Potential Unfairness Associated with the Development of Predictive Risk Models in the New Zealand Child Welfare Context

A thesis submitted to Auckland University of Technology in
fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Sahar Barmomanesh
School of Engineering, Computer and Mathematical Sciences

July 2025

Abstract

Over the past decade, predictive risk modelling, using machine learning techniques, has gained attention in the public sector, especially in child welfare systems, where it has shown potential in supporting decision-making processes, particularly towards identifying children at risk of maltreatment and recommending interventions. For example, Allegheny County in the United States has been using the Allegheny Family Screening Tool as an assistance system to enhance child welfare call screening. This system rapidly integrates and analyzes hundreds of data elements related to individuals involved in child maltreatment allegations and produces a *Family Screening Score* that supports decision-making by predicting the long-term likelihood of future child welfare involvement. A significant concern, however, raised by several authors, is that poorly designed models may result in biased outcomes, disproportionately impacting specific demographic groups. In the New Zealand care and protection system, for example, the over-representation of Māori children could be unintentionally exacerbated by these models, reinforcing cycles of bias and contributing to unfair decision-making.

While predictive tools in areas such as criminal recidivism and academic admissions have been widely scrutinized, the fairness of predictive models in child welfare has received far less attention. Research suggests that this is partly due to the limited availability of such tools, resulting in fewer being critically examined through the lens of algorithmic fairness. This thesis aims to address both of these gaps.

By attending to concerns of fairness and predictive bias, particularly regarding ethnicity, this research investigates predictive accuracy, fairness, and disparities in risk models within the New Zealand child welfare context. Data from the Statistics NZ Integrated Data Infrastructure are utilized, and a range of machine learning algorithms, such as logistic regression, LASSO, and XGBoost, are employed for predictive modelling. Fairness metrics, such as calibration, accuracy equity, statistical parity, and equalized odds, are also explored. Following the initial evaluation, an in-processing fairness-aware machine learning approach was implemented to address observed biases, focusing specifically on reducing disparities in error rates between Māori children and children from other ethnic groups. The results extensively highlight the inherent challenges of balancing predictive accuracy with fairness. Such challenges are influenced by data linkage strategies, modelling approaches, and variations in model performance across demographic groups.

Additionally, this research aims to provide critical insights into the development of fair, effective, and ethically responsible predictive models, contributing to the broader debate on how machine learning can support equitable decision-making in child welfare and beyond.

Disclaimer

Access to the data used in this study was provided by Statistics New Zealand (Stats NZ) under conditions designed to give effect to the security and confidentiality provisions of the Data and Statistics Act 2022. The results presented in this study are the work of the authors, not Stats NZ or individual data suppliers.

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. The opinions, findings, recommendations, and conclusions expressed in this work are those of the authors, not Stats NZ or individual data suppliers. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. For more information about the IDI please visit: https://www.stats.govt.nz/integrated-data/.

Attestation of Authorship

I affirm that this submission is solely the result of my own efforts. To the best of my knowledge and belief, it does not contain any previously published or authored material by another individual (except where explicitly defined in the acknowledgments). Furthermore, I confirm that the content of this work has not been submitted for the attainment of any other degree or diploma from a university or other institution of higher learning.

Acknowledgements

First and foremost, my sincere gratitude to my main supervisor, Dr Victor Miranda Soberanis, whose support, invaluable guidance, and scholarly insights have been vital throughout this journey. His profound insights not only directed the course of my research but also infused greater meaning into it. I am equally thankful to my secondary supervisor, Professor Jiling Cao who gave me the opportunity to undertake my Doctor of Philosophy (PhD) studies. His feedback and encouragement inspired me to do my best, and I'm grateful for the expertise he shared.

A special appreciation is reserved for Professor Pare Keiha, the Pro Vice Chancellor for Māori Advancement at Auckland University of Technology (AUT), whose guidance was crucial in crossing the path toward obtaining AUT Ethics Committee Approval. My deepest gratitude extends to Dr Valance Smith for providing helpful awareness into Treaty of Waitangi obligations, ethical considerations in using data about Whānau Māori, and the potential impact of my research on Tamariki and Whānau Māori. Their guidance has deepened my understanding and ensured a conscientious approach to my research initiatives.

Moreover, this thesis could not have been completed without the support from AUT who awarded me with the School of Engineering, Computer and Mathematical sciences Doctoral Scholarship, and without the support from Statistics New Zealand (Stats NZ) who provided the data and gave access to data lab facilities.

I am grateful to AUT colleagues, particularly those from the Department of Mathematical Sciences and the Postgraduate Research Programmes team in the School of Engineering, Computer, and Mathematical Sciences (ECMS).

Additionally, I am deeply thankful to my partner, Nader, and my son, Liam, whose love served as a powerful source of inspiration, driving me towards research dedicated to protecting children from abuse. His patience during moments when my attention was divided has been an invaluable support. Last but not least, I would like to thank my parents, my sister and niece, Samar and Liana, who are long away but always in my heart.

Contents

List of Figures

List of Tables

List of Formulae for Classification Metrics

Accuracy Metrics

Definitions

The terms used in the metrics presented are defined as follows:

- **True Positive (TP)**: The number of instances correctly classified as positive (e.g., correctly predicting a positive class).
- **True Negative (TN)**: The number of instances correctly classified as negative (e.g., correctly predicting a negative class).
- **False Positive (FP)**: The number of instances incorrectly classified as positive (e.g., predicting a positive class when it is actually negative).
- **False Negative (FN)**: The number of instances incorrectly classified as negative (e.g., predicting a negative class when it is actually positive).

1. Accuracy

Definition: Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

2. Precision (Positive Predictive Value)

Definition: Precision, also known as Positive Predictive Value (PPV), is the proportion of true positive results among all positive results predicted by the model.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

3. Recall (Sensitivity or True Positive Rate)

Definition: Recall, also known as Sensitivity or True Positive Rate (TPR), is the proportion of true positive results among all actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

4. F1 Score

Definition: The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

5. Specificity

Definition: Specificity, also known as True Negative Rate (TNR), is the proportion of true negative results among all actual negative cases.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

6. Negative Predictive Value

Definition: Negative Predictive Value (NPV) is the proportion of true negative results among all negative results predicted by the model.

$$\text{NPV} = \frac{TN}{TN + FN}. \tag{6}$$

7. False Positive Rate

Definition: False Positive Rate (FPR), also known as the probability of false alarm, is the proportion of false positive results among all actual negative cases. It is the complement of specificity.

$$\text{FPR} = \frac{FP}{FP + TN}. \tag{7}$$

8. False Negative Rate

Definition: False Negative Rate (FNR), also known as the miss rate, is the proportion of false negative results among all actual positive cases. It is the complement of recall.

$$\text{FNR} = \frac{FN}{FN + TP}. \tag{8}$$

9. Area Under the Curve (AUC)

Definition: The Area Under the Curve (AUC) refers to the area under the Receiver Operating Characteristic (ROC) curve. It measures the ability of the model to distinguish between positive and negative classes.

10. Matthews Correlation Coefficient (MCC)

Definition: The MCC is a measure of the quality of binary classifications, taking into account true and false positives and negatives. It is regarded as a balanced measure even if the classes are of very different sizes.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{9}$$

List of Abbreviations

AFST	Allegheny Family Screening Tool
AFST I	First Version of Allegheny Family Screening Tool
AFST II	Second Version of Allegheny Family Screening Tool
AI	Artificial Intelligence
AUC	Area Under the ROC Curve
CSDA	Centre for Social Data Analytics
DAT	Decision Aid Tool
DCDA	Douglas County Decision Aid
DCFS	Department of Children and Family Services
DHS	Department of Human Services
DRT	Decision Response Tool
CFA	Child and Family Assessment
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CP	Care and Protection
CPS	Child Protective Services
CWS	Child Welfare System
CYF	Child, Youth and Family
ECE	Equalized Calibration Error
ERSFT	Eckerd Rapid Safety Feedback Tool
FGC	Family Group Conference
FN	False Negative

FNR False Negative Rate

FP False Positive

FPR False Positive Rate

FWA Family Whānau Agreement

GDPR General Data Protection Regulation

IDI Integrated Data Infrastructure

LASSO Least Absolute Shrinkage and Selection Operator

LDAT Larimer Decision Aid Tool

MCC Matthews Correlation Coefficient

MSD Ministry of Social Development

NCANDS National Child Abuse and Neglect Data System

NFA No Further Action

NFAR No Further Action Required

NPV Negative Predictive Value

PPV Positive Predictive Value

PR Partnered Response

RBF Radial Basis Function

RCT Randomized Controlled Trial

ROC Receiver Operating Characteristic

SWS Social Welfare System

TN True Negative

TNR True Negative Rate

TP True Positive

TPR True Positive Rate

Chapter 1

Introduction

1.1 Introduction

Risk assessment plays a crucial part in child welfare agencies, also known as child protective services (CPS), to identify high-risk cases requiring child protection intervention (Van der Put et al., 2017). Over the last three decades, child protective services have increasingly relied on risk assessment tools to improve their decision-making processes (Cuccaro-Alamin et al., 2017). Categorized as either consensus-based or actuarial, the aim is to identify family risk factors and resources to classify referrals, investigations, and cases into various levels of risk for potential future child maltreatment (McNellan et al., 2022).

Algorithmic risk assessment tools have been developed in recent years, in a continuous adjustment as population risk factors change (Putnam-Hornstein et al., 2022; Vaithianathan, Dinh, et al., 2019; Vaithianathan et al., 2017). Such algorithmic approach to risk assessment follows the empirically based approach of actuarial risk assessment and makes use of advanced data-driven methods, such as machine learning, to arrive at the final risk score by an analysis of large, administrative datasets[1] (Drake et al., 2020). While studies indicate the superiority of actuarial over consensus-based tools in predicting adverse outcomes (Baird & Wagner, 2000; D'Andrade et al., 2008; Van der Put et al., 2017), emerging evidence suggests that algorithmic versions, developed through machine learning approaches, may surpass both (Putnam-Hornstein et al., 2018; Thurston & Miyamoto, 2018).

Prompted by a study conducted by Vaithianathan (2012), indicating that a predictive risk model can

[1] Administrative data refers to data collected by the government or other organizations for non-statistical reasons. Administrative data includes records for the organization's routine operations and is often used to evaluate how well an organization is achieving its' expected goals.

be developed and validated for New Zealand (NZ) children, the NZ government considered the possibility of incorporating predictive risk models into their child welfare system (Bennet, 2012). According to NZ government's "White Paper for Vulnerable Children" (Bennet, 2012), a model of this kind has the potential to link various characteristics of the child, young person, and their family effectively and consistently to generate a single measure of overall risk. Subsequently, a series of studies aiming to evaluate the technical feasibility and ethical risks of the government's proposal to the prevention of child maltreatment started to appear in the literature (Blank et al., 2015; Dare, 2013; Ministry of Social Development, 2014; Rea & Erasmus, 2017; Wilson et al., 2015). In this direction, studies evaluating the technical feasibility and predictive validity of predictive risk models developed for potential use within the NZ child welfare system, also referred to as NZ care and protection system, have shown optimistic results (Rea & Erasmus, 2017; Vaithianathan et al., 2013; Wilson et al., 2015). Additionally, the NZ government's "White Paper for Vulnerable Children" (Bennet, 2014) acknowledged that predictive risk modelling appears promising based on preliminary research, but takes ethical risks and requires cautious, staged feasibility study, and trials.

The most common concerns mentioned in ethical reviews and staged feasibility studies which are also the focus of this thesis, include the accuracy and fairness of the models being developed (Blank et al., 2015; Dare, 2013; Keddell, 2015, 2019; Rea & Erasmus, 2017). Fairness in machine learning involves addressing algorithmic bias or predictive bias in automated decision processes driven by predictive models. Decisions made by these models can be considered unfair,[2] if they rely on sensitive variables like *gender*, *race*, *ethnicity*, *sexual orientation*, or *disability* (Ruggieri et al., 2010).

While various international studies have explored or tested predictive risk modeling within the child welfare system, with some actively employed by the United States (U.S.) child protective services to assist with screening-in children upon receiving reports of concern (intake decision-making), there remains a significant gap in NZ's exploration of alternative approaches to predictive risk modeling in the child welfare context (Vaithianathan et al., 2021). Prior research emphasizes the need for heightened attention to enhance the accuracy and fairness of child maltreatment predictive risk models, a dimension that has been comparatively overlooked in the NZ context, particularly in relation to its impact on ethnic minorities (see Section 3.4 for details).

The NZ Ministry of Social Development most recent project, as reported in (Rea & Erasmus, 2017), marked the latest effort in this field. Since this project, particularly, no further studies in NZ have

[2]The use of terms like bias, discrimination, and unfairness is interchangeable throughout this work, aligning with common practice in the algorithmic fairness literature.

explored the enhancement and performance of predictive risk models which could potentially be applied in the initial intake decision-making process by child welfare agencies. One study, by A. James et al. (2019), examined the effect of incorporating information about a child's immediate family network on the effectiveness of predictive risk models developed for potential use by child protective services during the initial intake decision-making process. Although A. James et al. (2019) showed that including information about a child's family network can improve the efficacy of predictive risk models by having fewer predictor variables than other published models, the primary objective was not to construct an operational model or to surpass the accuracy and fairness of the one developed in the NZ Ministry of Social Development's project

Rea and Erasmus (2017) suggest that enhancing the predictive power of predictive risk models can be achieved by incorporating additional predictor variables. In the Ministry of Social Development's project, the final model includes variables related to the child's mother characteristics, suggesting the possibility of creating similar variables for the father's characteristics. Furthermore, they recommend for future attempts that utilizing data from other organizations may contribute to create new significant predictor variables. For instance, the Department of Corrections data could be employed to verify whether a child resides in the same household as an adult recently released from prison for a family violence-related offense. Moreover, in the evaluation of the model's performance across various ethnic groups, Rea and Erasmus (2017) found that the number of *Māori* children and young people referred to the site (intake decision outcome) using the model was 13% higher than those referred under the existing intake decision-making practice. This fact highlights the need for further investigation to better understand these disparities and to ensure that the model accurately identifies true risks without contributing to any over-representation of Māori within the NZ child welfare system.

There has been debate regarding the use of *race* in the U.S. context and *ethnicity* in the NZ context as predictor variables, due to concerns surrounding racial-stereotypes or race-based allocation of interventions (Vaithianathan et al., 2013, 2017). A notable example is the ROC*ROI algorithm, developed in NZ in the late 1990s, which used logistic regression to estimate the risk of imprisonment and included *ethnicity* as a predictor variable. This practice was legally challenged in the Waitangi Tribunal case (WAI1024), which ultimately led to the removal of *ethnicity* from the model (Johnston, 2021). Consequently, NZ studies have discarded the factor *ethnicity* as a predictor, yet they employed it to test the model's performance across various ethnic groups (Rea & Erasmus, 2017; Vaithianathan et al., 2013; Wilson et al., 2015). However, the idea of *fairness through unawareness* has proven ineffective due to existing correlated variables with the sensitive variable e.g. *race* (Lum & Johndrow,

2016). This approach may lead to indirect or unintentional discrimination, a phenomenon also known as the *redlining effect* (Žliobaitė, 2017). For example, using *zip codes* as predictors may inadvertently reflect racial characteristics due to residential patterns. Additionally, institutionalized racial bias, such as criminal justice history, introduces other highly correlated predictors with *race*, emphasizing that this remains a significant factor (Vaithianathan, Kulick, et al., 2019). Žliobaitė (2017) recommends the inclusion of protected characteristics, like *ethnicity*, in the model development process to actively ensure non-discrimination.

In response to the growing interest in understanding how child maltreatment predictive risk models might affect minority groups, this thesis takes a crucial step forward in the context of the NZ child welfare system. While previous research has highlighted concerns about possible unfairness against specific groups, notably Māori, a significant gap remains in the exploration of fairness-aware machine learning approaches to tackle accuracy-related concerns, especially in terms of fairness. This research aims to address this limitation by focusing on predictive risk models employed during the initial intake decision-making process, a domain that has yet to receive comprehensive exploration in NZ studies.

Moreover, our work seeks to extend existing literature, particularly the research conducted by Rea and Erasmus (2017). The approach in this thesis involves generating a novel research dataset using data accessible in Statistics NZ (Stats NZ)[3] Integrated Data Infrastructure (IDI) database system (see Section 4.1). This dataset integrates information from the NZ Ministry for Children (Oranga Tamariki), linking it with Census data and administrative data sourced from other government organizations, such as the Department of Corrections (Ara Poutama Aotearoa), the Ministry of Social Development (Te Manatū Whakahiato Ora), and the Ministry of Health (Manatū Hauora). By incorporating diverse data sources, this study aims to explore the potential for improving predictive model performance through the inclusion of expanded variables, thus contributing to the advancement of predictive risk modeling in the NZ context. However, the primary focus is not only on improving model performance in terms of accuracy but also on investigating the impact of these models on the Māori population. This analysis is not restricted to simply assess model performance; it investigates the potential for unfairness or discrimination. We aim to uncover subtle details that traditional evaluation metrics might ignore, offering a more profound insight into how predictive risk models impact various demographic groups. Additionally, we explore an alternative fairness-aware machine learning method to mitigate

[3]Statistics NZ Tatauranga Aotearoa is an official data agency and a government department that collects information from people and organisations through censuses and surveys to provide data and statistics about various aspects of NZ.

negative impacts on fairness and equity, adopting a more comprehensive approach than simply removing the *ethnicity* variable from the set of predictors.

1.2 Motivation

In recent years, the surge in the development of advanced predictive analytics tools through machine learning, coupled with the emerging field of Artificial Intelligence (AI), has sparked a significant increase in interest and research focused on *algorithmic fairness* or *fair machine learning* in the public sector. A pivotal moment in the evolution of this field was the groundbreaking investigative journalism by ProPublica on the COMPAS[4] recidivism risk score (Angwin et al., 2016). The COMPAS tool, widely used in U.S. courts, came under scrutiny for its alleged unfair treatment of African-American defendants by incorrectly identifying them as high-risk at a higher rate than their white counterparts (Angwin et al., 2016). Instances of inadvertent discrimination by algorithms have raised concerns among civil rights unions, governments, regulatory authorities, and researchers, stimulating calls to address the potential discriminatory effects of algorithmic decision-making (Barocas & Selbst, 2016; Gavighan et al., 2019; Goodman & Flaxman, 2017). Legal frameworks, such as *Recital 71 of the European General Data Protection Regulation (GDPR)*, emphasize the need to prevent discriminatory effects in algorithmic decision-making processes (G. D. P. Regulation, 2018).

Discrimination refers to the unfair treatment of an individual based on their membership in a certain group rather than individual merits (Ruggieri et al., 2010). Discrimination is expressly prohibited by national and international legislation, including *NZ Human Rights Act 1993*, which aims to ensure fair treatment in accordance with United Nations (UN) agreements. Moreover, the anti-discrimination laws in NZ extend to decisions made based on predictive models (Gavighan et al., 2019). As emphasised by M. K. Lee (2018), breaches of anti-discrimination legislation or ethical standards in organizational and managerial decisions made based on these models may expose institutions to legal consequences and have a significant impact on public perceptions. Recognizing that individual rights against discrimination are crucial but may not be entirely reliable, there is a need for proactive measures. As recommended by a recent report on government use of AI in NZ, institutions should test potential discrimination in models before deployment (Gavighan et al., 2019). Consequently, prevention of discrimination is a crucial goal in the machine learning process. While predictive analytics tools relating to criminal recidivism and academic admissions have received much attention, the tools

[4]Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.

of child protective services are far from this point. This is partly due to the limited number of such tools and the lack of scrutiny on their fairness from an algorithmic perspective.

In this thesis, both gaps are addressed. First, a variety of classification models are developed to predict adverse outcomes within the NZ child welfare context, with a particular focus on variations of logistic regression. These models assess the risk for children at the time of notification and identify those at higher risk of future adverse outcomes. Next, fairness-aware procedures are applied to address potential predictive bias, with detailed reasoning provided for each decision to ensure transparency and equity in predictive outcomes.

Prior studies across domains such as recidivism risk, credit risk, or employment have introduced diverse approaches to fairness-aware machine learning. These approaches are typically classified into three methods: pre-processing, in-processing, and post-processing, depending on where the emphasis lies in addressing discrimination (see Section 3.6). However, two general issues arise with these existing methods. First, there is a lack of consensus on how to assess the fairness of algorithms, likely due to the evolving nature of the field, which has placed researchers in an exploratory phase. This diversity, driven by the absence of a unifying framework, has resulted in the proposal of sometimes inconsistent measures, making it difficult to assess the effectiveness and applicability of current approaches. Second, each discrimination prevention method is often tailored to specific domains, limiting their generalizability to other variables or grounds of discrimination. As a result, there is no widely accepted method for preventing discrimination across different contexts.

This thesis also aims to identify and examine fairness metrics and notions specifically relevant to child welfare data and predictive models. By systematically exploring discrimination prevention approaches, we seek contributions to the development of fairer predictive risk models in the child welfare domain, addressing a critical gap in the current literature on algorithmic fairness.

1.3 Research Objectives

Our overall purpose is to utilize linked administrative data at the individual level to explore options for adjusting and creating a more accurate and fair predictive risk model for assessing the likelihood of future child maltreatment. The ultimate goal is to propose a model suitable for adoption in the NZ child welfare system with emphasis on conducting systematic inquiries into the potential discriminatory risks associated with developing child maltreatment predictive risk models using machine learning algorithms. We particularly focus on the Māori population.

This study is expected with a significant benefit to the research community, child welfare systems, and children. Firstly, it aims to provide practical solutions and awareness measures in response to the concerns raised about potential unfairness effects associated with using predictive risk modeling within the NZ child welfare system. By addressing these concerns, there is an opportunity for the government to progress towards the implementation phase, ensuring a more equitable and accurate model. The insights generated are expected to enhance the understanding of researchers and policymakers regarding the potential risks of discrimination in government use of algorithms in decision-making.

Secondly, our study offers a technical explanation of the measures that can be applied to address unfairness in algorithms. By communicating the degree of fairness achieved through machine learning approaches, we aim to foster greater trust in the utilization of these methods. Although the focus of this research is on discrimination or unfairness prompted during the mechanism of developing models for predicting the risk of child maltreatment, the analysis in this study can be translated to many other risk prediction settings.

Next, developing a more accurate and fairer predictive risk model will improve the ability of child protection staff to make more efficient and consistent decisions. It can assist child welfare agencies to avoid unnecessary investigations, which are costly for the system, and troublesome for the families. For example, in cases where children from a specific ethnic group are incorrectly labeled as high risk, due to the discriminatory behavior of the model. Finally, it will have an impact on the lives of children and their families who are at risk of maltreatment by identifying their risk score more accurately and preventing severe future outcomes.

1.4 Research Questions

This study aims to investigate the following areas of inquiry using administrative data from the NZ population, with a particular focus on Māori children:

1. Factors affecting the accuracy of risk models in the child welfare domain.
2. Factors contributing to bias in algorithms predicting adverse child welfare outcomes.
3. Development of more accurate and less discriminatory predictive risk models.
4. The extent to which child welfare authorities can mitigate discriminatory algorithmic behaviors.
5. Technical solutions for improving algorithmic fairness.

By exploring these inquiries, this study seeks to provide valuable insights into the intersection of predictive risk modeling, fairness, and child welfare in NZ, with particular emphasis on the experiences of Māori children as the Indigenous people of *Aotearoa*.

1.5 Outline

This thesis consists of nine chapters. Chapter 2 provides essential background information, offering a comprehensive overview of the NZ care and protection system. It covers the historical development of the NZ child welfare system, current assessment processes, and the use of screening tools such as the Decision Response Tool (DRT).

Chapter 3 reviews the relevant literature on predictive risk modeling in the child welfare domain, highlighting the increasing use of machine learning algorithms and the associated challenges of accuracy and fairness. This chapter identifies key gaps in the current body of work and establishes the need for further research into ethical concerns and bias mitigation in predictive modeling.

Chapter 4 introduces the data sources and software tools utilized throughout this research, including the development of a novel dataset from the Stats NZ IDI.

Chapter 5 outlines the research methodology, detailing the data preparation, sample selection, and analytical techniques employed to develop and evaluate the predictive models. It also describes the fairness-aware machine learning approach applied to address disparities in model performance between Māori children and children from other ethnic groups.

Chapter 6 presents an in-depth analysis of the empirical results, thoroughly evaluating model performance and fairness metrics across subgroups, with particular emphasis on addressing disparities between Māori children and other ethnic groups.

Chapter 7 interprets the findings in relation to the research questions and the broader child welfare context, summarizing the key contributions, implications, and recommendations for future research.

Chapter 8 offers two supplementary analyses prompted by the primary findings of this thesis, providing additional insights and implications.

Finally, Chapter 9 addresses the ethical considerations associated with conducting research on predictive risk modeling within the NZ child welfare context and outlines the steps taken to meet the requirements of the AUT Ethics Committee (AUTEC).

Chapter 2

Background

2.1 Introduction

The purpose of this chapter is to provide background information relevant to this work and the subsequent chapters by presenting an overview of the current state of the assessment system within the NZ care and protection system. It is important to acknowledge that the information provided in this chapter is gathered at the time of writing this thesis and may undergo changes due to policy and practice adjustments within the NZ care and protection system.

This chapter commences with Section 2.2, offering a comprehensive overview of the evolution of the child welfare system in NZ over recent decades, tracing its origins to the present day. Section 2.3 outlines the current child protection assessment process, including the three phases of initial, core, and full assessment, along with considerations within each phase. Section 2.4 describes current screening or intake decision-making process upon receiving notifications, exploring the utilization of an intake Decision Response Tool (DRT). In Section 2.5, the current international status regarding the use of predictive risk models by child protective services is explored, concluding with a brief summary in Section 2.6.

2.2 New Zealand Child Welfare System

The concept of child welfare refers to the provision of care for children who are suffering from inadequacies in their household environment or parenting. Among the primary objectives of early NZ governments was to improve the health of children and mothers as a means to promote reproduction, stabilize communities, and increase the country's population by providing child welfare services (M.

Baker & Plessis, 2018).

For a considerable period, this responsibility was one of the most important aspects of **Department of Education** operations until it was formally revised as *Child Welfare Act 1925* (Archives New Zealand, 2022). As a response to this development, a dedicated branch called the **Child Welfare Branch** was established within the Department of Education (Archives New Zealand, 2022; New Zealand Government, 2023). After its establishment in 1926, the branch took responsibility for all children's welfare, regardless of care setting. A superintendent of child welfare was then appointed, reporting to both the minister of education and the minister in charge of welfare (New Zealand Government, 2023).

The child welfare system in NZ has undergone various reforms and changes over the years, reflecting evolving societal perspectives and approaches to child welfare and protection (Keddell, 2018). By virtue of the *Child Welfare Amendment Act 1948*, the child welfare branch of the Education Department became an independent department with its own minister (Archives New Zealand, 2022), hence renamed as the **Child Welfare Division** (New Zealand Government, 2023). Later on April 1, 1972, the division was unified with the **Social Security Department** to form the **Department of Social Welfare** (Ministry of Social Development, n.d.). Accordingly, children came under the care of the Department of Social Welfare. However, residential special schools for hearing impaired, maladjusted, and backward children remained under the jurisdiction of the Department of Education (New Zealand Government, 2023).

On May 1, 1992, the Department of Social Welfare was restructured into business units, including the **Children and Young Persons Service** (Ministry of Social Development, n.d.). Following the merger of the NZ Children and Young Persons Service and the NZ Community Funding Agency in 1999, this service was renamed as the **Children, Young Persons, and their Families Agency** (New Zealand Government, 2023). Later in the year, this agency became the stand-alone Department of Child, Youth and Family Services known as **Child, Youth and Family (CYF)**, and finally, on July 1, 2006, the CYF became a service line of the Ministry of Social Development (Ministry of Social Development, n.d.).

In 2015, an Expert Advisory Panel was formed to review the NZ child welfare system with the aim of enhancing the care and protection of children in NZ. This panel actively sought input from *iwi* (tribe) and Māori, as well as *tamariki* (children) and *rangatahi* (young people), along with *whānau* (extended family) and caregivers (Oranga Tamariki, 2022). Through this inclusive approach, participants were able to share their personal experiences and outcomes. Based on the insights gained from these

conversations and in alignment with commitments under *Te Tiriti o Waitangi* (The Treaty of Waitangi), **Oranga Tamariki—Ministry for Children** was established in 2017 (Oranga Tamariki, 2022). Since then, Oranga Tamariki has been responsible for overseeing the welfare of children and young people throughout NZ.

2.3 New Zealand Child Protection Assessment System

According to Section 15 of *Oranga Tamariki Act 1989*, any person who believes that a child or young person has been, or is likely to be harmed, abused (whether physically, emotionally, or sexually), neglected, or who has concerns about the well-being of a child or young person, may report the matter to the chief executive or constable. Notifications can come in the form of calls, emails, in person, fax, or an automated record such as in the case of police family violence referrals, from a range of notifiers, including family members, members of the wider community, healthcare providers, schools, or legal entities (Oranga Tamariki, 2023d).

Following the receipt of a notification, social workers assigned to the case assess the matter and determine the appropriate course of action for *te tamaiti* (the child) and their whānau or family. According to Oranga Tamariki, assessment is an ongoing process of building understanding to inform whānau or family and professional decision-making. This understanding is built across three phases: initial assessment, core assessment, and full assessment (Oranga Tamariki, 2023g). Throughout each phase, a key decision must be made, which determines the purpose of that phase of assessment and whether the next phase of assessment is necessary. Additionally, the information collected in the preceding phases is used to guide the decision-making process for the current phase (Oranga Tamariki, 2023g). Figure 2.1 depicts the assessment process phases.[5]

2.4 Intake Decision Making Process

Social workers follow a structured decision-making process to evaluate if the level of concern reported regarding a child meets the threshold for Oranga Tamariki involvement, and if it does, how quickly they may need to respond.

The intake decision, as represented in Figure 2.1, is mainly based on the information collected during

[5]This map was created based on Oranga Tamariki's internal intakes map, shared with us by their research team. The original intake map is exclusively for internal use within Oranga Tamariki and cannot be included here.

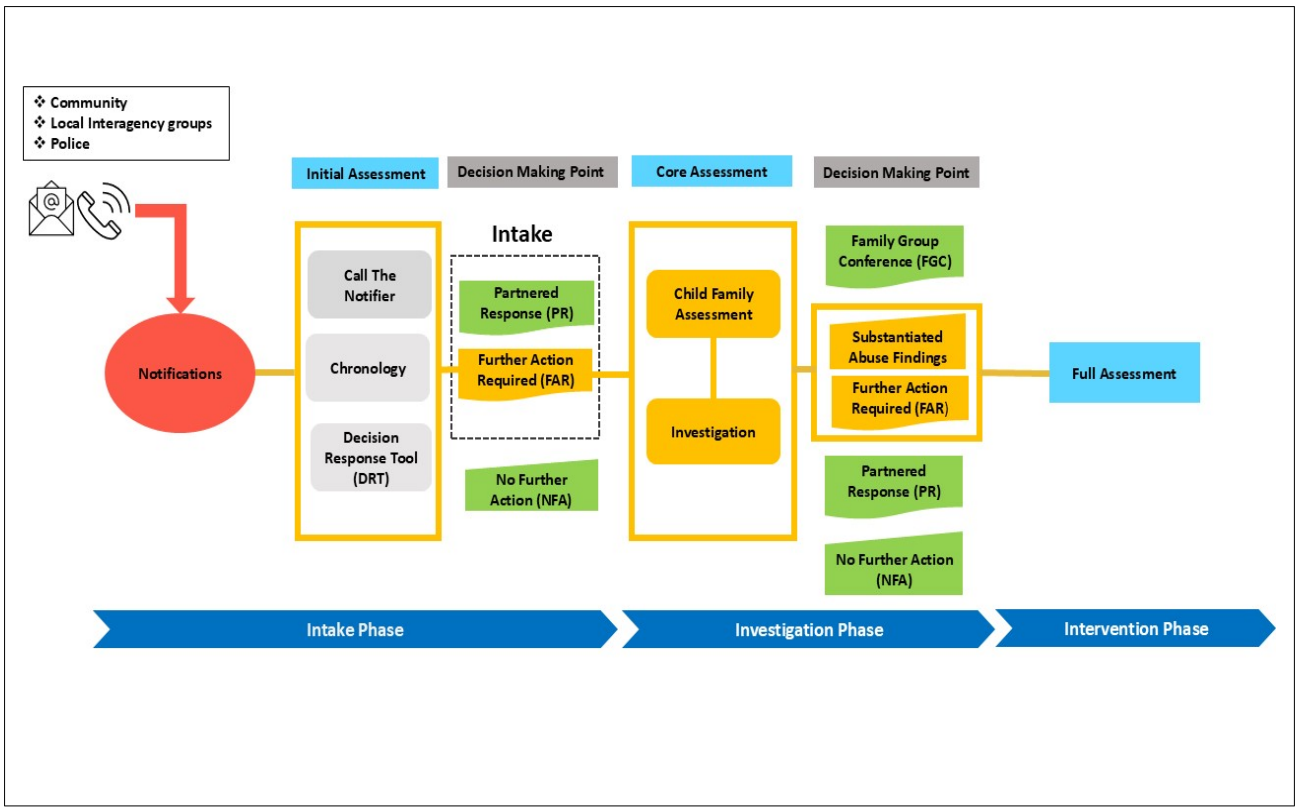

Figure 2.1: Short description of the Oranga Tamariki assessment process.

the initial assessment phase. The purpose of this phase of assessment is to gather quality information and develop a chronology to assist in deciding the best response and appropriate level of support for te tamaiti and their whānau or family (Oranga Tamariki, 2023f). In this phase, the critical sources of information are considered, specifically the initial report, the notifier, and a chronology combining relevant information that would provide a picture of te tamaiti and their whānau or family. The chronology highlights harmful cumulative patterns and previous responses to reports of concern. This process allows for the recognition of significant events that have influenced both te tamaiti and their whānau or family (Oranga Tamariki, 2023f). Social workers may also engage in direct communication with te tamaiti and their whānau, or seek information from external agencies, including schools, healthcare professionals, early childhood educators, non-governmental organizations, and iwi who know them. Additionally, a standardized intake Decision Response Tool (DRT) is used to guide social workers in analyzing key factors, including age, disability, the notifier's relationship to the child, cumulative harm, trauma, and the child's care status (Oranga Tamariki, 2021). This tool assists social workers in determining the appropriate response and time frame, which can range from *no further action* to *investigation*. Several other factors are considered when responding to reports, including the family's responsiveness, protective measures in place, relationship dynamics, child vulnerability, willingness for assistance, the availability of local services, and any previous offenses committed by the child (Oranga Tamariki, 2023c).

The initial assessment determines the level of concern and the urgency of involvement with Oranga Tamariki, leading to four potential pathways: *no further action*, *referral to services* (also known as *partnered response*), *child family assessment*, or *investigation*. If the social worker makes an assessment concerning serious abuse or neglect, and a joint approach between the Police and Oranga Tamariki is required, then an investigation is likely to be recommended. In cases where less serious abuse or neglect is suspected, they recommend a child family assessment (Oranga Tamariki, 2023h). Social workers usually determine the time frame for completing the safety and risk screening based on the severity of the situation. If a child is seriously harmed or is at immediate risk of serious harm, Oranga Tamariki must take action within 24 hours to ensure their safety. If protective factors are present, this time frame is extended to 48 hours. For all other cases, the assessment is completed within 10 working days (Oranga Tamariki, 2023i). Since circumstances can change rapidly, any developments or changes that may occur are addressed to ensure that the appropriate actions are taken in a timely manner. The referral to services pathway is recommended when support from another agency, iwi, or cultural social service is likely to achieve positive outcomes (Oranga Tamariki, 2023f). Table 2.1 provides additional information on the pathways and urgency response categories available to intake decision-makers.

Moreover, as outlined in Table 2.1, children who receive an intake outcome during the initial assessment phase are referred for investigations, child family assessments, or services (partnered response). The subsequent core phase of assessment builds on the initial assessment outcome and is undertaken for children referred for a child family assessment or investigation (Oranga Tamariki, 2023h). During this phase, the social worker draws conclusions on the safety, strengths, vulnerabilities, and needs of te tamaiti and their whānau or family. The outcome may involve *further action*, *referral to services*, or *no further action*, depending on the identified concerns and the level of harm experienced by te tamaiti (Oranga Tamariki, 2023e).

Cases with *substantiated abuse findings* or *further action required* outcome from the core assessment proceed to the full assessment phase, where the goal is to achieve a comprehensive understanding of te tamaiti and their whānau or family situation. This involves identifying their complete range of needs, ensuring long-term physical and psychological safety, evaluating the family's ability to meet those needs, and recognizing the child's potential. Social workers will use the information collected during this phase to inform the family at the *Family Group Conference (FGC)*[6] and to improve

[6]**Family Group Conference (FGC)** is a formal meeting where child protective services and the extended family of children work together to develop a plan to address any care and protection concerns, needs or well-being issues relating to the child.

Table 2.1: Pathway and urgency response categories.

Pathway/Urgency	Description
Intake - Further Action Required (FAR)	
Investigation 24 hours, 48 hours	When there is an allegation relating to sexual abuse, physical abuse, and neglect, or cases where the actions of an adult may constitute a criminal offence against the child, as defined in the Child Protection Protocol (CPP) (New Zealand police & Oranga Tamariki, 2021).
Child Family Assessment (CFA) 24 hours, 48 hours, 10 working days	If the child has undergone or is at risk of enduring severe harm, and/or if the concerns are significantly affecting their development, safety, health, or overall well-being, without necessarily indicating abusive actions that might constitute a criminal offense.
Partnered Responses (PR) No urgency	When positive outcomes can be achieved through support from another agency, iwi, or cultural social service, and the needs of the child can be addressed or minimized with the assistance of other professionals or services, provided there are indications of the family being receptive to such support.
No Further Action (NFA)	
Contact Record (CR) No urgency	When the report lacks substance, and/or the concerns do not indicate harm to the child, or the concerns are being adequately addressed by others.

quality planning for te tamaiti. Collaboration with core professionals, involving relevant ethnic community and spiritual leaders, also contributes to a full understanding of te tamaiti and their whānau or family. In this phase, a *safety plan* will be in place to support the family to take care of the child until a FGC plan is developed. The safety plan is regularly reviewed with all involved parties to assess changing needs and ensure the ongoing safety and well-being of te tamaiti (Oranga Tamariki, 2020).

2.5 Use of Predictive Analytics Tools for Intake Decision Making

Internationally, predictive risk models to assist social workers in identifying children at high risk of future maltreatment have been initiated by various governmental welfare authorities. Several states across the U.S. are actively testing or implementing predictive tools within their child welfare systems. Many other authorities worldwide are also interested in the implementation of these tools (Glaberson, 2019). Table 2.2 provides an overview of predictive risk modeling applications by child welfare agencies in the U.S.

The Eckerd Rapid Safety Feedback Tool (ERSFT) is one of the earliest tools developed by "Eckerd

Connects," a national multi-service child welfare agency in Hillsborough County, Florida. It worked toward mitigating child fatalities and serious injuries in Hillsborough County, employing predictive risk modelling and key risk factors to identify cases at high risk of adverse outcomes (Eckerd, 2014). Initiated in 2014, the tool saw statewide implementation and adoption in multiple states by early 2017. Eckerd reported a reduction in repeat maltreatment (7.09% to 5.58%) and zero child abuse-related fatalities three years after ERSFT implementation. However, limited information on the tool's model development process and performance measures raises concerns about its transparency. Blank et al. (2015) criticised the ERSFT, claiming that the algorithm has commercial purposes and therefore is not transparent.

Several other predictive risk models have been generated to support social workers during the initial screening or intake decision-making process. Examples of such models include the Allegheny Family Screening Tool (AFST), the Douglas County Decision Aid (DCDA), and the Larimer Decision Aid Tool (LDAT) (Allegheny County, n.d. Centre for Social Data Analytics, 2022, n.d.). Other models have been designed to improve the allocation of support and services as part of an early intervention prevention strategy (Allegheny County Analytics, 2023; Putnam-Hornstein et al., 2022).

Allegheny County in Pennsylvania developed the AFST to assist hotline workers in deciding whether screening a child in and carrying out an investigation is necessary. Unlike ERSFT, Drake et al. (2020) believe that AFST is perhaps the most carefully assessed predictive risk model in use by a Child Welfare agency. This tool was developed over several years and involves a series of publications describing its nature (Chouldechova et al., 2018; Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2017), ethical reviews on its use (Dare & Gambrill, 2017), the process evaluation (Hornby Zeller Associates, 2018), and a full impact evaluation on Allegheny County's decision-making process (Goldhaber-Fiebert & Prince, 2019). The AFST is developed using an integrated database system that incorporates data sources from child welfare, corrections, and health. In Allegheny County, reports of concern or referrals are received through emails or calls, and social workers are responsible for making screening or intake decisions. This process is very similar to the intake decision-making process in the NZ child welfare system, as represented in Figure 2.1.

In 2014, the NZ Ministry of Social Development initiated a project to explore the application of predictive risk modelling in the intake decision-making process within their child welfare agency known as CYF at the time. The primary goal was to assess whether predictive risk models could enhance decision-making in cases concerning a child or young person. This project involved a model development process, pre-testing of trial materials, and a trial at CYF National Contact Centre. The Ministry

of Social Development believed that consolidating relevant administrative data into a systematic measure, termed a "Background Risk Indicator," could benefit social workers in making intake decisions. The project's overall finding suggests that this tool has the potential to improve care and protection intake decisions. The trial results indicated that when successfully employed, the Background Risk Indicator influences social workers' decision-making in a safe and accepted manner, indicating potential positive outcomes for NZ children, young people, and their families (Rea & Erasmus, 2017). However, ongoing efforts are needed to enhance the models' predictive accuracy and fairness. The proposal to integrate predictive risk modelling into the NZ child welfare system was temporarily halted in 2015 due to ethical concerns and validation issues (Vaithianathan et al., 2021). The NZ government is currently refraining from employing predictive models in its intake decision-making process within the child welfare context, pending further research to resolve these issues. For a complete report on this project, refer to (Rea & Erasmus, 2017).

2.6 Conclusion

This chapter provided a background on the evolution of the child welfare system in NZ, tracing its development from its earliest origins up to the establishment of Oranga Tamariki—Ministry for Children in 2017. To facilitate the comprehensive understanding of the research design in this thesis, the current child protection assessment process, including its phases and considerations, was outlined. This background knowledge serves as a foundational guide for navigating the subsequent chapters, particularly in understanding the process of constructing the research dataset, developing predictive risk models, and their potential application at specific decision points, especially intake decision-making.

The discussion extended beyond the national context to explore international efforts in predictive risk modeling within child welfare, highlighting tools such as the AFST. This exploration highlights the potential effectiveness of these models as support tools, contingent upon the resolution of ethical and technical challenges.

While promising results have been observed in NZ regarding the feasibility of predictive risk modeling in the child welfare context, continuous efforts are essential to address concerns related to the accuracy and fairness of these models. The decision to postpone implementation in NZ emphasizes the need to resolve these issues through ongoing research before moving towards integration. This cautious approach ensures a robust foundation for potential implementation, aligning with the importance of upholding ethical standards, and validation precision when utilizing predictive risk models during the intake decision-making process.

The following chapter will review existing studies on predictive risk modeling in child welfare, emphasizing the methodologies used to address ethical concerns and potential biases. This review will provide a foundation for understanding the challenges and opportunities in creating equitable predictive models that are both effective and fair.

Table 2.2: Examples of predictive risk modelling applications by U.S. child welfare agencies.

Application	Jurisdiction	Usage Area	Status	References
Eckerd Rapid Safety Feedback Tool (ERSFT)	Colorado, Connecticut, Hampshire, Maine, Louisiana, New Ohio, and Oklahoma child welfare authorities.	Identification of open cases with a high risk of abuse.	Alaska, Connecticut, Illinois, Louisiana, Ohio, and Oklahoma jurisdictions discontinued the use of ERSFT approach.	(Eckerd, 2014) (Parker et al., 2022)
Allegheny Family Screening Tool (AFST)	Allegheny County Department of Human Services (DHS), Pennsylvania.	Assisting decision-making during the initial referral intake process.	Implemented.	(Allegheny County, n.d.)
Hello baby	Allegheny County Department of Human Services (DHS), Pennsylvania.	Prioritizing families with newborns for support within the high-needs category and priority tiers of the "Hello Baby" prevention program.	Implemented.	(Allegheny County Analytics, 2023)
Douglas County Decision Aid (DCDA)	Douglas County Department of Human Services (DHS), Colorado.	Assisting decision-making during the initial referral intake process.	Launched as a year-long randomized controlled trial in February 2019, with ongoing evaluation to assess its impact on long-term outcomes.	(Centre for Social Data Analytics, n.d.)

Table 2.2 Continued from previous page

Application	Jurisdiction	Usage Area	Status	References
Los Angeles (LA) County Risk Stratification Model	Los Angeles (LA) County Department of Children and Family Services (DCFS), California.	Improving the allocation of supervision and management resources to protect children and ensure families receive vital services during and after a maltreatment investigation.	After an 18-month planning period, launched in three regional offices: (1) Belvedere, (2) Lancaster, (3) Santa Fe Springs.	(Putnam-Hornstein et al., 2022)
Northampton County “Decision Aid Tool” (DAT)	Northampton County Department of Human Services (DHS), Pennsylvania.	Supporting county caseworkers in assessing a child’s risk for abuse and neglect, and determining the necessity of an investigation.	Conducted community and stakeholder consultations, now in the process of deploying an initial tool.	(Wandalowski & Vaithianathan, 2023)
Larimer Decision Aid Tool (LDAT)	Larimer County Department of Human Services (DHS), Colorado.	Assisting the RED team (Review, Evaluate, and Direct team) in decision-making during the referral process.	As of the Larimer County Board of Social Services meeting on November 13, 2023, Larimer concluded a Randomized Control Trial (RCT) research study to assess the impact of using LDAT.	(Centre for Social Data Analytics, 2022)

Chapter 3

Literature Review

3.1 Introduction

Child welfare agencies are increasingly using predictive risk models developed through machine learning algorithms to predict potential adverse outcomes and assist in their screening decision-making processes (Allegheny County, n.d. Centre for Social Data Analytics, n.d. Putnam-Hornstein et al., 2022). However, recent, yet growing evidence indicates a potential for these tools to inadvertently amplify biases (Barocas & Selbst, 2016; Pedreschi et al., 2009; Zliobaite, 2015). Ethical reviews and feasibility studies on the use of predictive risk modeling in the child welfare context consistently highlight concerns, with the accuracy and fairness of the developed tools being the most prevalent (Blank et al., 2015; Dare & Gambrill, 2017; Ministry of Social Development, 2014). Consequently, exploring how researchers have addressed ethical concerns, disparities, bias, and the performance of models in their design and evaluation of algorithmic risk assessment tools within the child welfare domain is imperative.

