J. Ashford Signature:

Percentage of contribution: 80 % Date: 17 / 01 / 2022

Brief description of contribution to the 'paper' and your central responsibilities/role on project:

Completed data cleaning, linkage, and analysis. Reviewed appropriate literature and wrote complete manuscript. Revised manuscript and made changes in line with feedback from supervisors and co-authors.

Second Author

Name: Benjamin L. Spivak Signature:

Percentage of contribution: 10% Date: 17 / 01 / 2022

Brief description of your contribution to the 'paper':

Involved in the conceptualisation and review of the manuscript. Assisted with data analysis questions.

Third Author

Name: James R. P. Ogloff Signature:

Percentage of contribution: 3% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in reviewing the manuscript.

Fourth Author

Name: Stephane M. Shepherd Signature:

Percentage of contribution: 7% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in the conceptualizing and review of manuscript.

Principal Supervisor:

Name: Stephane M. Shepherd Signature:

Date: 17 / 01 / 2022

In the case of more than four authors please attach another sheet with the names, signatures, and contribution of the authors.

Swinburne Research

Authorship Indication Form

For HDR students

NOTE

This Authorship Indication form is a statement detailing the percentage of the contribution of each author in each submitted/published 'paper.' This form must be signed by each co-author and the Principal Supervisor. This form must be added to the publication of your final thesis as an appendix. Please fill outa separate form for each published paper to be included in your thesis.

DECLARATION

We hereby declare our contribution to the publication of the 'paper' entitled:

Increasing the Cross-Cultural Fairness of the LS/RNR and Interpretability of Statistical Learning Methods

First Author

Name: Linda J. Ashford Signature:

Percentage of contribution: 80 % Date: 17 / 01 / 2022

Brief description of contribution to the 'paper' and your central responsibilities/role on project:

Completed data cleaning, linkage, and analysis. Reviewed appropriate literature and wrote complete manuscript. Revised manuscript and made changes in line with feedback from supervisors and co-authors.

Second Author

Name: Benjamin L. Spivak Signature:

Percentage of contribution: 10% Date: 17 / 01 / 2022

Brief description of your contribution to the 'paper':

Involved in the conceptualisation and review of the manuscript. Assisted with data analysis questions.

Third Author

Name: James R. P. Ogloff Signature:

Percentage of contribution:___3% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in reviewing the manuscript.

Fourth Author

Name:__Stephane M. Shepherd_________________________Signature:

Percentage of contribution:___7% Date: 17 / 01 / 2022

Brief description of your contribution to the ‘paper’:

Involved in the conceptualisation and review of the manuscript.

Principal Supervisor:

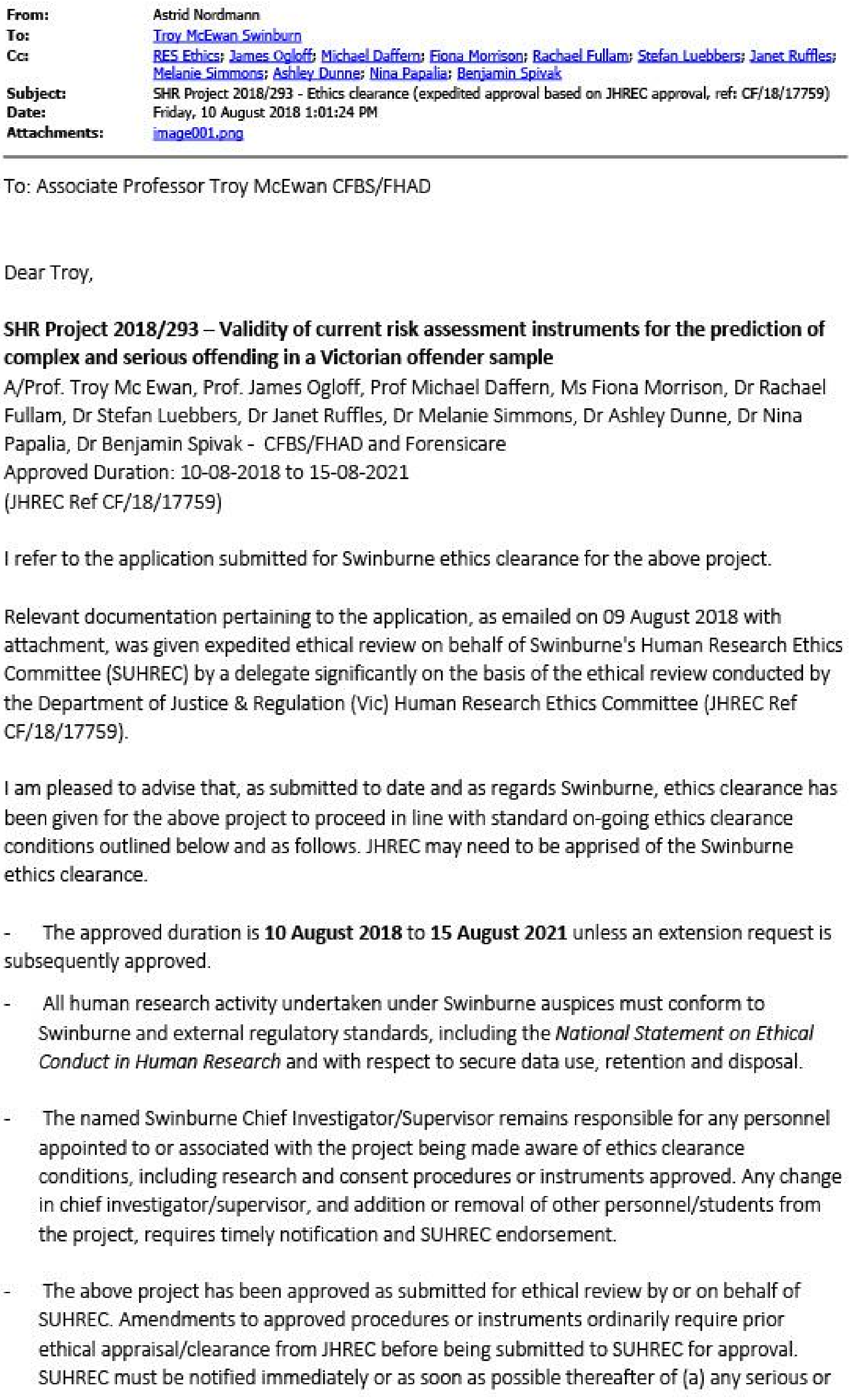
Name:__Stephane M. Shepherd____________________Signature:________________________

Date: 17 / 01 / 2022

In the case of more than four authors please attach another sheet with the names, signatures, and contribution of the authors.