Predictive risk models are developed using historical data and machine learning algorithms (Kuhn & Johnson, 2013). In the child welfare context, the data used by predictive risk models to predict an adverse outcome are often extracted directly from the child welfare agency's database system and include records of interactions with the children and their families. Depending on the jurisdiction, these records may be linked to government-collected administrative data from various agencies, including health, criminal justice, and education (Centre for Social Data Analytics, n.d. Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2017; Wilson et al., 2015). Subsequently, the strategic selection of data sources and the encoding of predictor variables to develop a predictive risk model rely on the recognition of child maltreatment risk factors. Thus, a comprehensive understanding of these

risk factors is crucial to identify variables that offer more accurate predictions of adverse outcomes such as child maltreatment. This topic is covered in Section 3.2 and Section 3.3 with a focus on the prevalent child maltreatment risk factors found in literature.

Having outlined the importance of recognizing child maltreatment risk factors and their role in developing predictive models, the subsequent sections, starting with Section 3.4, provide an overview of technical methodologies employed in previous studies, including a critical examination of how ethics, fairness, and bias are addressed in their methodological practices. Section 3.5 explores factors influencing the accuracy and fairness of predictive risk models during development and Section 3.6 delves into fairness-aware machine learning techniques, offering a deep understanding of the relevant notions of fairness, particularly in the context of predictive risk models for child maltreatment. This chapter plays a vital role as it examines the methods currently used to address unfairness and bias within the machine learning process. Crucially, this structured literature review seeks to provide a thorough understanding of predictive risk modeling in the child welfare context, covering a broad spectrum from identifying risk factors to exploring methodological considerations and ethical implications.

3.2 Correlates and Risk Factors

In risk-outcome research, the terms *correlate* and *risk factor* have been frequently used inconsistently and imprecisely. This misuse has led to inaccurate conclusions and recommendations in policy and research, lacking empirical support (Kraemer et al., 1997).

A risk factor refers to a characteristic, experience, or event that increases the likelihood (risk) of a specific negative outcome when present (Kazdin et al., 1997; Kraemer et al., 1997). Although a correlate is similarly associated with an increased likelihood of an outcome, there are two key differences between a risk factor and a correlate. Firstly, risk factors precede the outcome, whereas correlates are measured at the same time or after the outcome. For example, while exposure to parental neglect (a risk factor) may lead to behavioral problems in a child, low academic performance (a correlate) may be a consequence of those behavioral problems rather than a direct cause. Therefore, correlates may be the result of the outcome's effects. Secondly, risk factors enable the classification of populations into high-risk and low-risk subgroups, with the probability of the outcome being significantly higher in the high-risk group than in the low-risk group (G. S. K. Chung, 2021). Consequently, risk factors can function as predictor variables in predictive modeling, helping to estimate the likelihood of specific outcomes (Cuccaro-Alamin et al., 2017).

According to the definition provided by Kazdin et al. (1997), three types of risk factors have been identified. The first type is a fixed marker, representing a risk factor that cannot be changed, such as *gender*, *date of birth*, and *historical events*. The second type is a variable marker, where the risk factor can be demonstrated to change (e.g., *age*) or can potentially be altered (e.g., *parenting skills*). The third type is a causal risk factor, which directly influences the likelihood of the outcome when altered. Empirical evidence supports the mechanisms that explain how this type of risk factor contributes to the outcome. For instance, research has demonstrated that harsh parenting increases the risk of child physical abuse, making it a clear example of a causal risk factor (S. J. Lee et al., 2014).

Distinguishing between these types of risk factors and correlates is crucial, as it directly influences intervention targets and the accuracy of risk assessments in child welfare settings. For example, if a correlate is mistaken for a causal risk factor, interventions may focus on characteristics that do not directly impact the outcome, resulting in ineffective or misguided policies. Causal risk factors, however, can be used as both predictors and valuable intervention targets, as modifying them directly affects the likelihood of an adverse outcome. While non-causal risk factors can also serve as predictors, they may be less effective, and correlates are typically poor predictors and ineffective intervention targets (Franklin et al., 2017).

Although this thesis does not focus on investigating specific characteristics as risk factors for child maltreatment, understanding these factors through a review of the literature contributes to identifying information that can be used to develop predictors from administrative data to forecast adverse outcomes, such as child maltreatment.

3.3 Child Maltreatment Risk Factors

Child maltreatment risk factors have been extensively documented over the past few decades. A significant portion of these studies draws upon *Belsky's developmental-ecological model of child maltreatment*, as illustrated in Figure 3.1. Belsky (1980) suggests that the socio-ecological model (the Ecology of Human Development) proposed by Bronfenbrenner (1977) can be used to conceptualize the various factors contributing to the etiology of child maltreatment. The model consists of a series of nested environments that capture principal contexts related to human health and development, including the individual, interpersonal, community, and societal levels. The model demonstrates that an individual's immediate environment (i.e., individual and interpersonal levels) is influenced by the greater conditions surrounding the individual (i.e., the community and societal levels) (Bronfenbrenner, 1977).

According to Belsky's model (Figure 3.1), child maltreatment risk factors are organized into four levels: *Micro*, *Meso*, *Exo*, and *Macro*. The first level, Micro, focuses on individual factors. Some of these factors are associated with parents, while others are related to the child. For the Meso level, attention is directed towards family dynamics and communication within the family. This involves interactions between parents and their children or between parents themselves. The third level, Exo, examines the family's social system, including informal support, formal support based on government policies, and community support. Finally, the fourth level, Macro, considers cultural or social norms, laws, and policies. For instance, using physical force as a disciplinary method is a cultural risk factor (Macro). Additionally, specific government policies that increase the risk of negative outcomes for children and families fall under the Macro level.

However, it's important to acknowledge that certain child maltreatment risk factors may belong to multiple levels. Therefore, understanding the complex relationships between these risk factors across different levels is crucial. For example, a mother's unemployment can be seen both as an individual risk factor (Micro), concerning her job-seeking challenges, and as a community factor (Exo), influenced by the overall unemployment rate in the community.

In Belsky's model, the Micro level involves individual factors concerning both parents and children. To clarify the review of child maltreatment risk factors, we subdivided this level into two groups: *child level* and *caregiver-level* risk factors. Additionally, since the Meso level aligns with family characteristics, we included another group called *family level* risk factors. Finally, risk factors at the Exo level are grouped into *community level* risk factors. It's important to note that societal-level (Macro) risk factors are beyond the scope of this study and will not be considered in this thesis. Consequently, this review categorizes child maltreatment risk factors into four distinct groups: *child level*, *caregiver level*, *family level*, and *community level*.

3.3.1 Child Level Risk Factors

Although no child is responsible for their experiences of abuse or neglect, there are certain characteristics that may increase the risk of maltreatment. Some common factors that have been identified in the literature include *age* (A. E. Austin et al., 2020; Avdibegović & Brkić, 2020; G. Chung et al., 2022; Hindley et al., 2006; Walker & Wamser-Nanney, 2022; O. G. White et al., 2015), *gender* (Moody et al., 2018), and *special healthcare needs* or *disabilities* (Byrne, 2018; Hibbard et al., 2007; R. White et al., 1987).

A child's *disability* is often believed to place greater demands on their caregivers, which may exceed

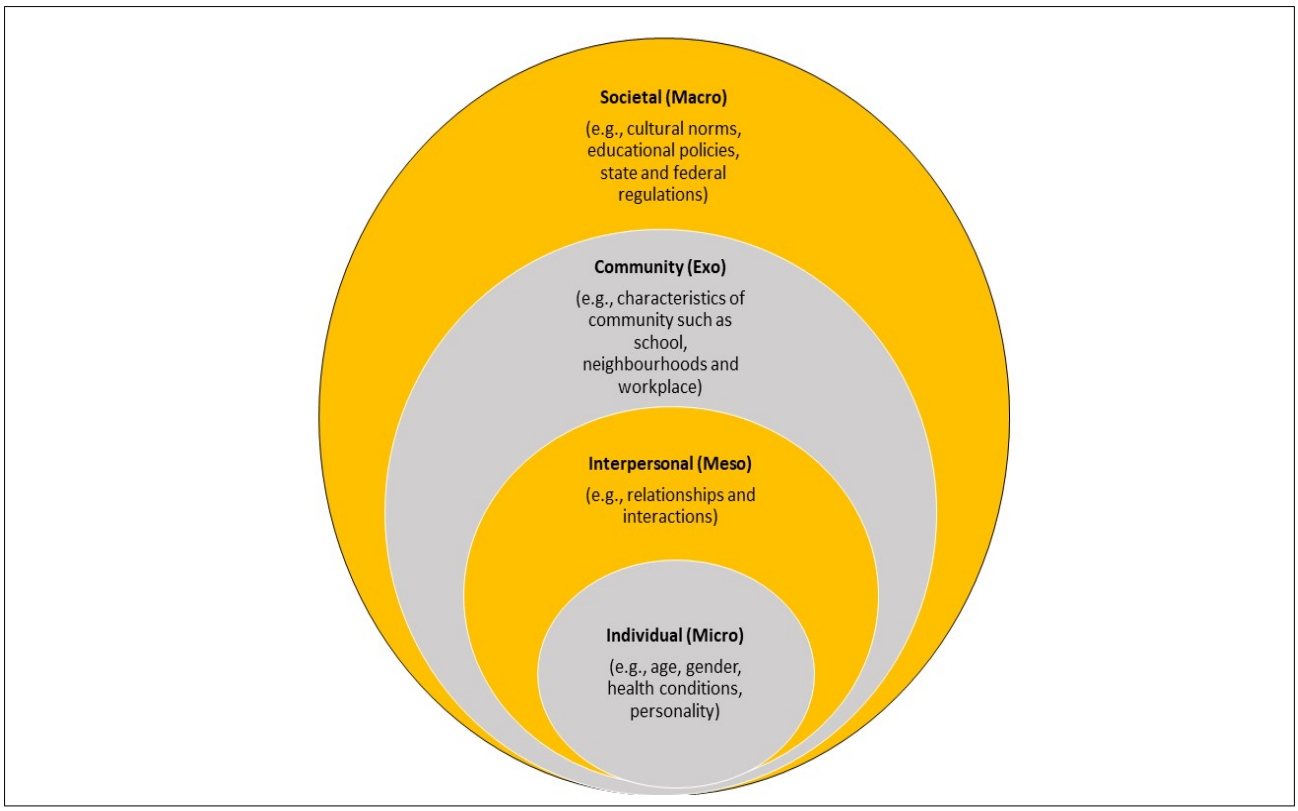

Figure 3.1: Belsky's developmental-ecological model of child maltreatment.

their capacity to meet the child's needs (G. S. K. Chung, 2021). In certain situations, when stress levels rise, there's a higher chance that a child might experience abuse (Hibbard et al., 2007).

With respect to *gender*, Moody et al. (2018) reviewed 337 study reports which provided prevalence rates of child maltreatment based on self-reports from either adults or children. The findings indicated that girls had higher rates of maltreatment than boys across all categories of abuse. Previous studies also highlighted gender differences in the prevalence of sexual abuse (Assink et al., 2019; Meinck et al., 2015).

A. E. Austin et al. (2020) suggest that *ethnicity* is often perceived as a risk factor for child maltreatment at the individual level, with 'African-American' and 'American Indian native children' experiencing higher rates of reports of concern, substantiated maltreatment, and out-of-home placements compared to 'white' children. However, it is not the child's *ethnicity* that increases their risk of maltreatment, but rather structural racism, systemic discrimination, and biases that lead to disproportionate reporting to child protective services. These systemic factors may also contribute to other risk factors for maltreatment among families and communities of these groups, such as *parental substance abuse*. Some studies suggest that *race* or *ethnicity* is not related to child maltreatment after controlling for other relevant individual, family, and extra familial factors (Ards et al., 2003; Freisthler et al., 2007).

A history of previous maltreatment is also consistently linked to the likelihood of experiencing maltreatment again (Littell & Schuerman, 2002; Rittner, 2002; Swanston et al., 2002). English et al. (1999) identified the number of prior reports of concern as the most significant factor contributing to the recurrence of maltreatment. Several studies indicate that the risk of recurrent maltreatment rises with each prior occurrence, and the time between episodes shortens as the number of maltreatment episodes increases (DePanfilis & Zuravin, 2001; Fluke et al., 1999).

3.3.2 Caregiver Level Risk Factors

Children's well-being relies on both their physical and mental health, with caregivers playing a crucial role in nurturing it. However, within the context of caregiving, some factors may have long-lasting and concerning effects on children. A number of well-established caregiver level risk factors for maltreatment include *level of education* (G. S. K. Chung, 2021; Thornberry et al., 2014; Younas & Gutman, 2022), *disabilities* or *physical problems* (Assink et al., 2019; G. S. K. Chung, 2021), *mental health problems* (A. E. Austin et al., 2020; Ayers et al., 2019; Mulder et al., 2018), *substance or alcohol use disorders* (A. E. Austin et al., 2020; Hindley et al., 2006; Huckle & Romeo, 2023), and a *history of childhood maltreatment* (Avdibegović & Brkić, 2020; G. S. K. Chung, 2021; Mulder et al., 2018; Younas & Gutman, 2022).

Thornberry et al. (2014) conducted a study examining risk factors during adolescence for involvement in child maltreatment in adulthood. The findings revealed that adolescents who demonstrated disengagement from school, poor academic performance, and low aspirations for college were more susceptible to showing maltreating behaviours later in life. Additionally, active involvement in antisocial behaviours, particularly problematic use of marijuana and alcohol during these formative years, demonstrated a significant correlation with subsequent instances of maltreatment (Thornberry et al., 2014).

Additionally, in a previous cohort study, Widom et al. (2009) observed that individuals who experienced childhood maltreatment, when controlled for their demographic characteristics, were more likely to meet the criteria for a borderline personality disorder (BPD) diagnosis in adulthood. This suggests a potential link between a *history of childhood maltreatment in parents* and an *increased likelihood of developing BPD*, which may increase the risk of inter-generational transmission of child maltreatment. Studies, such as Mulder et al. (2018), have also identified a *history of perpetrating harm to the child or other children*, along with a *history of criminal offending*, as significant caregiver level risk factors for child maltreatment.

3.3.3 Family Level Risk Factors

Focusing on family factors, some well-established family level risk factors for maltreatment include *domestic violence* (Avdibegović & Brkić, 2020; G. S. K. Chung, 2021), *poverty* (G. S. K. Chung, 2021; Moody et al., 2018), *family structure* (Assink et al., 2019; Littell & Schuerman, 2002; Younas & Gutman, 2022), and *family size* (G. S. K. Chung, 2021; Sedlak, 2014).

Avdibegović and Brkić (2020) state that *domestic violence* often becomes a risk factor when victimized mothers struggle to cope, focusing on the violent partner rather than meeting their children's basic needs. In addition to its direct effects, *domestic violence* can alter family dynamics, creating an indirect pathway to child abuse and neglect (Cox et al., 2003).

Poverty is widely recognized as a significant factor contributing to child maltreatment, often measured in research through indicators such as annual household income or participation in public benefits programs (Conrad-Hiebner & Byram, 2020). However, recent studies have shifted focus towards material hardship, directly assessing challenges in meeting basic needs like housing, food, utilities, and medical care, which are strongly associated with an increased risk of child maltreatment (Pelton, 2015; Yang, 2015). Additionally, families facing housing instability, utility shut-offs, food insecurity, and a higher count of hardships are more prone to child welfare investigations (Yang, 2015). Furthermore, housing instability is often linked to increased self-reported maltreatment behaviors by mothers, while food insecurity is associated with higher instances of physical and psychological aggression from mothers toward their children (Marcal, 2018).

With respect to *family structure*, Assink et al. (2019) found that non-nuclear family structures, like those with stepfathers or single-parent families, elevate the risk of child abuse victimization. Single-parent households often increase the likelihood of neglect and physical abuse (Oliver et al., 2006; Sedlak, 2014). Stepfather involvement, particularly for girls, is linked to a higher risk of sexual abuse (Assink et al., 2019). Finkelhor (2008) suggests that correlated features, such as exposure to unrelated individuals and dysfunctional interpersonal patterns, contribute to child maltreatment risks. Disrupted families may expose children to conflict, aggression, and violence, heightening the risk of victimization. Additionally, children in disrupted families may have less control over their environment, making them more susceptible to high-risk situations and victimization (Finkelhor, 2008).

3.3.4 Community Level Risk Factors

At the community level, most research on child maltreatment has focused on factors in the child's and family's neighbourhood that may contribute to their experiences of maltreatment. According to Avdibegović and Brkić (2020), the community in which a family resides significantly influences the behaviours of its members. Factors such as a *disadvantaged environment*, *low socio-economic status in the neighbourhood*, *lack of community support*, *limited social services*, and *prevalence of neighbourhood crime, violence, and alcohol consumption* can all contribute to instances of child abuse and neglect (Akehurst, 2015; Freisthler et al., 2007; Moody et al., 2018; Parkinson, 2017). The hypothesis is that these factors increase the risk of child maltreatment by rising parental and family stress while weakening social networks and community organization (Daley et al., 2016; Morris et al., 2019; Thurston et al., 2017).

3.4 Predictive Risk Modelling in the Child Welfare Context

With a rapid rise in the amount of information and data available, predictive analytics tools have emerged to assist organizations in making better decisions. These tools efficiently use data, identify patterns, and predict outcomes more accurately than humans can. The process of developing these kinds of tools has evolved across several fields such as computer science, statistics, and mathematics, and is referred to as predictive modeling (Kuhn & Johnson, 2013). Predictive models are developed using historical data and machine learning algorithms. Machine learning algorithms look for patterns in data and construct predictive models that assist in decision making by predicting outcomes (Binns, 2018). Predictive risk models are a type of predictive model that generate a risk score for the occurrence of an adverse event, allowing those events to be avoided through a more calculated delivery of services (Vaithianathan et al., 2017).

In specific contexts like child welfare, where social workers handle referrals, risk prediction models that use administrative data are proven valuable (Centre for Social Data Analytics, n.d. Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2017). When a referral is made, social workers follow practices and policies by gathering information about the parents and children associated with the referral. Even though child welfare agencies have access to administrative data, it's difficult to efficiently use the information about all the children and adults on a single referral. Risk prediction models can help social workers make better use of data and identify cases that are more likely to result in an adverse outcome (Vaithianathan, 2012).

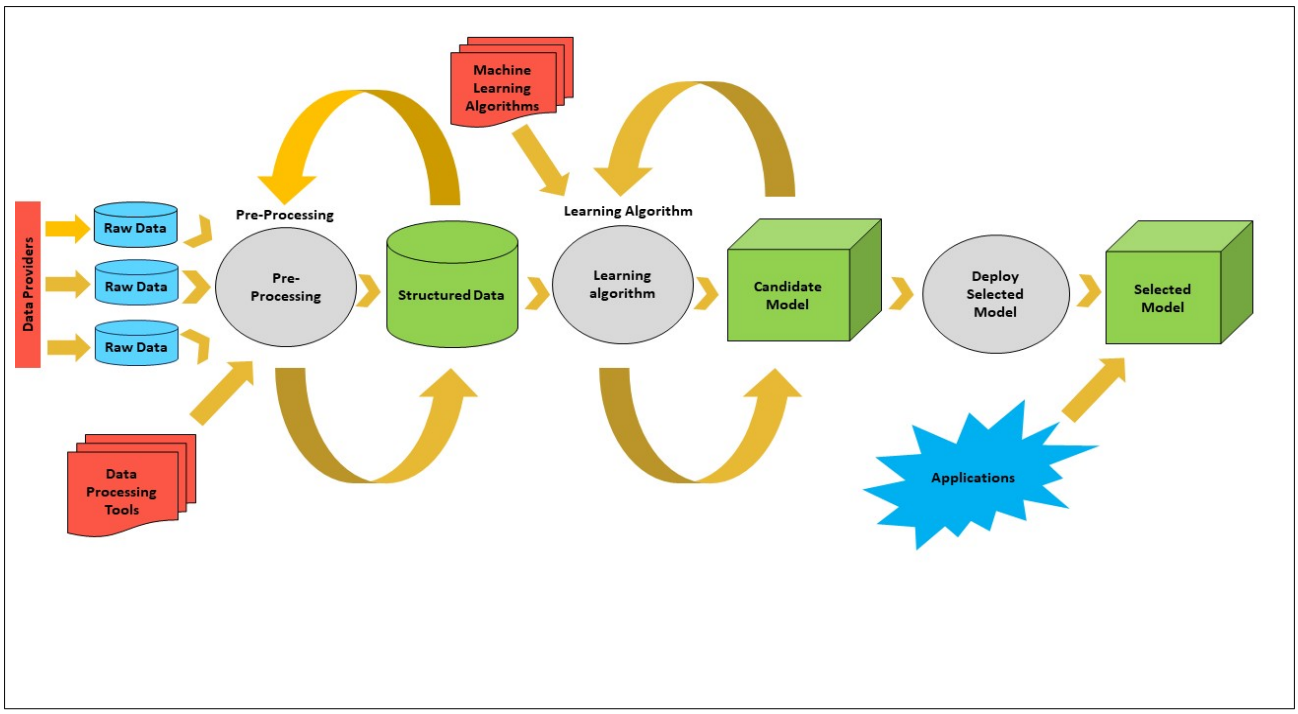

Figure 3.2: A visual guide to the standard machine learning process.

In child welfare settings, the training dataset used by these models are often generated through extracting information directly out of the child welfare agency's database systems and include records of interactions with the children and their families. Depending on the governing authority or agency responsible for data sharing policies, these datasets may be linked to data collected by other government agencies such as public hospitals, birth records, public benefits, criminal justice, education records, and more.

The development of predictive risk models using machine learning techniques follows a standard multi-step procedure. The majority of child welfare predictive risk models have adhered to these steps, as illustrated in Figure 3.2. The following subsections offer a comprehensive review of relevant studies that explored the development of predictive risk models within the child welfare domain, concerning their methodological approaches and evaluation processes.

3.4.1 Overview of International Studies

This section gives an overview of international literature on predictive risk modelling in the child welfare context. Methods and model evaluation processes for these studies are briefly summarized in Table 3.1. Studies in this table are classified by authors, country of origin, and purpose. The purpose of a study is labeled as "applied" if the model in that study was developed for a practical deployment within a child welfare setting, and "theoretical" if the study offers recommendations without an implementation plan for the developed predictive risk model. The table also includes sample size,

data sources, machine learning algorithms, and performance metrics, with a focus on *Area Under the Curve (AUC)*, *Cohen's kappa* (κ), and *Matthew's Correlation Coefficient (MCC)*.

AUC refers to the area under the Receiver Operating Characteristic (ROC) curve, a commonly used metric to evaluate the performance of binary classifiers. AUC measures the ability of a model to distinguish between positive and negative classes. Specifically, it represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance by the model (Kuhn & Johnson, 2013). Cohen's κ, first introduced by Cohen (1960), is a statistic used to assess inter-rater reliability for qualitative data. It is considered more robust than a simple percent agreement, as κ accounts for the likelihood that agreement may occur by chance (Kuhn & Johnson, 2013). Matthews Correlation Coefficient (MCC) is a metric used to evaluate the quality of binary classifications. It considers all elements of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and is particularly useful when the classes are imbalanced (Matthews, 1975).

The evaluation of model performance primarily relies on the AUC metric with values categorized as excellent (≥ 0.90), good ($0.80 \leq AUC < 0.90$), fair ($0.70 \leq AUC < 0.80$), or poor (< 0.70) based on guidelines from Kuhn and Johnson (2013). For studies with no AUC reported, Cohen's (κ) and MCC are used. An MCC or κ ranging from 0.76 to 1.00 is considered excellent, 0.51 to 0.75 as good, 0.26 to 0.50 as fair, and below 0.26 as indicative of poor performance, following Landis and Koch (1977) criteria.

A consistent trend emerges from the findings in Table 3.1: models trained on smaller datasets ($n < 1000$) reveal poorer performance, often falling below the standard for typical machine learning dataset results, when compared to those with larger samples (Benesh, 2017; Horikawa et al., 2016; Thurston & Miyamoto, 2018). Larger datasets are crucial for training complex, nonlinear algorithms (e.g. Neural Network, Random Forest, Support Vector Machine, etc). The resulting poor performance may signal under-fitting, suggesting a need for extended training or additional variables (Sen, 2021). Consequently, studies training nonlinear algorithms on small datasets, may have lacked sufficient data to accurately predict the outcome (Benesh, 2017; Thurston & Miyamoto, 2018).

The majority of studies with superior model performance have used linked data from diverse sources, including child welfare, social welfare, criminal justice, and demographic records. This integrated approach enables researchers to incorporate a wide range of risk factors for child maltreatment, consequently enhancing the predictive accuracy of the models. Examples include studies by Jolley (2012),

Putnam-Hornstein et al. (2022), Schwartz et al. (2017), Vaithianathan, Dinh, et al. (2019), Vaithianathan, Kulick, et al. (2019), and Vaithianathan et al. (2020), where models trained on cohesive datasets consistently outperform those relying on limited predictors (Benesh, 2017; Rodriguez et al., 2019).

In applied U.S. studies, the main focus is on predicting the likelihood of placement or removal as adverse outcomes for children referred to child protective services. Thus, the outcome variable is typically defined to classify children into high-risk and low-risk categories based on whether a removal or placement event occurs within a specified time frame (Putnam-Hornstein et al., 2022; Vaithianathan, Dinh, et al., 2019; Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2020). Chouldechova et al. (2018) has addressed challenges such as racial bias and the over-representation of African-American children which can stem from the practical definition of outcome variables, particularly within the development process of models utilized in AFST. The initial version of the tool (AFST I) employed logistic regression to predict re-referral or placement outcomes for children referred to Allegheny county's child protective services(Vaithianathan et al., 2017). However, a number of complications arose during the validation of the re-referral model, as it tended to assign higher scores to children involved in custody disputes or with recurrent reports. This matter raised concerns about potential racial bias implanted in the initial incoming referrals. Therefore, a model predicting future referrals tended to over represent African-American children and disadvantage their families compared to white children. In response, the researchers restricted the model in the second version (AFTST II) to predict significant safety issues that usually lead to a court ordering the placement of a child (Vaithianathan, Kulick, et al., 2019).

Moreover, in these applied studies, LASSO logistic regression is consistently preferred due to its simplicity and efficiency in deployment, despite alternative algorithms like random forest, support vector machines, and XGBoost being explored in training. Vaithianathan, Kulick, et al. (2019) state that random forest and XGBoost, consist of a sequence of linked trees that present greater challenges, particularly in debugging the intricacies of the deployed algorithm. Researchers, particularly those from the Vaithianathan's team (CSDA), favor LASSO because of its unique ability to set certain predictor weights to zero, facilitating feature selection and regularization for more interpretable and accurate models (Putnam-Hornstein et al., 2022; Vaithianathan, Dinh, et al., 2019; Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2020).

Table 3.1: Overview of international studies on child welfare predictive risk modeling.

Authors	n	Country/Purpose	Data Sources	Outcome	ML Algorithms	Performance
(Jolley, 2012)	6,747	US/Theoretical	- Demographics datasets - Child welfare system - Social welfare system - Criminal justice system - Juvenile justice system	Substantiated maltreatment recurrence.	- Logistic regression - Neural network *** - Random forest - Tree based models	AUC = 0.81 (Good)
(Horikawa et al., 2016)	716	Japan/Theoretical	- Child welfare system	Maltreatment substantiation recurrence within one year of initial maltreatment finding.	- Step-wise multiple logistic regression***	AUC = 0.69 (Poor)
(Amrit et al., 2017)	13,170	Netherlands/Applied	- Unstructured and structured medical data provided by the child health department (JGZ) of the largest public health organization in the Netherlands, the GGD Amsterdam.	Alleged maltreatment.	- Naïve bays - Random forest*** - Support vector machine	AUC = 0.91 (Excellent)
(Schwartz et al., 2017)	78,394	US/Theoretical	- Child welfare system (both private and public) - Criminal justice system	1.Substantiated finding of maltreatment. 2.Type and intensity of services delivered.	- Decision trees (C5 and CHAID) with ensemble learning and boosting.	1. AUC = 0.87 2. AUC = 0.81 (Good)

Table 3.1 continued from previous page

Authors	n	Country//Purpose	Data Sources	Outcome	ML Algorithms	Performance
(Benesh, 2017)	727	US/Theoretical	- National Child Abuse and Neglect Data System (NCANDS)	Future placement type for foster youth at 18 months.	- Random Forest	RQ1: κ = 0.35, MCC = 0.38 (Fair) RQ2: κ = 0.16, MCC = 0.1 (Poor)
(Thurston & Miyamoto, 2018)	700	US/Theoretical	- Child welfare system - Social welfare system - Criminal justice system	Substantiated finding of serious maltreatment.	- Model based recursive partitioning decision trees	κ = 0.1-0.16, MCC = 0.11-0.19 (Poor)
(Elgin, 2018)	233,633	US/Theoretical	- Adoption and Foster Care Analysis and Reporting System	Whether foster children would achieve legal permanency.	- Boosted trees - Classification trees - Elastic net regression - Logistic regression - Lasso regression - Multivariate adaptive regression splines - Neural networks*** - Partial least squares discriminant analysis - Random Forest*** - Support Vector machines	AUC=0.99 (Excellent) κ = 0.87, MCC = 0.87 (Excellent)
(Rodriguez et al., 2019)	12,017	US/Theoretical	- National Child Abuse and Neglect Data System (NCANDS)	Unsubstantiated maltreatment.	- Random forest	κ = 0.43, MCC = 0.43 (Fair)

Table 3.1 continued from previous page

Authors	n	Country//Purpose	Data Sources	Outcome	ML Algorithms	Performance
(Vaithianathan, Kulick, et al., 2019)	82,211	US/Applied (AFST2)	- Demographics datasets - Child welfare system - Criminal justice system - Juvenile justice system - Health system	Out of home placement within two years.	- LASSO regression*** - Logistic regression - Random forest - XGBoost	AUC = 0.76 (Fair)
(Vaithianathan, Dinh, et al., 2019)	221,519	US/Applied (DCDA)	- Demographics datasets - Child welfare system - Criminal justice system - Juvenile justice system - Social welfare system	Removal and placement between a childbirth and their 3rd birthday.	- Boosted trees - LASSO regression*** - Random forest - Risk slim	AUC = 0.924 (Excellent)
(Vaithianathan et al., 2020)	52,520	US/Applied (Hello baby)	- Child welfare system - Social welfare system	Removal and placement within next 24 months.	- LASSO Regression*** - Random forest - XGBoost	AUC = 0.804 (Good)
(Putnam-Hornstein et al., 2022)	341,428	US/Applied (LA Risk stratification model)	- Child welfare system - Case management system	Removal and placement in foster care within 24 months.	- LASSO regression*** - Random forest - XGBoost	AUC = 0.83 (Good)

3.4.2 Overview of New Zealand Studies

In NZ, limited research has investigated the use of predictive risk modeling within the child welfare sector. This initiative originated from a study conducted by Vaithianathan (2012) and was further supported by the NZ government to explore its feasibility (Bennet, 2012). Table 3.2 provides a summary of studies conducted in NZ with a focus on predictive risk modeling in the child welfare context. The table details the data sources used, encoded variables, machine learning algorithms tested, and the performance of candidate models, as measured by AUC, in these studies.

Vaithianathan (2012) and Wilson et al. (2015) contributed valuable insights to predictive risk modeling in the NZ child welfare system. The former focused on young children within the public benefits system, aiming to predict substantiated maltreatment findings before the age of five for every child entering the public benefit system by age two. In this study, the encoding of the dichotomous outcome variable involved longitudinally tracking these children to determine whether maltreatment occurred before they reached the age of five. The predictor variables were derived from record linkages between NZ's social welfare system and child welfare system, based on data availability and their potential to proxy known risk and protective factors for child maltreatment. The final model, developed through step-wise probit regression, demonstrated effectiveness by achieving an AUC of 0.76 which indicates a fair performance measure.

In addition, Wilson et al. (2015) explored a population-wide predictive risk model, centred on identifying newborns at significant risk of future maltreatment. The study aimed to predict whether these children would experience maltreatment by age five. Similar to Vaithianathan (2012), predictor variables in this study were encoded based on literature-identified risk factors, with multiple modeling techniques tested and compared (see Table 3.2). Despite similar predictive performances (with an AUC of approximately 0.87) across all machine learning algorithms tested, step-wise logistic regression was selected for its transparency. Wilson et al. (2015) concluded that predictive risk models based on administrative data have the potential to prioritize preventative services for newborns yet emphasized their supplementary role to professional judgment. The authors warned about the risk of not being able to identify all high-risk children under this approach, advocating for predictive risk models as complements, not replacements, in identifying high priority children.

Following comprehensive ethical reviews and feasibility studies on the integration of predictive risk modelling in the NZ child welfare system, the Ministry of Social Development launched the *Enhancing Intake Decision Making Project*. The primary objective of this project was to explore the use

of predictive risk modeling within intake decision-making process, with focus on refining decision-making rather than specifically targeting newborn or young children for intervention and prevention strategies. A detailed report of this project is available in the work by Rea and Erasmus (2017). Four modelling algorithms were tested and for all the models, variables selection was applied based on the contribution of variables to the predictive power of the model. Despite similar performance ($AUC \approx 0.75$) across all machine learning algorithms tested, logistic regression was chosen as the preferred model. The project's conclusion emphasized the potential improvement in child welfare intake decisions through predictive risk modeling. However, the study suggested that incorporating external data could enhance the model's predictive capability, providing an example of using Department of Corrections data to verify a child's living situation.

Three additional studies in NZ were also identified, which explored the utilization of predictive risk models in the NZ child welfare system (A. James et al., 2019; Vaithianathan et al., 2018; Walsh et al., 2020). However, the primary focus of these studies found to be different from the aim of developing a deployable predictive risk model.

For instance, A. James et al. (2019) developed a model as a proof-of-concept to emphasize the importance of incorporating family network data into predictive risk models, rather than aiming to create a model suitable for practical risk prediction purposes. The dataset for this study was constructed by linking birth records with data from the NZ child welfare and social welfare systems, and the analysis employed multivariate logistic regression. The outcome variable in this study was defined based on the established system outlined in the study by Rea and Erasmus (2017).

Vaithianathan et al. (2018) investigated the effectiveness of predictive risk modeling in the identification of children at high risk of injuries or even death during early childhood. Using a cohort of children born in NZ in 2010, the authors trained a logistic regression model to predict a child's risk of substantiated finding of maltreatment by age two. The logistic regression model was then used to assign risk scores to the 2011 birth cohort. Children in the top 10% of predicted risk scores were flagged as very high risk, and those in the top 20% were flagged as high risk. Through linking these children with health system records and comparing the incidence of injury and mortality rates between high-risk children identified by the model and the rest of the birth cohort, the authors revealed that children identified at both 10% and 20% risk thresholds had higher post-neonatal mortality rates (4.8 times and 4.2 times greater, respectively) and a greater relative risk of hospitalization (2 times higher and 1.8 times higher, respectively).

Employing data from the Growing Up in NZ (GUiNZ)[7], Walsh et al. (2020) aimed to uncover protective factors within families identified as high-risk by predictive risk models. Using predictors observed at or before birth, a predictive risk model was trained using logistic regression to identify children with two or more *adverse childhood experiences by 54 months*. The study designated the top 20% (n = 790) with the highest predicted risk as the high-risk group. Within this group, researchers identified children with zero observed *adverse childhood experiences* and examined 749 potential protective factors across five domains: *community and neighborhood* (36%), *family finances* (23%), *parent-child relationship* (18%), *parent health and wellness* (14%), and *mother-partner relationship* (9%). The analysis included estimating separate logistic models for each factor, with significant ones further analysed. More details are available in the Appendix by Walsh et al. (2020).

While the potential value of predictive risk modelling in decision-making is acknowledged worldwide, research indeed warns against risks associated with the use of such tools in the child welfare domain (Dare, 2013; Drake et al., 2020; Gillingham, 2016; Glaberson, 2019; Keddell, 2015, 2019). In addition to careful selection of outcome variables and the assurance of their consistent registration (Gillingham, 2016), there are moral and ethical challenges that require consideration (Keddell, 2015). Particularly ethics, fairness, and bias. These key aspects must be addressed in methodological practices to ensure that models are not only effective but also ethically sound, fair (Gillingham, 2016). For example, for historical data containing instances of discriminatory decisions, a trained algorithm not only replicates but may even amplify these biases (Barocas & Selbst, 2016; Edelman & Luca, 2014; Žliobaitė, 2017). Researchers aiming to mitigate such bias should explore strategies that trade-off precision in their models for increased fairness, such as optimizing for ethnicity equity (Section 3.6).

[7]Growing Up in NZ is NZ's largest longitudinal study of child health and well-being, following the lives of more than 6000 children and their families. Parents were enrolled and interviewed during pregnancy and at several points including at 9, 24, and 54 months after birth.

Table 3.2: Overview of New Zealand studies on child welfare predictive risk modeling.
Note: The term "full regression" refers to a regression model that includes all available or selected predictor variables, without simplification, variable selection, or regularization.

Authors	Outcome	Data Sources	Predictors	ML Algorithms	AUC
(Vaithianathan, 2012)	Substantiated maltreatment by the age of five including neglect, emotional, physical, and sexual.	- Child welfare system - Social welfare system	**Child level** • History of interaction with CPS: - Findings of abuse or neglect, - CP notifications, - Investigations, - Family Group Conferences (FGC), - CP assessments. • History of interaction with SWS: - Proportion of time on a benefit. **Caregiver level (Mother)** • Characteristics of care-giver at the start of benefit spell: - Gender, - Age, - Education. • History of interaction with CPS before the age of 16. • History of interaction with SWS in adulthood. **Family level** • History of interaction with CPS for other children, • Family characteristics: - Single versus dual caregivers, - Number of younger and older children, - Birth intervals to next youngest and oldest child, - Multiple birth children, - Maternal age of the caregiver regarding oldest child and subject child.	Probit regression***	0.76 (Fair)

Table 3.2 continued from previous page

Authors	Outcome	Data Sources	Predictors	ML Algorithms	AUC
(Wilson et al., 2015)	Substantiated finding of maltreatment (emotional,physical, or sexual abuse or neglect) by age two.	- Demographics datasets (Registrar-General of Births, Deaths, and Marriages) - Social Welfare system - Child Welfare System - Criminal justice system - Health system	***Child level*** • Gender, • Low birth weight or pre-term (Yes, No and Unknown). ***Caregiver level (Mother)*** • Age, • Mental health problems in the last 5 years, • Substance abuse issues in the last 5 years, • Corrections history in the last 5 years, • Behavioural or relationship difficulties as a child, • Time on benefit in the last 5 years, • History of interaction with CPS in childhood. ***Family level*** • History of domestic violence, • Whether benefit caregiver is a birth registration parent, • Benefit address changes in the last year, • History of Interaction with CPS for other children. ***Community level*** - CYF site.	- Decision tree - DMINE regression - Full regression - Gradient boosting - Step-wise logistic regression*** - Multilevel model - Neural network - Partial least squares - Regression with backward elimination	0.87 (Good)

Table 3.2 continued from previous page

Authors	Outcome	Data Sources	Predictors	ML Algorithms	AUC
(Rea & Erasmus, 2017)	*Estimated care and protection concern* within two years of the initial notification.	- Social welfare system - Child welfare system	***Child level*** • Age, Gender • History of interaction with CPS: - Prior risk and safety assessment, - Number of previous CP notifications, - Number of days since last Section 15 intake, - Number of previous substantiated findings of maltreatment, - Open phase case, - Prior custody or guardianship spell. • History of interaction with SWS including: - On a main benefit at the time of notification, - Level of contact with SWS and CPS. **Level 1.** No previous SWS or CPS contact. **Level 2.** Previous CPS contact, no previous SWS Contact. **Level 3.** Previous SWS contact, no previous CPS contact. **Level 4.** Previous SWS and CPS contact. ***Caregiver level*** • History of Interaction with CPS: - Level of interaction with CPS in childhood. **Level 1.** None. **Level 2.** Reported. **Level 3.** Finding of maltreatment. **Level 4.** Placement. - Whether Mother is known to CPS. ***Family level*** • Number of siblings at the time of notification. • Number of contact records for siblings. ***Community level*** • New Zealand deprivation index 2013. ***Others*** • Notifier role type	- Decision tree - Gradient boosting - Logistic Regression*** - Random forest	0.75 (Fair)

Table 3.2 continued from previous page

Authors	Outcome	Data Sources	Predictors	ML Algorithms	AUC
(Vaithianathan et al., 2018)	Substantiated finding of maltreatment by age two.	- Demographics datasets (birth records) - Child welfare system - Social welfare system - Criminal justice system - Health system	***Child level*** • Gender, • Pre terms birth indicator. ***caregiver level (Mother)*** • Maternal age, • Marital status, • Receipt of public income support, • History of mental health or substance abuse, • Criminal records in the last 5 years, • History of child maltreatment allegations to CPS as a child. ***Family level*** • Parenting demand: - Indication of having more than 3 children in the family, - Indication of having multiple birth children, - Indication of having multiple children under the age of 2, • History of child maltreatment allegations to CPS for siblings.	Logistic regression***	0.88 (Good)

Table 3.2 continued from previous page

Authors	Outcome	Data Sources	Predictors	ML Algorithms	AUC
(A. James et al., 2019)	*Estimated care and protection concern* within two years of the initial notification.	- Demographics datasets (birth records) - Social welfare system - Child welfare system	***Child level*** • Age at the time of notification, • History of interaction with CPS including: - Number of previous CP notifications, - Number of prior targets events, - Number of notifications in the last 2 years, - Number of target events in the last 2 years. ***Family level*** • Network constructed using relationships recorded prior to the notification: - Number of individuals in network, - Number of individuals in network with a prior notification, - Number of individuals in network with a prior target event, - Number of known abusers in network. • Network constructed using whole-life relationships (parent, child, sibling, half-sibling relationships only, backdated to the date of birth of the younger of the two individuals in the relationship: - Number of individuals in network with a prior notification, - Number of individuals in network with a prior target event, - Number of known abusers in network.	Multivariate Logistic Regression***	0.672 (Poor)
(Walsh et al., 2020)	Odds of more than two adverse childhood experience by age 54 months.	Data collected by growing up in NZ study (GUiNZ)	Kindly refer to Appendices 2 and 3 in the paper for an exhaustive list of predictors that couldn't be accommodated in this table.	Logistic regression***	0.76 (Fair)

3.4.3 Ethics, Fairness, and Bias

Ethical reviews and feasibility studies on the use of predictive risk modeling in the child welfare context consistently highlight concerns, with the accuracy and fairness of these tools being among the most prevalent (Blank et al., 2015; Dare & Gambrill, 2017; Ministry of Social Development, 2014). As a result, researchers are required to address these concerns during the development process of models intended for use by child welfare authorities..

Ethical considerations include examining the impact of algorithmic decision-making on children and their families (Keddell, 2015). Fairness issues relate to potential surveillance, stigmatization, or unfair distribution of resources based on demographic characteristics (Chouldechova et al., 2018), while bias refers to how a social worker's personal beliefs and attitudes might influence decision-making practices (Gillingham, 2016).

In the following sections, a qualitative review of studies on the use of predictive risk modeling in the child welfare domain is conducted to investigate how these studies have considered issues of ethics, fairness, and bias (see Tables 3.1 and 3.2). Particular attention is given to concerns raised by Gillingham (2016) and Keddell (2015) regarding how inequities in child welfare may be shaped or exacerbated by algorithms.

3.4.3.1 Considerations by International Studies

The U.S. applied studies highlighted in Table 3.1 serve as prime examples of a strong commitment to addressing ethical concerns. Before implementing these models, researchers and project leaders took proactive steps by consulting independent experts and seeking ethical reviews, which provided equity-related recommendations (Dare & Gambrill, 2017; Drake & Jonson-Reid, 2018; Veale, 2019).

For instance, in the *AFST project* (Vaithianathan, Kulick, et al., 2019; Vaithianathan et al., 2017), an ethical review overseen by Dare and Gambrill (2017) was conducted. Regarding the potential introduction of unfairness or bias based on *race* through the utilization of predictive risk modeling, the reviewers approved the inclusion of *race* as a predictor variable, provided it had a significant effect on the model's predictive performance. Additionally, they mandated the non-disclosure of predicted risk scores to workers handling screened-in calls to mitigate the risk of bias. To assess fairness in the AFST project models, Vaithianathan, Kulick, et al. (2019) compared AUC metrics across racial subgroups to investigate variations in performance concerning *race* (*accuracy equity* notion of fairness). They also examined placement rates relative to predicted risk among different racial groups to

estimate potential overestimations or underestimations based on *race* (*calibration* notion of fairness). For more details on these notions, see Section 3.7. Additionally, the researchers analyzed potential biases in decisions made by hotline social workers by comparing the rates at which they overrode algorithm recommendations across different levels of perceived risk. The findings indicated that the workers' decisions were more influenced by their subjective assessments than by the algorithm's recommendations, even at varying levels of perceived risk. This suggests that workers tended to rely more on their own judgment rather than trusting the algorithm's suggestions.

In the *Hello Baby project* (Vaithianathan et al., 2020), the variable *race* was excluded from the predictor variables due to concerns about potential bias raised during a community meeting. While testing *race* as a predictor, it was found that this exclusion did not significantly impact the model's predictive power. However, the researchers emphasised that this approach does not negate the possibility of *race* playing a role through existing correlated variables.

Moreover, highlighting the importance of defining fairness across demographic groups, Vaithianathan et al. (2020) considered the nature of actions taken in response to the predictive risk model, whether punitive or beneficial. In the context of the Hello Baby program, *predictive parity* was identified as the most suitable measure (see Section 3.7). This measure ensures equal risk for White and non-White priority children within each eligibility category, even if different proportions of each racial subgroup are captured. The analysis regarding predictive parity revealed that the rate of placements among the White population in the priority group is higher than that for the non-White population (248.3 per 1,000 versus 160.3 per 1,000), while the measures of adverse outcomes for children (404.3 per 1,000 versus 382.1 per 1,000) and mortality (20.0 per 1,000 versus 18.3 per 1,000) are generally similar. These findings suggest that White and non-White children who are prioritized are at similar risk of harm. However, whether the program will be equally effective for both subgroups remains an open question, and as stated by the research team, an impact evaluation study must be conducted to answer this.

Similar approaches were taken by Vaithianathan, Dinh, et al. (2019) and Putnam-Hornstein et al. (2022). Despite coding *race* to examine model performance for different subgroups, *race* was not included as a predictor in building the model, following feedback from community partners. An analysis of disparities in the DCDA model showed that the tool is well-calibrated across different subgroups of *race*, *age*, and *gender*.

Regarding theoretical studies, Schwartz et al. (2017) assessed unethical decision-making in the child

welfare system by focusing on referrals leading to unproven allegations. The study discovered that 40% of these reports were unproven and were linked to a 175% higher chance of the family facing repeated involvement with child welfare services. Their model showed better accuracy (90% and 93%) in identifying cases for court referral and court-involved services compared to the existing method with 60% accuracy. Although the study didn't explicitly consider ethics or fairness during the model's initial design, the researchers suggested adoption of their model could significantly reduce the risk of unsubstantiated referrals for families reported for potential maltreatment.

Rodriguez et al. (2019) and Jolley (2012) underscored the imperative of minimizing unfair or unethical predictions, yet their evaluations of efforts to mitigate unfairness were incomplete. Rodriguez et al. (2019) focused on addressing surveillance and stigmatization of marginalized communities in the child welfare system, developing a model that predicted unsubstantiated maltreatment by emphasizing protective factors rather than focusing solely on risks. Meanwhile, Jolley (2012) criticized actuarial risk assessments for relying on linear assumptions, which were believed to lead to inaccurate predictions of child maltreatment risk. Both studies prioritized ethical considerations in their approach to predictive modelling.

Horikawa et al. (2016) mentioned two key limitations within their study's methodology. First, they noted the challenge of relying on subjective judgment due to the absence of established scales for numerous variables. This absence introduces inherent subjectivity in data collection, which emphasizes the potential influence of individual perspectives. Second, they pointed out that the overstretched nature of staff handling maltreatment cases often results in the assessment and recording of many items as dichotomous data. This approach, driven by resource constraints, oversimplifies complex variables, thereby limiting the statistical analyses' effectiveness. Acknowledging these limitations highlights the importance of interpreting study results with caution, considering the potential impact of subjective judgments and resource constraints on data robustness. None of the remaining studies addressed issues of unfairness or bias. However, in instances where these concerns were identified, most did not attempt to address them by adjusting the algorithmic formula.

3.4.3.2 Considerations by New Zealand Studies

Predictive risk models developed in (Rea & Erasmus, 2017; Vaithianathan, 2012; Wilson et al., 2015) aimed to evaluate both the technical feasibility and predictive validity of implementing predictive risk modelling in the NZ child welfare system. These studies were the subject of comprehensive ethical reviews (Blank et al., 2015; Dare, 2013) and technical feasibility assessments (Ministry of Social

Development, 2014). The combined findings from these studies and their evaluations suggest that predictive risk models utilizing administrative data can effectively identify high-risk children in NZ for preventive services. However, it is essential to recognize these models as supplements to professional judgment rather than substitutes. These studies also highlight risks associated with predictive modelling, including concerns about bias and potential adverse impacts on Māori communities. For example, while Vaithianathan (2012) did not specifically investigate bias or assess the model's fairness regarding its impact on Māori, the authors emphasized the importance of an ethical framework if their model were to be applied. In a subsequent publication by Vaithianathan et al. (2013), building on earlier results, the authors argued against using *ethnicity* as a predictor variable due to concerns about reinforcing racial stereotypes.

Wilson et al. (2015) examined fairness by comparing how their model identified Māori children at risk of maltreatment with the observed substantiation rates. They found that the algorithm's predictions were not proportionate to the percentage of Māori children in the maltreated population. In an effort to address this, the researchers tried creating two separate algorithms, one for Māori children and another for other ethnic groups. However, this strategy did not yield the intended results.

Similarly, Rea and Erasmus (2017) explored the disparities between existing referral rates and those produced by the model across different ethnic groups. The findings revealed that the model overrepresented Māori children compared to the current practice. The reasons behind this elevated referral rate remain undisclosed, prompting the researchers to advocate for additional investigation. Rea and Erasmus (2017) emphasized the need to understand how the model could be utilized in the future to ensure accurate risk identification and prevent unintentional bias or unfairness, particularly concerning ethnic groups such as Māori.

A. James et al. (2019) decided not to include *gender* and *ethnicity* as predictor variables due to ethical concerns around stereotyping. The authors emphasised the importance of evaluating algorithmic fairness across various demographic groups. They assessed predictive bias by examining statistical calibration and comparing weighted error rates across gender and ethnic groups. Their models were well-calibrated, with similar weighted error rates across these groups (A. James et al., 2019).

Walsh et al. (2020) added protective factors to their model without explicitly addressing fairness across demographic groups as a motive, and Vaithianathan et al. (2018) did not mention anything regarding bias and fairness in their paper.

While these studies collectively underscore the significance of ethical considerations, fairness evaluations, and a comprehensive understanding when implementing predictive risk models in child welfare systems, a notable gap remains in the NZ literature. To our knowledge, no study in the country has explored fairness-aware machine learning approaches to address potential unfairness or bias in these models, which could be valuable for child welfare agencies.

This thesis takes a distinctive approach by concentrating on fairness-aware machine learning approaches to develop a model tailored for potential use in intake decision-making. As the literature review progresses, the focus will shift towards enhancing these specific aspects in predictive risk models. It is important to note that ethical considerations, while acknowledged as crucial, are deliberately outside the scope of this thesis.

3.5 Determinants of Unfairness in Child Welfare Predictive Modelling

In the predictive modelling process, a number of factors can cause disproportionately adverse outcomes for specific groups. Several factors or mechanisms are identified in the literature as potential contributors to unfairness or errors in child welfare predictive risk models, including but not limited to, characteristics of the training data, the definition of the outcome variable used for classification, and the selection of features or predictors. Subsequent subsections will review these mechanisms.

3.5.1 Training Data

Predictive models are developed based on training data. If the data contains errors, it may affect the model's output (Barocas & Selbst, 2016). Specifically, if an error changes an important characteristic of a child, it can lead to inaccuracies in the model's risk assessment for that child (Glaberson, 2019). For instance, the Illinois Department of Children and Family Services conducted an experiment with ERSFT that initially showed success. However, this initiative was terminated after the unfortunate deaths of two young children within a month of each other. Subsequent investigations by Illinois officials revealed that neither of these children had been identified as high-risk by the model. These flawed predictions stemmed from errors in the data provided to the model (Drake et al., 2020; Glaberson, 2019). While errors in data may be unavoidable in government administrative systems where human data entry is involved (Dare, 2013; Ministry of Social Development, 2014), this incident addresses the importance of linking children's records across different databases, particularly when using administrative data.

Another well-known determinant of bias is machine learning algorithms that learn from historical data containing prejudice (Barocas & Selbst, 2016; Edelman & Luca, 2014; Žliobaitė, 2017). For example, when algorithms are trained on datasets that reflect biased decisions made by social workers, such as instances of racial discrimination against specific groups, they have the potential to inadvertently replicate those biases in their outcomes (Chouldechova et al., 2018). As a result, the new model may exhibit discriminatory behavior towards the same group (Chouldechova et al., 2018). This risk highlights the potential for predictive modeling approaches to magnify bias in data and in the risk assessments made by child welfare professionals.

Biased outcomes may also occur in cases where the data is unbiased but is not well-sampled (Calders & Žliobaitė, 2013). For instance, over-representation of a certain group in the dataset can lead to disproportionate adverse effects on those groups (Barocas & Selbst, 2016). Examples are given in studies by Chouldechova et al. (2018), Drake et al. (2020), and Glaberson (2019). In particular, in NZ, the over-representation of Māori in the care and protection system has been stressed in the past decade (Keddell & Davie, 2018; Keddell & Hyslop, 2019) and according to Rea and Erasmus (2017), this over-representation can be a reason for potential discriminatory effects on Māori.

Official statistics and recent quantitative exploration of disparities for Māori children in the NZ care and protection system acknowledge the existence of disparities between the involvement of Māori children and children of NZ European and other ethnic groups with the child welfare system (Oranga Tamariki, 2023a). A study conducted by Rouland et al. (2019) identified substantial ethnic differences in child maltreatment and child protection involvement in NZ. Specifically, Māori children faced significantly higher rates of involvement with child welfare agencies compared to other children. They were more likely to be reported, substantiated as victims, and placed in out-of-home care. Māori children exhibited involvement rates over twice those of Pacific Island children and more than three times those of NZ European children, across all tiers of the child welfare system. As a result, predictive risk models are at risk of intensifying the over-representation of Māori, and their use in decision-making could encourage a cycle of bias that, in turn, could lead to the disadvantage or discrimination of Māori (Rea & Erasmus, 2017). However, there is evidence suggesting that these sources of bias can be controlled more effectively by using a model that is more accurate and carefully developed (Dare, 2013; Ministry of Social Development, 2014).

3.5.2 Outcome Variable Used for Prediction

In predictive modeling, the outcome variable is crucial because it determines what the model is trying to predict. For the model to be effective, this outcome variable needs to be well-defined and accurately reflect the event or phenomenon of interest (Glaberson, 2019). Without a well-defined outcome variable that correlates closely with reality, predictive models will struggle to make accurate predictions. However, it's important to recognize that how the outcome variable is defined can have different impacts on different groups (Barocas & Selbst, 2016; Glaberson, 2019). For instance, considering the first version of AFST (Vaithianathan et al., 2017), the models were developed to predict two outcomes: child re-referral and child placement. The re-referral model aimed to predict whether a child would be referred again, while the placement model aimed to predict whether a child would be placed in out-of-home care. However, it was discovered that the re-referral model tended to give high scores to children involved in custody disputes or other circumstances where there were frequent calls about the same issue. Eventually, it did not appear to strongly correlate with outcome measures such as serious abuse and neglect. Additionally, there were concerns that racial biases or prejudices were embedded in the initial referrals, further complicating the accuracy and fairness of the predictive models. This fact suggested that a model predicting future referrals tended to over represent African-American children compared to white children. In this line, the second version (AFST II) was limited to predicting significant safety issues that usually result in a court order for placement.

Considering the NZ population, findings from the study by (Rea & Erasmus, 2017) showed that the model would refer more Māori children and young people to the site than under current practice. Therefore, the over-representation of these groups in the child welfare system must be considered, and proactive measures are required to define the outcome variable in a way that does not reflect an over-representation inconsistent with their actual share of risk.

3.5.3 Variable Selection

Variable selection is the process by which developers decide which variables to use in their analysis. These decisions may have significant consequences for the treatment of certain groups if the factors that are a better representation of these groups are poorly represented in the set of selected variables (Barocas & Selbst, 2016). This fact would lead to systematically less accurate classifications or predictions about those members. A possible explanation according to Glaberson (2019) is that the absence of that variable, regardless of its significance for the outcome, will cause other correlated

variables to take on a weight that does not explain the influence of the missing variable while obscuring its significance. Additionally, the necessary information to achieve accurate outcomes might exist at a level of detail that the selected features fail to achieve or uncover critical points of disparity (Barocas & Selbst, 2016).

The use of *race* or *ethnicity* as a predictor variable has been controversial due to concerns regarding racial stereotypes and the allocation of interventions according to race. Particularly, NZ studies have avoided the inclusion of *ethnicity* as a predictor variable but tested the final model against different ethnic groups involving such factor (Johnston, 2021; Rea & Erasmus, 2017; Vaithianathan et al., 2013; Wilson et al., 2015). However, the concept of *fairness through unawareness* is flawed because other variables may be correlated with the sensitive one (Lum & Johndrow, 2016). This implies that even if a developer avoids using the sensitive variable directly, its effect may still reflect through other related variables. This phenomenon, referred to as the redlining effect in the literature often leads to indirect discrimination (Zliobaite, 2015). The use of zip codes as a predictor is one example. It is known that some suburbs may have a higher proportion of specific racial groups than others, thus providing a strong correlation between race and zip code. In addition, due to institutionalized racial bias (e.g., criminal justice history), it may be possible to find other highly correlated predictors with race that suggest race is still a significant factor (Vaithianathan et al., 2017). Zliobaite (2017) advises that the inclusion of protected characteristics such as race are needed in the model development process to actively ensure the resulting model is fair. Section 3.6 provides a review of the approaches which can be used for to develop a fair predictive model while including *ethnicity* as a predictor.

3.6 Fairness-aware Machine Learning

Fairness-aware machine learning approaches usually fall into one of three categories: pre-processing, in-processing, or post-processing. The classification of these approaches depends on where the emphasis lies in terms of addressing discrimination within the data or model.

With the pre-processing approach, developers adjust the training data so that unexplained disparities between protected and unprotected groups are minimized before developing the model using standard machine learning algorithms. A simple approach is to remove the sensitive variable and all variables correlated with it from the learning process (Calders et al., 2009). This method might lead to better results in terms of fairness but achieving fairness is not guaranteed. One downside of this method is the loss of information about the outcome in the available data. In the context of fairness in machine learning models, it's crucial to recognize that removing correlated variables might

not address all sources of unfairness. This is because unfairness can stem from complex interactions between variables that may not be fully captured by linear correlations alone. As mentioned by Radovanović and Ivić (2021), there's a risk that these intricate interactions contributing to unfairness may go unnoticed. In other pre-processing approaches, developers choose to modify the target variable (Fish et al., 2015; Kamiran & Calders, 2009; Luong et al., 2011; Mancuhan & Clifton, 2014; Mancuhan & Clifton, 2012), modify the input data (Calders et al., 2009; Feldman et al., 2015; Johndrow & Lum, 2019; Kamiran & Calders, 2010; Lum & Johndrow, 2016) or modify both the target variable and input data (Hajian & Domingo-Ferrer, 2012). While modifying the data appears to be a potential solution based on these studies, it raises concerns regarding data accuracy, a fundamental requirement emphasized by regulations such as the General Data Protection Regulation (GDPR) (Goodman and Flaxman (2017)). Consequently, employing such approaches, particularly in sensitive contexts like child welfare settings, warrants careful consideration due to potential ethical and legal implications.

Methods that fall under the category of in-processing involve modifying the learning algorithm to maximize both predictive accuracy and fairness. For instance, by modifying the splitting criteria in decision tree learning (Kamiran et al., 2010; Kamishima et al., 2012). Most recent in-processing methods add regularizers to the objective function to control for fairness or enforce fairness constraints during the model learning process to turn it into an optimization problem. More details can be found in (Agarwal et al., 2018; Berk et al., 2021; Calders et al., 2013; Corbett-Davies et al., 2009; Hu & Chen, 2020; K. D. Johnson et al., 2016; Nabi & Shpitser, 2018; Radovanović & Ivić, 2021; Radovanović et al., 2020; Zafar et al., 2019; Zafar et al., 2017).

In post-processing approaches, the standard model is generated and then adjusted to comply with non-discrimination constraints (Calders & Verwer, 2010; Hajian & Domingo-Ferrer, 2012; Hardt et al., 2016; Kamiran et al., 2010; Wu & Wu, 2016). One common practice is to change the labels of some leaves in a decision tree (Calders & Verwer, 2010; Kamiran et al., 2010). Some other approaches involve removing selected rules from the set of discovered decision rules (Hajian & Domingo-Ferrer, 2012) or in general adjusting predictions to be as fair as possible (Chzhen et al., 2019; Hardt et al., 2016; Wu & Wu, 2016).

Although some methods have already been proposed for each of the above approaches, discrimination prevention remains a relatively unexplored area of research. Typically, each solution is customized for a specific setting and discrimination situation and barely generalizes to other types of variables, or other grounds of discrimination.

The examination of algorithmic fairness in predictive analytic tools within the public sector has gained significant attention, particularly in areas like criminal recidivism and academic admissions. However, the predictive analytics tools of child welfare jurisdictions have received considerably less attention. This is partly because comparatively few such instruments exist and because even fewer have been scrutinized through the lens of algorithmic fairness.

Non-discrimination regulations often specify the sensitive features or the groups of people who must be protected against discrimination in a particular setting. This point raises the question of how a predictive model can be deemed fair or unfair for these groups. The answer to this question depends on the definition of fairness that is sought to be attained. The complexities surrounding fairness in predictive models necessitate a deeper understanding of the various fairness definitions that can guide these assessments. In the next section, these definitions are reviewed in detail to identify which are most relevant for the child welfare context.

3.7 Definitions of Fairness

Various fairness concepts have been proposed in the last decade, yet there is no clear consensus on their application in specific situations. To determine the most appropriate category of algorithmic fairness definitions for child welfare settings, three primary categories were initially examined: *individual-level*, *group-level*, and *causal reasoning-based* definitions, following the structure and content presented by Verma and Rubin (2018).

Within individual-level definitions of fairness, the main idea is that observations close to one another in feature space should also be close to one another in prediction space (Purdy & Glass, 2023). Although this concept is appealing in theory, it is challenging to implement in practice (Berk et al., 2021). One key issue is that two feature vectors may appear close due to bias rather than genuine underlying similarity (Purdy & Glass, 2023). For example, two children reported with care and protection concerns might have very similar allegation histories, but these similarities could be the result of reporting bias within the community rather than any true commonality between the children and their families. In contrast, group-level definitions categorize individuals based on one or more protected attributes, such as *ethnicity*, *gender*, or *age*. A specific metric, typically related to predictive performance, is then calculated and compared across these groups. If this metric is consistent across all groups, the algorithm is considered fair (Verma & Rubin, 2018). Although this approach may overlook potential unfairness related to non-protected attributes and similar observations may still receive significantly different predictions, as noted by (Purdy & Glass, 2023), group-level definitions appear to be

the most practical for this context. This approach is particularly useful for identifying and addressing systematic disparities that might not be apparent at the individual level. By focusing on group-level comparisons, it is possible to ensure that the model's performance does not disproportionately benefit or disadvantage any particular demographic group, making this method practical and effective for applications where social equity is a priority (Purdy & Glass, 2023).

Causal reasoning-based definitions of fairness focus on understanding and addressing the causal relationships between variables in a model, particularly how protected attributes influence outcomes (Verma & Rubin, 2018). Despite their theoretical appeal, causal reasoning-based definitions were dismissed due to the impracticality of constructing and testing a causal graph with the extensive set of features involved in these models. Definitions like *Counterfactual Fairness* and *No Proxy Discrimination* offer theoretically ideal approaches, but require the construction of a causal graph that accurately represents the relationships between the outcome and the many features used in the algorithm (Verma & Rubin, 2018). Given the challenge of creating such a graph, especially with a large number of features, and the difficulty in testing a classifier against causal definitions of fairness, this approach was considered impractical for this context. Taking these factors into account, group-level definitions were determined to be the most viable option. While this category has its limitations, it provides valuable and practical tools for assessing fairness within the specific constraints of the use case. Acknowledging the shortcomings of this approach, it is nonetheless seen as a significant improvement over current standards in algorithmic fairness within the child welfare domain (Purdy & Glass, 2023).

Table 3.3 outlines group-level definitions relevant to the child welfare context, along with their mathematical definitions, while Table 3.4 presents a summary of notations used to formulate them. Additionally, Table 3.5 provides detailed narrative descriptions of these notions, offering a clearer understanding of their conceptualization and implications. The fairness definitions outlined in Table 3.3 are not exhaustive of the definitions proposed in the literature. These are likely to be appropriate for measuring discrimination and comparing the fairness of the models developed for potential implementation within child welfare systems. Additionally, these definitions can be utilized to improve the fairness of models by transforming them into constraints and enforcing them during the learning process of machine learning algorithms, as seen in (Barmomanesh & Miranda-Soberanis, 2023; Radovanović & Ivić, 2021; Radovanović et al., 2020).

The choice of fairness notion to prioritize, however, depends on the dataset available and the fairness beliefs and principles of relevant stakeholders. For instance, researchers responsible for the

development of the AFST and DCDA evaluated fairness based on calibration and accuracy equity (Chouldechova et al., 2018). Accuracy equity is achieved when both protected and unprotected groups share equal prediction accuracy, meaning that the AUC is equal across groups (Berk et al., 2021). Calibration, on the other hand, is met when, for any predicted probability score (r), individuals in both protected and unprotected groups possess an equal probability of truly belonging to the positive class. This definition closely resembles statistical parity but extends it by considering the fraction of correct positive predictions for any predicted probability or risk score r (Chouldechova & G'Sell, 2017). The well-calibration definition builds upon the calibration concept, asserting that, for any predicted probability score r, individuals in both protected and unprotected groups should not only share an equal probability of truly belonging to the positive class but also ensure that this probability equals r (Kleinberg et al., 2016).

With respect to beliefs and principles of relevant stakeholders, Cheng et al. (2021) conducted a study to understand stakeholders' perspective of algorithmic fairness in the child welfare context. The results of this study showed that Equalized Odds is the more preferred group fairness approach for child maltreatment risk assessment; however, there is no agreement about individual fairness comparisons. Equalized Odds combines predictive equality and equal opportunity definitions and is satisfied if protected and non-protected groups have equal TPR and equal FPR (Hardt et al., 2016).

No research similar to that conducted by Cheng et al. (2021) appears to have been undertaken in NZ. Previous studies have revealed noteworthy findings regarding predictive risk models' classification patterns for different ethnic groups. For instance, Wilson et al. (2015) found that the developed predictive risk model classified Māori children as high risk at a higher rate relative to their share of known maltreatment. Additionally, the model in (Rea & Erasmus, 2017) referred fewer NZ European and Pacific Island children for further investigation compared to current practice. These findings suggest a potential for predictive risk models to exhibit disparate referral rates among ethnic groups, characterized by over-referral for Māori children (FPR) and under-referral for children from other ethnic groups (FNR). The notion of fairness is confined to the evaluation results of the developed predictive risk model in terms of accuracy and fairness, as discussed further in Chapter 5 on methodology.

Table 3.3: Considered group-level definitions of fairness in the child welfare context.

Definition	Mathematical Definition	Reference
Statistical Parity	$\mathbb{P}(\hat{y}\|s=0) = \mathbb{P}(\hat{y}\|s=1)$	(Dwork et al., 2012)
Conditional Statistical Parity	$\mathbb{P}(\hat{y}=1\|s=0, X_L) = \mathbb{P}(\hat{y}=1\|s=1, X_L)$	(Corbett-Davies et al., 2009)
Predictive Parity	$\mathbb{P}(y=1\|\hat{y}=1, s=0) = \mathbb{P}(y=1\|\hat{y}=1, s=1)$	(Chouldechova & G'Sell, 2017)
Predictive Equality	$\mathbb{P}(\hat{y}=1\|y=0, s=0) = \mathbb{P}(\hat{y}=1\|y=0, s=1)$	(Corbett-Davies et al., 2009)
Equal Opportunity	$\mathbb{P}(\hat{y}=0\|y=1, s=0) = \mathbb{P}(\hat{y}=0\|y=1, s=1)$ *or* $\mathbb{P}(\hat{y}=1\|y=1, s=0) = \mathbb{P}(\hat{y}=1\|y=1, s=1)$	(Hardt et al., 2016)
Equalized Odds	$\mathbb{P}(\hat{y}=1\|y=1, s=0) = \mathbb{P}(\hat{y}=1\|y=1, s=1)$ *and* $\mathbb{P}(\hat{y}=1\|y=0, s=0) = \mathbb{P}(\hat{y}=1\|y=0, s=1)$	(Hardt et al., 2016)
Accuracy Equity	$\mathbb{P}(\hat{y}=y, s=0) = \mathbb{P}(\hat{y}=y, s=1)$	(Berk et al., 2021)
Calibration	$\mathbb{P}(y=1\|r, s=0) = \mathbb{P}(y=1\|r, s=1)$	(Chouldechova & G'Sell, 2017)
Well-Calibration	$\mathbb{P}(y=1\|r, s=0) = \mathbb{P}(y=1\|r, s=1) = r$	(Kleinberg et al., 2016)

Table 3.4: Summary of notations employed in Table 3.3.

Symbols	Description
s	Sensitive variable also referred to as protected variable for which fairness should be recognised, with $s=1$ representing the protected group (disadvantaged group), and $s=0$ representing the unprotected group (privileged group).
X	Unprotected variables within the dataset .
X_L	A subset of variables which represent legitimate risk factors. Legitimate risk factors refer to characteristics or features that are relevant to decision-making processes and may legitimately affect outcomes.
y	The actual outcome as presented in the dataset, with $y=1$ representing the outcome of interest and $y=0$ representing the lack of such feature.
r	Predicted probability or risk score for a certain outcome i, denoted by $\mathbb{P}(y=i\|s,x)=r$.
$\hat{y}$	The predicted value of the label variable, where $\hat{y}$ is normally obtained from r. For example $\hat{y}=1$ when r is greater than a specified threshold.

Table 3.5: The main concepts of the fairness definitions outlined in Table 3.3.

Notion of Fairness	Concept
Statistical Parity	Satisfied if the decision made based on the predicted value is independent of a sensitive attribute (e.g., *race*, *gender*, *ethnicity*). It is also referred to as *Group Fairness* and *Equal Acceptance Rate*.
Conditional Statistical Parity	Satisfied if the decision made for individuals sharing same legitimate risk factors is independent of a specified sensitive variable.
Predictive Parity	Satisfied if both protected and unprotected group have equal positive predictive value (PPV).
Predictive Equality	Satisfied if both protected and unprotected group have equal false positive rate (FPR). It is also known as *False Positive Error Rate Balance*.
Equal opportunity	Satisfied if both protected and unprotected group have equal false negative rate (FNR). Mathematically a model with equal FNR will also have equal true positive rate (TPR). It's also known as *False Negative Error Rate Balance*.
Equalized Odds	It combines *Predictive Equality* and *Equal Opportunity* definitions and is satisfied if protected and unprotected group have equal TPR and equal FPR.
Accuracy Equity	Satisfied if both protected and unprotected groups have equal prediction accuracy. In other words, accuracy equity is satisfied if AUC is equal across groups.
Calibration	Satisfied if for any predicted probability score r, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class. This definition is similar to *Statistical Parity* except that it considers the fraction of correct positive predictions for any value of r.
Well-Calibration	It extends the calibration definition stating that, for any predicted probability score r, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to r.

Note: A preface in this thesis, titled *List of Formulae for Classification Metrics*, has been created, providing a detailed outline of the mathematical formulations used for various classification metrics in machine learning.

Chapter 4

Data Sources and Software

4.1 Overview of the Integrated Data Infrastructure (IDI)

This thesis utilised data from the IDI, managed by Stats NZ. The IDI is a large research database containing microdata about individuals and households from a range of NZ government agencies, Stats NZ surveys, and non-government organizations. Stats NZ collects data from different sources and links them together to create the IDI (Stats NZ, 2022c). Figure 4.1 displays the types of data available in the IDI.

Stats NZ grants access to the data on a case-by-case basis, provided that the research project meets specific access criteria and demonstrably contributes to the welfare and well-being of the general public and society as a whole. The project must meet several criteria to be considered. It should serve a statistical purpose and benefit the public good. The research must be conducted by a credible team with support from their organization, using the required data for the project, provided it is available in the IDI. Stats NZ must be able to enforce an agreement, and the research findings must be released publicly.

Additionally, researchers need to have the necessary skills to work with the data, such as intermediate SQL coding skills, with most researchers using SQL, SAS, R, or Stata. The Stats NZ application review process involves assessment by subject matter, legal, and methodology teams, possibly with input from data suppliers. Referees will be contacted, and additional information may be requested. Final approval is granted by the Government Statistician or a delegate. Following approval, confidentiality training is undertaken before access is granted.

Every dataset has had personal identifiers removed or encrypted; hence, the risk of disclosure of

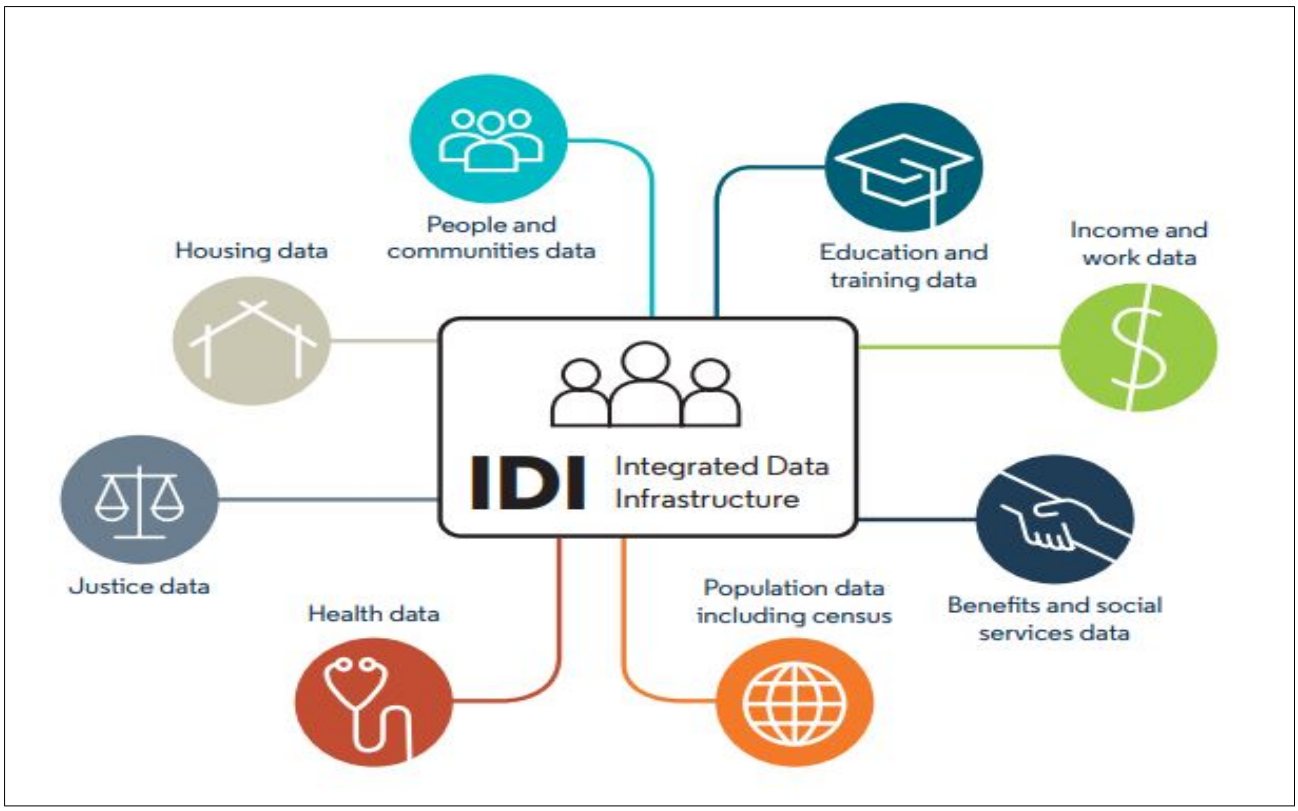

Figure 4.1: Data in Stats NZ Integrated Data Infrastructure (IDI).

Note: This figure is obtained from the Stats NZ website (Data in the IDI).

personal information is reduced to a minimum. There is an encrypted identifier for each identity in the IDI that is common across all datasets, allowing researchers to link variables from multiple sources to gain system-wide insights (Stats NZ, 2023). Data is subject to the *Data and Statistics Act 2022* and the *Privacy Act 1993* to protect the identities of people involved in the data they hold. Further, to protect the privacy of individuals, data can only be accessed through a secure virtual environment, and only in research facilities approved by Stats NZ. Additionally, Stats NZ conducts thorough checks on all outputs before their release from the Data Lab to ensure that no information identifies individuals (Stats NZ, 2022a). In this study, all findings and statistical analyses have been reviewed by qualified Stats NZ personnel.

4.2 Application Process for IDI Access

The application process commenced with a pre-application meeting with Stats NZ staff, where valuable advice was provided on the feasibility of the research and the most relevant data for the proposal. During this meeting, the staff discussed potential concerns to address in the application and offered guidance on demonstrating responsiveness to the *Te Tiriti o Waitangi* (The Treaty of Waitangi) and human rights obligations. In order to submit the application to Stats NZ, approval was required from

both the AUT Ethics Committee and the Ministry of Health's Health and Disability Ethics Committees. Following approval, participation in confidentiality training was mandatory, and a test was successfully completed before access was granted.

4.3 Datasets

Due to the nature of this research, a combination of Census data and administrative data was used. The administrative data utilized in this thesis are sourced from government agencies and are accessible through the Stats NZ IDI database. Every dataset comprises a collection of tables, each containing available records. For a detailed overview of the datasets utilized, their sources, and the tables available within each dataset, see Table 4.1.

4.3.1 Child, Youth and Family (CYF) Data

The CYF data is operational and is used by Oranga Tamariki—Ministry for Children to manage business operations. The tables in this dataset contain administrative records, focusing on children or young persons identified as people at risk of harm, ill-treatment, abuse, neglect, or deprivation. It also covers cases where a child or young person has allegedly committed an offense, or concerns are raised about their behaviour or security of care. A child's or young person's data is captured in the dataset when a report of concern is lodged with entities such as Oranga Tamariki, the Police, other law enforcement agencies, or the Youth and Family Courts. The CYF data includes details about person identities, characteristics, relationships, chronological facts, and information about specific events in a person's life such as proceedings, actions, measures, trials, procedures, dealings, occasions, and happenings (Stats NZ, n.d.).

The dataset development process in this work relies significantly on this dataset, as the majority of predictor variables and the outcome variable are encoded using information extracted from its existing tables. For a comprehensive list of the features extracted based on records available in the CYF data, see Table A.1 of Appendix A. Starting with the intake table, which identifies notifications received by Oranga Tamariki for constructing the sample cohort, and extending to the abuse and investigation tables for encoding the outcome variable, the CYF data serves as the basis for our research dataset development process.

As reviewed in Section 3.3.1 and Section 3.3.2, research indicates that a history of prior interactions

with child protective services, including the number of past reports of concerns and instances of maltreatment, is consistently linked to the likelihood of experiencing maltreatment (Littell & Schuerman, 2002; Rittner, 2002; Swanston et al., 2002). This correlation extends to the parents' or caregivers' experiences during their own childhood, and other children within the family (Widom et al., 2009). Consequently, this dataset also takes on a central role in this work, being utilized as the primary dataset for developing the model. The fundamental aim of the base model is to identify high-risk children by utilizing information extracted from Oranga Tamariki's administrative data. Subsequently, the encoded variables are linked to related datasets to explore enhancements in model performance when Oranga Tamariki's data is integrated with information from other organizations.

4.3.2 Children's Action Plan Data

The Children's Action Plan (CAP) data focuses on demographic and referral details of vulnerable children under CAP's care. While the CAP has been renamed *Children's Teams*, it is still referred to as CAP within the IDI.

Children's Teams are designed to support children and families with complex needs who do not qualify for statutory care and protection services. Statutory care and protection services are services that are legally required and typically involve formal governmental intervention to ensure the safety and well-being of individuals, particularly vulnerable populations such as children. Involvement with children's teams is voluntary, and families must agree to engage (Liston-Lloyd & Sun, 2019). Children's Teams support families through an integrated approach, where agencies, non-government organisations, and community members share information, collaboratively assess children and their family needs, develop a single plan of action, and facilitate access to necessary services (Oranga Tamariki, 2023b). The children's teams address various situations, such as children in homes with family violence, those facing school-related challenges, those with social or behavioural problems, unaddressed health issues, and families dealing with social or economic issues, requiring strengthened parenting capacity. In cases where concerns and risk factors persist, statutory intervention may be necessary (Liston-Lloyd & Sun, 2019)

Child maltreatment is a complex public health issue, influenced by a variety of risk and protective factors across different levels of the socio-ecological model. These factors can accumulate and interact in ways that either increase or decrease the likelihood of maltreatment occurring (Masten & Wright, 1998). Research has identified several community-level protective factors against child maltreatment. Key among these are the availability of services for parents and families, such as health,

social, and educational resources. Additionally, neighborhood processes like social cohesion (mutual trust among neighbors) and social control (willingness to intervene) play critical roles (A. E. Austin et al., 2020). Information derived from CAP data has been instrumental in identifying whether a child or young person has been under the care of the Children's Teams and has received social and community supports. The encoded predictor variables derived from this information have the potential to serve as protective factors against future child maltreatment. Furthermore, CAP data have been utilized to create a set of predictor variables indicating the history of children's complex needs based on the information captured in its existing tables. For more details on the features extracted from this dataset, see Table A.2 of Appendix A.

4.3.3 Personal Details Data

The Personal Details dataset, sourced from Stats NZ, encompasses the most up-to-date demographic information for individuals born in NZ. It comprises key details such as *gender*, *birth dates*, *death dates*, and *ethnicity*. Additionally, the dataset provides Stats NZ encrypted IDs for the parents, enabling linking of children and young persons under analysis with their respective parents. This source of information has been utilized in this study to encode predictor variables reflecting demographic characteristics of children in our sample cohort as well as their parents and siblings. For the complete list of features extracted from Personal Details data, see Table A.3 in Appendix A.

4.3.4 Benefit Dynamics Data

The Benefit Dynamics Data (BDD) is a comprehensive repository within the IDI. It includes detailed information about individuals who have received primary working-age social welfare benefits over defined periods referred to as *spells*. Collected since 1990 and regularly updated, this dataset provides essential insights into the demographic characteristics of beneficiaries. It precisely tracks their changing benefit status and other circumstances, including the start and end dates of their social welfare benefit periods.

BDD is primarily sourced from the Social Welfare Information for Tomorrow Today (SWIFTT) system administered by the Ministry of Social Development. In addition to the *benefit type* received, BDD covers the social welfare benefit histories of the partners and dependent children of the primary benefit recipient.

As reviewed in Section 3.3.3, in the domain of child maltreatment research, poverty is commonly measured through indicators such as annual household income or participation in public benefit programs

(Conrad-Hiebner & Byram, 2020; Mulder et al., 2018; Yang, 2015). By capturing benefit histories of parents and family members, BDD enables the determination of whether they were receiving benefits at the time of care and protection notification or had a prior history of involvement with the social welfare system.

The information extracted from the BDD serves a pivotal role in encoding predictor variables that are influential in constructing the training data required for the development of the predictive risk model. These variables are important as they provide a holistic understanding of the socio-economic dynamics surrounding the child and their family. For the list of features extracted based on the records available in the BDD dataset, see Table A.4 of Appendix A.

4.3.5 Sentencing and Remand Data

The Sentencing and Remand data tables contain administrative data collected since 1998 and focus on convicted adult offenders subject to either community sentences or imprisonment, as well as individuals remanded until the completion of their trials. The information related to legal warrants, sentences, and releases is sourced from the Department of Corrections, the Courts, Ministry of Justice, and Police. Essentially, the Sentencing and Remand data provides a chronological sequence of events and periods, offering insights into how each offender is managed by the Department of Corrections and providing a comprehensive overview of events associated with each managed period.

Previous studies, as discussed in Section 3.3.2, have recognized the criminal record of parents or caregivers as a significant risk factor for child maltreatment (Mulder et al., 2018). The information extracted from the tables in the Sentencing and Remand data facilitated the encoding of predictor variables to indicate whether a child is currently residing or has previously resided in the same household as an adult recently released from prison due to offenses related to family violence and other offences. For the list of features extracted from the Sentencing and Remand data, see Table A.5 in Appendix A.

4.3.6 Programme for the Integration of Mental Health Data

The Programme for the Integration of Mental Health Data (PRIMHD) is a comprehensive National Mental Health and Addiction information collection managed by the NZ Ministry of Health. It consolidates data on service activities and outcomes for individuals across the country. The primary purpose of PRIMHD is to report on the types of services offered, identify service providers, and assess the outcomes achieved within the Mental Health and Addiction sector. Healthcare entities, including

district health boards and non-governmental organizations, submit their activity and outcomes data electronically to the Ministry of Health's National Collections and Reporting.

As reviewed in Section 3.3.2, recognized risk factors for child maltreatment include mental health problems and substance or alcohol use disorders among caregivers (A. E. Austin et al., 2020; Ayers et al., 2019; Huckle & Romeo, 2023). In this context, the PRIMHD dataset becomes a valuable resource for generating predictor variables that may play a crucial role in determining whether parents or caregivers have a history of addiction or mental health problems, thereby contributing to a potential improvement in the predictive power of models being developed (Rea & Erasmus, 2017). For the list of features extracted from the PRIMHD data, see Table A.6 in Appendix A.

4.3.7 2018 Census

The NZ Census of Population and Dwellings is the official count of the number of people and dwellings in NZ on a specific Census night (e.g., 6 March 2018). This includes everyone on NZ soil, on a vessel in NZ waters, or on a passage between NZ ports. However, NZ residents who are abroad on Census night are not included in the Census. The first official Census was conducted in 1851, and since 1877, a Census has been carried out roughly every five years (Stats NZ, 2022d). It contains detailed information on individual characteristics, such as the year and month of birth (the day of birth is redacted in the IDI), age, gender, *ethnicity*, education, country of birth, income, etc.

In this thesis, data from the 2018 Census was used as it represented the most recent and up-to-date information available at the time of analysis and was closely aligned with the 2017 and 2018 Sample Cohorts used in this study. The Census information was recognized as valuable for encoding predictor variables related to significant risk factors associated with child maltreatment. These variables specifically focus on the family dynamic and parental characteristics of children and people in our sample cohort. For the list of features extracted from the 2018 Census data, see Table A.7 in Appendix A.

Table 4.1: Summary of datasets utilized in this work.

Data	Source	Coverage	Tables
CYF Data	Oranga Tamariki	1991-Ongoing	-Intake, -Investigations, -Safety and Risk screens, -Abuse, -Partnered Response, -Placements, -FGC, -FWA, -Legal Status, -Contact Record -Identity Cluster, -Sociocultural Characteristics, -Event from/to Date, -Staff Phase Allocation, -Gateway Client Needs, -Care Continuum Segmentation Options
Children's Action Plan (CAP)	Oranga Tamariki	2013-Ongoing	-Demographics, -Referrals
Benefit Dynamics Data (BDD)	Ministry of Social Development	1990-Ongoing	-Benefit Spells -Partnership Status of Primary Beneficiaries -Children in Benefit Entitlement -Incapacity Benefits Reason Codes -Benefit District Office Details, -Main Benefit Daily Rate (First Tier Expenditure) - Supplementary Benefit Daily Rate (Second Tier Expenditure) -Ad hoc Payments to Beneficiaries (Third Tier Expenditure)
Sentencing and Remand Data	Department of Corrections	1998-Ongoing	-Offender Aliases, -Offender Programme Completions, -Offenders, -Major Management Periods, -Major Periods, -Periods.
Program for the Integration of Mental Health Data (PRIMHD)	Ministry of Health	2008-Ongoing	-PRIMHD -PRIMHD Supplementary Consumer Record - PRIMHD Mental Health Information National Collection - PRIMHD Diagnosis Information

Table 4.1 Continued from previous page

Data	Source	Coverage	Tables
2018 Census	Stats NZ	6 March 2018	-Unoccupied Dwelling, -Absentee, -Extended Family, -Family, -Household, -Individual, -Dwelling, -Area, -Address.
Personal Details Data	Stats NZ	-	-Demographics.

4.4 Limitations and Challenges of Administrative Data

The utilization of administrative data for predictive modeling via machine learning algorithms offers several advantages, such as large sample sizes, cost-effectiveness, and real-world applicability (AsadZadehZanjani, 2022). However, its application is accompanied by inherent challenges. One such challenge lies in the potential for biases and insufficient representativeness, where certain social groups may be underrepresented in the training data (Parycek et al., 2023). Administrative data are initially entered by humans, making them susceptible to errors, such as inaccuracies in names or addresses, or missing information, crucially compromising the overall quality of the data (Glaberson, 2019). Additionally, administrative data may not accurately represent the general population since it mainly includes individuals who interact with specific systems or services. As a result, the ability of predictive models developed using this data to apply to broader populations may be limited. It is essential, consequently, to identify systematic biases during the development of predictive models, particularly when they are employed in decision-making processes involving individuals or legal entities (Parycek et al., 2023). For example, large datasets available in the IDI are not necessarily representative of the population of interest. Police data, while extensive and routinely collected, is a notable example: although it serves administrative and operational purposes, its use for research and policy must be approached with caution, as generalising from such data to the broader population can be misleading, depending on the context. Moreover, the 2018 New Zealand Census had a notably poor response rate, particularly among Māori, rural communities, and economically disadvantaged populations, further highlighting issues of underrepresentation in key national datasets.

Another significant limitation of applying machine learning methods to administrative data is that

the data cannot be readily analyzed in their raw form. This challenge arises when substantial pre-processing of the data is necessary to make it suitable for machine learning algorithms. The pre-processing step is crucial, as it significantly impacts the quality of the models applied across various domains, including child welfare. As discussed by Parycek et al. (2023), a critical aspect of model quality lies in the ability to extract relevant information from raw data through proper transformations, thereby enhancing predictive capabilities. Analyzing complex data, such as child welfare records, involves constructing a flat representation from multi-dimensional and highly temporal databases. The pre-processing aims to reduce data complexity and extract or construct relevant attributes for subsequent analysis (Kuhn & Johnson, 2013).

The majority of administrative data records are available with timestamps, offering a chronological sequence of events that serves as a vital input for machine learning models. In the context of child welfare, for instance, administrative data capture the chronology of events related to children and families under care. This temporal aspect is crucial for machine learning models, as it provides valuable insights into patterns and trends over time. For example, the frequency and duration of interactions with child welfare services can indicate the severity of a family's situation or the effectiveness of interventions. Furthermore, the time elapsed between significant events, such as placements or reunifications, can offer important context for understanding the dynamics of child welfare cases (DePanfilis & Zuravin, 2001; Fluke et al., 1999). As a result, it is imperative to establish optimal pre-processing techniques for administrative data in the time domain to maximize the performance of machine learning algorithms (Taib & Messier, 2024). To overcome these challenges, it's essential to carefully assess data preparation techniques and validation methodologies. Moreover, strict measures are needed to address privacy and confidentiality concerns, ensuring compliance with ethical guidelines and responsible handling of sensitive data (Stats NZ, 2018a). Despite these limitations, administrative data remains a valuable asset for predictive modeling in research. However, achieving its full potential requires interdisciplinary collaboration and methodological precision to enhance its effectiveness and trustworthiness.

4.5 Statistical Analysis Software

In this study, R was used as the primary software for statistical computing. R is a powerful programming language designed for statistical analysis and data visualization (R Core Team, 2022). All code for data analysis and model development was written and executed within RStudio, an integrated development environment specifically created for R (Posit team, 2022).

One of R's key strengths lies in its extensive collection of packages. These packages, which contain code, data, and documentation, can be easily installed by users, typically through central repositories like CRAN (the Comprehensive R Archive Network). A complete list of the R packages used to develop the models in this study is provided in Table 4.2. The datasets outlined in Table 4.1, which are essential for our analysis, were retrieved from the Stats NZ SQL Server database using SQL Server Management Studio (SSMS). These datasets were then imported into R for further analysis. Individuals within these tables were linked using Stats NZ encrypted IDs to extract the necessary information and records for encoding the predictor variables in this study. For further details on the linkage process and the development of the research dataset, see Figure 5.1.

Table 4.2: packages being used to train the candidate models in this thesis.

Method	package	Reference	Comments
-Logistic Regression -Support Vector Machine	caret	(Kuhn, 2008)	Provides a comprehensive suite of functions for building predictive models, including tools for data splitting, pre-processing, feature selection, and model tuning using resampling methods like cross-validation.
Regularized Logistic regressions	glmnet	(Friedman et al., 2021)	Fits generalized linear and similar models via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values (on the log scale) for the regularization parameter lambda. It fits linear, logistic and multinomial, Poisson, and Cox regression models. The value of hyper-parameters is obtained using inner ten-fold cross validation.
Random Forest	randomForest	(Liaw & Wiener, 2002)	Implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression. It can also be used in unsupervised mode for assessing proximities among data points.
Extreme Gradient Boosting	Xgboost	(Chen et al., 2024)	Implements gradient boosting framework and commonly used for classification, regression, and ranking tasks in machine learning.
Constrained Logistic regression	nloptr	(S. G. Johnson, 2008)	Enables the use of various optimization algorithms for solving nonlinear programming problems.

Chapter 5

Methodology

5.1 Introduction

Logistic regression serves as the central methodology in this thesis. Our research adopts an in-processing approach to fairness-aware machine learning, drawing inspiration from previous works as proposed by Zafar et al. (2019) and further extended by Radovanović et al. (2020). The former introduced a flexible, constraint-based framework to enable the design of fair boundary-based classifiers. Their framework was tested on real datasets from the employment and finance sectors, demonstrating to a large extent its efficiency in limiting disparate impact (also known as statistical parity) with minimal loss in accuracy, offering at the same time greater flexibility compared to other methods (Radovanović et al., 2020).