Appendix B

Ethics Approval Letters



From:	Astrid Nordmann
To:	Troy McEwan Swinburn
Cc:	RES Ethics; James Ogloff; Michael Daffern; Fiona Morrison; Rachael Fullam; Stefan Luebbers; Janet Ruffles; Melanie Simmons; Ashley Dunne; Nina Papalia; Benjamin Spivak
Subject:	SHR Project 2018/293 - Ethics clearance (expedited approval based on JHREC approval, ref: CF/18/17759)
Date:	Friday, 10 August 2018 1:01:24 PM
Attachments:	image001.png

To: Associate Professor Troy McEwan CFBS/FHAD

Dear Troy,

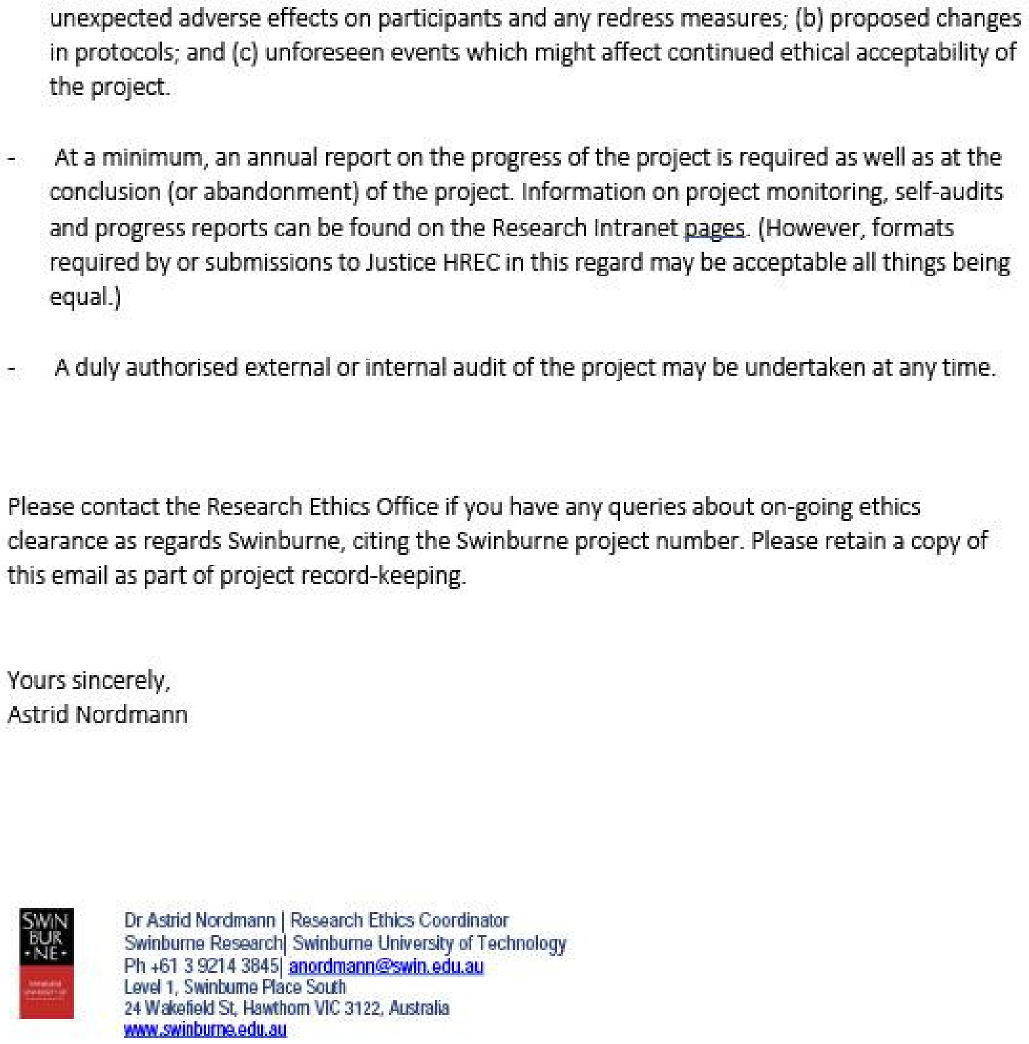
SHR Project 2018/293 – Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample
A/Prof. Troy Mc Ewan, Prof. James Ogloff, Prof Michael Daffern, Ms Fiona Morrison, Dr Rachael Fullam, Dr Stefan Luebbers, Dr Janet Ruffles, Dr Melanie Simmons, Dr Ashley Dunne, Dr Nina Papalia, Dr Benjamin Spivak - CFBS/FHAD and Forensicare
Approved Duration: 10-08-2018 to 15-08-2021
(JHREC Ref CF/18/17759)

I refer to the application submitted for Swinburne ethics clearance for the above project.

Relevant documentation pertaining to the application, as emailed on 09 August 2018 with attachment, was given expedited ethical review on behalf of Swinburne's Human Research Ethics Committee (SUHREC) by a delegate significantly on the basis of the ethical review conducted by the Department of Justice & Regulation (Vic) Human Research Ethics Committee (JHREC Ref CF/18/17759).

I am pleased to advise that, as submitted to date and as regards Swinburne, ethics clearance has been given for the above project to proceed in line with standard on-going ethics clearance conditions outlined below and as follows. JHREC may need to be apprised of the Swinburne ethics clearance.