Our fairness-aware machine learning approach integrates fairness constraints, derived from the fairness assessment of candidate models in this work, directly into the logistic regression learning phase. This methodology has been adopted to address unfairness in the NZ child welfare context for two reasons. First, logistic regression is preferred in this field due to its simplicity and interpretability, as evidenced in previous studies. Second, although constrained classification methods have not been previously employed in the child welfare context, they show potential for improving fairness in model predictions. Tables 3.1 and 3.2 provide an overview of studies conducted in the child welfare domain.

The statistical analysis in this thesis was conducted in two main stages. The primary objective at the initial stage was to address our first research question, which investigates the factors influencing the predictive accuracy of risk models intended for use by child protective services. We incorporated additional relevant features beyond those employed by Rea and Erasmus (2017), aiming to capture

significant information or interactions within the data.

The primary data source for this study was administrative datasets queried from the IDI, managed by Stats NZ (see Table 4.1). Substantial focus was placed on feature engineering, which involved incorporating domain knowledge to transform existing administrative records from various government agencies and NZ Census data into meaningful predictor variables. Additionally, ensuring data quality was a critical aspect of this process, involving the imputation of missing values, removal of outliers, and thorough data cleaning to maintain the integrity of the dataset.

Logistic regression models were then trained on these predictors to evaluate the potential enhancement in risk prediction resulting from the inclusion of new variables through data linkage. Both standard logistic regression and regularized methods such as Ridge, LASSO, and Elastic Net, were employed. These models are referred to as baseline models throughout this thesis.

Standard logistic regression was chosen as a baseline model due to its proven success in previous NZ studies and its application in the initial version of the Allegheny Family Screening Tool (AFST) (Vaithianathan et al., 2017). Regularized regression methods, such as LASSO, were considered based on their strong performance in recent U.S. studies, where LASSO was identified as the best model for overall prediction accuracy, particularly for high-risk groups. These studies demonstrated that LASSO achieved similar accuracy for both African-American children and children from other racial groups (Centre for Social Data Analytics, n.d. Putnam-Hornstein et al., 2022; Vaithianathan, Kulick, et al., 2019). Despite its success in the U.S., regularized logistic regression methods like LASSO have not been explored in NZ studies. Our goal in applying these methods is to evaluate whether the performance observed in the U.S. can be replicated in the NZ context, with a specific focus on comparing the accuracy for Māori children against children from other ethnic groups. In addition, more advanced predictive models, including random forest, support vector machine, and XGBoost, were assessed to examine the impact of different modeling approaches on risk predictions.

The second stage of our analysis aimed to address the remaining research questions, which were as follows:

(a) Investigating the factors contributing to the unfairness of algorithms used to predict the risk of adverse events, particularly those related to care and protection concerns in this thesis.

(b) Examining the feasibility of developing a predictive risk model that is both more accurate and fair.

(c) Exploring potential measures that child welfare authorities could implement to mitigate or prevent algorithms from exhibiting discriminatory behaviors.

This stage involved a detailed examination of the ethnic disparities in the predictions made by the candidate models. Through the application of specific pre-designed constraints in our fairness-aware machine learning approach, we addressed the sources of bias by enforcing fairness constraints during the model training phase. We then evaluated the impact of these constraints on both the accuracy and fairness of the predictions. Our goal here was to develop a framework that ensures equitable treatment to a large extent between Māori children and children from other ethnic groups, while also maintaining the predictive performance of the risk models.

The following sections provide a detailed examination of the methodologies employed, including data sources, pre-processing steps, model training procedures, and the evaluation metrics used to assess both predictive performance and fairness. By addressing these challenges, this study aims to contribute to the development of fairer and more effective predictive models that can be ethically integrated into child welfare decision-making practices.

5.2 Data

5.2.1 Outcome Variable

The outcome variable is derived from the NZ Ministry of Social Development's conceptualization of *estimated care and protection concern* (Rea & Erasmus, 2017). In practice, it is defined to predict, for each child notification, the probability that one or more of the events outlined in Table 5.1 will occur within two years of the initial notification. This probability could serve as a decision-assistance tool, guiding social workers in intake decision-making based on the risk thresholds established by the welfare system's screening process. The response outcome would then be either 'intake' or 'no intake.' Specifically, if the predicted probability exceeds a predefined threshold, the decision is likely to result in an 'intake,' meaning the case will be prioritized for further investigation or intervention. Conversely, if the probability falls below the threshold, the decision may result in 'no intake,' meaning the case will not be pursued further. Therefore, the outcome variable directly influences whether or not a child's case will be prioritized for intervention, based on the estimated level of risk.

While the method of estimating underlying care and protection-related concern based on events in Table 5.1 is a reasonable approximation, there are instances where it might not be accurate (A.

James et al., 2019; Rea & Erasmus, 2017). For example, the measure of *estimated care and protection concern* is not fully independent of NZ child protective services' practice. This measure is based on current practices, such as referrals to Family Group Conferences (FGC),[8] Family Whānau Agreements (FWA),[9] and substantiated findings of maltreatment[10] as well as on future intake decisions (Rea & Erasmus, 2017). Consequently, any changes in these practices may lead to misclassifications, especially considering the reforms in the NZ child protection system following the transition from CYF to Oranga Tamariki—Ministry for Children in April 2017 (New Zealand Family Violence Clearinghouse, 2017).

As an attempt to address these complexities, we extended our analysis beyond a two-year time frame in (A. James et al., 2019; Rea & Erasmus, 2017). We included subsequent events within the next three and four years for the population of children under analysis. This approach aimed to account for policy changes that may have potentially influenced decision-making practices by the current child welfare agency.

To formulate the outcome variables, thus, subsequent events within the next two, three, and four years were identified for our sample cohort. The outcome variables dichotomously coded, reflected children status as to whether they have experienced at least one of the events from Table 5.1 within the specified period of their initial notification.

The outcome variable (*estimated care and protection concern*) can be represented in mathematical form as:

$$\begin{aligned} &\text{If } y(Event\,1) = 1 \textit{ or } y(Event\,2) = 1 \textit{ or } y(Event\,3) = 1, \text{ then} \\ &\quad y(\textit{Estimated care and protection concern}) = 1, \text{ else} \\ &\quad y(\textit{Estimated care and protection concern}) = 0. \end{aligned} \tag{5.1}$$

To identify the outcome variable that most effectively predicts care and protection concern, separate LASSO logistic regression models were employed, utilizing a comprehensive set of predictor variables initially encoded based on recognized child maltreatment risk factors (see Section 5.2.3). The LASSO logistic regression was chosen as the primary methodology at this stage due to its efficacy in managing multicollinearity, performing feature selection and improving generalization (Akalin, 2020; Tibshirani, 1996). More details on LASSO logistic regression can be found in Section 5.3.2.2.

[8] **Family Group Conference (FGC)** is a formal meeting where child protective services and the extended family of children work together to develop a plan to address any care and protection concerns, needs, or well-being issues relating to the child.

[9] **Family/Whānau Agreement (FWA)** is an agreement between the family and child protective services that addresses concerns for their children and outlines how the NZ child welfare system will support them.

[10] **Substantiated findings of maltreatment** means that the social worker has obtained clear and sufficient evidence to determine that maltreatment has occurred.

Table 5.1: Care and protection-related events used to define the outcome variable (*estimated care and protection concern*).

Event 1	A substantiated finding of maltreatment including physical, sexual, emotional abuse, or neglect.
Event 2	A site social worker recommending a FGC or developing a FWA.
Event 3	The child or young person being the subject to a further notification which will be assessed as an intake.

In contrast to some earlier models that focused solely on substantiated findings of maltreatment to reflect high estimated concern (Vaithianathan, 2012; Vaithianathan et al., 2018; Wilson et al., 2015), our study adopts a broader definition. We include subsequent referrals and interventions in our definition of *estimated care and protection concern*, intending to capture the actions taken by child protective services to prevent maltreatment. Table B.1 in Appendix B presents the distribution of this outcome, along with related events, within the specified time frames for the sample cohorts analyzed in this thesis.

5.2.2 Sample Cohort and Sample Construction

This work focuses on care and protection notifications received between April 1, 2017 and March 31, 2018 under Section 15 of the *Oranga Tamariki Act 1989*. Two key considerations guided our selection of this cohort. First, the transition from CYF to Oranga Tamariki — Ministry for Children in April 2017, which signifies a reform in the NZ child welfare system. Selecting notifications received by child protective services post-reform allows us to consider policy changes that may have impacted the decision-making practices of the current child welfare agency. Secondly, utilizing the most recent data within the Stats NZ IDI database system during our analysis, i.e. the latest records from the Ministry for Children dated August 2022. Selecting a cohort of children reported in this period enabled us to reliably track these individuals up to a four-year period following the initial notification.

The process begins with the intake data from the CYF records, where historical notifications to the child welfare agency are documented. From the 110,058 notifications received between April 1, 2017 and March 31, 2018, we applied the inclusion criteria illustrated in Figure 5.1.

To be eligible for our analysis, notifications had to meet the following criteria: they must have been made within the specified period, under Section 15 of the *Oranga Tamariki Act*, and not have been submitted solely to provide additional information for a prior report of concern.

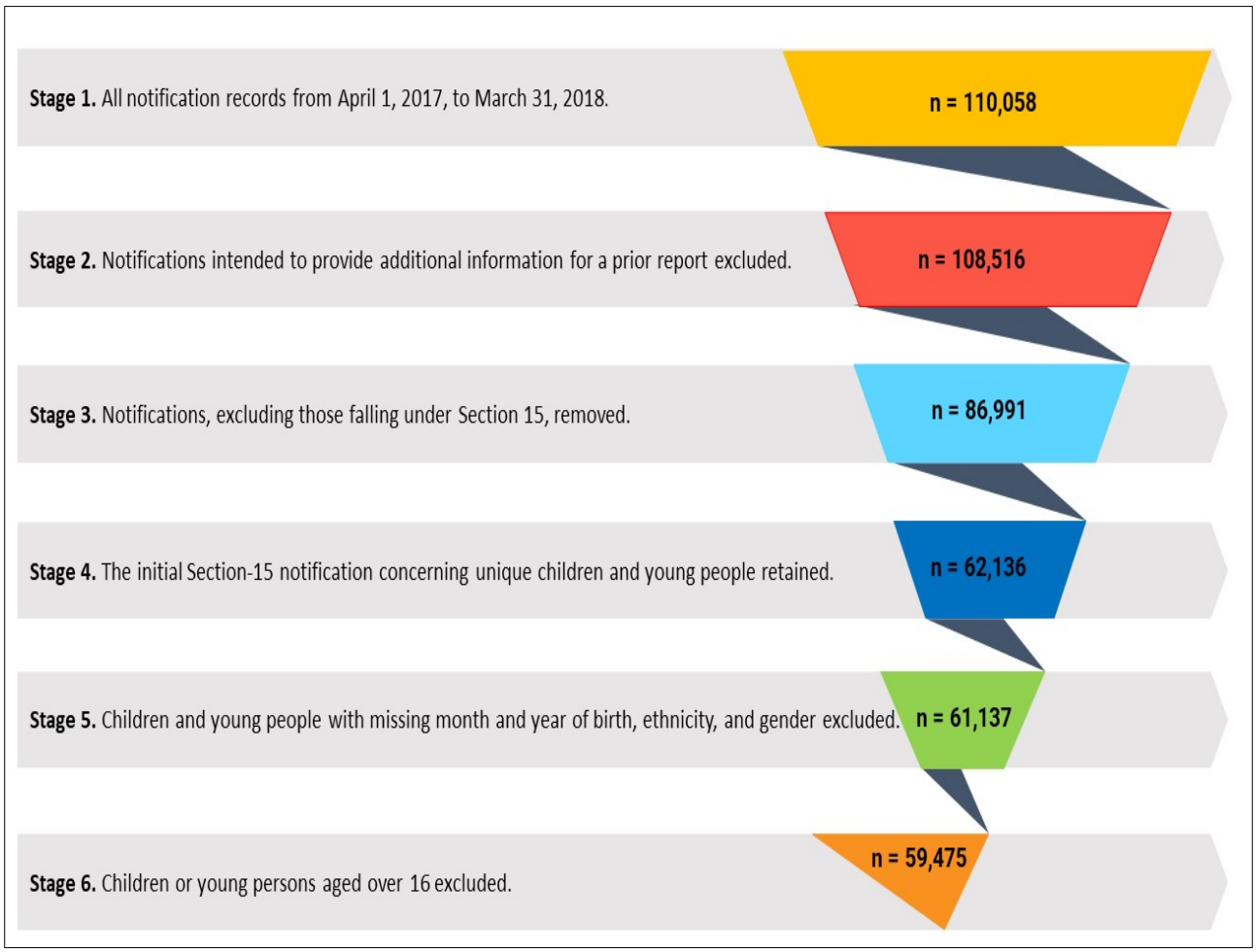

Figure 5.1: Summary of our final sample cohort construction process.

Recognizing the possibility of multiple notifications for the same child within this time frame, we prioritized the child's initial notification for inclusion in our sample. Additionally, notifications were excluded if key demographic information such as the child's *month and year of birth*, *ethnicity,* or *gender* was missing (n=999).

To ensure a consistent measurement of care and protection outcomes over the subsequent two, three, and four-year periods (Section 5.2.1), only children who were 16 years of age or younger at the time of their initial notification were included in the analysis (n=59,475). This criterion aligns with the focus of the NZ child welfare system, which predominantly captures records for individuals aged 18 or younger.

However, during the initial training of LASSO logistic regression models to find the optimal outcome variable, we tailored the sample cohort according to the age of the children. As illustrated in Figure 5.2, this means that when predicting the outcome (*estimated care and protection concern*) within a two-year time frame, we included children who were 16 years old or younger at the time of notification (n=59,475). For the three-year prediction window, children aged 15 years or younger were included

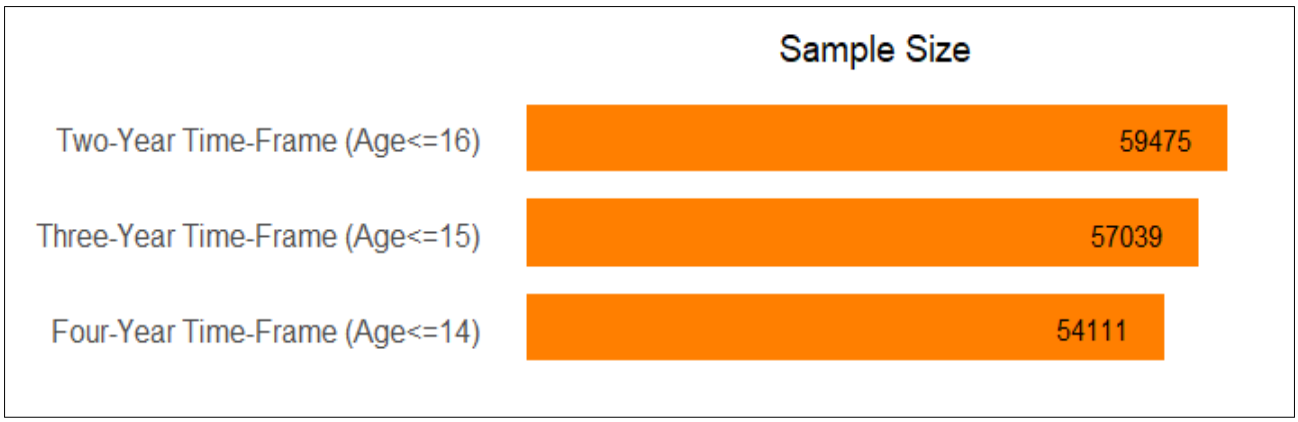

Figure 5.2: Sample size based on the outcome variable time frame and the children's age.

(n=57,039), and for the four-year prediction window, the analysis was restricted to children who were 14 years old or younger at the time of their initial notification (n=54,111).

5.2.3 Predictor Variables Encoding Process

The predictor variables for this study were mainly coded based on domain knowledge, and with the specific objective of enhancing the predictive power of the baseline logistic regression models. Figure 5.3 illustrates the development process of the research dataset, providing a visual representation of how various data sources were integrated to construct the final dataset used in this thesis. The diagram details the data linkage process, including the relevant tables and the Stats NZ encrypted IDs used to link the data.

The encoding process of predictor variables from raw administrative data and Census data was significantly influenced by the availability of records within the data outlined in Table 4.1, coupled with a comprehensive review of relevant literature and academic studies concerning the risk factors associated with adverse childhood outcomes, such as maltreatment (see Section 3.3). A thorough examination of the Stats NZ's IDI data dictionaries and classification files provided by government agencies facilitated the extraction of relevant information. These classification files contain descriptions of coded values within the administrative data of each respective agency.

As depicted in Figure 5.3, the research dataset development process began with the CYF data, specifically utilizing the Intake table, where historical records of notifications made to NZ child protective services are accessible. After identifying unique children associated with these notifications in our sample cohort, these were linked to their other records from the CYF data, Children's Action Plan data, and Personal Details data. This process allowed encoding of the relevant predictors to the history of interactions with child protective services, as well as the demographic characteristics of the

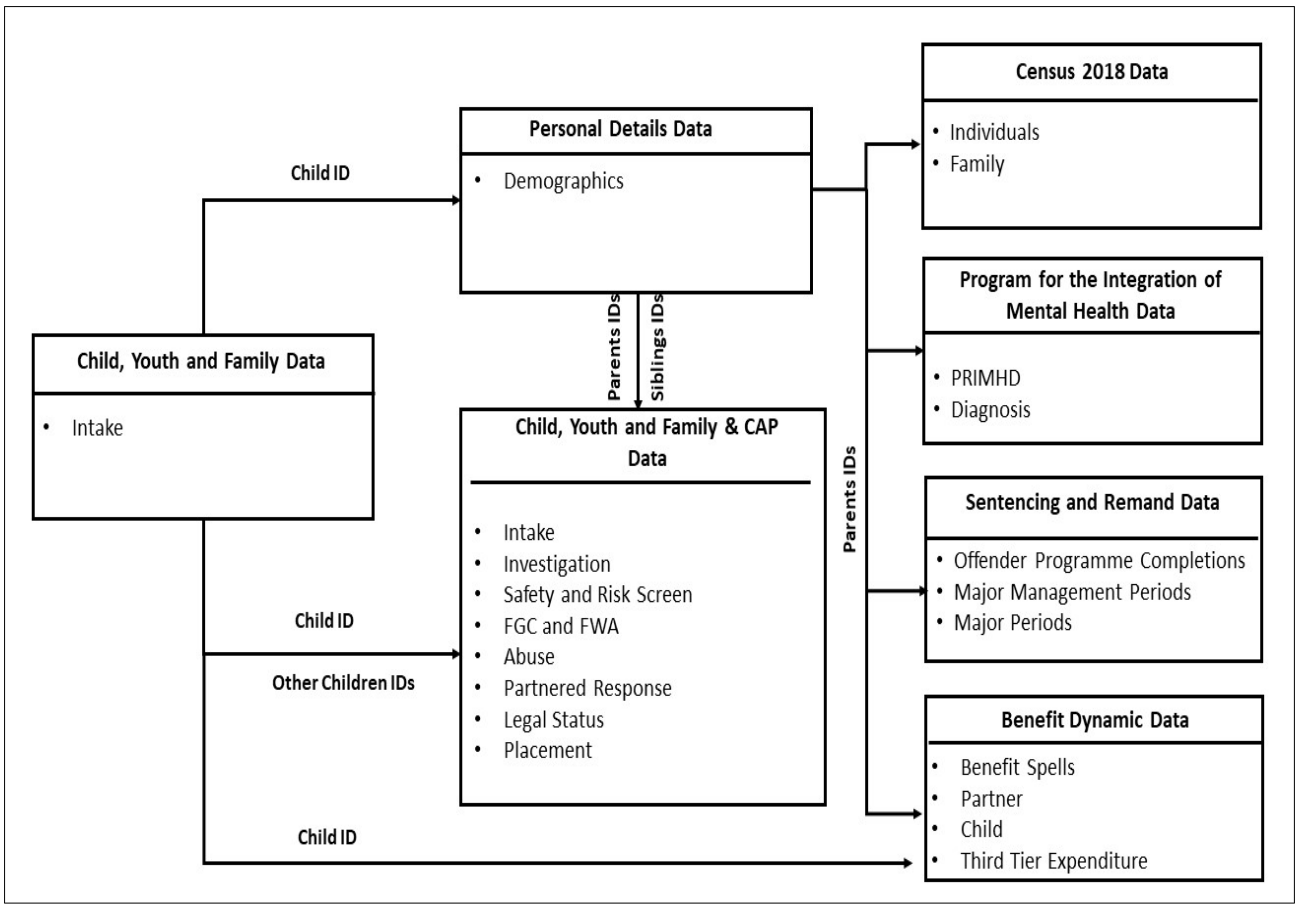

Figure 5.3: Stages of the research dataset development process through data linkage and integration.

children and their parents. Furthermore, the dataset was expanded by linking it to records available in Benefit Dynamics Data, Sentencing and Remand data, PRIMHD, and 2018 Census data to encode other predictor variables related to parents characteristics and family characteristics. Specifically, for each notification, we constructed data on the child and their family history, including information on the subject child, other children in the family, and parents.

For the complete list of initial features extracted from these data sources, categorized by their level and domain of risk factors, please refer to Tables A.1-A.7 in Appendix A. These features are systematically organized to provide a structured overview of the features included in the analysis. The *Feature Level* column of the tables indicates the specific entity to which each feature pertains, specifying whether it is related to the child, the family, or the parents. The *Domain* column categorizes each feature based on the type of information it represents, such as *demographic details*, *socioeconomic status*, *family structure*, *criminal history*, or other relevant risk factors. Crucially, the features listed in these tables are not the final predictor variables used to train our candidate predictive risk models. The final predictor variables were identified following an extensive pre-processing phase. This phase involved evaluating the predictive power of each feature and analyzing their interactions.

Furthermore, we introduced and integrated new relevant features, transformed existing ones, and incorporated domain knowledge to enrich the dataset. The final predictors are categorized into four groups: *child predictors*, *parents predictors*, *family predictors*, and *others*. Tables A.8-A.11 present the final list of predictors utilized for training the candidate predictive risk models in this thesis.

5.2.3.1 Ethnicity Classification and Prioritization

In the predictor variables encoding process, ethnicity was extracted from the Personal Details Data available in the IDI. Given the possibility for individuals to identify with multiple ethnicities, especially those overlapping between Māori and Pacific communities, we adopted the Oranga Tamariki approach to prioritize ethnicity (Oranga Tamariki, 2023a). Children and young people with unknown age ethnicity and gender are excluded from the analysis. Oranga Tamariki classifies ethnicity into four high-level categories:

- **Māori**: Children who identify Māori as one of their ethnicities.
- **Māori and Pacific**: Children who identify both Māori and Pacific as their ethnicities.
- **Pacific**: Children who identify as Pacific (but not Māori).
- **NZ European and Others**: Children who do not identify as Māori or Pacific. This category includes NZ European, European, Asian, Middle Eastern, Latin American, African, and other ethnicities.

However, since the primary focus of our analysis is on Māori, we adopted the ethnic classification method used for disparity analysis in (Oranga Tamariki, 2023a), as outlined below:

1. **Māori**: This group includes children from both the 'Māori' and 'Māori and Pacific' categories outlined above.
2. **Pacific**: This group includes children from the Pacific category.
3. **NZ European and Others**: This group includes children from the 'NZ European and Other' category, encompassing New Zealand European, European, Asian, Middle Eastern, Latin American, African, and other ethnicities.

5.3 Model Development Process

The process consisted of three stages, each with distinct objectives and methodologies focused on refining datasets and enhancing model performance.

The first stage focused on evaluating multiple data linkages formed by integrating information from various governmental sources. The goal was to identify the linkage that provided the highest predictive power (AUC) while minimizing ethnic disparities. Baseline logistic regression models were used to determine the most robust dataset for further modeling.

Building on the insights from the first stage, the second stage involved training and evaluating advanced models such as random forest, support vector machine, and XGBoost. This stage aimed to maximize predictive performance while addressing potential predictive bias, particularly concerning Māori children. A comprehensive evaluation framework was employed, incorporating traditional performance metrics along with fairness metrics such as calibration, accuracy equity, equalized odds, and statistical parity.

At both stages, two distinct logistic regression models were developed. The first model, termed "full logistic regression," included all predictor variables. The second model, called "refined logistic regression," was trained using only the predictors from the full model that met an entry criterion of p-values less than 0.1. It is crucial to emphasize that the models developed in this thesis are predictive, not causal. The decision to include predictor variables with a p-value threshold of 0.1 in the refined logistic regression model was guided by iterative experimentation with different cut-off levels to optimize model performance.

In the final stage, a fairness-aware machine learning approach was implemented to investigate whether the predictive bias identified against Māori children could be addressed, thereby developing models that are both technically sound and ethically responsible.

5.3.1 Data Linkage Formed for Modeling

In this work, five distinct datasets were considered to evaluate the potential enhancement in risk prediction through the integration of data available in Stats NZ's IDI. Each dataset was constructed using specific linkages between the data sources outlined in Table 4.1 and the process illustrated in Figure 5.3. These linkages are as follows:

- Link **1L**: Incorporates data from Child, Youth and Family, Children's Action Plan records, and

Personal Details data.

- Link **2L**: Utilizes data from Benefit Dynamics, alongside Child, Youth and Family, Children's Action Plan records, and Personal Details data.
- Link **3L**: Integrates Sentencing and Remand data with Benefit Dynamics, Child, Youth and Family, Children's Action Plan records, and Personal Details data.
- Link **4L**: Incorporates data from the Program for the Integration of Mental Health (PRIMHD) with Sentencing and Remand data, Benefit Dynamics, Child, Youth and Family, Children's Action Plan records, and Personal Details data.
- Link **5L**: Incorporates data from the 2018 Census with the Program for the Integration of Mental Health (PRIMHD), Sentencing and Remand data, Benefit Dynamics, Child, Youth and Family, Children's Action Plan records, and Personal Details data.

The baseline predictive models were trained and tested using predictor variables derived from each linkage (1L - 5L), where each linkage incorporated all predictors from the preceding one, thereby accounting for the cumulative effect of the added variables. With this approach, each subsequent stage not only introduced new predictors but also retained all previous ones, allowing for a comprehensive assessment of how the inclusion of new variables influenced the model's performance and providing a detailed understanding of their cumulative impact.

5.3.2 Model Training and Validation

After assembling required datasets, including both predictor variables and the outcome variable, and applying necessary pre-processing techniques, a standard machine learning process (see Figure 3.2) was used to train the models. The dataset was randomly partitioned into training and testing sets using a 70/30 split, with 70% of the records allocated for model estimation and 30% reserved for internal validation. For robust validation, final candidate models were tested on a separate cohort of children newly notified between April 1, 2018, and March 31, 2019 (external testing data). Due to data limitations and the required four-year follow-up period, evaluation was limited to this cohort.

The model's ability to generalize to new data was evaluated by cross-validation techniques, including k-fold cross-validation. K-fold cross-validation is a resampling technique used to evaluate the performance of a machine learning model on a limited data sample. It is commonly used to assess how well a model will generalize to an independent dataset (Kuhn & Johnson, 2013). Hyper-parameter tuning

was also critical for improving the model's AUC. In this work, we employed 5-fold cross-validation for hyperparameter tuning rather than the traditional 10-fold cross-validation. This choice was primarily influenced by the computational constraints posed by our large dataset, which comprises over 50,000 observations and numerous predictors. By reducing the number of folds from 10 to 5, we achieved a necessary balance between computational efficiency and model performance, ensuring that the tuning process was both feasible and reliable.

Baseline logistic regression models were initially developed and tested using predictor variables from each data linkage strategy in Section 5.3.1. The set of predictors that demonstrated the highest predictive power, as indicated by AUC, was then used to evaluate various candidate modeling approaches, including logistic regression (both full and refined), regularized logistic regression (Ridge, LASSO, and Elastic Net), random forest, support vector machine, and XGBoost. Each of these methods is discussed in detail in the subsequent subsections.

5.3.2.1 Logistic Regression

Logistic regression was selected as the baseline model for this work due to its well-established reliability and interpretability in binary classification tasks, and its proven effectiveness in predictive risk modeling studies within the NZ child welfare context (A. James et al., 2019; Rea & Erasmus, 2017).

Logistic regression is a supervised machine learning algorithm aiming to predict the probability of a binary outcome variable based on predictor variables. It is one of the simplest and most widely used algorithms for classification problems, primarily because its outputs are interpretable and easy to understand (G. James et al., 2013). The binary outcome is usually labelled as $y = 0$, $y = 1$. The model, in its sigmoid form, is given by:

$$\mathbb{P}(y = 1|X; w) = h_w(X) = \frac{1}{1 + e^{-w^T X}}, \quad so \tag{5.2}$$

$$\mathbb{P}(y = 0|X; w) = 1 - h_w(X) = \frac{e^{-w^T X}}{1 + e^{-w^T X}}, \tag{5.3}$$

$$Since \quad y \in \{0, 1\}, \quad \mathbb{P}(y|X; w) = h_w(X)^y (1 - h_w(X))^{(1-y)}, \tag{5.4}$$

Where **X** represents the predictor variables and **w** represents the weights associated with them. The logistic regression algorithm estimates these weights by minimizing the logistic loss function, also known as the negative log likelihood, as shown in Equation (5.5):

$$\text{Minimize } L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right), \tag{5.5}$$

Where y_i represents the actual observed value of the outcome variable for the i-th observation, and $\hat{y}_i = h_w(x_i)$ is the predicted probability estimated by the model.

This loss function, as shown in Equation (5.5), plays a crucial role in guiding the model's learning process, as it quantifies how well the predicted probabilities align with the actual outcomes. It comprises two main components: the contribution from correctly classified labels (true label) and the contribution from incorrectly classified labels (false label).

The true label contribution, represented by $y_i \log(\hat{y}_i)$, penalizes the model when it predicts a probability far from 1 for a positive outcome. Conversely, the false label contribution, represented by $(1 - y_i)\log(1 - \hat{y}_i)$, penalizes the model when it predicts a probability far from 0 for a negative outcome. Together, these components ensure that the model is penalized for incorrect predictions, thereby driving the optimization process to improve prediction accuracy.

By carefully balancing these penalties, the logistic loss function enables the model to iteratively adjust its weights to better discriminate between the two classes, ultimately enhancing its predictive performance (G. James et al., 2013). The logistic regression models in this thesis were trained using the *caret* package in R (Kuhn et al., 2020).

To implement our proposed in-processing methodology for fairness-aware machine learning using constrained logistic regression, the loss function in Equation (5.5) will be used as the objective function in an optimization problem. This objective function will be minimized while satisfying specific fairness constraints. The detailed theory and formulation of these constraints will be discussed further in Section 5.4.

5.3.2.2 Regularized Logistic Regression

In this thesis, regularized logistic regression methods were employed for their ability to handle high-dimensional data and mitigate the risk of overfitting (Akalin, 2020; Tibshirani, 1996). This is particularly important in child welfare research, where the relationships between numerous predictors and outcomes are often complex and interdependent.

Regularization systematically adjusts the learning algorithm to improve performance on unseen data,

including validation or entirely new datasets. It enhances traditional logistic regression by incorporating a penalty term into the loss function (see Equation (5.5)), which discourages the model from placing too much emphasis on predictors with limited contributions, effectively shrinking their coefficients towards zero (Akalin, 2020). This technique, known as "shrinkage" or "penalized regression", produces simpler models with smaller coefficients, promoting better generalization to new data. By controlling model complexity, regularization prevents overfitting, where the model might capture noise or random fluctuations in the training data rather than the true underlying patterns (Goodfellow et al., 2016).

While this increased generalizability often results in slightly higher training error, it is a worthwhile trade-off. Regularized models tend to be less precise on the training data but are more reliable when applied to new data, making them particularly suitable for real-world applications (Goodfellow et al., 2016). Consequently, regularization techniques like Ridge, LASSO, and Elastic Net are particularly effective in child welfare contexts, as they improve model stability and generalizability by penalizing excessive complexity and selectively retaining only the most relevant predictors (Akalin, 2020; G. James et al., 2013)

For training purposes, our regularized logistic regression models, utilize the *glmnet* package in R (Friedman et al., 2021). Regularization techniques such as Ridge, LASSO, and Elastic Net involve the hyperparameter (λ) that controls the complexity of the model (see Equations (5.6), (5.7), and (5.8)). This hyperparameter (λ) was optimized using 5-fold cross-validation to identify the model configuration that maximized the AUC. According to Probst et al. (2019), averaging the results from multiple repetitions of the entire cross-validation process yields more reliable results by reducing the variance of the estimation.

a) Ridge

Ridge regression, also known as L2 regularization, achieves shrinkage by adding a penalty term of the form :$\lambda \sum_{j=1}^{P} w_j^2$, to the logistic loss function. This penalty term is proportional to the sum of the squared coefficients. The updated loss function for Ridge logistic regression is:

$$L(y, \hat{y}) + \lambda \sum_{j=1}^{P} w_j^2, \tag{5.6}$$

Where $L(y, \hat{y})$ represents the logistic loss function (Equation (5.5)), j denotes the number of coefficients or weights in the model, and λ is the regularization parameter that controls the strength of the penalty.

In Ridge regression, as seen in Equation (5.6), the penalty term adds a cost to the loss function for large values of the coefficients. As λ increases, the model is forced to keep the coefficients small to minimize the overall loss function. This penalization effectively shrinks the coefficients towards zero, reducing their magnitude, and discouraging the model from giving too much weight to any single predictor.

When the penalty is added to the logistic loss function, the optimization process minimizes the overall loss by assigning smaller values to the coefficients. The λ parameter controls the strength of the penalty term where a higher λ value places more emphasis on the penalty, pushing more coefficients closer to zero (Akalin, 2020). This helps to prevent overfitting, especially when there are many correlated predictors in the model. However, the coefficients will not become exactly zero unless LASSO regularization is specifically applied.

b) LASSO

LASSO modeling algorithm which is derived from the term "Least Absolute Shrinkage and Selection", first introduced by Tibshirani (1996), has been widely used in various fields such as biological sciences, social sciences, and machine learning (Gustafsson et al., 2005; Muthukrishnan & Rohini, 2016; Vaithianathan, Dinh, et al., 2019; Vaithianathan, Kulick, et al., 2019).

The LASSO logistic regression method uses the absolute values of w_j as a penalty ($|w_j|$), as shown in Equation (5.7), instead of the squared values (w_j^2) used in Ridge regression (Equation 5.6). This penalty, known as the L1-penalty, allows LASSO to effectively select only the most important predictor variables by shrinking the coefficients of the least important variables to zero (Akalin, 2020). The loss function for LASSO is given by:

$$L(y, \hat{y}) + \lambda \sum_{j=1}^{P} |w_j|, \tag{5.7}$$

Both LASSO and Ridge regularized regression methods offer drawbacks and advantages (Friedman et al., 2010). To address these points, Zou and Hastie (2005) suggested a new method, called “Elastic Net”, which combines Ridge and LASSO regression.

c) Elastic Net

In this method some coefficients shrink toward zero (like Ridge regression) and some other coefficients shrink to exactly zero (like LASSO regression). The loss function for elastic net regression is given by:

$$L(y, \hat{y}) + \lambda \sum_{j=1}^{P} (\alpha w_j^2 + (1 - \alpha)|w_j|), \tag{5.8}$$

Where α controls the mix between L1 and L2 regularization, with α = 1 corresponding to LASSO and α = 0 corresponding to Ridge regression. Values of α between 0 and 1 result in a combination of both penalties, allowing the model to benefit from the sparsity of LASSO and the stability of Ridge. The choice of α can be guided by cross-validation to optimise predictive performance and model interpretability.

5.3.2.3 Support Vector Machines

Another machine learning algorithm considered in this thesis is the support vector machine (SVM). Originally developed by Cortes and Vapnik (1995), SVMs are supervised learning models designed for both classification and regression tasks, aiming to identify the optimal decision boundary that separates different classes within the data. These classes could represent different outcomes, such as "positive" and "negative" in a binary classification problem. Over time, SVMs have evolved significantly, becoming one of the most versatile and effective machine learning algorithms available (Vapnik, 2013).

The selection of support vector machine as a candidate machine learning algorithm in this thesis is motivated by several factors that align closely with the complexities of predictive modeling in the child welfare context. Firstly, SVMs are particularly adept at managing high-dimensional data (Bennett & Campbell, 2000; Cortes & Vapnik, 1995), a characteristic commonly found in child welfare datasets that encompass a wide range of socio-demographic and historical variables. Secondly, the robust regularization mechanisms inherent to SVMs effectively address the challenge of overfitting, thereby enhancing the model's ability to generalize to unseen cases (Burges, 1998; Schölkopf & Smola, 2002). This is especially critical in the child welfare domain, where the accuracy and reliability of predictions are paramount. Furthermore, SVMs possess the capability to capture complex non-linear relationships through the application of kernel functions (Schölkopf & Smola, 2002; Schölkopf et al., 1999), allowing them to model the intricate interactions between predictors that are often encountered in the child welfare field. These attributes collectively position SVM as a robust choice for developing predictive models that are both accurate and generalizable, making it an appropriate machine learning algorithm for supporting decision-making processes in this area.

The distinctive feature of SVMs lies in their foundation on the max-margin principle, which seeks to maximize the distance between the decision boundary (the hyperplane that best separates the classes) and the closest data points from each class, known as support vectors (Boser et al., 1992). This principle is key to how SVMs work: by maximizing the margin between classes, SVMs enhance

the model's ability to perform well on new, unseen data. A larger margin usually results in fewer errors when the model is applied to new data, making SVMs particularly effective for complex, high-dimensional datasets and cases where the separation between classes is not straightforward (Boser et al., 1992).

SVMs can efficiently perform a non-linear classification using the so-called kernel method or kernel trick that allows the model to produce extremely flexible decision boundaries (Schölkopf et al., 1999).

For an unknown sample u, the decision function for an SVM with a kernel ($K(\mathbf{x}_i, \mathbf{x})$) is expressed as:

$$D(\mathbf{u}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i \cdot \mathbf{u}), \tag{5.9}$$

where $D(\mathbf{u})$ is the decision function that determines the class label of the input vector $\mathbf{u}$. In this equation, y_i represents the class labels of the training data, α_i are the Lagrange multipliers, which are non-zero only for the support vectors, $\mathbf{x}_i$ denotes the support vectors (the critical data points from the training set), and β_0 is the bias term that adjusts the position of the decision boundary.

In the context of support vector machines this bias term β_0 (also known as the intercept), shifts the decision boundary away from the origin in the feature space, allowing the model to adjust the position of the separating hyperplane. Mathematically, the bias term ensures that the decision function $D(\mathbf{u})$ correctly aligns with the data, even when the optimal separating hyperplane does not pass through the origin (Schölkopf et al., 1999).

$K(\mathbf{x}_i \cdot \mathbf{u})$ in Equation (5.9) represents a kernel function applied to two vectors. In the linear case, this kernel function simplifies to the inner product $\mathbf{x}_i' \cdot \mathbf{u}$. However, SVMs can accommodate various other non-linear transformations through the kernel trick, enabling the model to capture complex relationships within the data. Some commonly used kernels include Polynomial Kernel, Radial Basis Function (RBF) kernel and Hyperbolic Tangent Kernel. The selection of kernel function parameters along with the cost value regulates the model's complexity and should be adjusted carefully to prevent overfitting of the training data (Kuhn & Johnson, 2013).

The SVM is implemented in the *caret* package for R (Kuhn et al., 2020). The RBF Kernel was identified as the most suitable method for our data. The RBF kernel is defined as:

$$\text{RBF Kernel: } K(\mathbf{x}, \mathbf{u}) = \exp(-\sigma \|\mathbf{x} - \mathbf{u}\|^2), \tag{5.10}$$

where **x** is a feature vector representing a data point in the input space and **u** a feature vector representing a support vector in the input space. σ is a parameter controlling the "width" of the RBF kernel which requires tuning.

The RBF kernel function in Equation (5.10) evaluates the similarity between data points based on their distance in the feature space, enabling the SVM to effectively capture complex patterns and non-linear relationships between features and classes (Kuhn & Johnson, 2013). This adaptability makes the RBF kernel particularly well-suited for our data.

During the training of the support vector machine (SVM) with a RBF kernel, we systematically optimized the σ and cost (C) parameters utilizing a 5-fold cross-validation resampling technique to identify the model with the highest testing AUC. The cost parameter C is not explicitly part of the kernel function or the decision function formula itself. It is a regularization parameter in the SVM optimization process that controls the trade-off between maximizing the margin (the distance between the decision boundary and the nearest data points of any class) and minimizing classification errors on the training data (Vapnik, 2013). According to Kuhn and Johnson (2013), the selection of the C value critically influences the model's propensity for underfitting or overfitting. For instance, a small C value yields a smoother decision surface, potentially leading to underfitting, while a large C value endeavors to classify all training instances accurately, leading to overfitting.

The σ parameter defines the influence of a single training example by controlling the width of the Gaussian function in RBF. It determines how quickly the influence of a support vector decreases as the distance from the input vector increases. A smaller σ value results in a narrower Gaussian (more localized influence), leading to a smoother decision boundary while a larger σ value results in a wider Gaussian (more global influence), potentially leading to a more complex decision boundary. Hence, employing a resampling technique to estimate these parameters effectively mitigates the risk of underfitting and overfitting, ensuring a robust model. We conducted a comprehensive grid search for both σ and C parameters to ascertain the optimal combination that maximizes performance in term of AUC.

Interestingly, despite the extensive tuning process, it was found that the default values of the C and σ parameters provided by the *caret* package yielded better performance on the testing set compared to the tuned values. This result underscores the robustness of the default settings in effectively balancing model complexity and generalization for this specific dataset. One possible explanation is that the default parameters are empirically chosen to perform reasonably well across a wide range

Table 5.2: Overview of the hyperparameters tuned for the SVM algorithm, along with the range of values considered during tuning and their typical default settings. Here, p represents the number of predictor variables in the dataset.

Hyperparameter	Description	Default Value	Considered Values
C	Cost parameter that balances the trade-off between maximizing the margin and minimizing classification errors.	1	$[10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}]$
σ	The σ parameter controls the width of the RBF kernel.	$\frac{1}{p}$	$[10^{-3}, 10^{-2}, 10^{-1}, 1, 10, \frac{1}{p}]$

of problems, and in this case, they aligned well with the structure and signal-to-noise ratio of the data. Additionally, it is possible that the tuning process led to slight overfitting on the training data, while the default settings retained better generalizability. Table 5.3 provides information on the tuned hyperparameters and the range of values considered during the tuning process.

5.3.2.4 Random Forest

Random forest is another machine learning algorithm considered in this thesis for its ability to handle the specific challenges posed by child welfare data, including high dimensionality, missing values, overfitting, and complex variable interactions (Kuhn & Johnson, 2013).

First introduced by Liaw and Wiener (2002), random forest is a powerful tree-based ensemble learning algorithm widely used in machine learning for predictive modelling. It includes a collection of decision trees, each trained on a random subset of the dataset drawn with replacement. This sampling process called *bootstrap aggregating or bagging* (Breiman, 1996). During the construction of the trees, the algorithm selects the best predictor from a randomly chosen subset of predictors at each split. This process introduces randomness into the model, ensuring that each tree in the forest captures different patterns in the data (Kuhn & Johnson, 2013). By considering only a subset of predictors at each split, the algorithm prevents any single predictor from dominating the model, which helps to reduce overfitting (Kuhn & Johnson, 2013). This diverse set of decision trees, each capturing a different aspect of the data, results in a model that is less sensitive to noise and specific data patterns in the training set. Consequently, the ensemble approach improves generalization by creating a more robust and stable model that performs well on unseen data, effectively capturing the underlying structure of the data rather than just memorizing the training examples (Zhang & Ma, 2012).

Moreover, this ensemble approach enhances prediction accuracy while maintaining interpretability, making random forest particularly effective for handling complex datasets with interactions between variables (Schönlau & Zou, 2020).

The random forest methodology is in the *randomForest* package for R (Liaw & Wiener, 2002). Although random forest models are generally known for their robust performance, tuning specific parameters can further enhance their accuracy in particular applications (Zhang & Ma, 2012). In this work, we tuned four parameters to find the model with the highest testing AUC. Table 5.3 provides information on the tuned hyperparameters and the range of values considered during the tuning process. We employed a 5-fold cross-validation approach to evaluate the performance of the algorithm with various hyperparameter values. A grid search strategy was used to explore all possible combinations within the specified discrete parameter spaces.

After assessing 900 different combinations of hyperparameter settings, the optimal set was determined to be:

- `ntrees`: 500
- `mtry`: $\sqrt{P} = 16$
- `maxnodes`: 10
- `nodesize`: 5

The remaining parameters were kept at their default values.

5.3.2.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable tree-boosting algorithm that has gained widespread popularity in machine learning due to its superior predictive accuracy and ability to handle complex data structures (Chen et al., 2024). It was considered in this study for its ability to effectively model complex, non-linear relationships and its robust performance in handling the intricate and often noisy nature of child welfare data, which includes a diverse range of features and a mix of categorical and numerical variables that complicate the predictive modeling process.

XGBoost employs an ensemble of decision trees that are trained sequentially, with each tree correcting the errors of its predecessors. This iterative learning process enables the model to capture intricate, non-linear relationships between features, making it particularly well-suited for predicting

Table 5.3: overview of the hyperparameters tuned for the random forest algorithm, along with the range of values considered during tuning and their typical default settings. Here, p represents the number of predictor variables in the dataset.

Hyperparameter	Description	Default values	Considered values
ntree	Number of trees in the forest	500	[100, 200, 300, 500, 750]
mtry	Number of variables randomly sampled as candidates at each split	$\sqrt{p}$	[0.1p, 0.25p, 0.5p, 0.7p, $\sqrt{p}$]
nodesize	Minimum number of observations required to split an internal node.	1	[1, 5, 10, 15, 20, 25]
maxnodes	Maximum number of terminal nodes (leaves) in each tree. If not given, trees are grown to the maximum possible (subject to limits by nodesize).	User-defined	[5, 7, 10, 12, 15, 20]

outcomes in child welfare, where risk factors and adverse outcomes often interact in complex ways (Kuhn & Johnson, 2013). Additionally, XGBoost incorporates regularization techniques to prevent overfitting, ensuring that the model generalizes well to new, unseen data. We implemented XGBoost using the *xgboost* package in R, an open-source library that provides a comprehensive suite of tools for model training and evaluation (Chen et al., 2024). To optimize the model's performance, we conducted a comprehensive hyperparameter tuning process using a 5-fold cross-validation approach. This method systematically evaluates the model's performance across different combinations of hyperparameters to identify the settings that yield the best results in terms of predictive accuracy and stability. The hyperparameters considered in this study, along with their tuned values, are detailed in Table 5.4. After evaluating various hyperparameter settings, the optimal configuration was determined to be:

- `nrounds`: 500
- `colsample_bytree`: 0.5
- `subsample`: 0.5
- `eta`: 0.01
- `max_depth`: 6
- `gamma`: 1

Table 5.4: overview of the hyper-parameters tuned for the XGBoost algorithm, along with the range of values considered during tuning and their typical default settings.

Hyperparameter	Description	Default values	Considered values
nrounds	The number of boosting rounds, which is equivalent to the number of trees to be built.	User defined	[100, 200, 300, 400, 500]
colsample_bytree	The sub-sample ratio of features used when constructing each tree to controls the proportion of features to sample.	1	[0.25, 0.5 ,0.75, 1.0]
subsample	The subsample ratio of the training instances to control the proportion of the training data to sample when building each tree.	1	[0.25, 0.5 ,0.75, 1.0]
eta	The learning rate, also known as the shrinkage parameter to scale the contribution of each tree.	0.3	[0.01, 0.05, 0.1, 0.2, 0.3]
max_depth	The maximum depth of the individual trees to control the maximum number of levels (depth) in each tree.	6	[3, 4, 5, 6, 7, 8, 9, 10]
gamma	The minimum loss reduction required to make a further partition on a leaf node of the tree.	0	[0, 0.1, 0.5, 1, 2, 5, 10]
min_child_weight	The minimum sum of instance weight (hessian) needed in a child to control the minimum number of samples required to create a new node in the tree.	1	[1, 3, 5, 7, 10]

- `min_child_weight`: 1

The remaining parameters were kept at their default values. This configuration was selected after assessing multiple combinations to achieve the highest possible testing AUC.

5.3.3 Model Evaluation process

In this work, the selection of a model required a careful balance between fairness and accuracy, particularly with respect to potential predictive bias against Māori children and the overall effectiveness in predicting outcomes.

Fairness refers to the model's ability to generate equitable predictions across different demographic groups, ensuring that no single group is disproportionately disadvantaged by the model's decisions. This means that the model should not systematically under-predict or over-predict outcomes for any specific group, thereby avoiding the reinforcement of existing inequalities.

Accuracy, on the other hand, relates to the model's ability to correctly predict outcomes across the entire dataset. This is typically assessed using metrics such as AUC, precision (PPV), recall (TPR), and overall prediction error rates. High accuracy ensures that the model is effective in identifying true outcomes, which is crucial for making informed decisions in the child welfare context.

Achieving a balance between fairness and accuracy requires evaluating and comparing different models using a comprehensive set of fairness and accuracy metrics, rather than selecting a model based solely on overall performance and assessing its impact on protected groups, such as Māori, only after development. It is crucial to address any identified unfairness proactively during the model development process.

This approach helps to identify a model that not only performs well in terms of accuracy but also strives to be fair, thereby reducing potential bias and supporting more effective and equitable decision-making. The following subsections provide an overview of the metrics and model evaluation methods used in this thesis.

5.3.3.1 Accuracy Metrics

During the model development process in this thesis, we prioritized AUC as a key measure of predictive accuracy to evaluate the performance of the tested algorithms. In the context of child welfare intake decision-making, AUC represents the probability that a randomly selected true positive notification will receive a higher predicted risk score than a randomly selected true negative notification (Drake et al., 2020). This measure is particularly effective in the child welfare context, as it evaluates the model's performance across all possible thresholds, providing a comprehensive view of its ability to distinguish between at-risk and non-at-risk children. The threshold independence of AUC is crucial, given the significant consequences of false positives and false negatives in child welfare, where

varying thresholds may be applied based on specific circumstances (Coohey et al., 2013).

In addition to AUC, this thesis reports several other key performance metrics, including Sensitivity (TPR), also referred to as Recall, Positive Predictive Value (PPV) or Precision, Negative Predictive Value (NPV), Accuracy, and the F1 score.[10] These metrics were selected for their relevance to the objectives of child protective services. TPR is emphasized due to its critical role in minimizing false negatives, which is particularly important in child welfare contexts where failing to identify an at-risk child could have severe consequences. PPV and NPV provide insights into the accuracy of positive and negative predictions, respectively, ensuring that both types of errors are adequately assessed. Accuracy is reported as a general measure of overall classification performance, while the F1 score is considered for its ability to balance precision and recall, offering a detailed assessment of the model's performance, especially in the presence of class imbalance.

We explored a range of modeling methodologies aimed at enhancing the AUC. As detailed in Section 5.3, these efforts involved integrating CYF data with additional government and Census datasets to assess the impact of incorporating broader variables on the models' predictive power. Additionally, we employed data transformation techniques and implemented resampling methods, such as cross-validation, across various machine learning algorithms to optimize model performance. Decisions concerning the methods adopted were guided by overall performance (AUC), accuracy for the specific high-risk group, and comparable levels of accuracy for Māori children versus others.

5.3.3.2 Fairness Metrics

Numerous group-level definitions of algorithmic fairness for binary classification tasks have been proposed in the literature, as detailed in Section 3.7. In this thesis, we explored various fairness concepts, including calibration, accuracy equity, disparate impact, and equalized odds. Ultimately, we prioritized addressing fairness concerns related to disparate impact and equalized odds, with a specific focus on mitigating disparities affecting the Māori group.

One method to ensure fairness and equal treatment of cases, regardless of the child's ethnicity, involves evaluating how well the models are calibrated across different subgroups (Centre for Social Data Analytics, n.d.). Specifically, Māori children and children from other ethnic groups who receive the same risk score should have an equal likelihood of experiencing an observed *estimated care and protection concern*. In this thesis, risk scores were derived by dividing the predicted probabilities

[10] A dedicated preface section, *Accuracy Metrics and Formulae*, is included to serve as a reference for the key evaluation metrics used throughout this analysis.

into 20 equally distributed risk ventiles, with a score of 20 representing the top 5% of predicted probabilities and a score of 1 representing the bottom 5%. Mathematically, this condition can be expressed as:

$$\mathbb{P}(y = 1 \mid r, s = \text{Māori}) = \mathbb{P}(y = 1 \mid r, s = \text{Non-Māori}) \tag{5.11}$$

Where y represents the observed outcome (*estimated care and protection concern*), s denotes the sensitive or protected variable (*ethnicity* in this case), and r is the risk score assigned based on the predicted probability.

In this study, we extended the assessment of calibration as a fairness measure beyond ethnicity to include disparities across subgroups defined by *gender* and *age*, with the aim of evaluating the equity and accuracy of predictive risk models for these demographic segments. As noted by Vaithianathan, Dinh, et al. (2019), the effectiveness of a predictive risk model in promoting equality and ensuring consistent treatment across cases is closely linked to its calibration performance across these diverse subgroups.

While calibration is often regarded as a primary, or even sole, criterion for predictive fairness (Chouldechova et al., 2018), it is crucial to acknowledge that calibration alone can be misleading. As pointed out by Corbett-Davies et al. (2009), risk scores can be adjusted to maintain calibration while exacerbating disparities in outcomes. Consequently, while miscalibration is a valid concern, it is not the only relevant issue (Chouldechova et al., 2018). To provide a more comprehensive evaluation of the models' fairness, it is necessary to also consider other metrics such as accuracy equity and error rate balance. This multifaceted approach offers a more comprehensive understanding of the model's fairness in practice, particularly in sensitive contexts like child welfare.

Accuracy equity assesses whether a model's predictive accuracy is consistent across different demographic groups (Berk et al., 2021). As defined by Dieterich et al. (2016), this concept extends fairness evaluation beyond fixed thresholds by considering the model's discriminative ability across all decision thresholds. In this thesis, accuracy is primarily evaluated using the AUC, which reflects the model's ability to distinguish between positive and negative cases. To assess accuracy equity, the AUC is calculated separately for each ethnic group, and the results are compared to identify any disparities. Ideally, the AUC should be approximately equal across all subgroups, indicating that the model performs consistently and fairly across different demographic segments (Chouldechova et al., 2018). Mathematically, this condition is expressed as:

$$\mathbb{P}(\hat{y} = y \mid s = \text{Māori}) = \mathbb{P}(\hat{y} = y \mid s = \text{Non-Māori}) \tag{5.12}$$

This equation indicates that the probability of the predicted outcome $\hat{y}$ matching the true outcome y should be the same, irrespective of the sensitive attribute, such as ethnicity (Berk et al., 2021).

Statistical parity, also known as disparate impact, demographic parity, or group fairness, requires that the probability of receiving a specific outcome, such as being referred for intervention or not, is equal across all ethnic groups (Dwork et al., 2012). Mathematically, this condition can be expressed as:

$$\mathbb{P}(\hat{y}|s = \text{Māori}) = \mathbb{P}(\hat{y}|s = \text{Non-Māori}) \tag{5.13}$$

While statistical parity is a formal fairness criterion requiring equal positive outcome rates across groups, this thesis uses the term disparate impact to refer more broadly to unjustified differences in model outcomes based on group membership.

Finally, we consider error rate balance, which is equivalent to equalized odds as defined in (Hardt et al., 2016). Error rate balance requires that both the FPR and FNR are equal across different levels of a protected attribute, such as ethnicity (Hardt et al., 2016). Since equal FNR implies equal TPR, this condition (Equalized Odds) can be mathematically expressed as

$$\mathbb{P}(\hat{y} = 1 \mid y = 1, s = \text{Māori}) = \mathbb{P}(\hat{y} = 1 \mid y = 1, s = \text{Non-Māori}) \tag{5.14}$$

and

$$\mathbb{P}(\hat{y} = 1 \mid y = 0, s = \text{Māori}) = \mathbb{P}(\hat{y} = 1 \mid y = 0, s = \text{Non-Māori}) \tag{5.15}$$

These equations indicate that Māori children and children from other ethnic groups should have an equal chance of receiving a positive outcome, whether the prediction is correct (TPR) or incorrect (FPR).

However, prioritizing one measure over another may come with certain trade-offs (Purdy & Glass, 2023). It is therefore essential to carefully navigate this trade-off. To address these risks, we evaluate the models' performance using all four fairness metrics. While each metric has its own strengths and limitations, together they provide a comprehensive framework for assessing fairness in algorithmic decision-making.

To measure the extent to which disparate impact exists between Māori children and children from other ethnic groups, we used the statistical parity measure of fairness, represented by Equation 5.16. This equation quantifies the balance of positive outcomes (intake) between Māori and non-Māori

groups by comparing the probability of receiving a positive outcome across the two groups. While statistical parity is a specific fairness metric, it is used here as an indicator of whether disparate impact may be present in the model's outcomes.

$$\text{Statistical Parity} = \min\left(\frac{\mathbb{P}(\hat{y}=1|s=\text{Māori})}{\mathbb{P}(\hat{y}=1|s=\text{Non-Māori})}, \frac{\mathbb{P}(\hat{y}=1|s=\text{Non-Māori})}{\mathbb{P}(\hat{y}=1|s=\text{Māori})}\right) \tag{5.16}$$

In an ideal situation of fairness, statistical parity should equal 1, ensuring that the probability of receiving a positive outcome (such as Intake) is uniform across all groups, irrespective of their sensitive attribute (*ethnicity* in this case). Values of statistical parity greater or less than 1 indicate disparities in how these outcomes are distributed among different groups. However, both Statistical Parity = 2 and Statistical Parity = 0.5 present the same level of discrimination, differing by the extent of their deviation from the ideal value of 1. A Statistical Parity value of 2 suggests that one group (e.g., Māori) has twice the probability of receiving a positive outcome compared to another group (Non-Māori). This indicates a form of favoritism or bias towards Māori. Conversely, a value of 0.5 implies that Māori are half as likely to receive a positive outcome compared to others, suggesting possible underrepresentation or disadvantage. To mitigate such disparities, the preference is given to the lower of these two values. To avoid confusion, we use the lower value of two possible values. Similarly, to quantify equalized odds, the equality of opportunity was calculated for both possible outcomes using Equation 5.17 and Equation 5.18.

$$\text{EOO (y=0)} = \min\left(\frac{P(\hat{y}=1 \mid s=\text{Māori}, y=0)}{\mathbb{P}(\hat{y}=1 \mid s=\text{Non-Māori}, y=0)}, \frac{\mathbb{P}(\hat{y}=1 \mid s=\text{Non-Māori}, y=0)}{\mathbb{P}(\hat{y}=1 \mid s=\text{Māori}, y=0)}\right) \tag{5.17}$$

$$\text{EOO (y=1)} = \min\left(\frac{\mathbb{P}(\hat{y}=1 \mid s=\text{Māori}, y=1)}{\mathbb{P}(\hat{y}=1 \mid s=\text{Non-Māori}, y=1)}, \frac{\mathbb{P}(\hat{y}=1 \mid s=\text{Non-Māori}, y=1)}{\mathbb{P}(\hat{y}=1 \mid s=\text{Māori}, y=1)}\right) \tag{5.18}$$

To claim that discrimination or unfairness does not exist, statistical parity or equality of opportunity must be equal to 1. An equality of opportunity value of 1 indicates that all ethnic groups have equal access to a given outcome for a particular class. By assessing equality of opportunity across classes, we can address equalized odds; that is, if equality of opportunity is achieved for each class, equalized odds is also met. However, it is unrealistic to expect models to achieve perfect statistical parity or equality of opportunity. Instead, we rely on the 80% rule, a widely accepted guideline in fairness-aware machine learning and anti-discrimination law. This rule suggests that the selection rate for a protected group should be at least 80% of the selection rate for the unprotected group (Biddle, 2017).

In the following section, we discuss the fairness-aware machine learning approach adopted in this thesis to address and mitigate the predictive bias identified through these metrics. This approach

aims to respond to the specific fairness concerns within the NZ child welfare context, ensuring that the use of predictive risk models promotes equitable outcomes across all demographic groups.

5.4 Fairness-aware Machine Learning

In this thesis, we employed an in-processing approach to address fairness in machine learning, specifically within the context of child welfare. This methodology involves embedding fairness constraints directly into the logistic regression algorithm during its learning phase, transforming the standard learning task into a constrained optimization problem. While the implementation of fairness constraints is well-documented in the literature (Agarwal et al., 2018; Berk et al., 2021; Calders et al., 2013; Corbett-Davies et al., 2009; Hu & Chen, 2020; K. D. Johnson et al., 2016; Nabi & Shpitser, 2018; Radovanović & Ivić, 2021; Radovanović et al., 2020; Zafar et al., 2019; Zafar et al., 2017), their application to predictive risk models in child welfare is novel (see Tables 3.1 and 3.2). By integrating relevant fairness constraints, we aimed to reduce bias and ensure equitable outcomes for vulnerable children, regardless of their ethnic background. To operationalize this fairness-aware approach, we focused on optimizing the model with respect to specific fairness metrics that are crucial in evaluating the equitable performance of predictive models.

5.4.1 Selected Fairness Measures for Model Optimization

This work focuses on two key fairness metrics: disparate impact and equalized odds. The decision to prioritize equalized odds as the primary fairness metric, and to correct for it through our fairness-aware machine learning approach, is based on both its ethical principles and the results of the fairness assessment conducted on the predictive models developed in this study, particularly in comparison to accuracy equity and calibration.

From an ethical standpoint, equalized odds is preferred because it ensures that predictions are independent of the protected attribute, such as *ethnicity*, when conditioned on the actual outcome (see Equations (5.14) and (5.15)). This criterion promotes the use of features that are directly related to the outcome, while preventing the sensitive attribute from being misused as a proxy in the prediction process (Hardt et al., 2016).

In contrast, accuracy equity focuses on achieving consistent predictive accuracy, such as AUC, across different ethnic groups (see Equation 5.16). While this measure aims to ensure fairness by providing equal predictive performance across all groups, it does not specifically prevent the potential misuse

of protected attributes in the prediction process (Hardt et al., 2016). Therefore, equalized odds aligns more closely with the values that stakeholders prioritize when defining algorithmic fairness in the child welfare context (Cheng et al., 2021; Purdy & Glass, 2023). While calibration is often regarded as a primary, or even sole, criterion for predictive fairness (Chouldechova et al., 2018), our assessment of the models developed in this study revealed significant disparities. Specifically, Māori children were incorrectly referred for further intervention at a substantially higher rate (FPR), while children from other ethnic groups experienced a higher rate of being overlooked for necessary intervention (FNR). This discrepancy could negatively impact both Māori children and those from other ethnic groups (Dare, 2013). Furthermore, the models consistently recommended Māori children for further intervention (intake) at a higher rate compared to children from other ethnic groups. These findings are consistent with previous research conducted in NZ by Rea and Erasmus (2017). According to the definitions of fairness outlined in Table 3.3, these issues pertain to disparate impact and equalized odds.

5.4.2 Constrained Logistic Regression

The logistic regression loss function, as defined in Equation (5.5) of Section 5.3.2.1, serves as the objective of our optimization problem, which is minimized under fairness constraints aimed at mitigating disparities in error rates across groups. To promote fairness in terms of equalized odds while preserving model accuracy, our analysis demonstrated that incorporating both *disparate impact* (statistical parity) and equalized odds as constraints during the logistic regression learning process effectively reduces disparities in error rates while maintaining statistical parity. This method achieves a balance between fairness and accuracy, with a reduction in error rate differences accompanied by a relatively modest trade-off in overall model performance. Building on the theoretical framework previously discussed, this section details the step-by-step implementation of these constraints in the logistic regression learning process. It includes the identification of protected and unprotected groups, the formulation of the corresponding constraints, and the construction of the constrained logistic regression optimization problem.

5.4.2.1 Contextualizing the Protected Variable

In our effort to address potential biases affecting Māori children within predictive models, we encoded the protected variable (s) as a binary indicator of *ethnicity*, where Māori children were assigned a value of 1, and all other ethnic groups were assigned a value of 0. The specific groupings were defined as follows:

Māori ($s = 1$)*:* Children identified as "Māori" or "Māori and Pacific".

Non-Māori ($s = 0$)*:* Children identified as "NZ European", "Pacific (excluding those also identified as Māori)", "Asian", "Middle Eastern", "Latin American", "African", or "Other ethnicity".

To streamline the analysis and maintain a focus on Māori children, the Pacific group was combined with the NZ European and Other category. Despite the under-representation of the Pacific group in the 2017 sample cohort (Table 6.8), a comprehensive assessment was performed to ensure that this aggregation did not obscure or exacerbate any disparities affecting Pacific children (Section 6.6.1.4). This additional analysis was critical in verifying that the model's performance and its implications for the Pacific group were adequately considered.

5.4.2.2 Logistic Loss Function Formulation

While the logistic loss function was previously defined in Equation (5.5), it is reiterated in Equation 5.19 with notations consistent with the constraints formulation and subsequent implementation in R. This ensures coherence between the theoretical framework and the practical application, facilitating a clear understanding of how the loss function integrates with the fairness constraints in the optimization process.

$$\begin{aligned} \text{Minimize } L(y, \hat{y}) &= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i + \epsilon) + (1 - y_i)\log(1 - \hat{p}_i + \epsilon)\right], \\ \text{where} \quad \hat{p}_i &= \frac{1}{1 + e^{-X_i\beta}}. \end{aligned} \tag{5.19}$$

In this formulation:

- y_i is the actual value of the outcome variable for the i-th observation, where $y_i \in \{0, 1\}$,
- $\mathbf{X}_i$ represents the feature vector for the i-th observation, representing the values of all the predictor variables for that observation,
- $\hat{p}_i$ is the predicted probability that the i-th observation belongs to the positive class, calculated using the sigmoid function,
- β represents the coefficient vector (or weight vector) associated with the features, and $\mathbf{X}_i\beta$ computes the linear predictor (logits), and
- ϵ is a small constant added to the probabilities in the logistic loss function to address a numerical stability issue in optimization problems.

In the implementation of constrained logistic regression as an optimisation problem, the logistic regression loss function (Equation (5.19)) and its gradient with respect to β (see Equation 5.20) were formulated in R to facilitate optimization. To avoid numerical instability, a small constant ϵ is added to the probabilities. The reason behind adding ϵ to the probabilities in the logistic loss function is that the logistic loss function involves taking the logarithm of predicted probabilities $\log(\hat{p}_i)$ and $\log(1-\hat{p}_i)$, as defined in Equation (5.19). If any predicted probability $\hat{p}_i$ is exactly 0 or 1, log(0) becomes undefined $(-\infty)$. This can cause the loss function to return `NaN` or `Inf` values, leading to convergence issues in the optimization process. The value of ϵ was chosen to be extremely small (e.g., 1×10^{-15}) so that it has a negligible impact on the value of the loss itself but was sufficient to avoid computational issues. The gradient function of the logistic loss function is given by:

$$\nabla_{\beta} L(\beta) = \frac{\partial L(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i) X_i, \tag{5.20}$$

where $(\hat{p}_i - y_i)$ is the difference between the predicted probability and the actual label, representing the error for the i-th observation. The gradient is averaged over all n observations, scaling the error by the corresponding feature values.

The loss function, defined in Equation (5.19), computes the negative log-likelihood of the predicted probabilities $\hat{p}_i$ with respect to the true labels y_i and the gradient of the logistic loss function in Equation (5.20), calculates the partial derivatives of the loss with respect to the coefficient vector β (Hastie et al., 2009). This gradient is essential for the optimization algorithm to update the coefficients and minimize the loss function efficiently (Hastie et al., 2009).

5.4.2.3 Disparate Impact and Equalized Odds Constraints Formulation

The formulation of the fairness constraints for logistic regression involved two key components: disparate impact (statistical parity) and equalized odds. Statistical parity is the most common notion of fairness and it can be ensured to some extent if both protected and unprotected groups have equal probability of the outcome occurring (Dwork et al., 2012). The mathematical formulation is:

$$p(\hat{y} \mid s = 1) = p(\hat{y} \mid s = 0). \tag{5.21}$$

Disparate impact arises when the rate of a favorable outcome (e.g., hiring, loan approval) is significantly lower, or when the rate of an unfavorable outcome (e.g., rejecting bail) is significantly higher for a protected group (e.g., a minority group) compared to others (Dwork et al., 2012).

In the context of child welfare, the implications of positive outcomes are nuanced. An intake decision, often considered a positive outcome, can have dual interpretations: it is favorable for high-risk children as it leads to necessary protection, but it can be unfavorable for low-risk children, as it results in unnecessary investigations that may be distressing for families. Thus, the challenge lies around ensuring that predictive models proportionately impact certain groups, balancing the need for protection with the avoidance of undue scrutiny.

To assess disparate impact, we often use the disparate impact ratio (DI_{ratio}) given by:

$$\text{DI}_{\text{ratio}} = \frac{p(\hat{y} \mid s = 1)}{p(\hat{y} \mid s = 0)}. \tag{5.22}$$

Considering the "intake decision" as the positive outcome, Equation (5.23) calculates the positive outcome rate for the protected group (e.g. Māori) and Equation (5.24) the positive outcome rate for unprotected groups (e.g. children from other ethnic groups), given by:

$$p(\hat{y} = 1 \mid s = 1) = \frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i}, \quad \text{and} \tag{5.23}$$

$$p(\hat{y} = 1 \mid s = 0) = \frac{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1 - s_i)}, \tag{5.24}$$

where:

- s_i is the indicator variable for the protected group (e.g., $s_i = 1$ if individual i is in the protected group, and $s_i = 0$ otherwise),
- $\hat{p}_i$ is the predicted probability for individual i, given by the logistic regression model: $\hat{p}_i = \frac{1}{1+e^{-X_i\beta}}$, and
- $\mathbb{I}(\cdot)$ is an indicator function, which is 1 if the condition inside ( ) is true and 0 otherwise.

Additionally, we have demonstrated that by adjusting the disparate impact ratio (DI_{ratio}), through a threshold δ, we can achieve improvements in terms of this fairness criterion.

Substituting Equations (5.23) and (5.24) in Equation (5.22) considering positive predicted outcome ($\hat{y} = 1$ or intake), the updated DI_{ratio} is given by:

$$\mathrm{DI}_{\text{ratio}} = \frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i}}{\frac{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1-s_i)}} \tag{5.25}$$

If this condition is formulated as a constraint to enforce a fairness threshold of δ in terms of statistical parity during the logistic regression learning process, the resulting disparate impact constraint, $C_{\mathrm{DI}}(\beta)$, is defined as:

$$\frac{\frac{\sum_{i=1}^{n} s_i \cdot I(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i}}{\frac{\sum_{i=1}^{n} (1-s_i) \cdot I(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1-s_i)}} - \delta \geq 0. \tag{5.26}$$

Equalized odds requires that both protected and unprotected group have equal TPR and equal FPR (Hardt et al., 2016). This condition is mathematically represented as follows:

$$p(\hat{y} = 1 \mid y = i, s = 1) = p(\hat{y} = 1 \mid y = i, s = 0), \quad i \in \{0, 1\}. \tag{5.27}$$

To enhance equalized odds, constraints were formulated to ensure that the TPR and FPR for the protected group are comparable to those for the unprotected group. Similar to the disparate impact constraint, these ratios are adjusted using a threshold δ. Separate constraints were specifically formulated for both TPR and FPR to rigorously enforce these conditions.

Also, the TPR for the protected group (e.g. Māori) and the TPR for unprotected groups (e.g. children from other ethnic groups) are given by:

$$p(\hat{y} = 1 \mid y = 1, s = 1) = \frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 1)}, \quad \text{and} \tag{5.28}$$

$$p(\hat{y} = 1 \mid y = 1, s = 0) = \frac{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 1)}. \tag{5.29}$$

In Equations (5.28) and (5.29), the symbol $\wedge$ within the indicator function $\mathbb{I}(\cdot)$ represents the logical term AND.

Balancing the TPR between protected and unprotected groups aligns with the equality of opportunity criterion for $y = 1$, denoted as $EOO(y = 1)$ (Hardt et al., 2016). Equality of opportunity ensures that

individuals who truly belong to the positive class ($y = 1$) are treated equally regardless of their membership in a protected or unprotected group. In mathematical terms, this means that the probability of a positive prediction given a positive ground truth ($y = 1$) should be the same for both protected and unprotected groups.

The corresponding constraint to enforce this balance, namely $C_{\text{EOO(y=1)}}(\beta)$, is designed to ensure that the TPR (i.e., the probability that $\hat{y} = 1$ given $y = 1$) is comparable across groups. This can be expressed as:

$$\frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=1)}}{\frac{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=1)}} - \delta \geq 0. \tag{5.30}$$

In Equation (5.30), the numerator represents the TPR for the protected group, while the denominator represents the TPR for the unprotected group. The threshold δ ensures that the TPRs are balanced within a specified tolerance. By enforcing this constraint, we ensure that individuals in the positive class are equally likely to be correctly identified as such, regardless of their group membership, thus achieving equality of opportunity.

Similarly for FPRs we have:

$$p(\hat{y} = 1 \mid y = 0, s = 1) = \frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 0 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 1)}, \quad \textit{and} \tag{5.31}$$

$$p(\hat{y} = 1 \mid y = 0, s = 0) = \frac{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 0 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 1)}. \tag{5.32}$$

Balancing the FPR for protected and unprotected groups is equivalent to the equality of opportunity notion of fairness for $y = 0$. This ensures that individuals who truly belong to the negative class ($y = 0$) are treated equally, irrespective of their group membership. In mathematical terms, this means that the probability of a positive prediction given a negative ground truth ($y = 0$) should be the same for both protected and unprotected groups.

The corresponding constraint to enforce this balance, $C_{\text{EOO(y=0)}}(\beta)$, is designed to ensure that the FPR (i.e., the probability that $\hat{y} = 1$ given $y = 0$) is comparable across groups. This can be expressed as:

$$\frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=0 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=0)}}{\frac{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=0)}} - \delta \geq 0. \tag{5.33}$$

By enforcing this constraint, we ensure that individuals in the negative class are equally likely to be

correctly identified as such, regardless of their group membership, thus achieving equality of opportunity.

The constraints for enforcing fairness, such as disparate impact (Equation5.26) and equalized odds (Equations (5.30) and (5.33)), were implemented as custom functions in R. In addition to the constraints themselves, their gradients were also formulated as separate functions to enable efficient optimization. These gradient functions are essential for the optimization process, as they provide the necessary information to adjust the coefficients β to minimize the loss function while satisfying the specified fairness constraints (Nocedal & Wright, 2006). For example, the disparate impact constraint function calculates the ratio of intake rates between Māori and non-Māori groups and compares it to a specified threshold δ, while the corresponding disparate impact gradient function computes the gradient of this constraint. Similarly, separate functions were developed for the TPR and FPR constraints, along with their respective gradients, to rigorously enforce equalized odds during the model training process.

5.4.2.4 Optimization Problem Formulation

Based on the above theories, we modified the problem of learning a fair algorithm in terms of statistical parity and equalized odds to an optimization problem where the logistic regression loss function (Equation (5.34)) is minimized subject to constraints defined based on disparate impact (Equation 5.35) and equalized odds (Equations (5.36) and (5.37)) measures of fairness.

$$\begin{aligned} \mathrm{L} &= \min_{\beta} \left(-\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{p}_i + \epsilon) + (1 - y_i) \log(1 - \hat{p}_i + \epsilon)] \right), \\ \text{where} \quad \hat{p}_i &= \frac{1}{1 + e^{-X_i \beta}}, \\ &\text{subject to} \end{aligned} \tag{5.34}$$

$$\delta \leq \frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i}}{\frac{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(\hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1 - s_i)}} \leq \frac{1}{\delta} \tag{5.35}$$

$$\delta \leq \frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i = 1)}}{\frac{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 1 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1 - s_i) \cdot \mathbb{I}(y_i = 1)}} \leq \frac{1}{\delta} \tag{5.36}$$

$$\delta \leq \frac{\frac{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=0 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} s_i \cdot \mathbb{I}(y_i=0)}}{\frac{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=0 \wedge \hat{p}_i \geq 0.5)}{\sum_{i=1}^{n} (1-s_i) \cdot \mathbb{I}(y_i=0)}} \leq \frac{1}{\delta} \tag{5.37}$$

In our optimization problem, we included both lower and upper bounds for constraints to ensure a balanced and fair treatment of different demographic groups. The lower bound, defined by a user-specified fairness threshold (δ), ensures that the protected group (e.g., Māori children) is not unfairly disadvantaged compared to the unprotected group (e.g., children from other ethnic groups). To complement this, we introduced an upper bound, defined as the inverse of the threshold δ, which ensures that the protected group does not receive an undue advantage over the unprotected group. By incorporating both constraints, we aim to maintain a balanced ratio of positive outcomes across different groups, thus promoting fairness and preventing significant disparities in the treatment of any particular demographic group, such as ethnic minorities. This dual constraint approach aligns with the principles of equitable treatment and helps mitigate the risk of reverse discrimination while striving for a fair predictive model. Additionally, our approach is adaptable to different binary classification thresholds, though the current implementation utilizes the conventional 50% threshold.

The choice of δ in this work is informed by the 80% rule, a widely recognized standard in fairness-aware machine learning literature and anti-discrimination law (Biddle, 2017). The 80% rule states that the selection rate for a protected group should be at least 80% of the selection rate for the unprotected group. This rule serves as a practical and enforceable threshold to detect and address potential biases in decision-making processes (Feldman, 2015). By adopting the 80% rule as our threshold, we align our model's fairness constraints with established literature. However, this value can be adjusted based on stakeholders' requirements, acknowledging that such adjustments may come at the cost of reduced accuracy.

To achieve a balance between predictive performance and fairness, two regularization parameters, λ_{DI} (for disparate impact) and λ_{EO} (for equalized odds), are introduced to balance the trade-off between these objectives. These regularization parameters enable the model to adjust its emphasis on accuracy versus fairness, providing a flexible framework to address the inherent tensions between these goals. The resulting loss function, incorporating these parameters, is formulated as follows:

$$L(\beta) + \lambda_{\text{DI}} \left(C_{\text{DI}}(\beta)\right)^2 + \lambda_{\text{EO}} \left(C_{\text{EOO(y=0)}}(\beta)^2 + C_{\text{EOO(y=1)}}(\beta)^2 \right), \tag{5.38}$$

where:

- $L(\beta)$ represents the logistic loss function,
- $C_{\text{DI}}(\beta)$ is the constraint formulated to reduce disparate impact, controlled by the regularization parameter λ_{DI},
- $C_{\text{EOO(y=0)}}(\beta)$ and $C_{\text{EOO(y=1)}}(\beta)$ are the constraints formulated to enforce equality of opportunity for both positive and negative outcomes, controlled by the regularization parameter λ_{EO}. Together, these constraints define equalized odds,

Having established the optimization framework to incorporate fairness constraints into logistic regression, the subsequent section will outline the practical steps for implementing this framework in R, including the algorithms used, parameter tuning process, and evaluation methods.

5.4.2.5 Implementation and Evaluation

To solve the optimization problem (constrained logistic regression) defined in previous section, we employed the *Sequential Least Squares Programming (SLSQP)* algorithm using the *nloptr* package in R (S. G. Johnson, 2008). This algorithm is well-suited for handling non-linear constraints and optimizing complex functions. The logistic loss function (Equation (5.34)), along with the fairness constraints for disparate impact (Equation (5.35)) and equalized odds (Equations (5.36) and (5.37)), were implemented as custom functions in R, along with their corresponding gradients. The optimization process returns a vector of coefficient values (β) that define the model, indicating the effect of each predictor on the outcome while satisfying the fairness constraints.

A grid search method was employed to tune λ_{DI} and λ_{EO} by systematically evaluating combinations of these parameters over a predefined range of $[0, 0.1]$. The primary objective was to identify the parameter values that best balance the fairness constraints (equalized odds and disparate impact) with overall model performance (e.g., AUC). In predictive risk modeling, the trade-off between fairness and performance is usually inevitable (Purdy & Glass, 2023; Radovanović et al., 2020). A higher δ value enforces stricter fairness constraints, potentially reducing disparities between groups by ensuring more equitable treatment. However, this stringent fairness criterion may negatively impact overall model performance, as the model might need to sacrifice some accuracy to meet the fairness requirements. Conversely, lower λ values relax the fairness constraints, potentially allowing for better overall performance but at the risk of increasing disparities between groups. By meticulously tuning these parameters, we aimed to strike an optimal balance where the model maintains high performance while satisfying fairness constraints to a reasonable extent.

Through this rigorous tuning process, we identified the optimal parameter values: $\lambda_{\text{DI}} = 0.03$ and $\lambda_{\text{EO}} = 0.06$. These values were found to best balance the competing objectives of fairness and model performance, ensuring that the predictive model operates effectively while mitigating unfair biases.

To evaluate the effectiveness of the constrained logistic regression model, its performance was assessed using the metrics defined in Section 5.3.3. Disparities across various subgroups, including gender and age groups, were also examined. Additionally, the model was tested on the Sample Cohort 2018 to assess its generalization capabilities with new data. To determine the effectiveness of our predictive risk model in supporting social workers, it was also crucial to assess its impact on enhancing the existing intake decision-making process. Consequently, we also compared the model's results with existing decision-making practices to evaluate its potential advantages in real-world applications.

5.5 Conclusion

In this chapter, we have outlined the methodological framework employed to investigate fairness-aware machine learning within the context of child welfare. This approach integrates fairness constraints directly into the logistic regression algorithm during its learning phase, transforming the standard learning task into a constrained optimization problem.

Our methodology was guided by several of our research questions such as: What factors influence the predictive accuracy of risk models intended for use by child protective services? What contributes to the unfairness of algorithms used to predict the risk of adverse events, specifically care and protection concern-related events? How feasible is it to develop a predictive risk model that is both more accurate and fair? What measures can Child Welfare authorities implement to mitigate or prevent algorithms from adopting discriminatory behaviors?

The methodological process was conducted in two main stages. The first stage was focused on answering the first research question by developing baseline logistic regression models that exhibit acceptable predictive power. We accomplished this by utilizing a novel research dataset, which we constructed by extracting records from both administrative and Census data sources outlined in Table 4.1. In this stage, we constructed a sample cohort and coded relevant outcome and predictor variables to facilitate the development process. Furthermore, we expanded our analysis beyond logistic regression by exploring other machine learning algorithms such as LASSO, Ridge, and Elastic Net,

collectively known as regularized logistic regression models. Additionally, models including random forest , support vector machine, and XGBoost were trained, and their performance compared in terms of fairness and accuracy metrics. Regularization techniques and cross-validation methods were employed to mitigate overfitting and enhance model generalization, with hyper-parameter tuning playing a critical role in optimizing the model's AUC.

The second stage aimed to address the remaining research questions by investigating the factors contributing to algorithmic unfairness and exploring measures to mitigate these biases. We integrated fairness constraints into the logistic regression learning phase to address racial disparities in the predictions made by the baseline models. Disparate impact and equalized odds constraints were formulated to ensure equitable treatment across different demographic groups while maintaining predictive performance. A comprehensive grid search over key parameters (λ_{DI} and λ_{EO}) was conducted to identify the optimal balance between predictive performance and fairness.

The fairness-aware machine learning methodological process began with the careful definition and preparation of our outcome variable and sensitive variable, converting the necessary data into appropriate formats. We then constructed the predictor matrix using relevant features extracted from the training data, ensuring a robust basis for subsequent analysis. The logistic loss function and its gradient were clearly defined to facilitate the optimization process, providing a rigorous foundation for the development of our predictive model. A significant aspect of our methodology was the incorporation of fairness constraints, specifically targeting disparate impact and equalized odds. The disparate impact constraint ensured that the positive outcome rate for the protected group was at least δ times the positive outcome rate for the unprotected group, promoting statistical parity. The equalized odds constraint aimed to ensure that the TPR and FPR for the protected group was at least δ times of that for unprotected group, thereby enhancing fairness in terms of both TPR and FPR. We meticulously derived the gradients of these constraints to enable their integration into the optimization algorithm.

The optimization process effectively handled the constrained optimization problem, ensuring that the model adhered to the predefined fairness criteria while maintaining high predictive accuracy. The evaluation of the final model included calculating a range of performance metrics and plotting ROC curves to assess both overall performance and fairness across different ethnic groups. This thorough evaluation confirmed that our model not only achieved higher accuracy compared to existing decision making process based on our data but also adhered to fairness principles, reducing potential unfairness to some degrees against the protected group.

The next chapter presents the results of our analysis, exploring the insights and trends that emerged from the data. This will be followed by a detailed discussion on how these findings relate to the existing literature and theoretical frameworks, providing a deeper understanding of the implications of our methodological choices. Through this comprehensive approach, we aim to contribute to the development of fairer and more effective predictive models for child welfare decision-making, addressing critical fairness issues and ensuring equitable treatment across different demographic groups.

Chapter 6

Empirical Results

6.1 Introduction

In this chapter, we present the empirical findings derived from the statistical analysis to address our research questions as stated in Section 1.4. These findings are structured in accordance with the methodological framework from Chapter 5, providing a cohesive narrative of our investigative process. The primary goal of this chapter is to offer essential insights for a deeper understanding of the potential unfairness or predictive bias associated with the development of predictive risk models in the NZ child welfare context.

6.2 Outcome Variable time frame Analysis

In our pursuit of identifying a suitable outcome variable for predicting care and protection-related events (*estimated care and protection concern*) as defined in Table 5.1, we explored different time frames. Specifically, we evaluated whether the target should be events occurring within two, three, or four years following the initial notification. With this aim, three separate LASSO logistic regression models were trained, each corresponding to one of the time frames. These models were developed using a comprehensive set of 322 features derived from multiple linked data sources, including Child, Youth and Family, Children's Action Plan, Personal Details, Benefit Dynamics, Sentencing and Remand, PRIMHD, and the 2018 Census (see Tables A.1-A.7 in Appendix A). These models will be referred to as the Two-Year, Three-Year, and Four-Year models, allowing us to assess their predictive performance and disparities across different time periods and providing insights into their effectiveness for short-term and long-term risk predictions in the context of child welfare.

Using the 70/30 rule, the models were trained on randomly selected 70% of children from the Sample Cohort 2017 and internally tested on the remaining 30%. In the initial training of LASSO logistic regression models to find the optimal outcome variable, we adjusted the sample according to the age of the children. For the two-year prediction period, we included children who were 16 years old or younger at the time of notification (n=59,475). For the three-year prediction window, children aged 15 years or younger were included (n=57,039), while for the four-year prediction window, the analysis was restricted to children who were 14 years old or younger at the time of their initial notification (n=54,111). Section 5.2.2 provides more details on this approach.

The outcome variable across the models, as illustrated in Table B.1 of Appendix B, has a relatively balanced distribution, with proportions of 47% vs 53% for the two-year model, 54% vs 46% for the three-year model, and 58% vs 42% for the four-year model, where the first percentage corresponds to the number of positive cases (*estimated care and protection concern* = 1),

The following subsections provide a comparative analysis of the performance and disparities across our trained models. Importantly, we ensure that the selected outcome variable not only predicts risk more accurately but also provides more equitable and reliable results across diverse demographic groups. This thorough comparison serves as a guide for selecting the most robust outcome variable for practical applications in this context.

6.2.1 Accuracy and Calibration Analysis

To evaluate the effectiveness of these models, we analyzed their performance using *estimated care and protection concern* as the outcome variable across designated time frames. The results based on specific accuracy metrics (Section 5.3.3.1) calculated at the standard 50% threshold for binary classification, are presented in Table 6.1. Based on this threshold, probabilities greater than or equal to 50% were classified as a positive outcome (intake), while those below 50% were classified as a negative outcome (no intake).

The Four-Year model consistently outperforms the Two-Year and Three-Year models across almost all metrics, particularly in terms of AUC (0.7213), classification accuracy (0.6743), and F1 score (0.7352). This model also returned the highest TPR (0.7792), suggesting its effectiveness in correctly identifying true positive cases. However, the slight decrease in NPV should be considered. This reduction in NPV can be attributed to the model's broader ability to capture long-term positive cases, which naturally increases the likelihood of false negatives as more uncertain cases from the earlier years manifest over time. This trade-off reflects the inherent complexity of balancing predictive

Table 6.1: Predictive performance of LASSO logistic regression on internal testing data at the 50% threshold for binary classification, considering various time frames for care and protection-related events (*estimated care and protection concern*).

Model	n	AUC	AUC 95% CI		PPV	NPV	TPR	Accuracy	F1
			Lower	Upper					
Two-Year	17,843	0.6963	0.6887	0.7040	0.6432	0.6488	0.5733	0.6464	0.6063
Three-Year	17,112	0.7199	0.7124	0.7275	0.6769	0.6470	0.7126	0.6639	0.6943
Four-Year	16,233	0.7213	0.7134	0.7291	0.6959	0.6343	0.7792	0.6743	0.7352

performance across extended time horizons in child welfare contexts. However, in the realm of predictive risk modeling, particularly within the child welfare context, the primary objective is to enhance decision-making accuracy to mitigate severe outcomes. This means that, the optimal model is typically the one that maximizes predictive accuracy. From our results, the Four-Year model emerges as the most suitable option, offering a slight, yet meaningful advantage in accurately identifying high-risk children.

In addition to accuracy measures, our models underwent calibration analysis based on the predicted probabilities from the LASSO logistic regression model. In classification modeling, it is essential to ensure that the estimated class probabilities accurately reflect the true underlying probabilities within the sample (Steyerberg & Vergouwe, 2014). In other words, the predicted probabilities must be well-calibrated, which means that their probabilities should genuinely represent the true likelihood of the event of interest (Kuhn & Johnson, 2013).

One way to assess the quality of the class probabilities is using a calibration plot. For a given set of data, a calibration plot compares the predicted probabilities (bin midpoints) against the observed event rates, allowing an evaluation of how closely the predicted probabilities align with the actual outcomes. Figure 6.1 shows the calibration plots for the Two-Year, Three-Year, and Four-Year models.

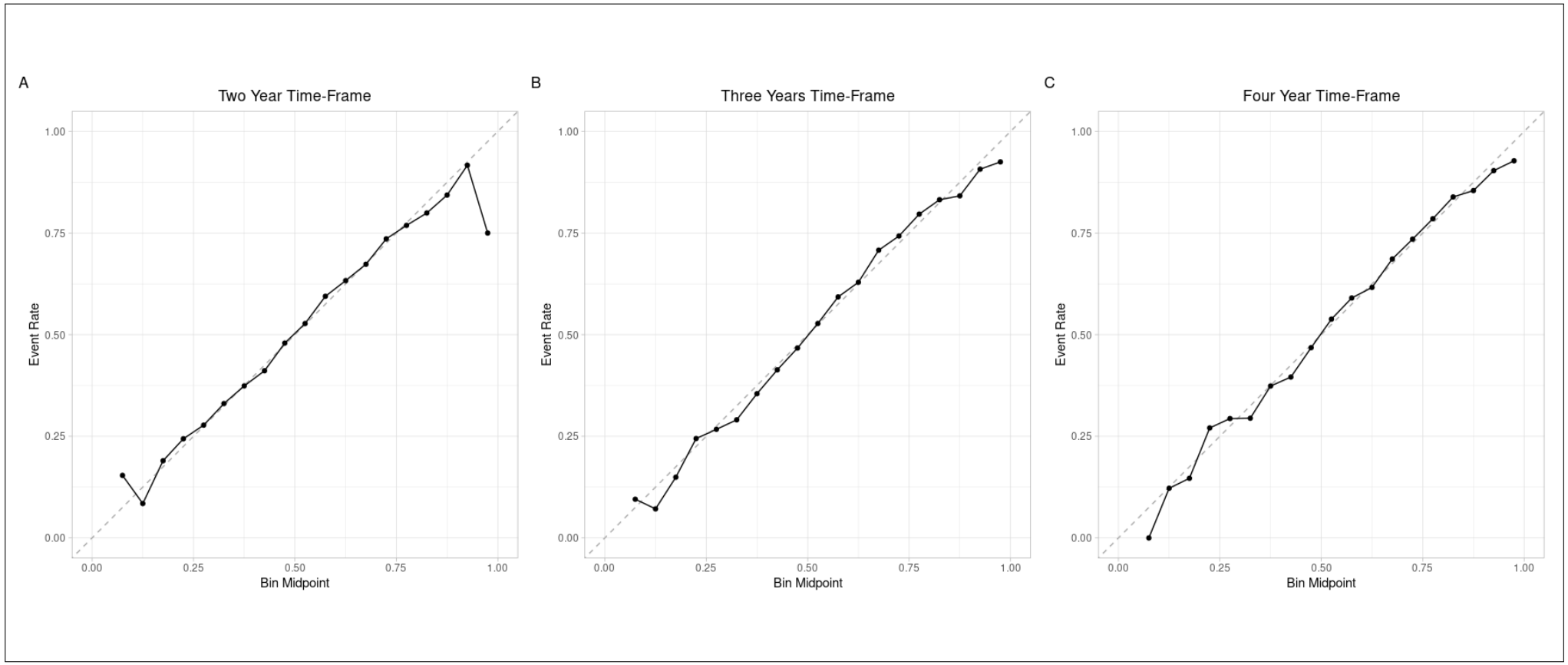

Figure 6.1: Calibration plots of LASSO logistic regression, considering various time frames for care and protection-related events (*estimated care and protection concern*).

In the Two-Year model, there is a notable deviation from the ideal calibration line, particularly at higher predicted probabilities, indicating a potential overestimation of risks. The Three-Year model exhibits improved calibration, with predictions more consistently aligning with the diagonal, reflecting better accuracy and reliability. The Four-Year model achieves a higher degree of calibration, with predictions closely mirroring actual outcomes across the entire probability spectrum.

Given the challenge of distinguishing calibration performance between the Three-Year and Four-Year models using calibration plots alone, we supplemented our analysis with Brier scores to provide a more precise validation and comparison (Brier, 1950). While calibration plots visually depict how well predicted probabilities align with actual outcomes, Brier scores offer a quantitative measure that enhances and complements these visual assessments.

For binary outcomes (where the outcome y_i is either 0 or 1), the Brier score for a set of predictions can be calculated as:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2, \tag{6.1}$$

where:

- n is the number of predictions.
- p_i is the predicted probability of the positive class for the i-th observation.
- y_i is the actual outcome (0 or 1) for the i-th observation.