- The approved duration is **10 August 2018** to **15 August 2021** unless an extension request is subsequently approved.
- All human research activity undertaken under Swinburne auspices must conform to Swinburne and external regulatory standards, including the *National Statement on Ethical Conduct in Human Research* and with respect to secure data use, retention and disposal.
- The named Swinburne Chief Investigator/Supervisor remains responsible for any personnel appointed to or associated with the project being made aware of ethics clearance conditions, including research and consent procedures or instruments approved. Any change in chief investigator/supervisor, and addition or removal of other personnel/students from the project, requires timely notification and SUHREC endorsement.
- The above project has been approved as submitted for ethical review by or on behalf of SUHREC. Amendments to approved procedures or instruments ordinarily require prior ethical appraisal/clearance from JHREC before being submitted to SUHREC for approval. SUHREC must be notified immediately or as soon as possible thereafter of (a) any serious or



unexpected adverse effects on participants and any redress measures; (b) proposed changes in protocols; and (c) unforeseen events which might affect continued ethical acceptability of the project.

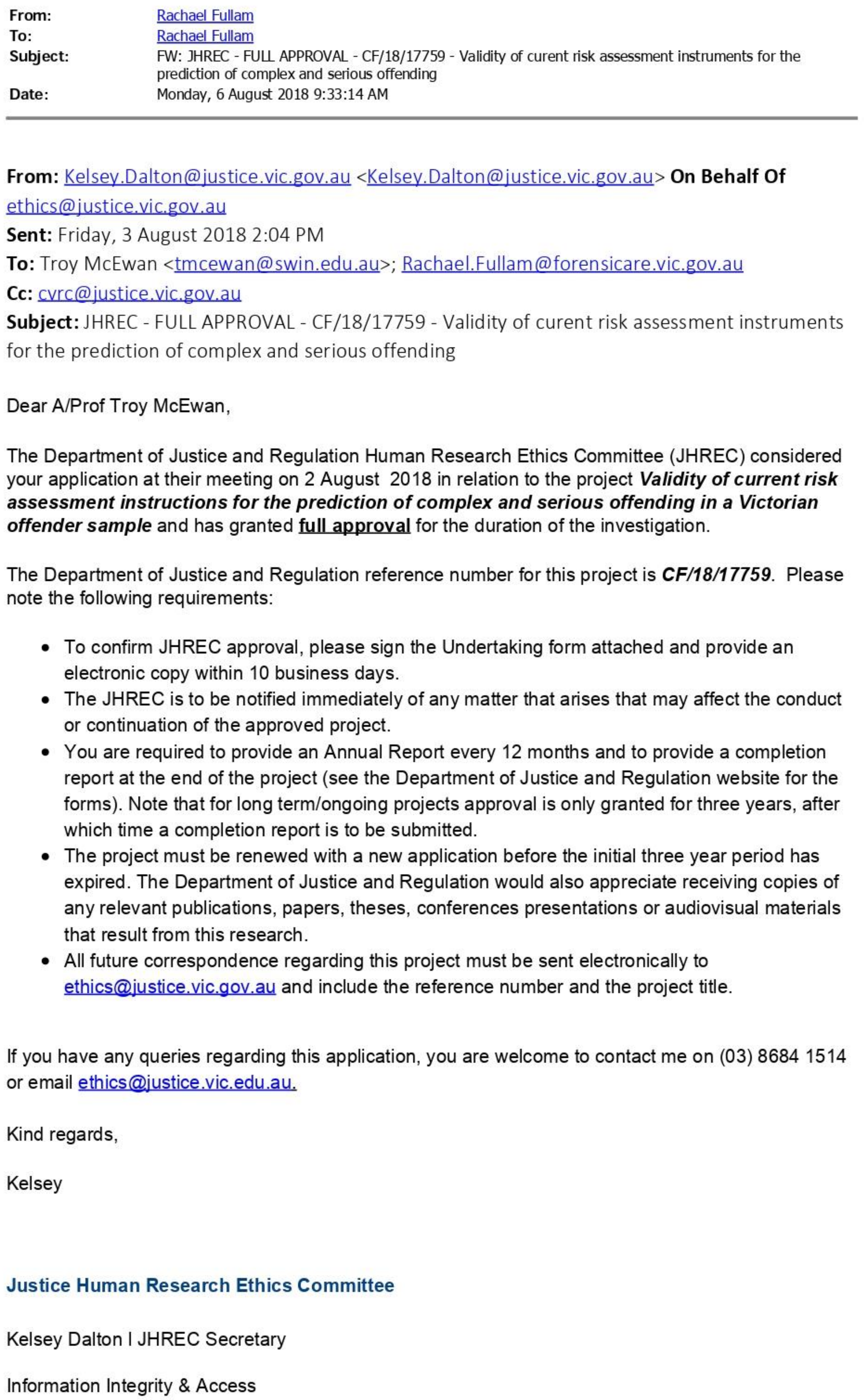
- At a minimum, an annual report on the progress of the project is required as well as at the conclusion (or abandonment) of the project. Information on project monitoring, self-audits and progress reports can be found on the Research Intranet pages. (However, formats required by or submissions to Justice HREC in this regard may be acceptable all things being equal.)

- A duly authorised external or internal audit of the project may be undertaken at any time.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance as regards Swinburne, citing the Swinburne project number. Please retain a copy of this email as part of project record-keeping.

Yours sincerely,
Astrid Nordmann

Dr Astrid Nordmann | Research Ethics Coordinator
Swinburne Research| Swinburne University of Technology
Ph +61 3 9214 3845| anordmann@swin.edu.au
Level 1, Swinburne Place South
24 Wakefield St, Hawthorn VIC 3122, Australia
www.swinburne.edu.au



From:	Rachael Fullam
To:	Rachael Fullam
Subject:	FW: JHREC - FULL APPROVAL - CF/18/17759 - Validity of curent risk assessment instruments for the prediction of complex and serious offending
Date:	Monday, 6 August 2018 9:33:14 AM

From: Kelsey.Dalton@justice.vic.gov.au <Kelsey.Dalton@justice.vic.gov.au> **On Behalf Of** ethics@justice.vic.gov.au
Sent: Friday, 3 August 2018 2:04 PM
To: Troy McEwan <tmcewan@swin.edu.au>; Rachael.Fullam@forensicare.vic.gov.au
Cc: cvrc@justice.vic.gov.au
Subject: JHREC - FULL APPROVAL - CF/18/17759 - Validity of curent risk assessment instruments for the prediction of complex and serious offending

Dear A/Prof Troy McEwan,

The Department of Justice and Regulation Human Research Ethics Committee (JHREC) considered your application at their meeting on 2 August 2018 in relation to the project ***Validity of current risk assessment instructions for the prediction of complex and serious offending in a Victorian offender sample*** and has granted **full approval** for the duration of the investigation.