We calculated Brier scores for the Two-Year, Three-Year, and Four-Year models using the internal testing data from the Sample Cohort 2017, with the *ModelMetrics* package in R (Mortimer et al., 2018). We applied bootstrap sampling to estimate the variability and robustness of the scores. The results are represented in Table 6.2.

Overall, the models show reasonable calibration, with the Four-Year model being the best calibrated. However, none of the models have a Brier score close to 0, which would indicate perfect calibration. Hence, while the models are not poorly calibrated, there may still be room to improve their accuracy and calibration. As shown in Table 6.2, the Four-Year model outperforms both the Two-Year and Three-Year models, as reflected by its lower Brier score. This lower score suggests that the Four-Year model provides more accurate and better-calibrated predictions. The confidence intervals for all models are narrow, suggesting that the Brier scores are reliable estimates of model performance.

Table 6.2: Brier Scores with 95% Confidence Intervals (CI) for LASSO logistic regression predictions.

Model	Brier Score	Lower 95% CI	Upper 95% CI
Two-Year	0.2201	0.2179	0.2225
Three-Year	0.2128	0.2105	0.2151
Four-Year	0.2085	0.2058	0.2112

6.2.2 Disparities Analysis Across Subgroups

In order to address equity and accuracy of models, we analyzed disparities across subgroups based on *gender*, *age*, and *ethnicity*. According to Vaithianathan, Dinh, et al. (2019), the ability of a predictive risk model to promote equality and to ensure that cases are treated consistently depends on how well the models are calibrated across these segments of the population.

We employed the Equalized Calibration Error (ECE) method suggested by Chawla et al. (2004) as a measure of calibration, to assess and compare the calibration of predictive model among these subgroups. The ECE is a metric that quantifies the difference between predicted probabilities and observed outcomes, averaged over multiple bins of predicted probabilities (Chawla et al., 2004). The formula for ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |\text{acc}(B_m) - \text{conf}(B_m)|\,, \tag{6.2}$$

where:

- M is the number of bins (or groups) into which predicted probabilities are divided.
- B_m represents the set of indices of samples whose predicted probabilities fall into the m-th bin.
- $|B_m|$ is the number of samples in bin B_m.
- n is the total number of samples.
- $\text{acc}(B_m)$ is the accuracy or the average of true labels for samples in bin B_m.
- $\text{conf}(B_m)$ is the average predicted probability for samples in bin B_m.

Table 6.3: Gender-based disparities in mean Equalized Calibration Error (ECE) for LASSO logistic regression predictions.

Model	Females	Males	Disparity (Females-Males)
Two-Year	0.0161	0.0141	0.0021
Three-Year	0.0200	0.0193	0.0007
Four-Year	0.0217	0.0211	0.0006

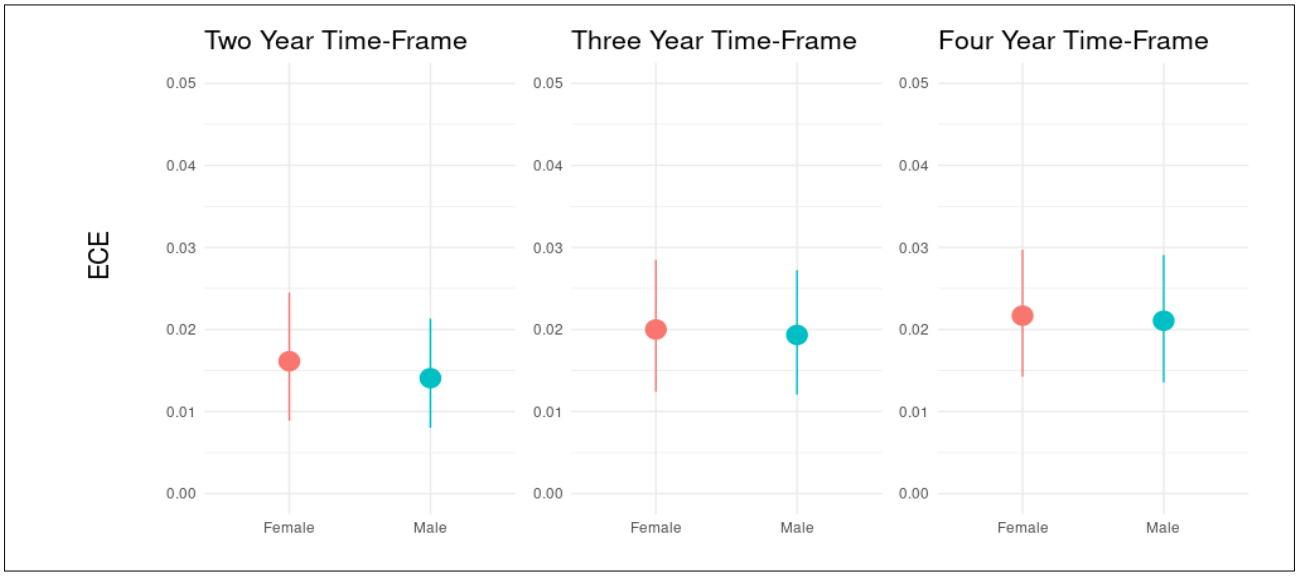

Figure 6.2: Bootstrapped Equalized Calibration Error (ECE) by gender.

To ensure robustness in our conclusions, we employed a bootstrapping approach, which involved repeatedly resampling the data with replacement and recalculating the ECE for each group (1000 bootstraps). This method allowed us to estimate the variability and 95% confidence intervals around the ECE, providing a reliable assessment of calibration disparities. Overlapping confidence intervals suggest similar calibration errors across groups, meaning that the model is equally well-calibrated. In contrast, non-overlapping intervals would imply potential disparities, with the model performing better for some groups than others.

Figures 6.2, 6.3, and 6.4 illustrate the bootstrapped ECE for subgroups across the Two-Year, Three-Year, and Four-Year models, while Tables 6.3, 6.4, and 6.5 present the mean ECE values along with disparities between groups, highlighting the differences in calibration across these models.

6.2.2.1 Gender Disparities

As shown in Figure 6.2 and Table 6.3, in the Two-Year model, the ECE for females is slightly higher at 0.0161 compared to 0.0141 for males, resulting in a disparity of 0.002. This indicates that while the model demonstrates relatively low calibration errors overall, there is a noticeable discrepancy

Table 6.4: Age-based disparities in mean Equalized Calibration Error (ECE) for LASSO logistic regression predictions.

Model	0_Newborn	One-Four	Five-Eight	Nine-Twelve	Over 12	Disparity (Highest-Lowest)
Two-Year	0.0483	0.0235	0.0174	0.0245	0.0304	0.0309
Three-Year	0.0616	0.0305	0.0238	0.0223	0.0303	0.0393
Four-Year	0.0453	0.0337	0.0189	0.0198	0.0321	0.0265

based on gender, with the model being less well-calibrated for females. As the outcome variable time frame extends to three years, both genders experience an increase in ECE, with values of 0.0200 for females and 0.0193 for males. However, the disparity between genders narrows to 0.0007, implying

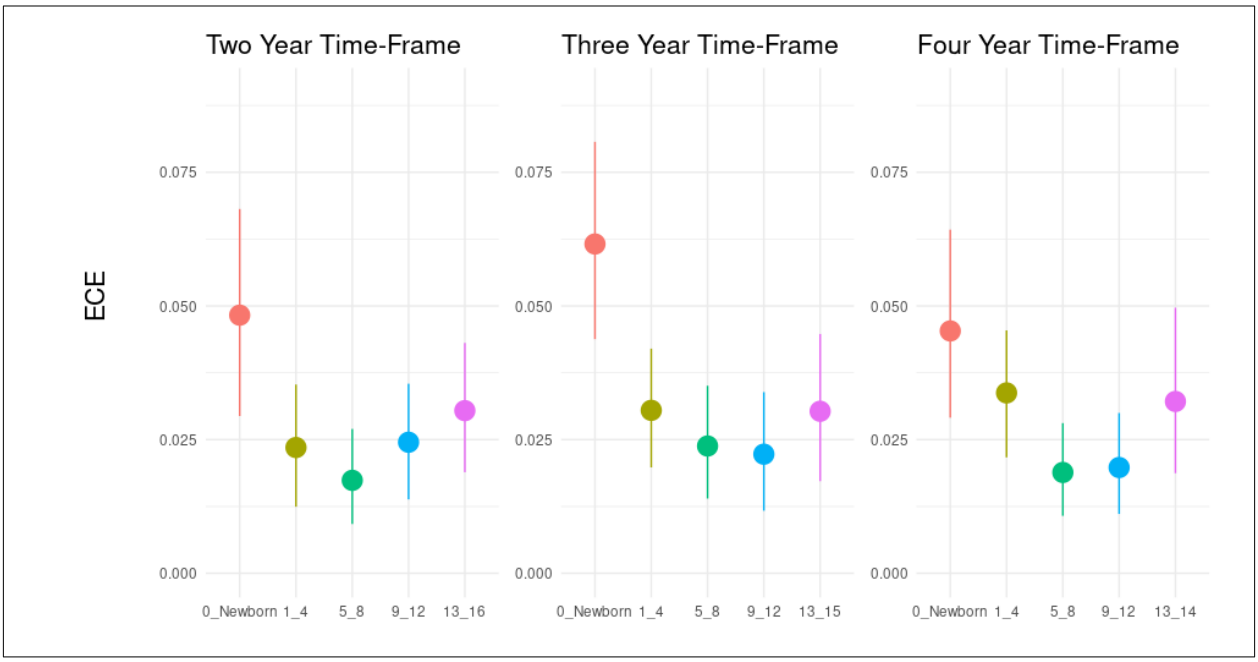

Figure 6.3: Bootstrapped Equalized Calibration Error (ECE) by age group.

that the model becomes more balanced in its predictive accuracy across genders, even as the overall calibration error increases. In the Four-Year model, the ECE further increases to 0.0217 for females and 0.0211 for males, but the disparity reduces slightly to 0.0006, marking the smallest difference across the three time frames.

This observed trend suggests that, although the model's calibration error generally worsens over longer time frames, it becomes more equitable in its performance between genders. Thus, while the Two-Year model exhibits lower ECE values, it also has the largest gender disparity. Conversely, the Four-Year model, despite its higher ECE, offers a more consistent calibration across genders, highlighting the trade-off between overall model calibration and equity in predictions based on gender.

Table 6.5: Ethnic group disparities in mean Equalized Calibration Error (ECE) for LASSO logistic regression predictions.

Model	NZ European and Others	Māori	Pacific	Disparity (Highest-Lowest)
Two-Year	0.0295	0.0126	0.0384	0.0259
Three-Year	0.0259	0.0189	0.0306	0.0118
Four-Year	0.0126	0.0187	0.0317	0.0191

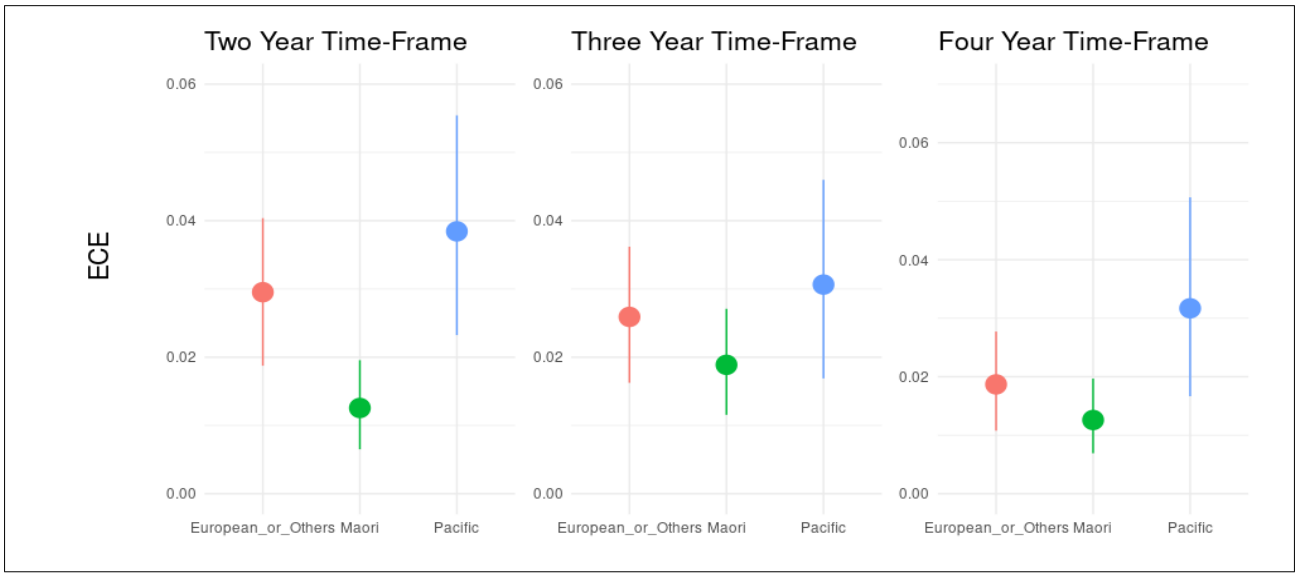

Figure 6.4: Bootstrapped Equalized Calibration Error (ECE) by ethnicity.

6.2.2.2 Age Group Disparities

Across all models, the 0–Newborn age group consistently shows the highest ECE, with the Three-Year model reaching a peak value of 0.0616 (Table 6.4). Despite this, disparities across age groups stabilize in the Four-Year model, which has the lowest variation (0.0265). Although some differences remain, particularly between the 0–Newborn and Over 12 age groups, these disparities are less pronounced compared to the Two-Year and Three-Year models, suggesting a more balanced calibration across age groups in the Four-Year model (Figure 6.3).

6.2.2.3 Ethnic Group Disparities

As illustrated in Figure 6.4 and Table 6.5, the highest ECE is observed for the Pacific group across all models, followed by NZ European and Others, with the lowest ECE for the Māori group. The largest disparity between ethnic groups is evident in the Two-Year model, with a significant difference between the Pacific and Māori groups. The disparity decreases in the Three-Year model, but increases slightly in the Four-Year model, particularly between the Pacific and Māori groups.

6.2.3 Summary of Findings

Our analysis, utilizing separate LASSO logistic regression models based on care and protection-related events across different time frames, revealed an enhancement in predictive accuracy (Table 6.1). Extending the period covered by the outcome variable improved AUC and overall accuracy by approximately 3% (Table 6.1), with the Four-Year model achieving the highest AUC (0.7213) and overall accuracy (0.6743). This indicates better performance in identifying high-risk children. Moreover, the Four-Year model consistently outperformed the others in terms of calibration across age groups and genders. Its calibration was more closely aligned with observed outcomes, particularly in long-term risk prediction.

While the Three-Year model exhibited slightly better performance regarding equity across ethnic groups, the Four-Year model still provides fair calibration and equitable predictions across all demographics, with only minor disparities that could potentially be mitigated through further feature refinement and the use of alternative modeling techniques.

Overall, the Four-Year model, with its better predictive accuracy and relatively balanced calibration across different demographic groups, emerges as the most suitable choice for defining the outcome variable in this study. Consequently, the outcome variable is defined to predict for each child notification, the probability that one or more of the events outlined in Table 5.1 will occur within four years of the initial notification. This is referred to as the '*estimated care and protection concern* within four years' throughout this study. This selection enables a more comprehensive assessment of risk over an extended period, providing a robust foundation for subsequent analyses.

6.3 Predictor Variables

After a comprehensive review of the Stats NZ IDI data dictionaries and classification files provided by government agencies, we extracted relevant information from datasets listed in Table 4.1. The extracted information from these administrative datasets was then used to construct the initial predictor features, detailed in Tables A.1-A.7 of Appendix A.

The final set of predictor variables used for modeling in this thesis was determined following a rigorous pre-processing phase, which involved evaluating the predictive power of each feature based on AUC and examining potential interactions among them.

During the data pre-processing phase, some integer predictors were transformed into categorical

variables by binning (Kuhn & Johnson, 2013). This process involved grouping integer values into discrete categories based on predefined ranges, allowing for more interpretable and manageable feature sets. The binning thresholds were selected based on domain knowledge and exploratory data analysis, ensuring that each bin represented meaningful distinctions within the data (e.g., *age*, *number of siblings*, *proportion of time on benefit*, etc). Additionally, all categorical variables were converted into dummy variables prior to their inclusion in the models to ensure appropriate handling of categorical data (Kuhn & Johnson, 2013). As a result, the total number of predictors includes dummy variables, which contributed to a larger number of 250 predictors. Given the predictive nature of the models developed in this thesis, the large number of predictors (p=250) ensures comprehensive coverage of factors influencing child welfare outcomes.

We categorized the final set of predictors into four groups: *child predictors*, *parents predictors*, *family predictors*, and *other predictors*. A detailed list of these predictors, along with the number of features in each category, can be found in Tables A.8-A.11 of Appendix A.

Parents predictors were encoded consistently for both the mother and the father to reflect their characteristics as risk factors. There was only one exception where the predictor was defined solely for the father due to a lack of available records for the mother. This exception is a binary variable indicating whether the father completed a sexual offense program while in custody within the past 5 years.

Other key considerations involved in encoding predictor variables and selecting the final set of features are outlined in the following subsections.

6.3.1 Encoding Considerations

Predictor variables related to the child's history of interactions with child protective services were encoded across multiple time frames, including the past 3, 6, 12 months, 2 years, or earlier, by considering the temporal nature of the data and for the following reasons:

Given that most administrative data are timestamped, they capture trends and changes over time (Chen et al., 2024). By considering children's interactions with the child welfare system over these different periods, the model can evaluate varying levels of relevance to current risk. For instance, recent interactions, such as those within the last 3 or 6 months, may signal more immediate concerns, while interactions spanning the past 2 years or longer provide important context for chronic or recurring issues.

Incorporating multiple time frames also enhances the model's robustness and generalization by allowing it to capture both short-term and long-term risk factors. This approach reduces the impact of missing data and provides a more complete picture of a family's interactions with the child welfare system, thereby increasing the model's suitability across diverse cases. Additionally, this temporal encoding enables the model to adapt to broader social and policy shifts that might influence reporting rates and patterns, ensuring that it remains relevant even as external conditions change (Melz et al., 2023).

The frequency of events within these time frames also offers crucial insights; a child reported 10 times in a week presents a vastly different risk profile than a child reported once in a year. This difference in frequency is reflected in the model by encoding both, the total number of events and binary indicators of whether certain events occurred within specific time frames, allowing the model to distinguish between high-frequency and low-frequency cases (Chen et al., 2024).

Likewise, we encoded variables related to a child's interaction with the social welfare system, as well as variables related to parental criminal history, mental health, or addiction issues, across the past 1, 2, 3, 4, 5 years, or earlier. This approach is consistent with the methodologies applied in the development of the DCDA and LA Risk Stratification models (Putnam-Hornstein et al., 2022; Vaithianathan, Dinh, et al., 2019).

6.3.2 Selection Considerations

Initially, predictor variables related to the child's history of interactions with child protective services were encoded to capture both, the number of events within the past 3, 6, 12 months, 2 years, or earlier, as well as binary indicators for whether specific events occurred within each time frame. To determine the most effective format for each variable, either as a binary indicator or as a count of events, both sets of features were evaluated separately for their predictive power in estimating the outcome variable (*estimated care and protection concern* within four years). Separate logistic regression models were trained for each set, and variables with higher AUC values were then selected for inclusion in the final predictor set.

In instances where the AUC values were comparable across different sets of predictors, the Akaike Information Criterion (AIC) was considered to inform the selection process. AIC is particularly useful in logistic regression, as it evaluates model quality by accounting for both goodness of fit and model complexity, so helping to mitigate overfitting while ensuring the model captures the data patterns effectively (Kuhn & Johnson, 2013). Although the Bayesian Information Criterion (BIC) is known to be

Table 6.6: Selected variable types related to children's history of interactions with the child welfare system.

Extracted Information	Selected Variable Type
Substantiated finding of emotional abuse	Number of events
Substantiated finding of physical abuse	Binary indicator
Substantiated finding of sexual abuse	Number of events
Substantiated finding of neglect	Number of events
Substantiated finding of maltreatment	Number of events
Section 15 report of concern	Number of events
Section 15 report of concern with NFAR outcome	Number of events
Intake	Number of events
Investigation	Number of events
Risk and safety assessment	Number of events
Placement	Binary indicator
Custody guardianship spell	Number of events
Family whānau Agreement (FWA)	Binary indicator
Family Group Conference (FGC)	Number of events

more conservative than AIC due to its stronger penalty for model complexity, it was not used systematically in this study. AIC was prioritised because the primary aim was to improve predictive accuracy, and AIC is often preferred when prediction is the goal. Predictor variables and their selected type based on this assessment are presented in Table 6.6. Additionally, simple logistic regressions were conducted for each predictor variable independently against the outcome variable to evaluate their individual contributions. This analysis provided insights into the strength of association between each predictor and the outcome, as well as the variance each predictor explained on its own. Predictors that did not demonstrate a statistically significant effect on the outcome were excluded from further analysis.

6.4 Risk Prediction Enhancement through Data Linkage

This section presents our empirical findings from systematically investigating the impact of incorporating additional predictor variables on the predictive power of baseline logistic regression models by linking CYF data with information from other government agencies and NZ Census records (see Section 5.3.1).

We developed and trained baseline logistic regression models, including regularized variants such as LASSO, Ridge, and Elastic Net, using 70% of the Sample Cohort 2017. These models were subsequently validated on the remaining 30% of the sample, and their performance across various metrics at a 50% threshold is outlined in Table 6.7. The regularization parameter (λ) was tuned using a 5-fold

cross-validation resampling technique for regularized logistic regressions to identify the model with the highest testing AUC (see Section 5.3.2.2).

The probability of the outcome (*estimated care and protection concern* within four years) was obtained from our models. To assess the effect of incorporating additional predictors at each linkage level, we evaluated the models' performance in terms of both accuracy and potential predictive bias or unfairness based on these probabilities.

6.4.1 Evaluating Accuracy Across Data Linkages

As shown in Table 6.7, the highest AUC values are achieved by LASSO (AUC: 0.7193), Elastic Net (AUC: 0.7193), and Ridge (AUC: 0.7189), with the LASSO logistic regression model consistently and slightly outperforming the others in terms of AUC and overall accuracy. Specifically, the LASSO model exhibits the highest AUC values across all data linkage approaches (1L-5L, as described in Section 5.3.1), indicating better discriminatory power in distinguishing between positive and negative cases. As additional predictor variables were incorporated through each data linkage level (1L-5L), the performance metrics generally improved, with the most significant enhancement observed in the 5L data linkage approach (from 0.7073 to 0.7193).

As the regularized logistic regression variants (LASSO, Ridge, and Elastic Net) demonstrated comparable predictive performance, our analysis focused on probabilities or predictions obtained from three models moving forward: LASSO logistic regression, full logistic regression, and refined logistic regression. At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$).

To provide a more comprehensive evaluation of model performance across different risk groups, we also assessed cumulative accuracy, which allowed us to examine how well the models performed not only at a fixed threshold but also across varying levels of risk.

Figure 6.5 presents cumulative accuracy plots for these three candidate models, offering a detailed comparison of model performance across various risk ventiles for linkages 1L and 5L. Each plot in Figure 6.5 displays cumulative accuracy (%) as a function of ventiles (5% risk groups). The solid black line represents models trained on predictors derived from CYF data, Children's Action Plan, and Demographic Details data (1L data linkage approach), while the red dotted line corresponds to models trained on an expanded set of predictors that includes additional data from Benefit Dynamics, Sentencing and Remand, PRIMHD, and the 2018 Census (5L data linkage approach). It is evident

that models trained on both 1L and 5L predictors exhibit a decreasing trend in cumulative accuracy from the highest risk groups (top 5%) to the lowest risk groups (lower 5%). This trend highlights the models' heightened accuracy in predicting outcomes for higher-risk individuals.

Within each logistic regression model, the one trained on the expanded set of predictors (5L) consistently outperforms the models trained solely on predictors from Child, Youth and Family, Children's Action Plan, and Demographic Details data (1L) across all risk cutoffs. This superiority is particularly pronounced in the highest risk groups, where the cumulative accuracy remains above 85%, compared to approximately 80% for models trained on predictors from the 1L linkage approach. The performance gap between the models narrows as we move towards the lower-risk groups, indicating a convergence in accuracy for individuals with lower risk scores. This comparative analysis suggests that incorporating additional predictors from other organizations enhanced the model's predictive power, particularly in high-risk populations.

The overall results suggest that models utilizing predictors from the 5L data linkage approach consistently outperform those trained on the more limited set of predictors (1L), particularly in the highest risk groups. This indicates that incorporating additional predictors from other government data sources enhances the model's ability to identify high-risk individuals, leading to enhanced predictive performance.

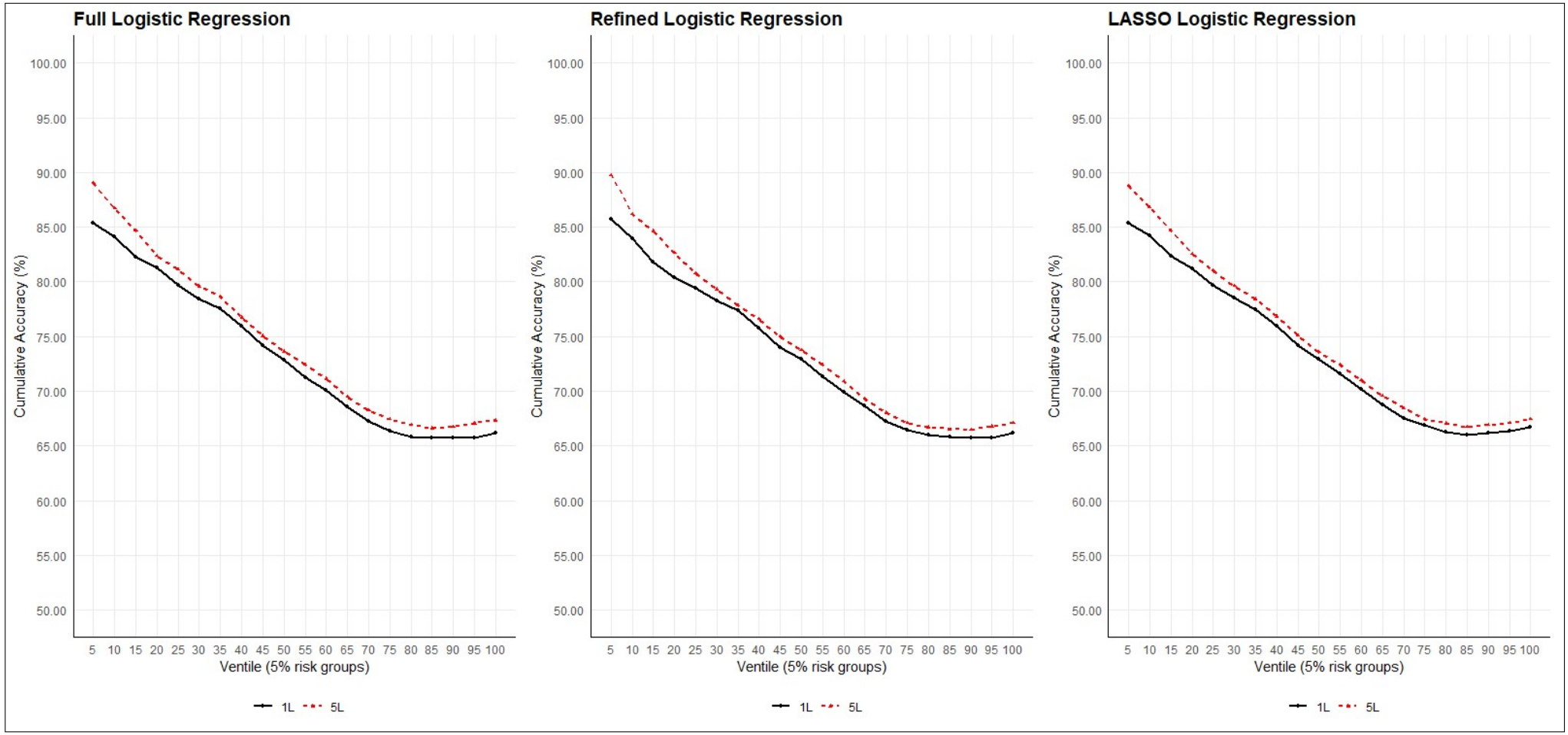

Figure 6.5: Cumulative accuracy gain for candidate logistic regression models across 5% risk groups (ventiles).

Table 6.7: Predictive performance results of baseline logistic regression models using a standard 50% threshold for binary classification, evaluated on 30% of the Sample Cohort 2017 across data linkages 1L-5L

Linkage	Model	AUC	AUC 95% CI		PPV	NPV	Accuracy	TPR	F1
			Lower	Upper					
1L	Full Logistic Regression	0.7055	0.6975	0.7135	0.6866	0.6160	0.6618	0.7675	0.7248
	Refined Logistic Regression (p<0.1)	0.7037	0.6957	0.7118	0.6865	0.6160	0.6618	0.7682	0.7251
	LASSO Logistic Regression (λ = 0.00045)	0.7073	0.6993	0.7152	0.6866	0.6167	0.6621	0.7688	0.7254
	Elastic Net Logistic Regression (λ = 0.000901)	0.7056	0.6976	0.7136	0.6865	0.6165	0.6620	0.7689	0.7253
	Ridge Logistic Regression (λ = 0.010905)	0.7053	0.6973	0.7133	0.6849	0.6147	0.6604	0.7685	0.7243
2L	Full Logistic Regression	0.7149	0.7070	0.7229	0.6954	0.6279	0.6713	0.7720	0.7317
	Refined Logistic Regression (p<0.1)	0.7142	0.7062	0.7221	0.6944	0.6278	0.6709	0.7733	0.7317
	LASSO Logistic Regression (λ = 0.000495)	0.7152	0.7073	0.7231	0.6943	0.6283	0.6710	0.7739	0.7319
	Elastic Net Logistic Regression (λ = 0.000821)	0.7152	0.7073	0.7231	0.6945	0.6281	0.6710	0.7736	0.7319
	Ridge Logistic Regression (λ = 0.010905)	0.7149	0.7070	0.7228	0.6941	0.6295	0.6714	0.7755	0.7326
3L	Full Logistic Regression	0.7157	0.7078	0.7236	0.6959	0.6266	0.6710	0.7694	0.7308
	Refined Logistic Regression (p<0.1)	0.7150	0.7071	0.7229	0.6946	0.6278	0.6709	0.7723	0.7314
	LASSO Logistic Regression (λ = 0.000451)	0.7160	0.7081	0.7239	0.6951	0.6273	0.6710	0.7717	0.7314
	Elastic Net Logistic Regression (λ = 0.000821)	0.7160	0.7081	0.7239	0.6951	0.6271	0.6709	0.7710	0.7311
	Ridge Logistic Regression (λ = 0.011968)	0.7156	0.7077	0.7235	0.6950	0.6285	0.6714	0.7729	0.7319
4L	Full Logistic Regression	0.7171	0.7092	0.7250	0.6969	0.6280	0.6722	0.7701	0.7317
	Refined Logistic Regression (p<0.1)	0.7161	0.7082	0.7240	0.6964	0.6284	0.6721	0.7713	0.7319
	LASSO Logistic Regression (λ = 0.000495)	0.7173	0.7094	0.7252	0.6973	0.6293	0.6729	0.7714	0.7325
	Elastic Net Logistic Regression (λ = 0.000989)	0.7173	0.7094	0.7252	0.6966	0.6289	0.6724	0.7713	0.7321
	Ridge Logistic Regression (λ = 0.010905)	0.7171	0.7092	0.7250	0.6970	0.6300	0.6731	0.7725	0.7328
5L	Full Logistic Regression	0.7190	0.7112	0.7269	0.6977	0.6298	0.6734	0.7717	0.7328
	Refined Logistic Regression (p<0.1)	0.7175	0.7096	0.7254	0.6948	0.6275	0.6709	0.7720	0.7313
	LASSO Logistic Regression (λ = 0.000596)	0.7193	0.7114	0.7272	0.6970	0.6330	0.6744	0.7764	0.7346
	Elastic Net Logistic Regression (λ = 0.001307)	0.7193	0.7114	0.7271	0.6966	0.6323	0.6739	0.7762	0.7342
	Ridge Logistic Regression (λ = 0.010905)	0.7189	0.7110	0.7268	0.6952	0.6314	0.6728	0.7768	0.7337

Note: At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as *refined logistic regression (p<0.1)*.

6.4.2 Evaluating Predictive Bias Across Data Linkages

This section provides a thorough analysis of the ethnic bias properties of models developed through each data linkage approach (1L-5L). The objective is to determine whether incorporating additional variables through linkages, as well as employing different modeling methods, reduces potential predictive bias across ethnic groups. The evaluation in this section focuses on Māori, Pacific, and NZ European and Others ethnic groups.

Table 6.8 presents the distribution of ethnic groups across the primary records used in developing predictive risk models in this thesis. Māori children are notably over-represented in records of past interactions with child protective services. For instance, Māori children in the Sample Cohort 2017 (n=54,111) comprise 60% of records for past Section 15 notifications, 65% of past intakes, and 65% of past substantiated findings of maltreatment, despite representing a smaller proportion of the general population. According to the 2018 New Zealand Census, 16.5% of the population identified as Māori, compared to 70.2% identifying as European, 15.1% as Asian, 8.1% as Pacific peoples, and 1.9% as Middle Eastern, Latin American, or African (MELAA). Because individuals could identify with more than one ethnic group, these percentages do not sum to 100% (Stats NZ, 2018b). Similarly, Māori represent 61% of the records in the Benefit Dynamics data (2L) and are significantly over-represented in the Sentencing and Remand data, making up 72% of the records for fathers and 74% for mothers (3L). Additionally, Māori constitute 65% of the records in the PRIMHD data for fathers and 62% for mothers.

In contrast, Pacific children constitute only 7% to 11% of these records, while NZ European and Others range between 19% and 31%. This discrepancy is also reflected in the 2017 Sample Cohort's outcome variable, where 61% of children with an observed *estimated care and protection concern* within four years are Māori, 11% are Pacific, and 29% are NZ European and Others ethnic groups (see Table B.2 in Appendix B).

The ethnic group imbalance across records raises important concerns about the potential for predictive models to reinforce systemic bias. The over-representation of Māori children in welfare and criminal justice datasets could lead to models disproportionately predicting higher risk for Māori, thereby perpetuating inequitable outcomes (Barocas & Selbst, 2016).

To assess predictive bias based on *ethnicity*, we specifically focus on three key fairness metrics: calibration, accuracy equity, and error rate balance, also referred to as equalized odds. The rationale behind this decision was discussed in Section 5.3.3.2.

Table 6.8: Distribution of ethnic groups (Māori, Pacific, and NZ European and Others) across key records in the Sample Cohort 2017.

Category	Record	Māori	Pacific	NZ European and Others
Past Section 15 notifications	36,375	60%	10%	30%
Past intake	31,785	65%	10%	25%
Past substantiated findings of maltreatment	19,080	65%	10%	25%
Benefit Dynamics data	30,480	61%	11%	28%
Sentencing and Remand data (Father)	14,541	72%	8%	20%
Sentencing and Remand data (Mother)	6,999	74%	7%	19%
PRIMHD (Father)	20,058	65%	8%	27%
PRIMHD (Mother)	23,175	62%	7%	31%
Total records (n)	54,111	29,841	18,296	5,974

Note: Since the number of records varies across different Census questions, the number of records for the 2018 Census data is excluded from this table. Each variable encoded from the *2018 Census* data includes an *unknown* category to account for missing or incomplete responses.

The results presented in the following subsections are derived from a validation sample of 16,223 children from the 2017 cohort, which was excluded from the training of the predictive models (internal testing data). The fairness evaluation in this section primarily focuses on three ethnic groups: Māori (n=8,988), Pacific (n=1,806), and NZ European and Others ethnic groups (n=5,436). For details on ethnic group classification, refer to Section 5.2.3.1.

6.4.2.1 Calibration

One way to address equity is by evaluating how well the models are calibrated across ethnic groups (Centre for Social Data Analytics, n.d.). Ideally, Māori, Pacific, and NZ European and Others children, when assigned the same risk score, should have an equal likelihood of experiencing at least one care and protection-related event (*estimated care and protection concern*) within four years. For assessment purposes, we conducted a comparative analysis of the three candidate models across data linkage levels (1L-5L), as shown in Figure 6.6. Each plot represents the rate of observed *estimated care and protection concern* within four years against the predicted risk score, stratified by ethnic group. A score of 20 represents the top 5% of predicted probabilities, while a score of 1 represents the bottom 5%.

Across models and linkages, the NZ European and Others group (red line) consistently displays smooth calibration, with fewer fluctuations compared to Māori and Pacific groups. The Pacific group,

in particular, exhibits the most calibration instability, especially in the full and refined logistic regression models, suggesting weaker calibration. However, the LASSO model demonstrates better calibration, particularly in higher linkage levels (4L and 5L), where the curves are more aligned across ethnic groups.

The full logistic regression model demonstrates progressively improved calibration as additional data is incorporated, specifically with the inclusion of Benefit Dynamics data, Sentencing and Remand data, and PRIMHD (4L). Despite these improvements, disparities persist for Māori and Pacific groups at lower linkage levels (1L and 2L), indicating a potential overestimation of risk. While the inclusion of the 2018 Census data (5L) slightly improves calibration overall, it appears to have negatively impacted the Pacific group, especially in full and refined logistic regression models. The LASSO model, however, maintains the best calibration across all linkage levels, particularly in 4L and 5L, providing the most equitable predictions by minimizing disparities between groups.

These findings are further supported by the assessment of Equalized Calibration Error (ECE), as discussed in Section 6.2.2, which provides a quantitative measure of calibration (Chawla et al., 2004). These measures are visually represented in Figure 6.7. In summary, incorporating more comprehensive data through linkages enhances calibration, with the LASSO model emerging as the most effective in producing fair predictions across ethnic groups. This suggests that regularization techniques like LASSO improve fairness by minimizing disparities between groups. However, calibration alone is not the sole criterion for fairness; metrics such as accuracy equity and error rate balance also need consideration to holistically evaluate the model's performance (Chouldechova et al., 2018; Corbett-Davies et al., 2009). This strongly suggests that our analysis should extend beyond calibration to assess the broader fairness implications of the predictive risk models.

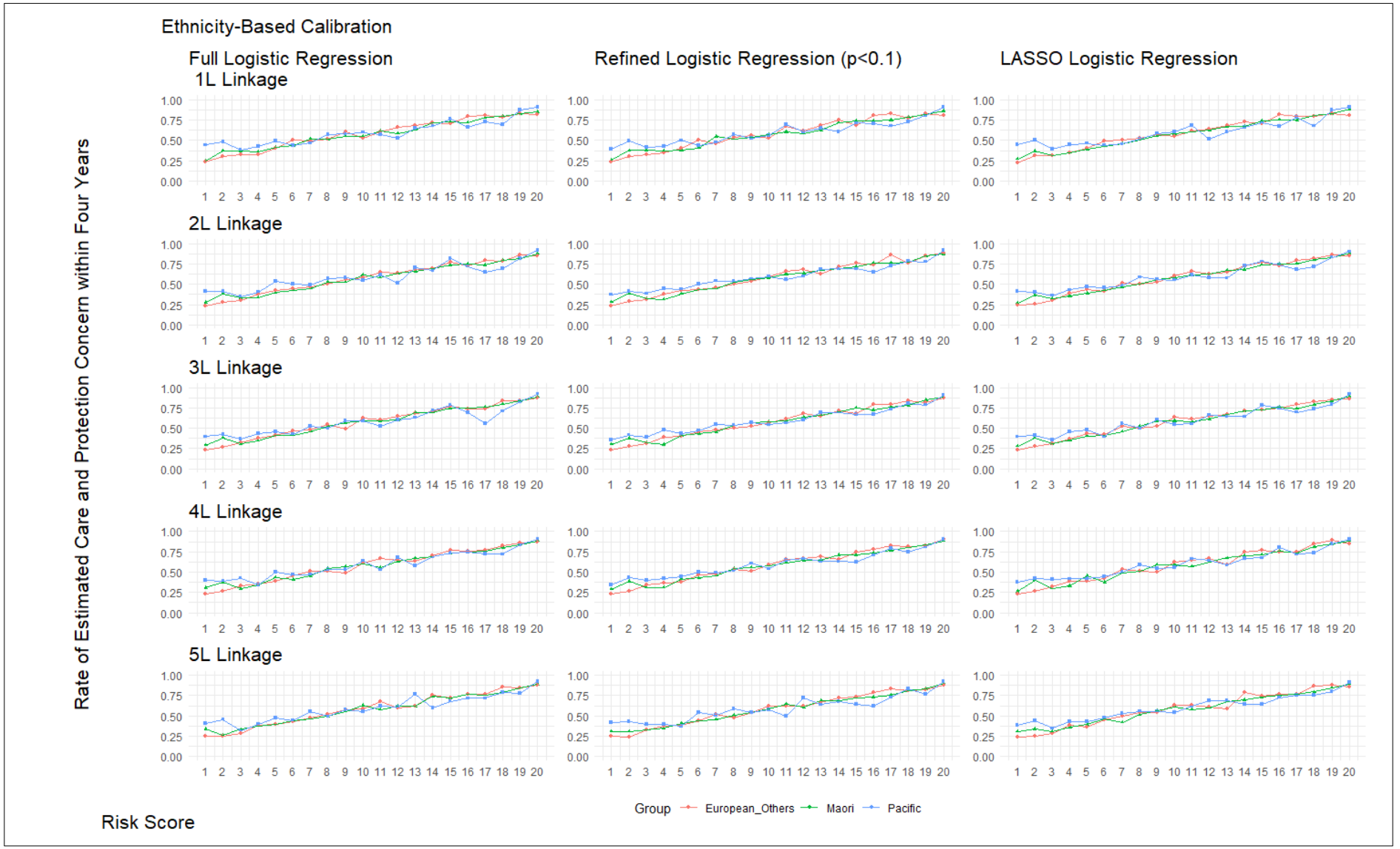

Figure 6.6: The rate of observed *estimated care and protection concern* within four years for Māori, Pacific and NZ European and Others ethnic groups across five linkage levels (1L-5L) for three logistic regression models: full logistic regression, refined logistic regression (p< 0.1), and LASSO logistic regression.

Note: At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as *refined logistic regression (p<0.1)*.

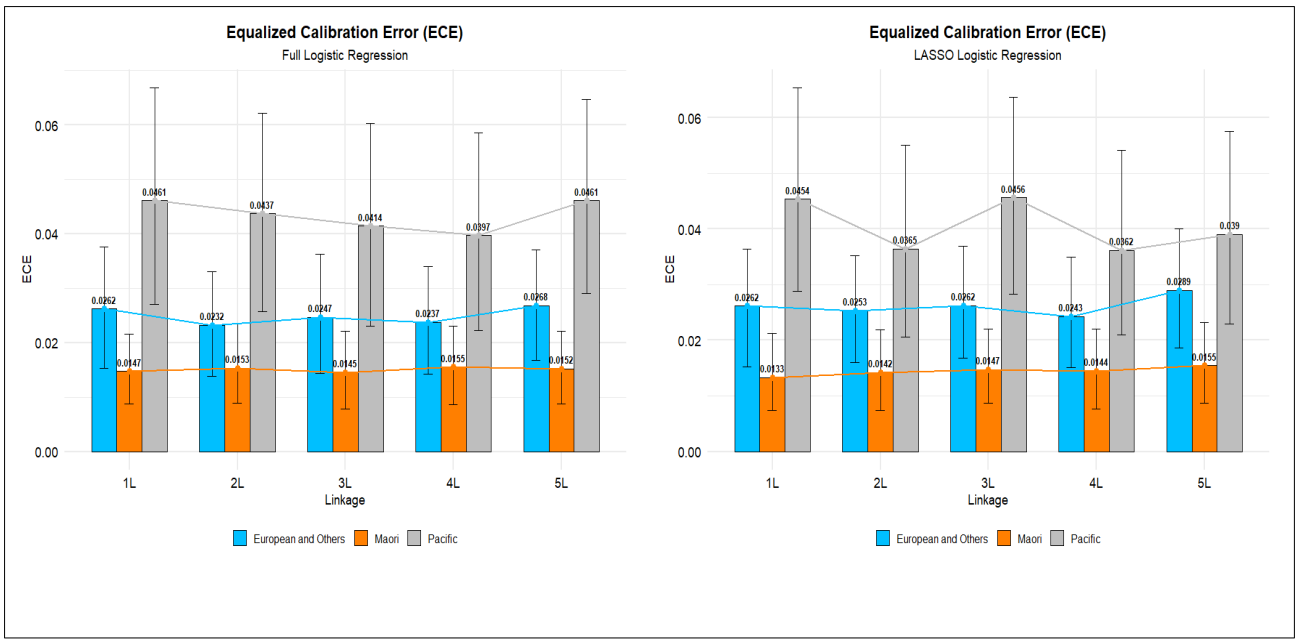

Figure 6.7: Equalized Calibration Error (ECE) across data linkages for full logistic regression and LASSO logistic regression models. The bars represent the mean ECE for each group, with error bars denoting the 95% confidence intervals (CI). Lines connecting the groups highlight the trends in ECE disparities as more comprehensive data is incorporated through higher linkage levels.

6.4.2.2 Accuracy Equity

In evaluating model performance across different data linkage strategies, a comparative analysis was conducted with a focus on accuracy equity across ethnic groups. Accuracy equity, as defined by Dieterich et al. (2016), extends the evaluation beyond fixed thresholds by considering the model's discriminative ability across the entire risk scale. This approach helps to establish fair assessment of the model's ability to distinguish between outcomes across all decision thresholds.

By assessing the difference in AUC values with respect to *ethnicity*, a comprehensive evaluation of accuracy equity is achieved. Table 6.9 presents AUC values for the three baseline logistic regression models: full logistic regression, refined logistic regression, and LASSO logistic regression across ethnic groups (Māori, Pacific, and NZ European and Others) and the five data linkage approaches (1L - 5L). This comparison reveals disparities in model performance across groups. Figure 6.8 visualizes these AUC gaps, illustrating the disparities in accuracy equity between pairs of ethnic groups for each model and linkage method.

Across all models and linkages, Māori and NZ European and Others show relatively small disparities in AUC values, particularly in Linkages 2L and 3L, where the inclusion of Benefit Dynamics data and Sentencing and Remand data results in more equitable predictive performance. For Māori, AUC values improve with more comprehensive linkages, increasing from 0.6993 in Linkage 1L to 0.7032

in Linkage 5L with the logistic regression model, and from 0.6872 to 0.7037 with the LASSO model. This suggests that more comprehensive data better captures predictive patterns for Māori children, reducing the gap between Māori and NZ European and Others, particularly with the LASSO model.

However, persistent disparities are observed for Pacific children. Their AUC values remain the lowest across all models and linkages, starting at 0.6405 in Linkage 1L and improving slightly to 0.6626 in Linkage 5L. Even with the LASSO model, where overall performance improves, Pacific children's AUC values remain significantly lower than those for Māori and NZ European and Others. Figure 6.8 highlights that the disparity in AUC values between Pacific and NZ European and Others is the largest, particularly in Linkage 1L. Although this gap narrows in higher-level linkages, the disparity persists, indicating that even with more comprehensive data, the models struggle to achieve accuracy equity for Pacific children.