The Department of Justice and Regulation reference number for this project is ***CF/18/17759***. Please note the following requirements:

- To confirm JHREC approval, please sign the Undertaking form attached and provide an electronic copy within 10 business days.
- The JHREC is to be notified immediately of any matter that arises that may affect the conduct or continuation of the approved project.
- You are required to provide an Annual Report every 12 months and to provide a completion report at the end of the project (see the Department of Justice and Regulation website for the forms). Note that for long term/ongoing projects approval is only granted for three years, after which time a completion report is to be submitted.
- The project must be renewed with a new application before the initial three year period has expired. The Department of Justice and Regulation would also appreciate receiving copies of any relevant publications, papers, theses, conferences presentations or audiovisual materials that result from this research.
- All future correspondence regarding this project must be sent electronically to ethics@justice.vic.gov.au and include the reference number and the project title.

If you have any queries regarding this application, you are welcome to contact me on (03) 8684 1514 or email ethics@justice.vic.edu.au.

Kind regards,

Kelsey

Justice Human Research Ethics Committee

Kelsey Dalton I JHREC Secretary

Information Integrity & Access

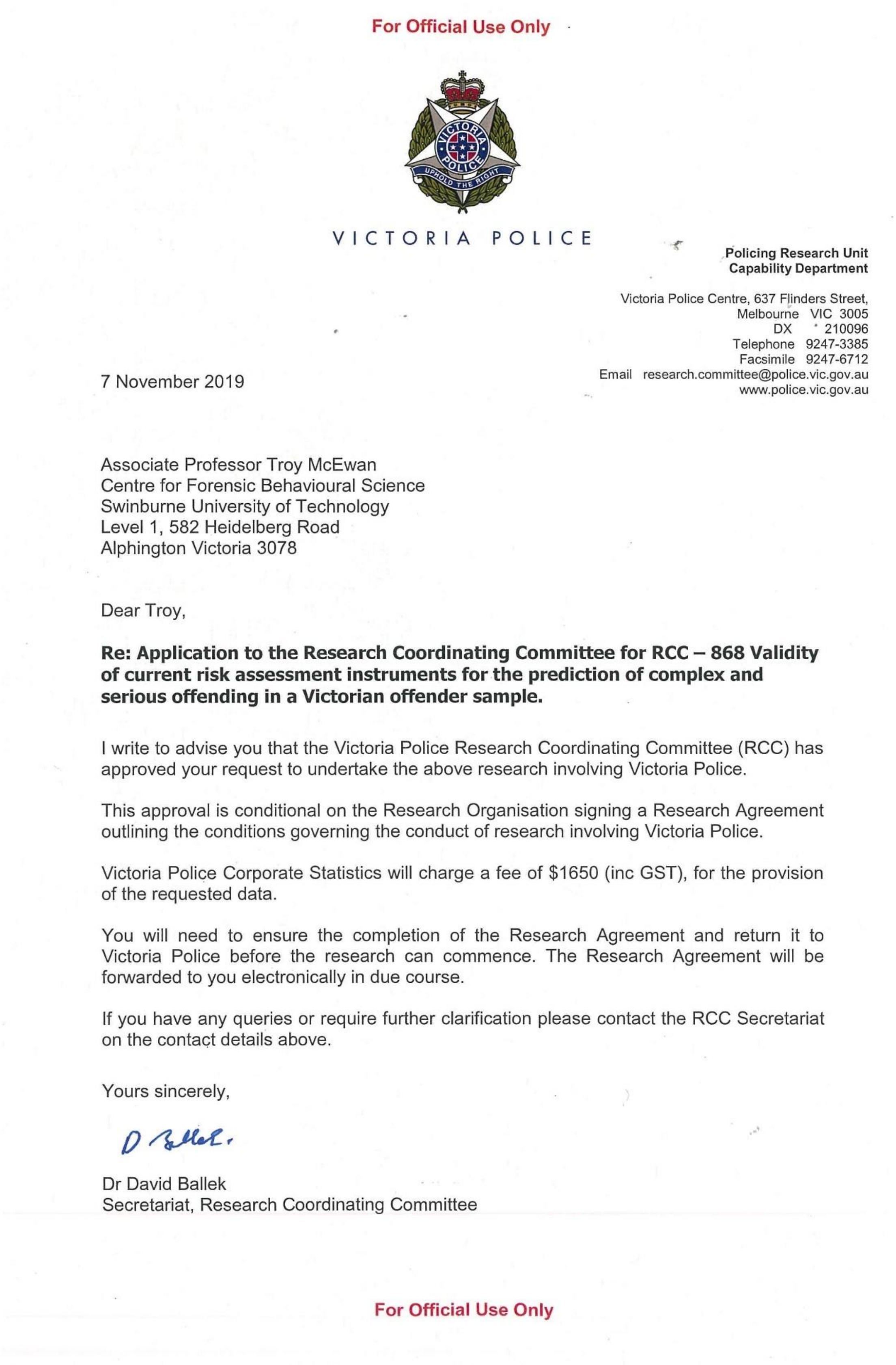
Department of Justice & Regulation
GPO Box 4356 / Level 24, 121 Exhibition Street
MELBOURNE VIC 3001

Phone: (03) 8684 1514

Email: ethics@justice.vic.gov.au
Web: http://www.justice.vic.gov.au/utility/data+and+research/making+a+jhrec+application

Please help to conserve paper by not printing this e-mail

PRIVATE & CONFIDENTIAL
The content of this e-mail and any attachments may be private and confidential, intended only for use of the individual or entity named. If you are not the intended recipient of this message you must not read, forward, print, copy, disclose, use or store in any way the information this e-mail or any attachment contains.
If you are not the intended recipient, please notify the sender immediately and delete or destroy all copies of this e-mail and any attachments.
Our organisation respects the privacy of individuals. For a copy of our privacy policy please go to our website or contact us.



For Official Use Only

VICTORIA POLICE

Policing Research Unit
Capability Department

Victoria Police Centre, 637 Flinders Street,
Melbourne VIC 3005
DX 210096
Telephone 9247-3385
Facsimile 9247-6712
Email research.committee@police.vic.gov.au
www.police.vic.gov.au

7 November 2019

Associate Professor Troy McEwan
Centre for Forensic Behavioural Science
Swinburne University of Technology
Level 1, 582 Heidelberg Road
Alphington Victoria 3078

Dear Troy,

Re: Application to the Research Coordinating Committee for RCC – 868 Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample.

I write to advise you that the Victoria Police Research Coordinating Committee (RCC) has approved your request to undertake the above research involving Victoria Police.

This approval is conditional on the Research Organisation signing a Research Agreement outlining the conditions governing the conduct of research involving Victoria Police.

Victoria Police Corporate Statistics will charge a fee of $1650 (inc GST), for the provision of the requested data.

You will need to ensure the completion of the Research Agreement and return it to Victoria Police before the research can commence. The Research Agreement will be forwarded to you electronically in due course.

If you have any queries or require further clarification please contact the RCC Secretariat on the contact details above.

Yours sincerely,

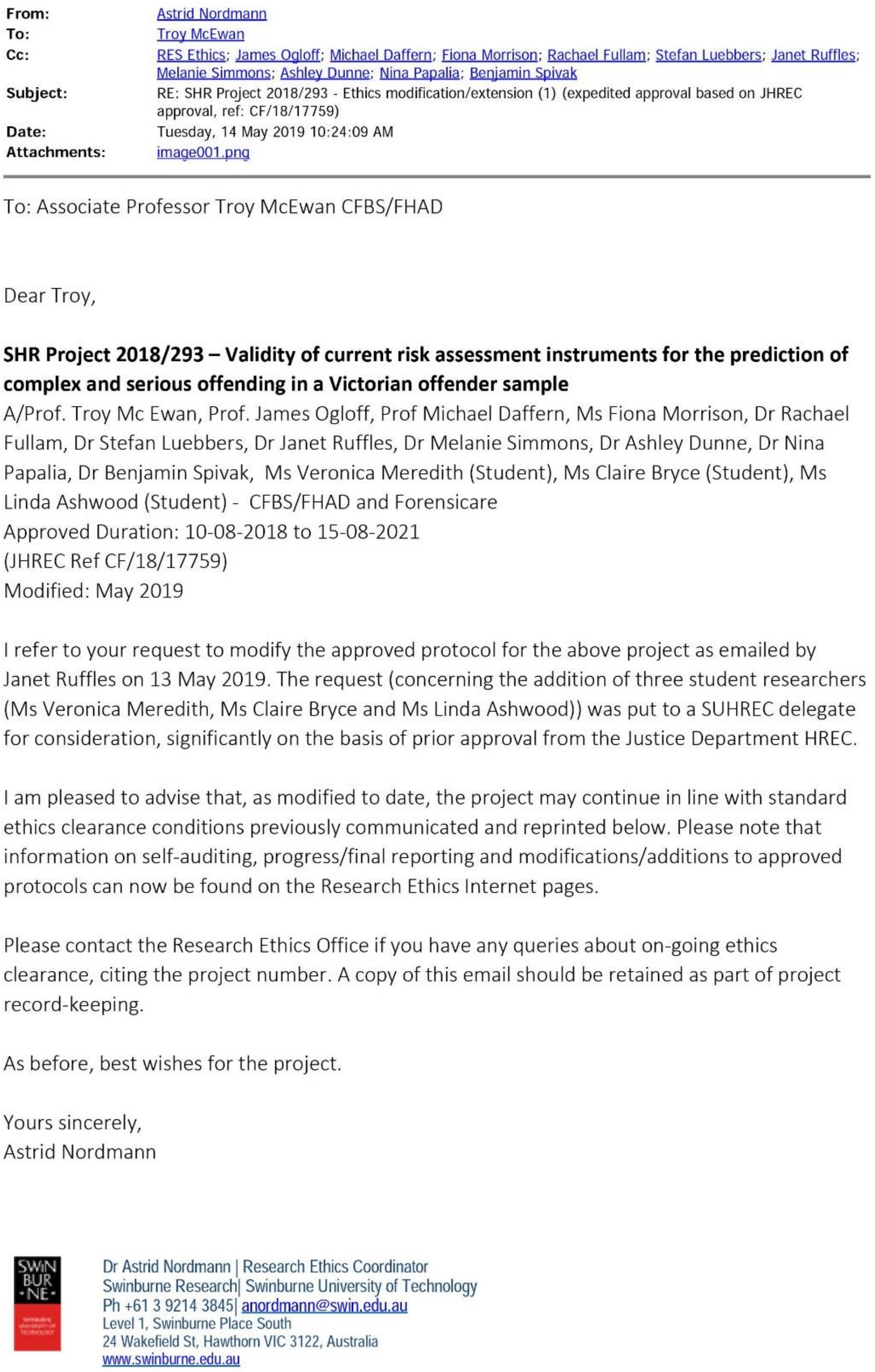
D Ballek

Dr David Ballek
Secretariat, Research Coordinating Committee

For Official Use Only

Appendix C

Ethics Amendment Approval Letters – Access to LS/RNR Data



From:	Astrid Nordmann
To:	Troy McEwan
Cc:	RES Ethics; James Ogloff; Michael Daffern; Fiona Morrison; Rachael Fullam; Stefan Luebbers; Janet Ruffles; Melanie Simmons; Ashley Dunne; Nina Papalia; Benjamin Spivak
Subject:	RE: SHR Project 2018/293 - Ethics modification/extension (1) (expedited approval based on JHREC approval, ref: CF/18/17759)
Date:	Tuesday, 14 May 2019 10:24:09 AM
Attachments:	image001.png

To: Associate Professor Troy McEwan CFBS/FHAD

Dear Troy,

SHR Project 2018/293 – Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample
A/Prof. Troy Mc Ewan, Prof. James Ogloff, Prof Michael Daffern, Ms Fiona Morrison, Dr Rachael Fullam, Dr Stefan Luebbers, Dr Janet Ruffles, Dr Melanie Simmons, Dr Ashley Dunne, Dr Nina Papalia, Dr Benjamin Spivak, Ms Veronica Meredith (Student), Ms Claire Bryce (Student), Ms Linda Ashwood (Student) - CFBS/FHAD and Forensicare
Approved Duration: 10-08-2018 to 15-08-2021
(JHREC Ref CF/18/17759)
Modified: May 2019

I refer to your request to modify the approved protocol for the above project as emailed by Janet Ruffles on 13 May 2019. The request (concerning the addition of three student researchers (Ms Veronica Meredith, Ms Claire Bryce and Ms Linda Ashwood)) was put to a SUHREC delegate for consideration, significantly on the basis of prior approval from the Justice Department HREC.