A potential factor contributing to this challenge is the unbalanced nature of the dataset, where Pacific children are notably under-represented. In the 2017 Sample Cohort, Pacific children make up only 11%, and across all linkages, they comprise just 7% to 11% of the available records, as shown in Table 6.8. This under-representation results in the models having fewer examples from which to learn patterns specific to Pacific children, which affects their ability to make accurate predictions for this group. As a result, even though the inclusion of more data sources through the data linkage approaches improves accuracy equity overall, the limited data for Pacific children constrains the model's capacity to generalize effectively for this group, leading to lower predictive performance.

The imbalance in the data highlights a broader issue in predictive modeling, where under-represented groups, like Pacific children, may not benefit equally from model improvements, even when additional data is incorporated. Addressing this imbalance, through adjustment techniques during model training, could be key to improving accuracy equity for Pacific children.

Table 6.9: AUC results with 95% confidence intervals stratified by ethnicity across 1L-5L linkages for candidate logistic regression models.

Linkage	Group	AUC [Lower CI , Upper CI]	AUC [Lower CI , Upper CI]	AUC [Lower CI , Upper CI]
		Full Logistic Regression	**Refined Logistic Regression (p<0.1)**	**LASSO Logistic Regression**
1L	Māori	0.6993 [0.6881 , 0.7104]	0.6852 [0.6738 , 0.6965]	0.6872 [0.6759 , 0.6984]
	Pacific	0.6405 [0.6152 , 0.6658]	0.6360 [0.6105 , 0.6614]	0.6400 [0.6146 , 0.6653]
	NZ European & Others	0.7158 [0.7022 , 0.7294]	0.7144 [0.7008 , 0.7280]	0.7164 [0.7028 , 0.7299]
2L	Māori	0.7005 [0.6893 , 0.7116]	0.6987 [0.6876 , 0.7099]	0.6997 [0.6886 , 0.7108]
	Pacific	0.6600 [0.6350 , 0.6849]	0.6534 [0.6283 , 0.6785]	0.6586 [0.6336 , 0.6836]
	NZ European & Others	0.7207 [0.7072 , 0.7341]	0.7203 [0.7068 , 0.7338]	0.7213 [0.7079 , 0.7348]
3L	Māori	0.7010 [0.6899 , 0.7121]	0.6998 [0.6887 , 0.7110]	0.7008 [0.6897 , 0.7119]
	Pacific	0.6605 [0.6356 , 0.6854]	0.6535 [0.6284 , 0.6785]	0.6594 [0.6345 , 0.6844]
	NZ European & Others	0.7211 [0.7076 , 0.7345]	0.7212 [0.7078 , 0.7347]	0.7219 [0.7084 , 0.7353]
4L	Māori	0.7032 [0.6921 , 0.7143]	0.7000 [0.6888 , 0.7111]	0.7013 [0.6902 , 0.7124]
	Pacific	0.6639 [0.6390 , 0.6887]	0.6571 [0.6321 , 0.6821]	0.6629 [0.6380 , 0.6878]
	NZ European & Others	0.7233 [0.7099 , 0.7367]	0.7235 [0.7101 , 0.7369]	0.7241 [0.7107 , 0.7375]
5L	Māori	0.7032 [0.6921 , 0.7143]	0.7026 [0.6915 , 0.7137]	0.7037 [0.6927 , 0.7148]
	Pacific	0.6626 [0.6377 , 0.6875]	0.6538 [0.6287 , 0.6789]	0.6611 [0.6362 , 0.6860]
	NZ European & Others	0.7260 [0.7126 , 0.7394]	0.7245 [0.7111 , 0.7379]	0.7263 [0.7129 , 0.7397]

Note: At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as refined logistic regression (p<0.1).

Figure 6.8: Ethnic disparities in AUC values for candidate logistic regression models: full logistic regression, refined logistic regression (p<0.1), and LASSO logistic regression.

Note: At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as refined logistic regression (p<0.1).

6.4.2.3 Error Rate Balance

This section presents the results from assessing error rate balance based on predictions obtained from the candidate baseline logistic regression models across the five data linkage approaches (1L-5L). This fairness metric evaluates whether the predictive models treat all groups equally in terms of prediction errors, indicating that no group disproportionately suffers from higher false positive or false negative rates. As explained in Section 5.3.3.2, achieving error rate balance, also known as equalized odds, requires that both FPRs and FNRs be similar across ethnic groups. Failure to achieve this balance could indicate potential bias, disproportionately affecting certain ethnic groups in predictive decision-making.

To quantify the error rate balance when the protected variable, ethnicity, has more than two levels, we followed the approach outlined by Purdy and Glass (2023), calculating pairwise error rate ratios for FPR and FNR across all subgroup pairings. The ratio is computed by dividing the smaller error rate by the larger one, producing values between 0 and 1. A ratio of 1 or close to 1 indicates equal error rates, while values closer to 0 reflect greater disparities. To highlight the most significant disproportionality, we focused on the minimum value of these ratios.

Table 6.10 shows the FNR and FPR for various ethnic groups across three models: logistic regression, refined logistic regression, and LASSO logistic regression, each applied to five data linkages (1L-5L). The left side of Table 6.10 provides the FNR and FPR for each ethnic group, while the right side presents the pairwise ratios of FNRs and FPRs across all subgroup pairings. Ratios are always constructed with the larger error rate in the denominator, resulting in values between 0 and 1. For example, the pairwise FNR ratio between Māori and NZ European and Others in Table 6.10 is 0.3747, calculated by dividing the smaller FNR (Māori) by the larger FNR (NZ European and Others), i.e., $\frac{0.1434}{0.3826} \approx 0.3747$. This value indicates that Māori children are incorrectly predicted to have no *estimated care and protection concern* (false negatives) at 37.47% of the rate of NZ European and Others. In this case, NZ European and Others children have a higher FNR and are more likely to be misclassified, meaning the model is more prone to missing care and protection concerns for NZ European and Others children compared to Māori children.

Similarly, the pairwise FPR ratio between Māori and NZ European and Others in Table 6.10 for the full logistic regression (1L) is 0.4456, calculated by dividing the smaller FPR (NZ European and Others) by the larger FPR (Māori), as $\frac{0.2918}{0.6547} \approx 0.4456$. This indicates that the false positive rate for NZ European and Others children is approximately 44.56% of the false positive rate for Māori children,

suggesting that Māori children are more likely to be incorrectly predicted to have care and protection concerns (false positives) compared to NZ European and Others children.

While each ratio highlights potential algorithmic unfairness at the average-risk threshold (the standard 50% threshold for binary classification), a comprehensive assessment requires considering all six pairwise error rate ratios across models and data linkage approaches. Ratios closer to 1 indicate smaller disparities, while lower ratios reflect greater disproportionality. To identify the model with the least disparity, the minimum error rate ratios across models and linkages are compared, with the highest minimum ratio signifying the least disparity. Based on this approach, the full logistic regression model trained on predictors through the 5L linkage achieves the most balanced error rates, exhibiting the lowest disparities across ethnic groups (Table 6.10).

The most significant finding from Table 6.10 is that the error rate ratios (both FNR and FPR) between Pacific and NZ European and Others groups are consistently higher, indicating less disparity in error rates for these two groups across all linkages and models (FNR ratio: 0.89-0.90 and FPR ratio: 0.64-0.65). This suggests that Pacific and NZ European and Others groups have more balanced error rates, with the smallest disparities, particularly in the more comprehensive linkages (4L and 5L). These results highlight that, compared to other ethnic groups, Pacific and NZ European and Others exhibit a more equitable distribution of misclassification errors.

Overall, examining error rate balance across data linkages suggests that while the incorporation of more comprehensive linkages (such as 4L and 5L) slightly improves error rate balance across ethnic groups, disparities still persist, particularly between Māori and NZ European and Others. Notably, all ratios fall below the 0.80 fairness threshold according to the 80% rule, with only the FNR ratio for Pacific and NZ European and Others exceeding it. This underscores the need for continued efforts to address these disparities in predictive models.

Table 6.10: Error rate ratios stratified by ethnicity across 1L-5L linkages for candidate logistic regression models: full logistic regression, refined logistic regression (p<0.1), and LASSO logistic regression.

Linkage	Group	FNR	FPR	Group Pair	FNR Ratio	FPR Ratio
Full Logistic Regression						
1L	Māori	0.1434	0.6547	Māori vs European and Others	0.3747	0.4456
	NZ European and Others	0.3826	0.2918	Māori vs Pacific	0.4227	0.6934
	Pacific	0.3392	0.4540	NZ European and Others vs Pacific	0.8865	0.6427
2L	Māori	0.1387	0.6259	Māori vs NZ European and Others	0.3638	0.4609
	NZ European and Others	0.3811	0.2885	Māori vs Pacific	0.4222	0.7048
	Pacific	0.3284	0.4412	NZ European and Others vs Pacific	0.8618	0.6539
3L	Māori	0.1418	0.6213	Māori vs NZ European and Others	0.3703	0.4620
	NZ European and Others	0.3830	0.2871	Māori vs Pacific	0.4279	0.7121
	Pacific	0.3314	0.4425	NZ European and Others vs Pacific	0.8652	0.6488
4L	Māori	0.1404	0.6201	Māori vs NZ European and Others	0.3652	0.4594
	NZ European and Others	0.3845	0.2849	Māori vs Pacific	0.4314	0.7094
	Pacific	0.3255	0.4399	NZ European and Others vs Pacific	0.8466	0.6476
5L	Māori	0.1429	0.6201	Māori vs NZ European and Others	0.3835	0.4548
	NZ European and Others	0.3725	0.2820	Māori vs Pacific	0.4324	0.7156
	Pacific	0.3304	0.4437	NZ European and Others vs Pacific	0.8869	0.6355
Refined Logistic Regression (entry criteria p<0.1)						
1L	Māori	0.1409	0.6578	Māori vs NZ European and Others	0.3665	0.4419
	NZ European and Others	0.3845	0.2907	Māori vs Pacific	0.4143	0.6940
	Pacific	0.3402	0.4565	NZ European and Others vs Pacific	0.8847	0.6367
2L	Māori	0.1359	0.6287	Māori vs NZ European and Others	0.3593	0.4595
	NZ European and Others	0.3781	0.2889	Māori vs Pacific	0.4017	0.7241
	Pacific	0.3382	0.4552	NZ European and Others vs Pacific	0.8944	0.6345
3L	Māori	0.1366	0.6247	Māori vs NZ European and Others	0.3562	0.4601
	NZ European and Others	0.3834	0.2874	Māori vs Pacific	0.4133	0.7410
	Pacific	0.3304	0.4629	NZ European and Others vs Pacific	0.8618	0.6209
4L	Māori	0.1392	0.6238	Māori vs NZ European and Others	0.3631	0.4556
	NZ European and Others	0.3833	0.2842	Māori vs Pacific	0.4151	0.7114
	Pacific	0.3353	0.4437	NZ European and Others vs Pacific	0.8747	0.6404
5L	Māori	0.1362	0.6293	Māori vs NZ European and Others	0.3578	0.4579
	NZ European and Others	0.3808	0.2881	Māori vs Pacific	0.3959	0.6990
	Pacific	0.3441	0.4399	NZ European and Others vs Pacific	0.9037	0.6550
LASSO Logistic Regression						
1L	Māori	0.1420	0.6559	Māori vs NZ European and Others	0.3707	0.4459
	NZ European and Others	0.3830	0.2925	Māori vs Pacific	0.4247	0.6960
	Pacific	0.3343	0.4565	NZ European and Others vs Pacific	0.8729	0.6407
2L	Māori	0.1373	0.6308	Māori vs NZ European and Others	0.3648	0.4637
	NZ European and Others	0.3763	0.2925	Māori vs Pacific	0.4155	0.6953
	Pacific	0.3304	0.4386	NZ European and Others vs Pacific	0.8781	0.6668
3L	Māori	0.1374	0.6250	Māori vs NZ European and Others	0.3585	0.4657
	NZ European and Others	0.3834	0.2910	Māori vs Pacific	0.4111	0.7079
	Pacific	0.3343	0.4425	NZ European and Others vs Pacific	0.8720	0.6578
4L	Māori	0.1385	0.6225	Māori vs NZ European and Others	0.3617	0.4582
	NZ European and Others	0.3829	0.2852	Māori vs Pacific	0.4192	0.6858
	Pacific	0.3304	0.4269	NZ European and Others vs Pacific	0.8629	0.6682
5L	Māori	0.1362	0.6259	Māori vs NZ European and Others	0.3683	0.4569
	NZ European and Others	0.3699	0.2860	Māori vs Pacific	0.4123	0.7069
	Pacific	0.3304	0.4425	NZ European and Others vs Pacific	0.8932	0.6463

Note: At each linkage level, the refined logistic regression was trained using only the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as refined logistic regression (p<0.1).

6.4.3 Summary of Findings

The analysis of incorporating additional predictor variables through various data linkage approaches (1L-5L) revealed that, across all models, the LASSO logistic regression consistently achieved the highest AUC values, indicating slightly better discriminatory power. The inclusion of additional predictor variables slightly enhanced the model's accuracy, with the most improvement observed in linkage 5L.

Cumulative accuracy plots in Figure 6.4 showed that models using the expanded set of predictors from 5L data linkage outperformed our competitor models using only Child, Youth and Family, Children's Action Plan, and Demographic Details data (1L), particularly in the highest risk groups, where the accuracy remained above 85% compared to approximately 80% for models with fewer predictors (1L). Incorporating more comprehensive data from various sources improved the model's overall accuracy by 1% from linkage 1L to linkage 5L. While this improvement may seem modest, it translates to correctly identifying the risks for approximately 162 additional children. In the context of child welfare, even small gains in accuracy are crucial due to the significant impact that incorrect risk assessments can have on children and their families. Ensuring the highest possible accuracy in our models is vital to minimize potential harmful effects and to provide the most reliable support and interventions.

Furthermore, the evaluation of calibration across ethnic groups revealed that more comprehensive data linkages (4L and 5L) effectively reduced disparities, with the LASSO model consistently demonstrating the best calibration performance. In terms of accuracy equity, the LASSO model, again, outperformed both the full and refined logistic regression models when applied to the 5L data linkage. Despite these improvements, disparities in accuracy equity persist, particularly with Pacific children, who continue to show the lowest AUC values. While error rate balance improved with the addition of more comprehensive data, significant disparities remain, especially between Māori and NZ European and Others.

In conclusion, incorporating additional predictors through data linkages had a positive, though modest, impact on both accuracy and fairness, particularly in predictive models trained on higher linkage levels. Models utilizing 4L and 5L linked data consistently outperformed those with fewer predictors. The LASSO logistic regression model, in particular, demonstrated better calibration and consistency across ethnic groups. Consequently, the 5L linkage will be adopted for the final model analysis moving forward, given its demonstrated ability to enhance model performance.

6.5 Predictive Risk Modeling

This section presents findings from employing advanced machine learning algorithms, including random forest, support vector machine, and XGBoost, alongside baseline logistic regressions, to evaluate their performance in terms of accuracy and potential predictive bias against Māori children. The rationale for incorporating these additional models is to determine whether they can better capture the unique risk factors and complexities associated with different demographic groups, ultimately enhancing both predictive accuracy and fairness (see Section 5.3.2). The fairness evaluation of the models in this phase focused on potential bias in predictions between Māori and non-Māori populations, reflecting the central concern of this thesis regarding ethnic disparities in child welfare predictions.

Building on the results from Section 6.2 and Section 6.4, the models developed were trained using the final set of 250 predictors derived from the 5L data linkage strategy (see Section 5.3.1). The models were trained on 70% of children randomly selected from the Sample Cohort 2017 to predict the probability of at least one care and protection-related event occurring within four years from the initial child notification (*estimated care and protection concern* within four years). These models were then internally tested on the remaining 30% of the sample (n=16,233). Drawing from the lessons learned during the development of the Allegheny Family Screening Tool and the issues identified with its model validation process, as outlined by Chouldechova et al. (2018), we took deliberate steps to prevent over-optimism in model performance. Specifically, we ensured that children reported at the same time (such as siblings) were not split between the training and testing sets, thus maintaining the integrity of the validation process.

Furthermore, to ensure robust validation, the candidate models were externally tested on an independent cohort of children who were newly notified between April 1, 2018, and March 31, 2019 (n=53,997). This external validation step was crucial for evaluating the robustness of the models beyond the initial training data.

6.5.1 Accuracy

Table 6.11 presents the performance results of the methods considered at this stage, based on accuracy metrics discussed in Section 5.3.3.1. Logistic regressions, including the full, refined, LASSO, Elastic Net, and Ridge, exhibit consistent performance with AUC values around 0.72 when tested on the randomly selected 30% of children from the 2017 sample cohort. These models also demonstrate similar TPR values, approximately 0.77-0.78, indicating a strong ability to correctly identify

at-risk cases. The F1 scores across these models are also consistent, typically around 0.73, reflecting a good balance between precision (PPV) and recall (TPR). However, when these models were applied to the 2018 cohort, a slight reduction in AUC to around 0.71 was observed, alongside a marginal decrease in sensitivity and F1 score (see Table tab12).

In contrast, the advanced machine learning algorithms, particularly XGBoost, consistently outperform the logistic regressions across all metrics. For the 2017 cohort, XGBoost achieved the highest AUC (0.74) and TPR (0.79), indicating that the model captures more at-risk cases. The F1 score for XGBoost is also superior, at 0.74, reflecting a more balanced model with better overall predictive accuracy. Even when tested on the 2018 cohort, XGBoost maintains strong predictive power with an AUC of 0.73 and a TPR of 0.77. Other advanced models, such as random forest and support vector machines, also outperformed the baseline logistic regression models, achieving AUC values approximately 1% higher.

The observed AUC decrease of 1% across all models when testing on children from the Sample Cohort 2018 is both expected and reasonable. Natural variations in data distribution or shifts in patterns and population characteristics between the two years are likely contributors (see Section 6.6.1.5 for more details). These variations may be influenced by demographic changes, reporting behaviors, or other external factors, which can affect the models' ability to generalize across different time periods (Guo et al., 2021). Despite this slight decline, the models demonstrate robust performance, maintaining fair predictive capabilities in the face of such natural fluctuations.

Table 6.11: Predictive performance results of candidate models on the Sample Cohort 2017 (internal testing data) and the Sample Cohort 2018 (external testing data), based on the standard 50% threshold for binary classification.

Model	AUC	AUC 95% CI		PPV	NPV	TPR	Accuracy	F1
		Lower	Upper					
Testing on 30% of the Sample Cohort 2017 (n=16,233)								
Full Logistic Regression	0.7190	0.7112	0.7269	0.6977	0.6298	0.7717	0.6734	0.7328
Refined Logistic Regression (p<0.1)	0.7175	0.7096	0.7254	0.6948	0.6275	0.7720	0.6709	0.7313
LASSO Logistic Regression	0.7193	0.7114	0.7272	0.6970	0.6330	0.7764	0.6744	0.7346
Elastic Net Logistic Regression	0.7193	0.7114	0.7271	0.6966	0.6323	0.7762	0.6739	0.7342
Ridge Logistic Regression	0.7189	0.7110	0.7268	0.6952	0.6314	0.7768	0.6728	0.7337
Random Forest	0.7259	0.7181	0.7337	0.6988	0.6347	0.7764	0.6760	0.7356
Support Vector Machine	0.7235	0.7156	0.7313	0.6995	0.6336	0.7741	0.6759	0.7349
XGBoost	0.7386	0.7310	0.7462	0.7012	0.6497	0.7920	0.6835	0.7439
Testing on Samp Cohort 2018 (n=53,997)								
Full Logistic Regression	0.7061	0.7017	0.7105	0.6710	0.6345	0.7759	0.6584	0.7197
Refined Logistic Regression (p<0.1)	0.7063	0.7020	0.7107	0.6704	0.6366	0.7797	0.6588	0.7209
LASSO Logistic Regression	0.7075	0.7032	0.7119	0.6706	0.6374	0.7803	0.6592	0.7213
Elastic Net Logistic Regression	0.7076	0.7032	0.7119	0.6706	0.6374	0.7803	0.6592	0.7213
Ridge Logistic Regression	0.7073	0.7029	0.7116	0.6712	0.6378	0.7801	0.6597	0.7216
Random Forest	0.7198	0.7155	0.7241	0.6956	0.6267	0.7698	0.6710	0.7308
Support Vector Machine	0.7191	0.7148	0.7234	0.6965	0.6250	0.7664	0.6707	0.7298
XGBoost	0.7275	0.7232	0.7318	0.6815	0.6421	0.7723	0.6674	0.7241

Note 1: The refined logistic regression was trained using the predictors from the full model that met the entry criterion of p-values less than 0.1 ($p < 0.1$). We refer to this as refined logistic regression (p<0.1).

Note 2: The results for the baseline logistic regression models (full, refined, LASSO, Elastic Net, and Ridge variants), tested on 30% of the 2017 sample cohort, are identical to those shown in Table 6.7 for the 5L data linkage approach.They have been included here again for clearer visual comparison.

6.5.2 Predictive Bias

To assess the predictive bias or fairness of the models trained in this stage, the analysis focused on three models: full logistic regression, LASSO, and XGBoost. These models were selected for further analysis based on their performance across different modeling approaches (see Table 6.11). LASSO was chosen as a candidate from the regularized methods, emerging as the top-performing model with higher AUC and F1 scores. Full logistic regression, while simpler, showed similar AUC and F1 scores, highlighting its reliability and interpretability, which are key factors in understanding the impact of various predictors on model outcomes. XGBoost was selected from the advanced machine learning algorithms, as it achieved the highest AUC, TPR, and F1 scores, making it the top performer in terms of predictive accuracy.

The fairness of these three models was evaluated using a set of fairness metrics, including calibration, accuracy equity, statistical parity, and equalized odds. The rationale behind selecting these measures is detailed in Section 5.3.3.2. Calibration is employed to assess whether the predicted probabilities accurately correspond to the observed occurrence of care and protection-related events across ethnic groups. This is critical to ensuring that the models are properly calibrated for both Māori and non-Māori children, thereby supporting equitable and reliable predictions across these populations. Accuracy equity examines whether the overall accuracy, as measured by AUC, is consistent across these groups. Statistical parity evaluates whether the models treat Māori and non-Māori children equally in terms of their predicted outcomes, specifically whether the referral rates for children with an intake outcome are similar across these groups. Finally, equalized odds are analyzed to determine if the models maintain comparable TPRs and FPRs across groups. This consideration is particularly important given the significant concerns surrounding the use of predictive risk modeling within the child welfare context.

The empirical findings from the fairness assessment of the models are presented in the following sections. These results are based on a validation sample of 16,233 children from the Sample Cohort 2017, which was not included in the training of the predictive models.

6.5.2.1 Calibration

Figure 6.9 presents the observed *estimated care and protection concern* rates for children across various risk scores assigned by the model, stratified by *ethnicity* (Māori vs. Non-Māori). Risk scores were derived by dividing the predicted probabilities produced by models into 20 equally distributed risk ventiles, with a score of 20 representing the top 5% of predicted probabilities and a score of 1

representing the bottom 5%.

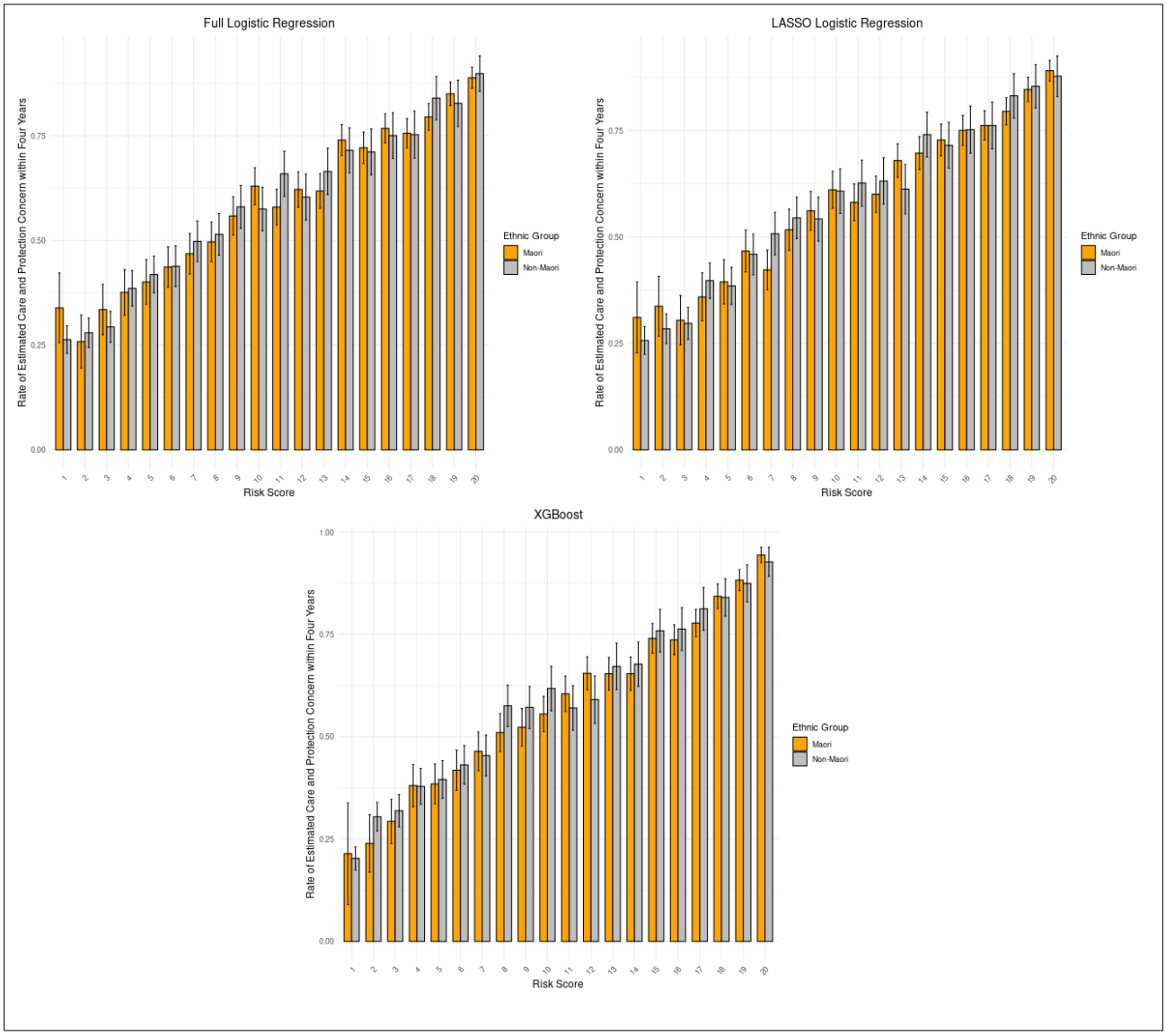

Figure 6.9: Observed *estimated care and protection concern* by risk ventile based on predicted probabilities from full logistic regression, LASSO logistic regression, and XGBoost, broken down by child's ethnic group (Māori vs. Non-Māori). Error bars correspond to 95% confidence intervals.

In evaluating the calibration of the models for Māori and Non-Māori children, the barplots indicate that, overall, the models are well-calibrated, as the observed care and protection concern rates closely match the predicted estimates for both groups. However, slight calibration differences emerge in the lower and middle score ventiles, particularly in the full logistic regression and LASSO models, suggesting potential biases that could result in unequal treatment. For instance, in the lowest ventile (score of 1) on the full logistic regression model, Māori children have an *estimated care and protection concern* rate of approximately 30%, compared to 25% for Non-Māori children. This indicates that the model may slightly underestimate the risk for Māori children in the lower risk ranges, potentially leading to discrepancies in how risk is assessed between the two groups. Similarly, miscalibrations

are observed in the middle ventiles (scores 10-13) in both the full logistic regression and LASSO, although the magnitude of these discrepancies is less pronounced in the LASSO model. These patterns suggest that while the models generally perform well, certain score ranges require further refinement to ensure calibration accuracy and fairness in risk assessments across ethnic groups.

In contrast, the XGBoost model exhibits better calibration across all score ventiles, with observed rates for Māori and Non-Māori children remaining closely aligned. However, even in XGBoost, slight miscalibration is detected in the highest ventiles (scores 19-20), where Māori children scoring 20 have an *estimated care and protection concern* rate of 94%, compared to 88% for Non-Māori children. While this difference is relatively small, it highlights that even advanced models require careful attention to calibration across all risk levels to prevent subtle biases from impacting decisions.

6.5.2.2 Accuracy Equity

Building upon the theoretical framework of accuracy equity discussed earlier in Section 5.3.3.2, we present the empirical findings from the candidate models across Māori and Non-Māori children. Figure 6.10 shows the ROC curves stratified by *ethnicity* for the models.

Across all models, Non-Māori children tend to have higher AUC values than Māori children, reflecting a slight but consistent discrepancy in accuracy equity. The XGBoost model demonstrates the highest performance in terms of reducing this disparity, but even in this advanced model, Non-Māori children benefit from slightly higher predictive accuracy.

To statistically evaluate these differences, we applied the DeLong test to compare the AUC values between Māori and Non-Māori children for each model (DeLong et al., 1988). The null hypothesis of the DeLong test is that there is no difference between the AUCs of the two ROC curves being compared. The test provides a D statistic and a p-value to assess the statistical significance of the observed differences in AUC. The results from implementing the test in R are presented in Table 6.12.

From Table 6.12, we observe that for the XGBoost model which shows the highest overall performance, Māori children have an AUC of 0.7254 compared to 0.7341 for Non-Māori children, with a D statistic of -1.098 and a p-value of 0.2722. This indicates that while there is a slight performance disparity in favor of Non-Māori children, the difference is not statistically significant. The LASSO logistic regression model shows similar results, with an AUC of 0.7037 for Māori children and 0.7135 for Non-Māori children, resulting in a D statistic of -1.192 and a p-value of 0.2333. Again, the difference

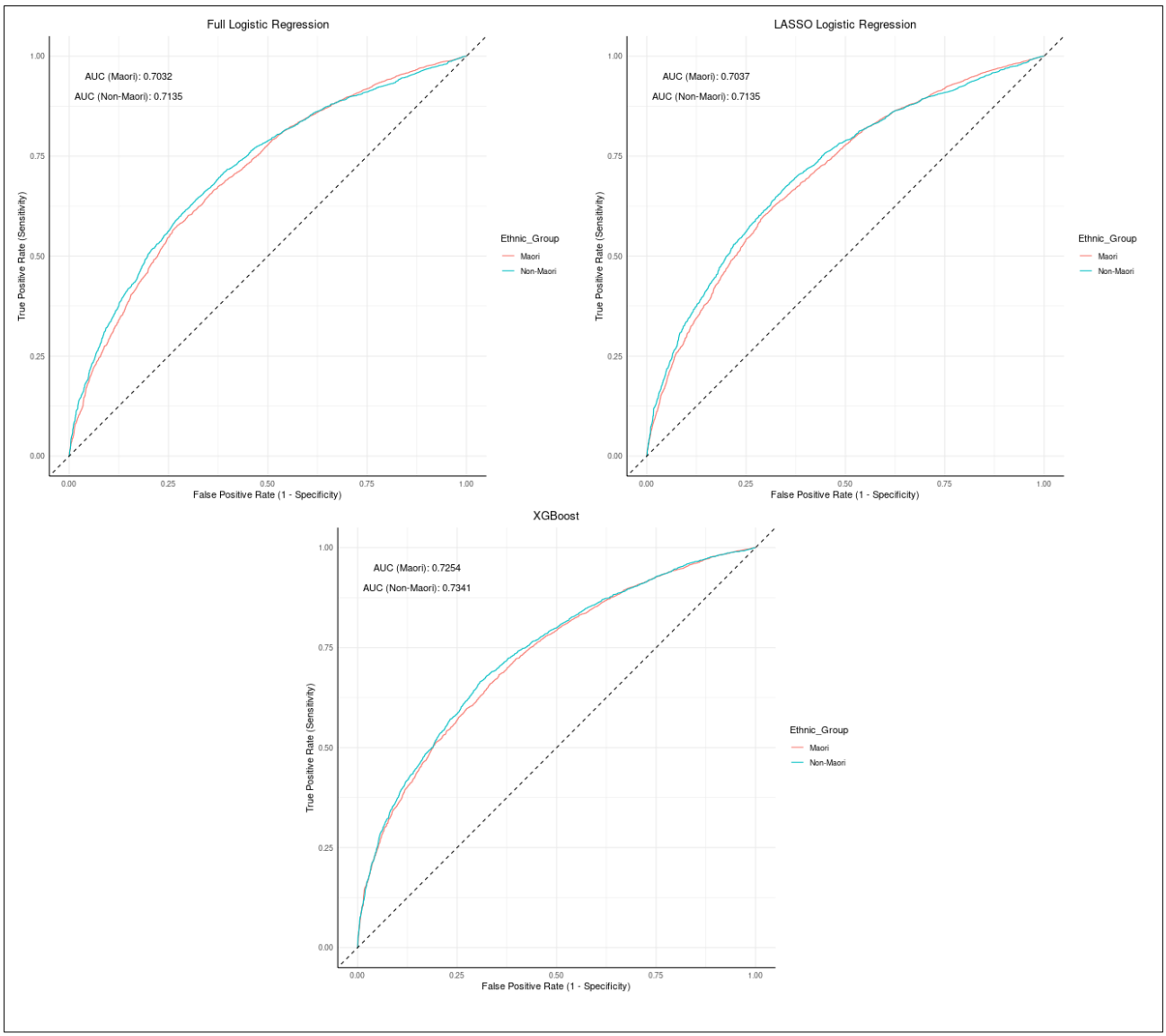

Figure 6.10: ROC curves stratified by ethnic group (Māori (red) vs. Non-Māori (green)) for full logistic regression, LASSO logistic regression, and XGBoost.

is small, yet not statistically significant.

The full logistic regression model yields an AUC of 0.7032 for Māori children and 0.7135 for Non-Māori children. Although the AUC difference is more pronounced in this model, the D statistic of -1.2581 and the p-value of 0.2084 suggest that the difference is not statistically significant.

Overall, the DeLong test results across all models indicate that although Non-Māori children consistently have slightly higher AUC values than Māori children, these differences are not statistically significant. The findings suggest that the models perform comparably across the two groups, meaning that all three predictive risk models maintain fairness toward Māori if accuracy equity was considered the main measure of fairness approved by stakeholders in the child welfare context.

Table 6.12: AUC and DeLong statistical test results for Māori and Non-Māori across candidate models.

Note: A p-value > 0.05 indicates no statistically significant difference in AUC values.

Model	Group	AUC [95% CI]	D	p-value
Full Logistic Regression	Māori	0.7032 [0.6921, 0.7143]	-1.2581	0.2084
	Non-Māori	0.7135 [0.7018, 0.7253]		
LASSO Logistic Regression	Māori	0.7037 [0.6927, 0.7148]	-1.192	0.2333
	Non-Māori	0.7135 [0.7018, 0.7253]		
XGBoost	Māori	0.7254 [0.7148, 0.7360]	-1.098	0.2722
	Non-Māori	0.7341 [0.7227, 0.7455]		

6.5.2.3 Disparate Impact and Equalized Odd

To assess fairness in addition to calibration and accuracy equity, we used two key metrics: disparate impact (DI) and equalized odds also referred to as error rate balance. These metrics offer a comprehensive view of how models perform across different groups, focusing on both bias and equity in predictions (Section 5.3.3.2).

Disparate impact (DI) specifically examines whether the model's intake decision rate (the rate of children referred for further intervention by the models) differ significantly across groups, leading to potential disparities in the decision-making process by these models. In binary classification, equalized odds is synonymous with equality of opportunity (EOO) as it assesses the positive predicted outcomes ($\hat{y} = 1$) for both true positive ($y = 1$) and true negative ($y = 0$) cases (Hardt et al., 2016). By measuring EOO for each outcome, EOO(y=0) for the negative outcome and EOO(y=1) for the positive outcome, we can effectively address the requirements of equalized odds. Subsequently, if equality of opportunity is maintained for both outcomes, the model satisfies the criterion of equalized odds.

Table 6.13 provides a detailed examination of intake rates, TPRs, and FPRs, stratified by ethnic groups, specifically Māori children versus Non-Māori, based on the predicted probabilities from full logistic regression, LASSO, and XGBoost models, all evaluated at a standard threshold of 50%. The data reveals a consistent pattern across all models, where Māori children are referred for intake by the models at substantially higher rates, approximately 77% to 78%, compared to children from other ethnic groups. This disparity suggests a potential bias in the models, as they systematically identify a larger proportion of Māori children as needing further intervention. The elevated TPRs for Māori, ranging from 0.857 to 0.871, indicate that the models are more sensitive in correctly identifying true positives within this group. However, this increased sensitivity comes at the cost of higher FPRs, which are also consistently greater for Māori across all models (0.620 to 0.631). Conversely, the

lower TPRs for Non-Māori (ranging from 0.639 to 0.669) could disadvantage children from these other ethnic groups, by under-identifying those who genuinely require intervention. This imbalance in both TPR and FPR across ethnic groups raises significant concerns about the fairness and equity of these predictive models in child welfare decision-making, potentially leading to both over-intervention for Māori children and under-intervention for children from other ethnicities.

Applying the two-proportion z-test revealed that these differences in intake rate, TPR and FPR are statistically significant between Māori children and children from other ethnic groups (p-value < 2.2e-16).

Table 6.13: Candidate models intake rate, TPR, and FPR for Māori and Non-Māori groups based on standard 50% threshold for binary classifications.

Model	Group	Intake	TPR	FPR
Full Logistic Regression	Māori	0.7711	0.8571	0.6201
	Non-Māori	0.4816	0.6391	0.3176
LASSO Logistic Regression	Māori	0.7774	0.8638	0.6259
	Non-Māori	0.4840	0.6410	0.3205
XGBoost	Māori	0.7841	0.8713	0.6311
	Non-Māori	0.4959	0.6692	0.3154

To quantify these disparities as measures of disparate impact and equalized odds, we employed the approach discussed in Section 5.3.3.2, where disparate impact (DI) is quantified using Equation (5.16) and equalized odds by evaluating EOO(y=0) using Equation (5.17), and EOO(y=1) using Equation (5.18). Table 6.14 presents these key fairness metrics across three different models: full logistic regression, LASSO logistic regression, and XGBoost. Disparate impact (DI) values are consistent across the models, with full logistic regression at 0.6246, LASSO at 0.6224, and XGBoost slightly higher at 0.6324. In terms of equality of opportunity for negative outcomes EOO(y=0), full logistic regression and LASSO perform similarly, with values of 0.5121 and 0.5077, respectively, while XGBoost trails slightly at 0.4997. These values suggest that Māori children are twice as likely to experience FPs compared to Non-Māori children. This result is crucial, if predictive models are more likely to falsely predict an intake decision for Māori children, which could be seen as a potential fairness issue. For positive outcomes EOO(y=1), XGBoost outperforms the others with a value of 0.7680, compared to 0.7458 for full logistic regression and 0.7414 for LASSO.

Overall, the models show comparable fairness metrics, with XGBoost's higher EOO(y=1) suggesting it may offer better equality of opportunity for positive outcomes. While these values indicate a reasonable level of fairness, they are still lower than the standard which is 80% in machine learning

Table 6.14: Fairness evaluation of candidate models based on statistical parity (disparate impact, DI) and equalized odds (EOO) for both negative outcomes EOO(y=0) and positive outcomes EOO(y=1).

Model	DI [95% CI]	EOO(y=0) [95% CI]	EOO(y=1) [95% CI]
Full Logistic Regression	0.6246 [0.6092, 0.6405]	0.5121 [0.4870, 0.5394]	0.7458 [0.7272, 0.7557]
LASSO Logistic Regression	0.6224 [0.6057, 0.6382]	0.5077 [0.4843, 0.5319]	0.7414 [0.7218, 0.7608]
XGBoost	0.6324 [0.6163, 0.6484]	0.4997 [0.4733, 0.5273]	0.7680 [0.7491, 0.7873]

literature (Feldman, 2015).

To claim that discrimination or unfairness does not exist based on these notions of fairness, statistical parity or equality of opportunity must be equal to 1. A value of equality of opportunity equal to 1 indicates that there is equal opportunity for every ethnic group for a given class. By measuring equality of opportunity for each class, we can assess equalized odds. Therefore, if equality of opportunity is appropriate for both the positive and negative classes, then equalized odds is satisfied. However, it is not expected that the models to achieve perfect statistical parity or equality of opportunity equal to 1. Instead, we aim to satisfy the 80% rule, which allows for some disparity as long as the outcomes for different groups do not fall below 80% of each other, ensuring a reasonable threshold for fairness

6.5.3 Summary of Findings

In this section, we examined both the predictive performance and fairness of various machine learning models, including full logistic regression, LASSO, and XGBoost, using predictors derived from the 5L data linkage strategy. While XGBoost consistently outperformed the logistic regression models in terms of AUC, TPR, and F1 scores, achieving the highest overall predictive accuracy, its improvement in fairness metrics was less significant.

Calibration and accuracy equity were not major concerns across the models, as the models showed relatively consistent performance for both Māori and Non-Māori groups. The observed differences in AUC between these groups were small and statistically insignificant, indicating that the models maintained a reasonable degree of accuracy equity (Table 6.12).

However, fairness issues arose when evaluating statistical parity and equalized odds. The results showed that, despite XGBoost's strong performance in predictive accuracy, it did not substantially improve fairness in terms of disparate impact or equalized odds. The models consistently exhibited disparities in TPR and FPR, with Māori children being identified at higher rates for intake compared to Non-Māori children, which raises concerns about potential bias in over-identifying Māori children for intervention. XGBoost demonstrated a marginal improvement in equality of opportunity for positive

outcomes EOO(y=1), but its performance on equality of opportunity for negative outcomes EOO(y=0) was comparable to or slightly worse than the logistic regression models.

Given that the logistic regression models, particularly full logistic regression and LASSO, offer similar levels of fairness while being more interpretable and easier to implement, they remain preferable, especially within the child welfare system, where simplicity and transparency are highly valued. Although advanced models like XGBoost offer higher predictive power, their added complexity does not translate into significant fairness improvements. Therefore, this research recommends continuing to prioritize logistic regression for predictive risk modeling in child welfare settings, as it balances predictive accuracy with interpretability and fairness.

6.6 Fairness-aware Machine Learning

This section presents the results of applying an in-processing fairness-aware machine learning approach, to tackle the unfairness issues identified during the predictive bias assessments of the candidate models in Section 6.5.2.

The analysis in Section 6.5.2.3 revealed persistent disparities in intake rates, with Māori children being referred for further intervention (intake) at significantly higher rates compared to children from other ethnic groups. They were also more frequently misclassified as having *estimated care and protection concern* within four years (FPR) across all models, including advanced algorithms like XGBoost (Table 6.13). These findings highlighted the need to explore alternative fairness approaches, as simply applying different models did not lead to significant improvements in fairness.

In response, our in-processing fairness-aware machine learning approach focused on addressing disparate impact and equalized odds, rather than calibration and accuracy equity, by integrating fairness constraints directly into the logistic regression model during the training phase. This approach involved reformulating the traditional learning task into a constrained optimization problem (see Section 5.4), ensuring that fairness considerations were embedded into the model's objective function from the outset.

While calibration and accuracy equity provide important insights into model performance across ethnic groups, they do not fully address the critical disparities in error rates, which seemingly disproportionately affect Māori children. Moreover, the models in this study did not exhibit significant issues with these fairness metrics, particularly in terms of accuracy equity (see Section 6.5.2).

Prioritizing calibration alone as a measure of fairness risks overlooking crucial differences in the distribution of error rates between ethnic groups. By emphasizing disparate impact and equalized odds, our approach aims to promote a more equitable distribution of false positives, false negatives, and true positives for Māori children and children from other ethnic groups. This focus is particularly important in addressing the systematic overestimation of risk for Māori children and the underestimation of risk for other ethnic groups. Addressing these disparities is crucial for promoting fairness in predictive model outcomes, especially within the sensitive context of child welfare decision-making, where fairness and equity are paramount.

The following subsections present the findings from our evaluation of the constrained logistic regression model, with a focus on its predictive performance and its effectiveness compared to existing decision-making practices. In addition, we examine the impact of this method on gender and age group disparities, as well as its effect on key fairness measures, particularly accuracy equity and calibration. A detailed analysis was also conducted to ensure that the aggregation of Pacific children with the NZ European and Others ethnic groups did not result in any negative consequences for the Pacific population. Finally, the model's robustness was assessed through external validation using the Sample Cohort 2018.

6.6.1 Implementation and Performance Evaluation

Initially, two distinct constrained logistic regression models were developed. The first model considered only a disparate impact constraint (constrained logistic regression (DI)). The second incorporated both disparate impact and equalized odds constraints (constrained logistic regression [DI & EOO(y=0) & EOO(y=1)]). For each model, we tuned the parameters to achieve an optimal balance between fairness and model predictive accuracy.

The optimization process was solved using the Sequential Least Squares Programming (SLSQP) algorithm, implemented through the *nloptr* package in R (S. G. Johnson, 2008). After parameter tuning and experimentation, for the constrained logistic regression (DI) model, $\lambda_{\text{DI}} = 0.1$ was found to provide the best balance, while in the model constrained by both disparate impact and equalized odds, the optimal values were $\lambda_{\text{DI}} = 0.03$ and $\lambda_{\text{EO}} = 0.06$. These values provided an effective balance between fairness and predictive accuracy, ensuring that the models effectively mitigated potential biases while maintaining robust predictive power in terms of AUC.

To evaluate the efficacy of the fairness-aware logistic regression models, we assessed their performance against the baseline logistic regression model on internal testing data (30% of the Sample

Cohort 2017). The results, detailed in Tables 6.15 and 6.16, demonstrate the trade-offs between model accuracy and fairness, illustrating how the inclusion of fairness constraints affects both model performance and the equitable treatment of different ethnic groups.

In terms of accuracy measures (Table 6.15), the integration of fairness constraints had minimal impact on the AUC values, which remained relatively stable across models, with only a slight reduction observed. Notably, Non-Māori children consistently exhibited marginally higher AUC values in all cases. These results reveal that the fairness-aware machine learning approach implemented in this work had a noticeable effect on key metrics such as intake rate, TPR, and FPR across Māori and Non-Māori children (Figure 6.11).

In the baseline logistic regression model, Māori children had significantly higher intake, TPR, and FPR values compared to Non-Māori children, indicating that the model was more likely to identify Māori children as at risk, but also more prone to generating false positives. This discrepancy highlights an inherent bias within the model's predictions, disproportionately affecting Māori children by misidentifying them as high-risk more frequently. This overestimation of risk has the potential to unfairly impact Māori children, subjecting them to higher rates of false identification.

However, the application of our proposed fairness-aware approach, incorporating both disparate impact (DI) and equalized odds [EOO(y=0) & EOO(y=1)] constraints, led to measurable improvements in equity across the key metrics. Specifically, the introduction of these fairness constraints resulted in a reduction of both intake and FPR for Māori children, indicating a moderation of the model's previous tendency to overestimate risk and reduce the frequency of false-positive classifications. These results suggest that the fairness-aware modifications were effective in mitigating some of the disparities observed in the baseline model, particularly by reducing the FPR for Māori children, while maintaining overall predictive performance.