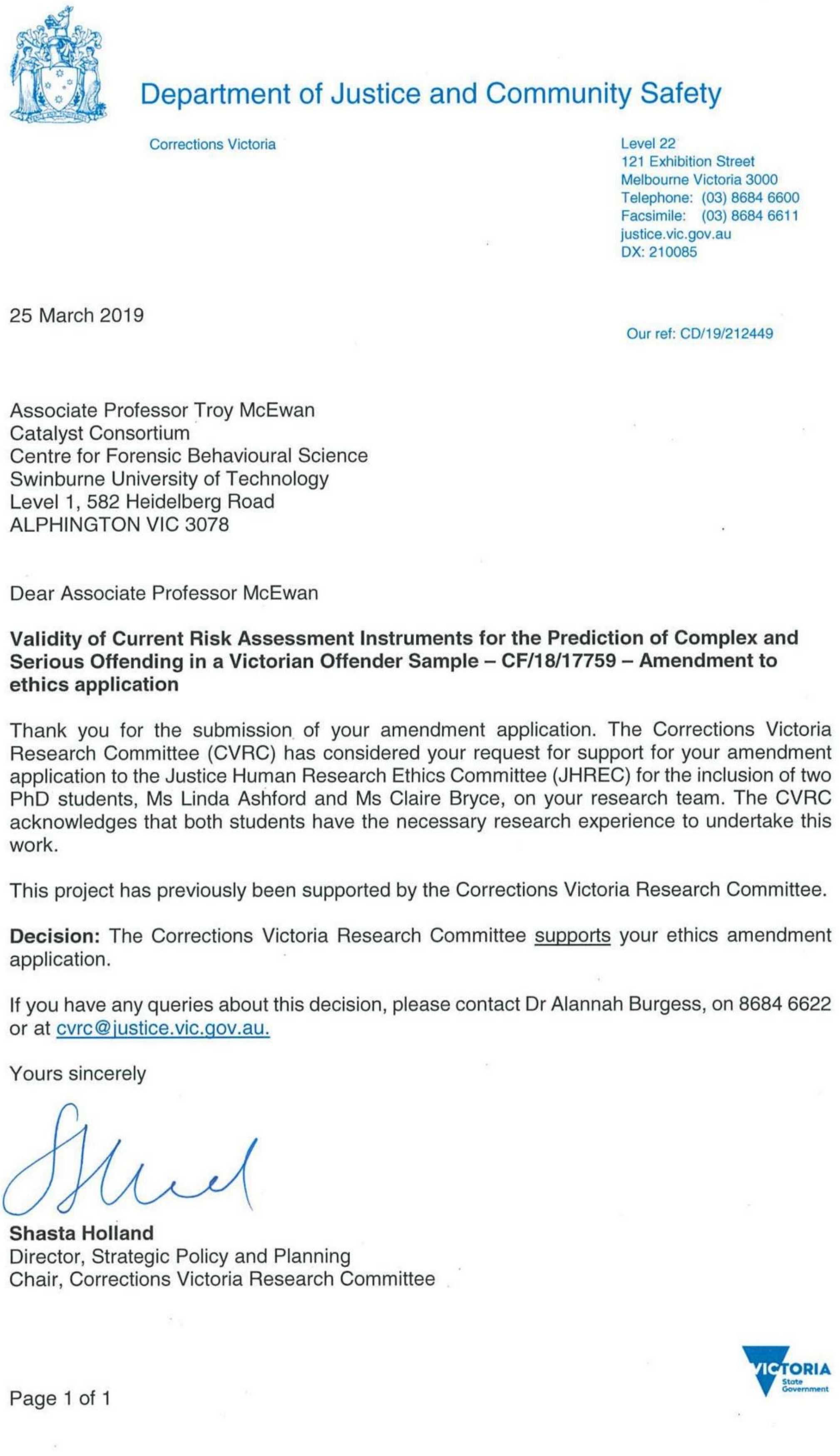
I am pleased to advise that, as modified to date, the project may continue in line with standard ethics clearance conditions previously communicated and reprinted below. Please note that information on self-auditing, progress/final reporting and modifications/additions to approved protocols can now be found on the Research Ethics Internet pages.

Please contact the Research Ethics Office if you have any queries about on-going ethics clearance, citing the project number. A copy of this email should be retained as part of project record-keeping.

As before, best wishes for the project.

Yours sincerely,
Astrid Nordmann

Dr Astrid Nordmann | Research Ethics Coordinator
Swinburne Research| Swinburne University of Technology
Ph +61 3 9214 3845| anordmann@swin.edu.au
Level 1, Swinburne Place South
24 Wakefield St, Hawthorn VIC 3122, Australia
www.swinburne.edu.au



Department of Justice and Community Safety

Corrections Victoria

Level 22
121 Exhibition Street
Melbourne Victoria 3000
Telephone: (03) 8684 6600
Facsimile: (03) 8684 6611
justice.vic.gov.au
DX: 210085

25 March 2019

Our ref: CD/19/212449

Associate Professor Troy McEwan
Catalyst Consortium
Centre for Forensic Behavioural Science
Swinburne University of Technology
Level 1, 582 Heidelberg Road
ALPHINGTON VIC 3078

Dear Associate Professor McEwan

Validity of Current Risk Assessment Instruments for the Prediction of Complex and Serious Offending in a Victorian Offender Sample – CF/18/17759 – Amendment to ethics application

Thank you for the submission of your amendment application. The Corrections Victoria Research Committee (CVRC) has considered your request for support for your amendment application to the Justice Human Research Ethics Committee (JHREC) for the inclusion of two PhD students, Ms Linda Ashford and Ms Claire Bryce, on your research team. The CVRC acknowledges that both students have the necessary research experience to undertake this work.

This project has previously been supported by the Corrections Victoria Research Committee.

Decision: The Corrections Victoria Research Committee supports your ethics amendment application.

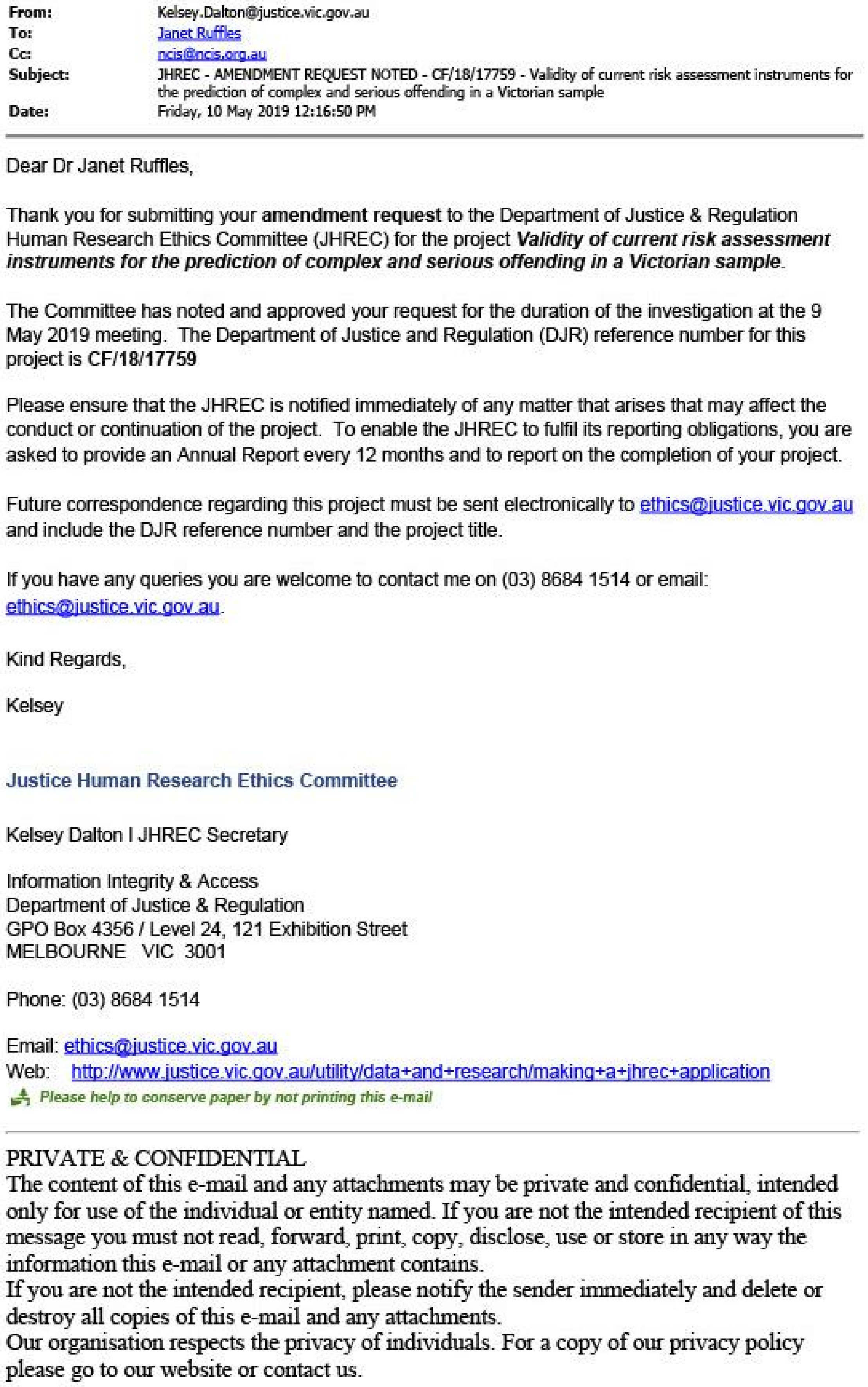
If you have any queries about this decision, please contact Dr Alannah Burgess, on 8684 6622 or at cvrc@justice.vic.gov.au.

Yours sincerely

Shasta Holland
Director, Strategic Policy and Planning
Chair, Corrections Victoria Research Committee

Page 1 of 1

VICTORIA
State
Government



From: Kelsey.Dalton@justice.vic.gov.au
To: Janet Ruffles
Cc: ncis@ncis.org.au
Subject: JHREC - AMENDMENT REQUEST NOTED - CF/18/17759 - Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian sample
Date: Friday, 10 May 2019 12:16:50 PM

Dear Dr Janet Ruffles,

Thank you for submitting your **amendment request** to the Department of Justice & Regulation Human Research Ethics Committee (JHREC) for the project ***Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian sample***.

The Committee has noted and approved your request for the duration of the investigation at the 9 May 2019 meeting. The Department of Justice and Regulation (DJR) reference number for this project is **CF/18/17759**

Please ensure that the JHREC is notified immediately of any matter that arises that may affect the conduct or continuation of the project. To enable the JHREC to fulfil its reporting obligations, you are asked to provide an Annual Report every 12 months and to report on the completion of your project.

Future correspondence regarding this project must be sent electronically to ethics@justice.vic.gov.au and include the DJR reference number and the project title.

If you have any queries you are welcome to contact me on (03) 8684 1514 or email: ethics@justice.vic.gov.au.

Kind Regards,

Kelsey

Justice Human Research Ethics Committee

Kelsey Dalton I JHREC Secretary

Information Integrity & Access
Department of Justice & Regulation
GPO Box 4356 / Level 24, 121 Exhibition Street
MELBOURNE VIC 3001

Phone: (03) 8684 1514

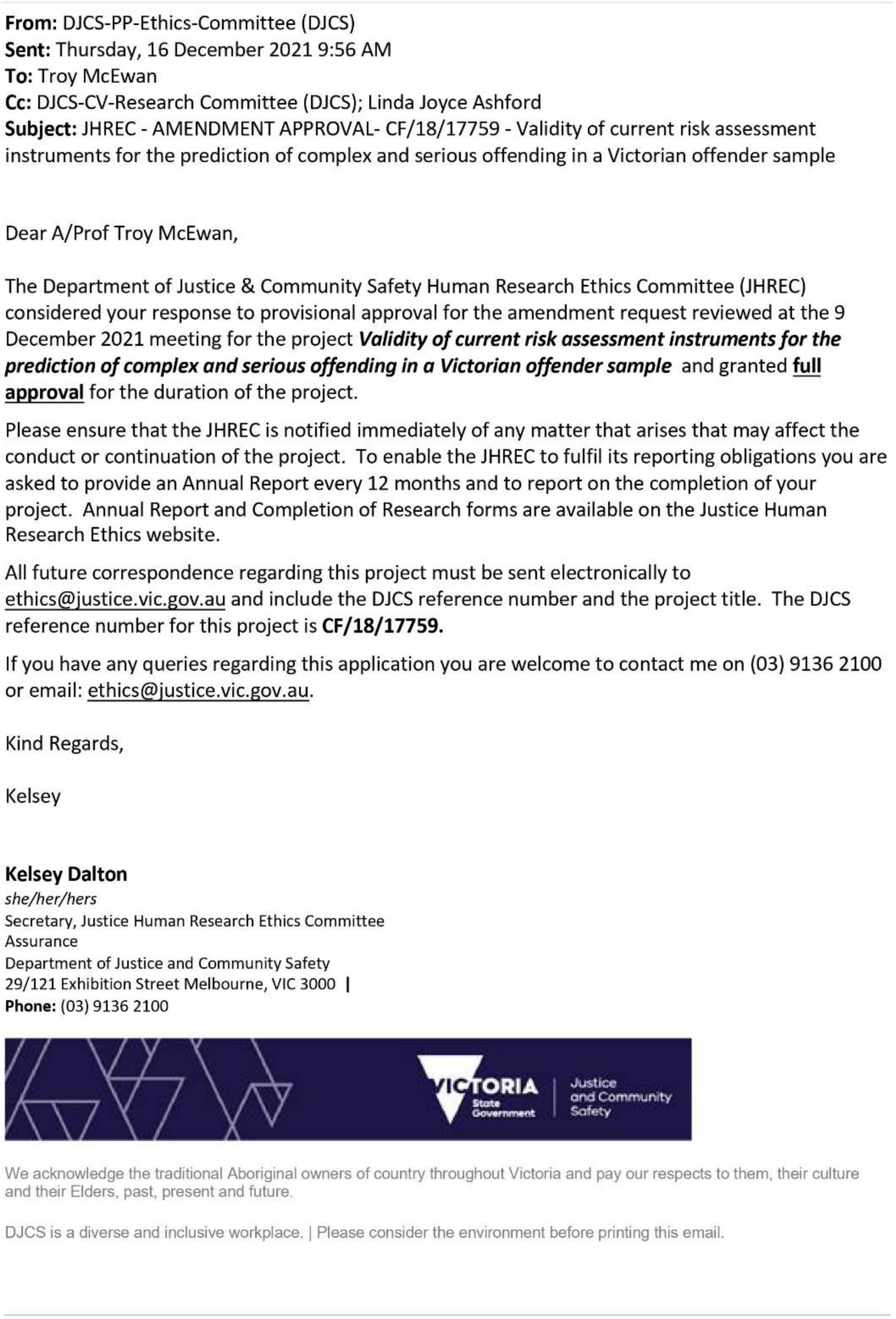
Email: ethics@justice.vic.gov.au
Web: http://www.justice.vic.gov.au/utility/data+and+research/making+a+jhrec+application

Please help to conserve paper by not printing this e-mail

PRIVATE & CONFIDENTIAL
The content of this e-mail and any attachments may be private and confidential, intended only for use of the individual or entity named. If you are not the intended recipient of this message you must not read, forward, print, copy, disclose, use or store in any way the information this e-mail or any attachment contains.
If you are not the intended recipient, please notify the sender immediately and delete or destroy all copies of this e-mail and any attachments.
Our organisation respects the privacy of individuals. For a copy of our privacy policy please go to our website or contact us.

Appendix D

Ethics Amendment Approval Letters – Empirical Studies Two and Three



From: DJCS-PP-Ethics-Committee (DJCS)
Sent: Thursday, 16 December 2021 9:56 AM
To: Troy McEwan
Cc: DJCS-CV-Research Committee (DJCS); Linda Joyce Ashford
Subject: JHREC - AMENDMENT APPROVAL- CF/18/17759 - Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample

Dear A/Prof Troy McEwan,

The Department of Justice & Community Safety Human Research Ethics Committee (JHREC) considered your response to provisional approval for the amendment request reviewed at the 9 December 2021 meeting for the project ***Validity of current risk assessment instruments for the prediction of complex and serious offending in a Victorian offender sample*** and granted **full approval** for the duration of the project.

Please ensure that the JHREC is notified immediately of any matter that arises that may affect the conduct or continuation of the project. To enable the JHREC to fulfil its reporting obligations you are asked to provide an Annual Report every 12 months and to report on the completion of your project. Annual Report and Completion of Research forms are available on the Justice Human Research Ethics website.

All future correspondence regarding this project must be sent electronically to ethics@justice.vic.gov.au and include the DJCS reference number and the project title. The DJCS reference number for this project is **CF/18/17759.**

If you have any queries regarding this application you are welcome to contact me on (03) 9136 2100 or email: ethics@justice.vic.gov.au.

Kind Regards,

Kelsey

Kelsey Dalton
she/her/hers
Secretary, Justice Human Research Ethics Committee
Assurance
Department of Justice and Community Safety
29/121 Exhibition Street Melbourne, VIC 3000 **|**
Phone: (03) 9136 2100

We acknowledge the traditional Aboriginal owners of country throughout Victoria and pay our respects to them, their culture and their Elders, past, present and future.

DJCS is a diverse and inclusive workplace. | Please consider the environment before printing this email.

PRIVATE & CONFIDENTIAL
The content of this e-mail and any attachments may be private and confidential, intended only for use of the individual or entity named. If you are not the intended recipient of this message you must not read, forward, print, copy, disclose, use or store in any way the information this e-mail or any attachment contains.
If you are not the intended recipient, please notify the sender immediately and delete or destroy all copies of this e-mail and any attachments.
Our organisation respects the privacy of individuals. For a copy of our privacy policy please go to our website or contact us.