While the fairness-aware approach was effective in mitigating some of the disparities observed in the baseline model, particularly by reducing the FPR for Māori children, these improvements came with certain trade-offs. Specifically, the fairness-aware approach also resulted in a reduction of the TPR for Māori children by approximately 7%. Despite this decline, the reduction in TPR was relatively modest when compared with the significant improvements in FPR and intake rate, which decreased by 11% and 8%, respectively (Table 6.15). For Non-Māori children, the fairness-aware approach led to increases in intake rate, TPR, and FPR, ultimately resulting in more balanced and equitable outcomes across both groups. This underscores the potential of fairness-aware methodologies to

Table 6.15: Predictive performance measures of the baseline logistic regression model and constrained logistic regression models, with 95% Confidence Intervals (CI) for AUC and the intake outcome. Intake refers to the rate of positive outcomes predicted by the models.

Group	Intake [95% CI]	AUC [95% CI]	Accuracy	PPV	NPV	TPR	FPR	F1
Logistic Regression								
All	0.6448	0.7190 [0.7112, 0.7269]	0.6734	0.6948	0.6298	0.7717	0.4626	0.7328
Māori	0.7711 [0.7629, 0.7800]	0.7032 [0.6921, 0.7143]	0.6839	0.7080	0.6025	0.8571	0.6201	0.7755
Non-Māori	0.4816 [0.4706, 0.4933]	0.7135 [0.7018, 0.7253]	0.6603	0.6769	0.6449	0.6391	0.3176	0.6575
Constrained Logistic Regression (DI)								
All	0.6427	0.7134 [0.7055, 0.7213]	0.6711	0.6956	0.6270	0.7704	0.4662	0.7311
Māori	0.7043 [0.6945, 0.7140]	0.7013 [0.6902, 0.7124]	0.6819	0.7263	0.5760	0.7628	0.5309	0.7628
Non-Māori	0.5663 [0.5550, 0.5772]	0.7123 [0.7006, 0.7241]	0.6577	0.6481	0.6702	0.6820	0.4067	0.6820
Constrained Logistic Regression (DI & EOO)								
All	0.64379	0.7117 [0.7030, 0.7122]	0.6682	0.6930	0.6233	0.7688	0.4709	0.7289
Māori	0.6901 [0.6811, 0.6992]	0.7030 [0.6919, 0.7141]	0.6786	0.7286	0.5673	0.7896	0.5159	0.7579
Non-Māori	0.5858 [0.5753, 0.5964]	0.7122 [0.7004, 0.7240]	0.6552	0.6410	0.6753	0.7366	0.4295	0.6855

Table 6.16: Fairness measures of the baseline logistic regression and constrained logistic regression models, based on disparate impact (DI) and equality of opportunity for both negative outcomes EOO(y=0) and positive outcomes EOO(y=1).

Model	DI [95% CI]	EOO(y=0) [95% CI]	EOO(y=1) [95% CI]
Logistic Regression	0.6246 [0.6092, 0.6405]	0.5121 [0.4870, 0.5394]	0.7458 [0.7272, 0.7557]
Constrained Logistic Regression (DI)	0.8040 [0.7837, 0.8223]	0.7660 [0.7281, 0.8040]	0.8959 [0.8741, 0.9196]
Constrained Logistic Regression (DI & EOO)	0.8489 [0.8298, 0.8686]	0.8324 [0.8034, 0.8835]	0.9335 [0.9121, 0.9554]

address biases in predictive models while carefully managing trade-offs in predictive performance.

Additionally, the results presented in Table 6.16 highlight the significant impact of incorporating fairness constraints into logistic regression models to achieve more equitable outcomes across ethnic groups. While the baseline logistic regression model achieves a solid AUC of 0.7190 and a high TPR of 0.7712, it falls short on fairness metrics, with a DI of 0.6246 and inconsistent EOO(y=0) and EOO(y=1) values across groups.

Introducing the constrained logistic regression model (DI) slightly reduced the AUC to 0.7134 but notably improved fairness, with DI increasing to 0.8040 and EOO(y=1) rising to 0.8959. However, this model does not address EOO(y=0) sufficiently to meet the 80% rule of fairness in the machine learning literature (Feldman, 2015). In contrast, the constrained logistic regression model that incorporates both DI and EOO constraints demonstrates a balanced improvement in fairness metrics without a significant loss in predictive performance. This model maintains a competitive AUC of 0.7117 while substantially improving DI to 0.8493, EOO(y=1) to 0.9329, and EOO(y=0) to 0.9335, all of which satisfy the 80% rule of fairness.

Improvements in fairness are further illustrated in Figure 6.12, which compares the mean predicted probabilities for Māori and Non-Māori children across the models. In the logistic regression model

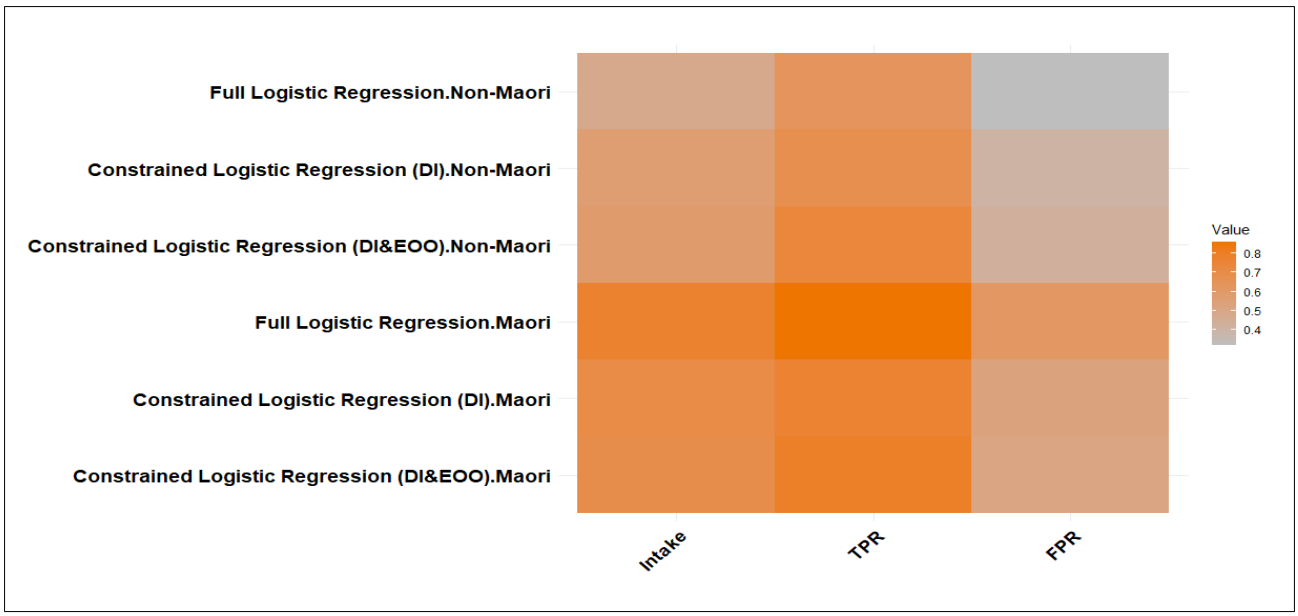

Figure 6.11: Intake rate, TPR, and FPR across the baseline logistic regression and constrained logistic regression models for Māori and Non-Māori groups. The heatmap illustrates the differences in intake, TPR, and FPR between models. The color gradient represents the magnitude of these metrics, with darker shades indicating higher values, showing how fairness constraints influence the model's performance across these groups.

(Panel A), Māori children consistently have higher predicted probabilities, regardless of the true outcome, indicating that the model systematically overestimates risk for this group. However, in the constrained logistic regression model with DI (Panel B), the gap in predicted probabilities narrows, particularly for positive outcomes, showing an improvement in equity. Finally, in the constrained logistic regression model with both DI and EOO constraints (Panel C), the predicted probabilities for Māori and Non-Māori converge closely, demonstrating the model's ability to reduce both false positives and disparities in risk prediction across ethnic groups. This visual alignment supports the improvements in fairness metrics observed in Table 6.16, highlighting that the combined disparate impact and equalized odds constraints lead to more equitable risk assessments while maintaining acceptable levels of predictive performance.

The results from the constrained logistic regression models indicate improvements in fairness across Māori and Non-Māori children when both disparate impact and equalized odds constraints are applied. While these fairness metrics suggest more equitable model performance, it is still essential to assess how these models behave across different risk groups. To further evaluate fairness, we divided the predicted probabilities into three categories: low risk (ventiles 1-6), medium risk (ventiles 7-14), and high risk (ventiles 15-20). The boxplots in Figure 6.13 display the distribution of predicted probabilities for each group across these risk levels. This breakdown helps identify how the models distribute risk scores between Māori and Non-Māori children, ensuring that fairness gains are reflected not only in overall metrics but also within specific risk categories. While improvements were

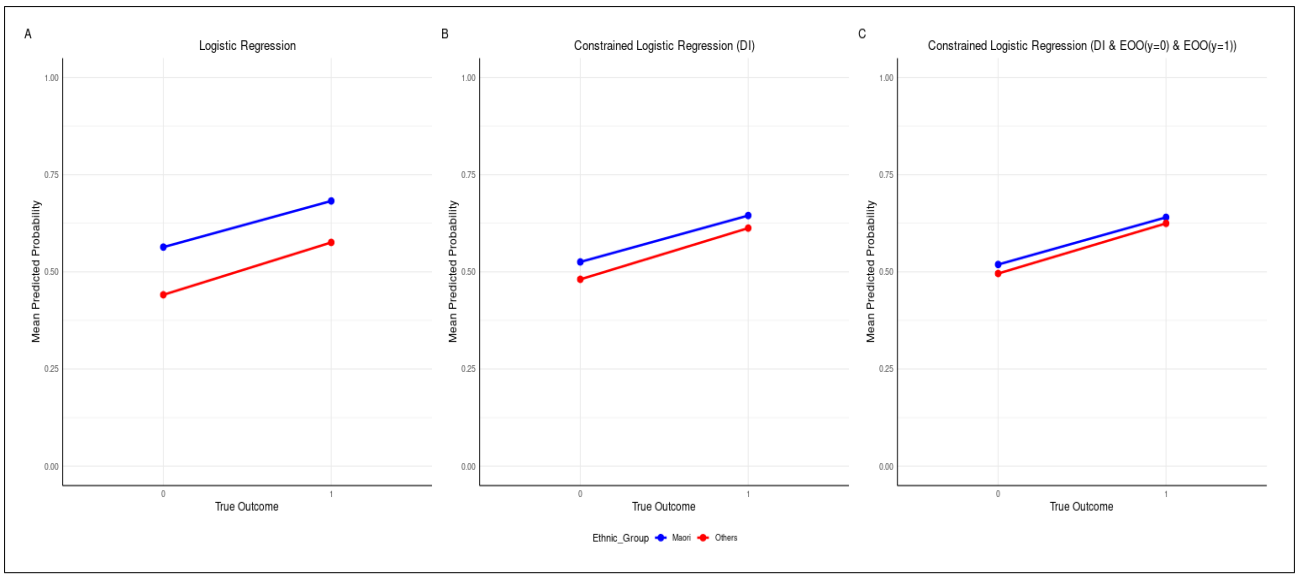

Figure 6.12: Mean predicted probabilities by true outcome for Māori and Non-Māori children across the baseline logistic regression and constrained logistic regression models.

observed, it remains crucial to monitor predicted probabilities to ensure fairness is maintained across all risk levels.

As seen in Figure 6.13, in the baseline logistic regression model, Māori children tend to receive higher predicted probabilities than Non-Māori across the low and medium risk groups, indicating a potential bias toward overestimating risk for Māori children. However, when fairness constraints are applied in the constrained logistic regression model, the predicted probabilities between Māori and Non-Māori become more aligned. This is most evident in the low and medium risk levels, where the medians of the predicted probabilities for both groups are very close, showing that the constrained model successfully mitigates the bias observed in the baseline model. At the high-risk level, the medians of predicted probabilities are closely aligned in both models, indicating that the fairness constraints do not significantly affect the predictions for high-risk cases. This suggests that the bias was most prominent in the low and medium risk levels in the baseline model, and the application of fairness constraints particularly helped address discrepancies in these groups. In summary, the constrained logistic regression model improves fairness by ensuring that the predicted probabilities are more consistent across ethnic groups, especially in the low and medium risk categories, while maintaining equitable predictions at the high-risk level.

6.6.1.1 Disparities Analysis Across Subgroups

In predictive risk modeling, it is critical to assess not only the overall performance of a model, but also the presence of any disparities in its predictions across different demographic groups (Vaithianathan,

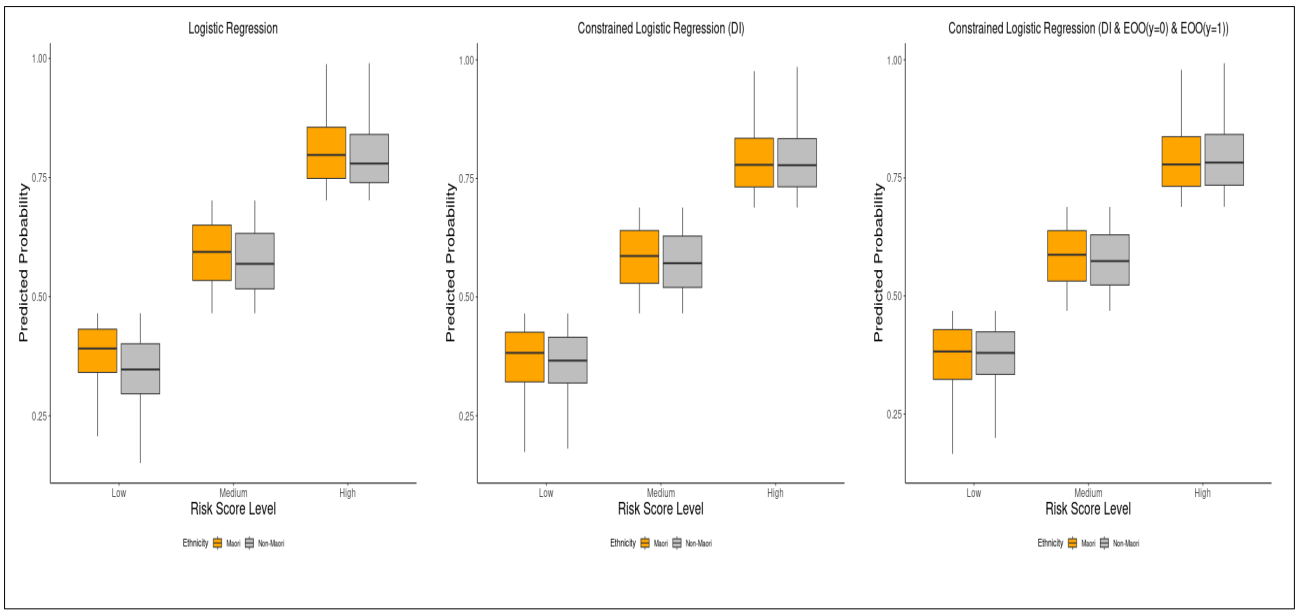

Figure 6.13: Distribution of predicted probabilities across risk score levels for the baseline logistic regression model and constrained logistic regression models. The boxplots represent the predicted probabilities for Māori (orange) and Non-Māori (gray) children across low (ventiles 1-6), medium (ventiles 7-14), and high (ventiles 15-20) risk score levels.

Dinh, et al., 2019). While the use of constrained logistic regression models has demonstrated improvements in fairness based on *ethnicity*, as detailed in Section 6.6.1, it is important to further explore how well the constrained models maintain equity in terms of calibration across key subgroups such as *gender* and *age*.

The gender-based calibration analysis, shown in Figure 6.14, highlights significant improvements in reducing gender disparities through the use of constrained logistic regression models. In the baseline logistic regression model, discrepancies between male (blue line) and female (red line) calibration are evident, particularly in the lower risk ventiles, where males tend to exhibit slightly higher rates of observed *estimated care and protection concern*. The observed disparities decrease as the risk scores increase, with the lines converging more closely in the higher ventiles (above 12). When the model is constrained to achieve demographic parity by applying a disparate impact constraint (constrained logistic regression (DI)), the calibration between males and females aligns more closely across all risk ventiles, especially in the lower ranges where disparities were previously more prominent. While minor discrepancies persist in some low to mid-range ventiles, the overall fairness between genders improves significantly compared to the baseline logistic regression model.

Further constraining the model to optimize both disparate impact and equalized odds (constrained logistic regression [DI & EOO(y=0) & EOO(y=1)]) nearly eliminates calibration differences across genders, resulting in the closest alignment between male and female calibration curves across the entire

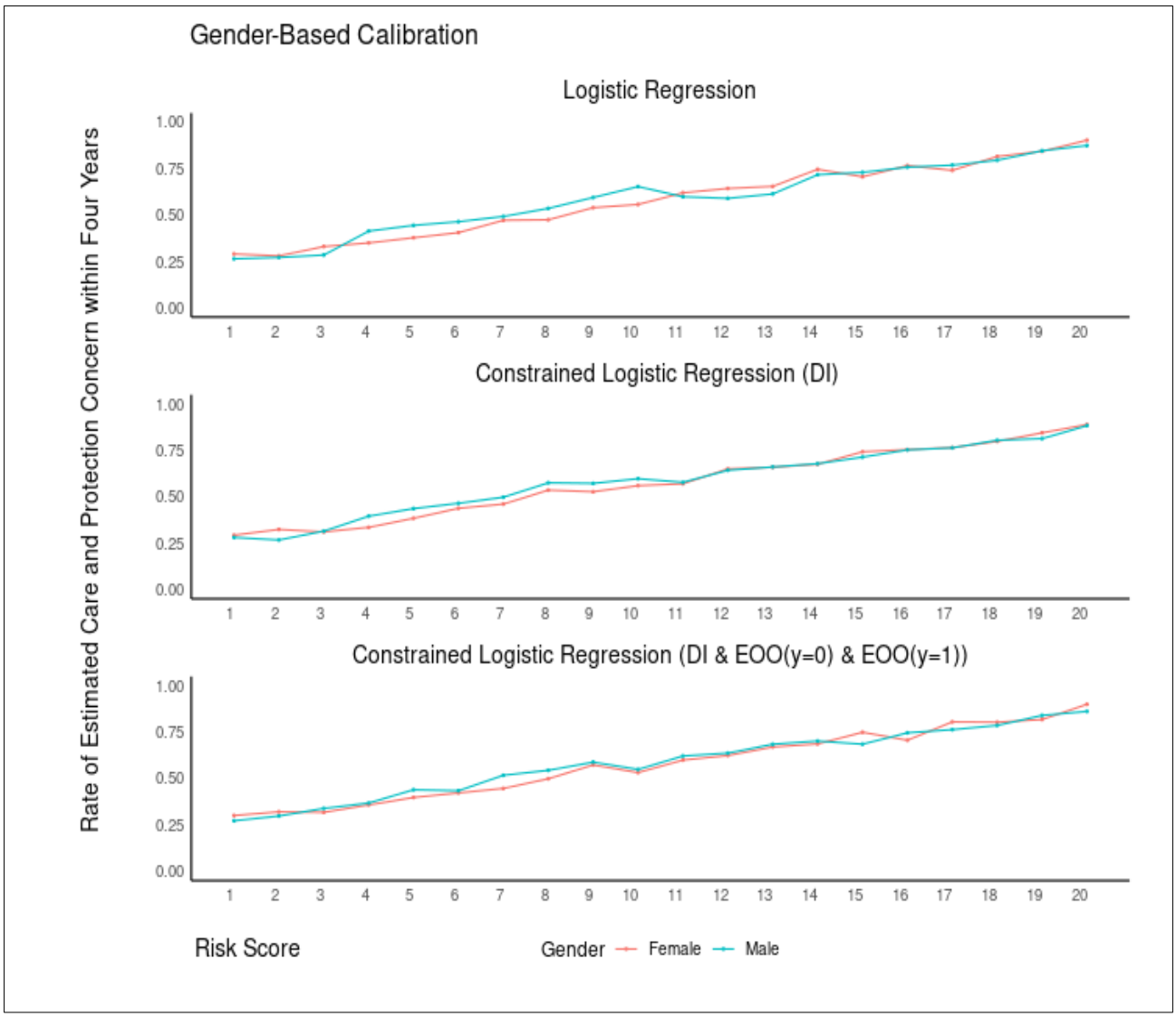

Figure 6.14: Gender-based calibration plots for the baseline logistic regression and constrained logistic regression models.

range of risk ventiles, with particular improvements in the lower ventiles. Overall, both constrained models demonstrate improved calibration, ensuring that predicted probabilities more equitably reflect observed outcomes across gender groups. This analysis underscores the effectiveness of fairness-aware techniques in mitigating gender disparities while also preserving fairness across ethnic groups, further strengthening the model's equity-focused approach.

Analyzing the age-group calibration in Figure 6.15, the baseline logistic regression model reveals considerable variation between age groups, particularly in the lower ventiles (1–6). For instance, children under five (red line) consistently exhibit lower rates of observed *estimated care and protection concern* in these lower risk ventiles. In contrast, in the higher ventiles (15–20), the rates for all age groups converge more closely, reducing visible disparities. When the model is constrained by applying the disparate impact constraint, the variability between age groups in the lower ventiles decreases,

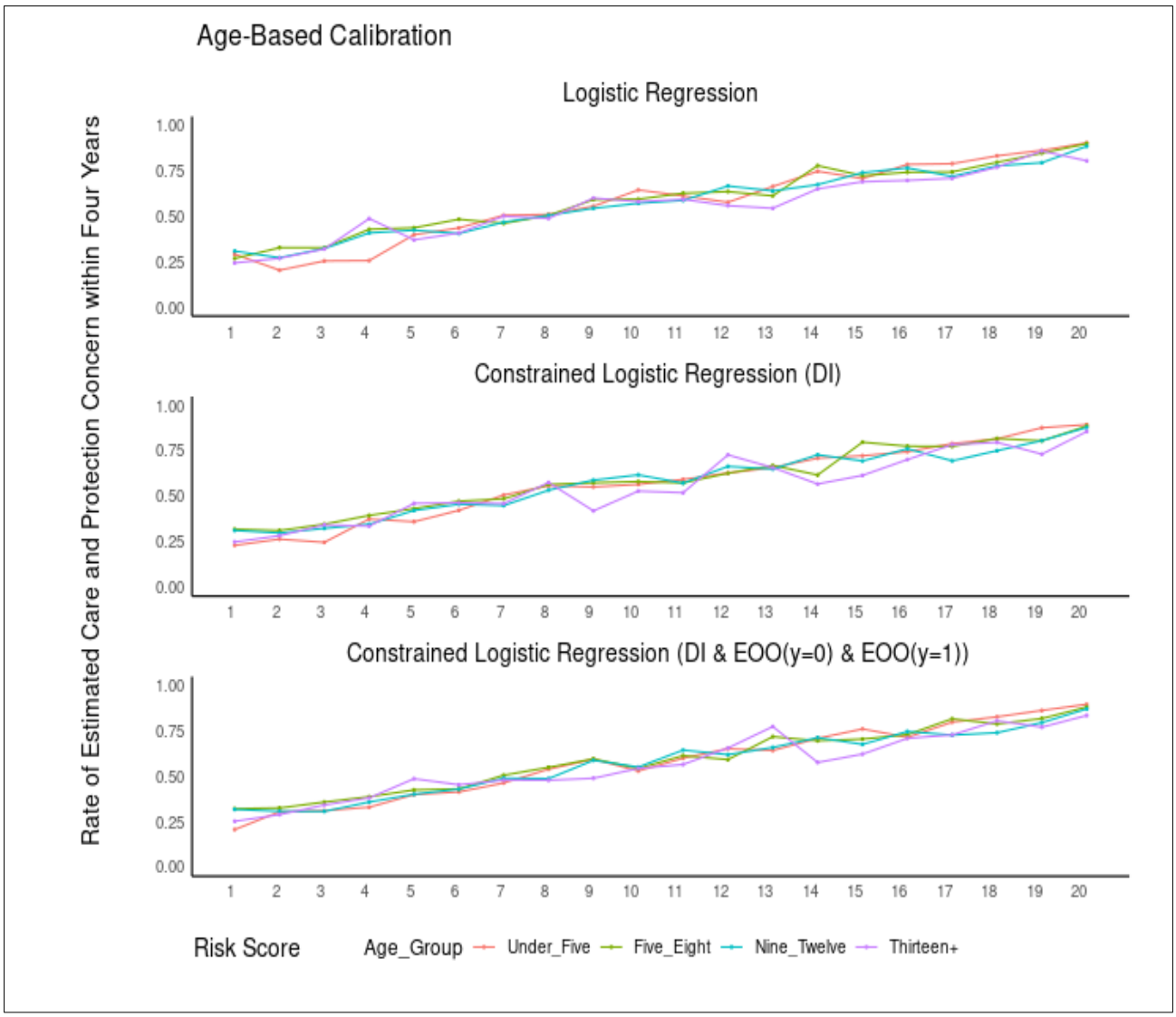

Figure 6.15: Age-based calibration plots for the baseline logistic regression and constrained logistic regression models.

although some disparities remain. The calibration lines for different age groups draw closer together, reflecting improved alignment across age groups, though notable divergence persists, particularly for children over 13 years old (purple line). The introduction of additional fairness constraints, incorporating both demographic parity and equalized odds, results in a much tighter alignment between age groups, including in the lower risk ventiles (1–6). The calibration lines overlap significantly more across all ventiles, especially in the lower and mid-range ventiles where disparities were previously more pronounced. This suggests that the additional fairness constraints substantially reduce most calibration differences between age groups. However, some divergence remains, particularly for children over 13 years old, indicating that further refinement may be needed to address this age-specific disparity fully.

6.6.1.2 Potential Effects on Ethnic Group Calibration and Accuracy Equity

While one of the primary focuses of this thesis has been on error rate balance or equalized odds, particularly in reducing outcome disparities between Māori and Non-Māori children, it is important to extend the fairness evaluation to include other dimensions such as accuracy equity and calibration. Accuracy equity ensures consistent predictive performance across ethnic groups (Chouldechova & G'Sell, 2017), while calibration assesses whether predicted probabilities align with actual outcomes (Berk et al., 2021). These concepts were previously discussed in Section 5.3.3.2 and examined in Section 6.5.2.

To provide transparency and a comprehensive fairness analysis, we evaluated the impact of disparate impact and equalized odds constraints on accuracy equity and calibration. This evaluation allows us to assess whether improvements in error rate balance lead to trade-offs in these other dimensions. The analysis in Section 6.5.2.2 revealed that accuracy equity is achieved in the baseline logistic regression model (Table 6.12). In this section, we examine whether the disparate impact and equalized odds constraints maintain this equity. Figure 6.16 represents the ROC curves for these models, while Table 6.17 presents the results of statistically testing AUC differences between Māori and Non-Māori children.

The constrained logistic regression models, both with disparate impact and the combined disparate impact and equalized odds constraints, demonstrate minimal performance differences in AUC between Māori and Non-Māori. DeLong statistical tests confirm that these differences are not statistically significant, with both p-values greater than 0.05 (Table 6.17), supporting the conclusion that accuracy equity is preserved between Māori children and children from other ethnic groups, despite the application of fairness constraints.

However, as shown in Figure 6.17, the application of these constraints introduces a trade-off between improving fairness in terms of error rates and maintaining calibration accuracy. While these constraints are designed to mitigate disparities in error rates between Māori and Non-Māori by balancing TPRs and FPRs, they affect the model's ability to align predicted probabilities with observed outcomes for these ethnic groups.

In the baseline logistic regression model, the observed rates of *estimated care and protection concerns* increase consistently with higher risk scores for both Māori and Non-Māori. However, some disparities are evident, particularly in the lower and middle ventiles (e.g., Māori children have an estimated rate of 30% in the lowest ventile, compared to 25% for Non-Māori). Similar discrepancies are

Table 6.17: AUC and DeLong statistical test results for Māori and Non-Māori for constrained logistic regression with disparate impact (DI) and constrained logistic regression with both disparate impact (DI) and equalized odds constraints [DI EOO(y=0) EOO(y=1)].

Model	Group	AUC [95% CI]	D	p-value
Constrained Logistic Regression (DI)	Māori	0.7013 [0.6902, 0.7124]	-1.3414	0.1798
	Non-Māori	0.7123 [0.7006, 0.7241]		
Constrained Logistic Regression (DI & EOO)	Māori	0.7030 [0.6919, 0.7141]	-1.1178	0.2637
	Non-Māori	0.7122 [0.7004, 0.7240]		

Note: A p-value > 0.05 indicates no statistically significant difference in AUC values.

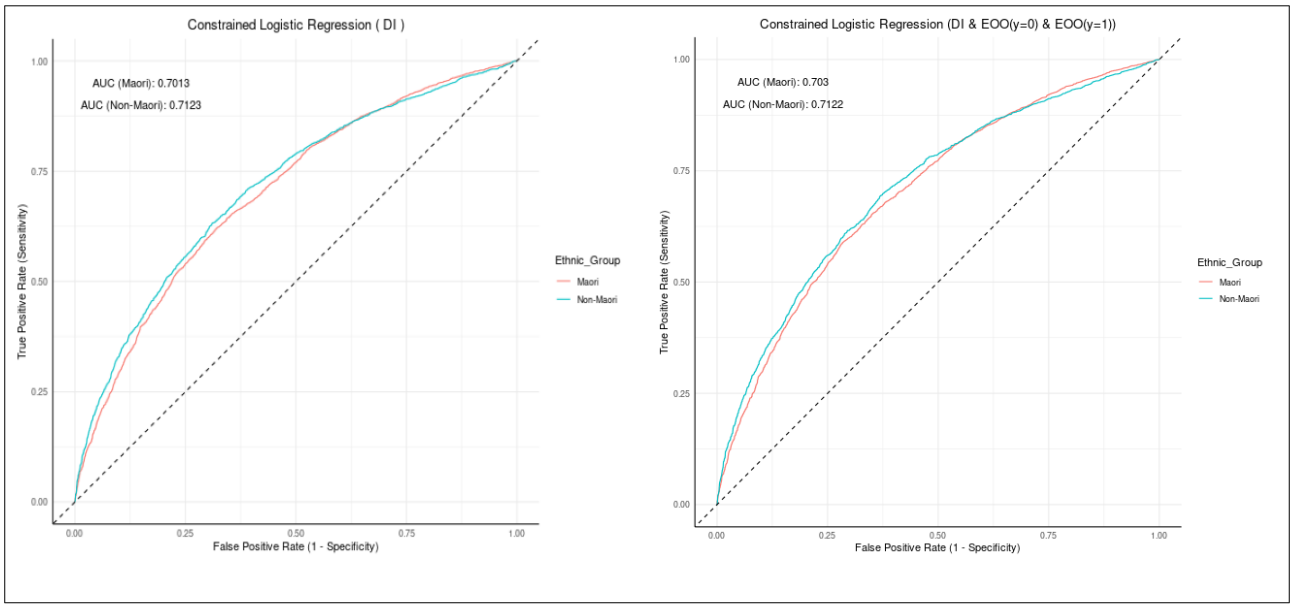

Figure 6.16: ROC curves stratified by child's ethnic group (Māori vs. Non-Māori) for constrained logistic regression with disparate impact (DI) and constrained logistic regression with both disparate impact (DI) and equalized odds constraints [DI & EOO(y=0) & EOO(y=1)].

present in the mid-range ventiles, though these are more pronounced in the constrained models.

Calibration, which ensures that predicted probabilities align with actual event rates, is compromised as the model adjusts its predictions to achieve fairness in error distribution. This is evident in the observed divergence between the predicted and actual outcomes for Māori and Non-Māori at certain risk levels, particularly in the constrained models. Despite the constraints improving fairness by reducing classification bias, the cost is a reduction in the accuracy with which the predicted risks reflect true outcomes for each group. These findings highlight a fundamental trade-off between fairness in terms of equalized odds and fairness in terms of calibration in predictive modeling.

Overall, the combination of disparate impact and equalized odds constraints proved to be the most effective strategy for reducing predictive bias in outcomes without significantly compromising predictive performance. Consequently, this approach is the preferred method for addressing fairness in this

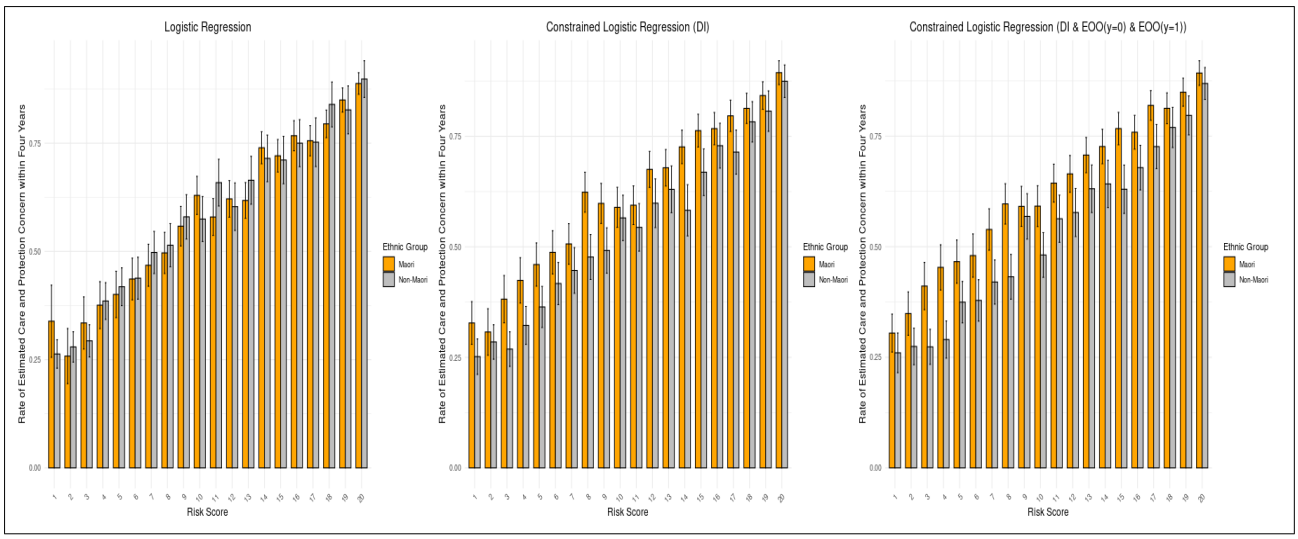

Figure 6.17: Observed *estimated care and protection concern* by risk ventile based on predicted probabilities from baseline logistic regression model and constrained logistic regression models for Māori (orange) and Non-Māori(gray) children. Error bars correspond to 95% confidence intervals (CI).

context. However, it remains critical to compare the results of these predictive models with the existing decision-making accuracy in child welfare to assess whether their use provides tangible benefits over current practices. Such a comparison would offer a more comprehensive understanding of the value that predictive modeling can bring to decision-making processes and help determine whether the improvements in fairness and accuracy achieved by the models translate into meaningful gains in real-world child welfare interventions.

6.6.1.3 Performance Comparison with Existing Decision Making

To assess the accuracy of our constrained logistic regression model with both disparate impact and equalized odds constraints, we directly compare accuracy measures against the existing intake decision-making accuracy.[11] For consistency, this analysis was conducted on the randomly selected 30% of the data drawn from the Sample Cohort 2017, which was also used for internal testing of the models. The process for obtaining these samples is outlined in Section 5.2.2. Our investigation involved determining whether a child in our sample cohort received an intake outcome following the initial assessment of the notification and tracking subsequent events recorded for that child within four years of the initial notification in the CYF data (see Section 2.4). These tracked events had to align with the care and protection-related events used to define the outcome variable (see Table 5.1).

We categorized the accuracy measures into the following four possible outcomes:

Outcome 1. A child has an intake outcome at the time of notification, and the concerns related to

[11] The statistics presented in Table 6.1 are derived from a research dataset we constructed and should not be considered official statistics. We do not make any judgments regarding the decision-making process of the data provider.

care and protection are subsequently found to be true (TP).

Outcome 2. A child doesn't have an intake outcome at the time of notification, and the concerns related to care and protection are subsequently found to be false (TN).

Outcome 3. A child has an intake outcome, but the concerns related to care and protection are later found to be false (FP).

Outcome 4. A child doesn't have an intake outcome, but the concerns related to care and protection are subsequently found to be true (FN).

Enhancing the identification of true positives among children referred to child protective services results in more children receiving the necessary assistance. Conversely, improving the identification of true negatives reduces unnecessary assessments, thereby enabling social workers to allocate more time to children and young people in need (Kearney et al., 2023).

A critical measure for evaluating decision-making accuracy is the percentage of decisions that prove to be 'true.' Specifically, this means that children who did not receive an intake outcome also did not experience any care and protection-related events within four years of the initial notification. Similarly, children who received an intake outcome experienced at least one of these events. This percentage corresponds to the model's classification accuracy.

To evaluate the performance of the current intake decision-making process, we applied the same accuracy metrics used throughout this work to assess the predictive performance of the models, ensuring consistency in the evaluation. These results are outlined in Table 6.18.

The comparison between the existing intake decision-making process by NZ child protective services, for the period between April 1, 2017, and March 31, 2018, and the constrained logistic regression model demonstrates potential advantages of our predictive model. Our constrained logistic regression consistently outperformed the existing system across key performance metrics. With a higher True Positive Rate (77% vs. 62%), the model more accurately identifies children with actual care and protection concerns, potentially enhancing decision-making and ensuring that those in need receive timely support. Our results also show that our model improves precision (PPV) and negative predictive value (NPV), increasing both the accuracy and reliability of intake decisions. The F1 score, which balances precision and recall, was also notably higher for the constrained model (73% vs. 65%), indicating a more efficient process with fewer unnecessary assessments.

Additionally, the constrained logistic regression model resulted in a higher overall intake rate (64% vs. 53%). While this increase in intake could raise concerns about service burden, it may also be

Table 6.18: Accuracy measures of the existing intake decision-making process, baseline logistic regression, and constrained logistic regression with both disparate impact and equalized odds constraints [DI & EOO(y=0) & EOO(y=1)].

Proportion of Children	**Existing Decision-making**	**Logistic Regression**	**Constrained Logistic Regression**
Referred with an intake outcome (**Intake**)	0.53	0.64	0.64
Received an intake outcome where there was *estimated care and protection concern* (**PPV**)	0.67	0.70	0.69
Did not receive an intake outcome where there was no *estimated care and protection concern* (**NPV**)	0.53	0.63	0.62
Had *estimated care and protection concern* and received an intake outcome (**TPR**)	0.62	0.77	0.77
Did not have *estimated care and protection concern* and received an intake outcome (**FPR**)	0.42	0.46	0.47
Received an outcome that correspond with the level of *estimated care and protection concern* (**Accuracy**)	0.60	0.67	0.67
Had *estimated care and protection concern* and received an intake outcome, reflecting a balance between precision and recall (**F1**)	0.65	0.73	0.73

related to the improvements seen in other metrics. For example, flagging more children for intervention can lead to a higher number of appropriate assessments, reflected in improved precision and a greater proportion of correct classifications (both true positives and true negatives), without a significant rise in false positives. This suggests that while more resources may be needed to handle the increased number of flagged cases, the model is more effectively identifying children who require support, potentially leading to better-targeted interventions and outcomes. Overall, it suggests a balance between increasing intake and improving the quality of decisions being made, where the benefits may outweigh the concerns about the additional service load.

Additionally, we assessed the predictive performance of the baseline and constrained logistic regression models in comparison to the existing decision-making process across key metrics for both Māori and Non-Māori children. The results are depicted in Figure 6.18.

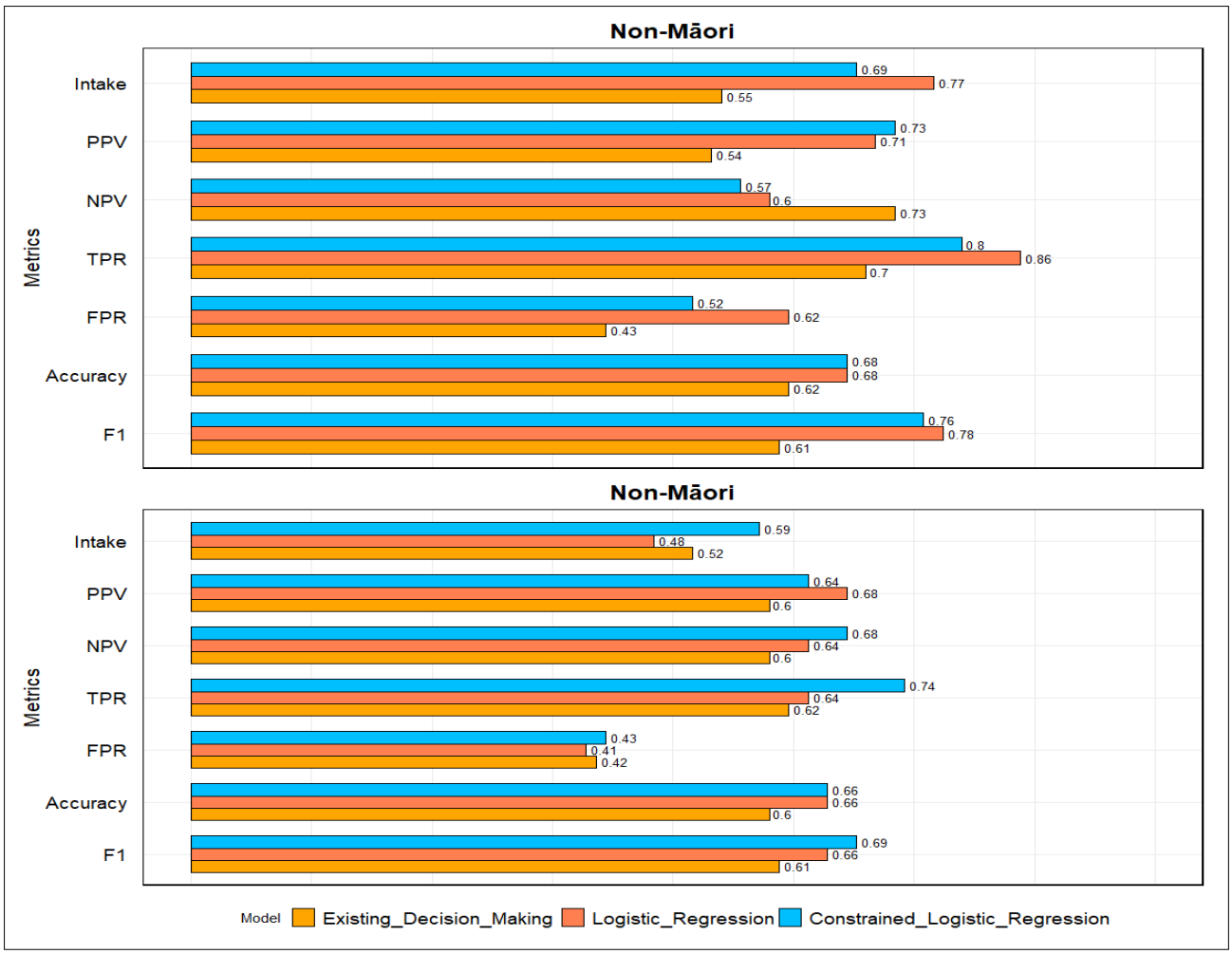

Figure 6.18: Predictive performance comparison of existing decision-making process, logistic regression, and constrained logistic regression [DI & EOO(y=0) &EOO(y=1)] for Māori children and children from other ethnic groups (Non-Māori)

The predictive models, particularly the constrained logistic regression, show potential to enhance decision-making by improving the identification of true cases of children with care and protection concerns. For Māori children, the constrained model demonstrates a higher TPR of 0.8, compared to 0.7 for the existing system. This suggests that the model could support more accurate identification of children requiring timely intervention. Additionally, the constrained model achieves a PPV of 0.73, significantly higher than the 0.54 observed in the existing system, indicating greater precision in identifying those with actual care and protection needs.

Moreover, while the FPR of the constrained model is higher than the existing system (0.52 vs. 0.43), its overall accuracy remains higher at 0.68, suggesting improved reliability in identifying true cases with *estimated care and protection concern*. Similarly, for Non-Māori children, the constrained model shows a notable increase in TPR (0.74 vs. 0.62) and F1 score (0.69 vs. 0.61), reflecting better balance between precision and recall compared to the existing decision-making process.

Importantly, the constrained logistic regression model demonstrates more balanced performance across key metrics when compared to the unconstrained logistic regression model. While the unconstrained model achieves a higher TPR, it also produces more false positives. The constrained model, by contrast, manages to reduce the FPR while maintaining high TPR and overall accuracy, aligning fairness with predictive performance. This balance suggests that the model holds promise for enhancing decision-making in child protective services, offering a method for identifying children in need of intervention while addressing fairness across different groups, including Māori.

6.6.1.4 Potential Effect on Pacific Ethnic Group

While the primary aim of this thesis was to assess disparities in predictions between Māori children and children from other ethnic groups, it is also crucial to evaluate the potential impact of the model on the Pacific group specifically, as their integration with other ethnic groups may influence how effectively the model captures distinct patterns or outcomes unique to this subgroup.

For this purpose, to assess the potential effect of combining the Pacific group with NZ European and Others in the fairness-aware machine learning approach, we focus on the way this group's inclusion within a larger non-Māori category may influence their representation and the model's ability to predict outcomes specific to them (Table 6.19).

The analysis of our results for Pacific children (n=1,806) highlights the potential benefits of the constrained logistic regression model, particularly in improving the identification of those in need of care and protection services. The TPR for the constrained model is significantly higher at 77%, compared to 73% in the existing decision-making process and 67% in the logistic regression model, indicating that the constrained model is more effective in ensuring that Pacific children requiring intervention are correctly identified. Furthermore, while the FPR increases in the constrained model (0.56 vs. 0.44 in logistic regression), it remains below the rate observed in the existing system (0.58).

The constrained model also performs well in terms of NPV, with an improved NPV of 0.59 compared to 0.55 in the existing system, suggesting it better identifies children not in need of care. Importantly, the constrained model's accuracy is consistent at 62%, outperforming the existing decision-making process (59%) while balancing precision (PPV: 0.64) and recall (F1: 0.76). This indicates that the constrained model enhances decision-making by improving the identification of true positives without a significant increase in false positives. The increased intake rate of 68% in the constrained model reflects its more proactive approach, ensuring more children receive timely assessments, although this may place additional demands on child protective services.

Overall, the constrained logistic regression model shows a significant enhancement in decision-making for Pacific children, providing a more accurate and equitable system compared to both the existing process and baseline logistic regression. While these metrics indicate an overall improvement in decision-making for Pacific children, it is crucial to examine how the constrained logistic regression model affects Pacific children across different risk groups. In this context, we categorized the predicted probabilities into three risk levels: low risk (ventiles 1-6), medium risk (ventiles 7-14), and high risk (ventiles 15-20). Figure 6.19 presents boxplots that provide a clearer view of how the model assigns risk scores among Māori, Pacific, NZ European, and other children. To a great extent, performance gains are observed not only in accuracy metrics but also within specific risk categories.

The analysis of predicted probabilities across risk levels for Pacific children also demonstrates the impact of applying fairness constraints in the constrained logistic regression model. In the baseline logistic regression, Pacific children exhibit slightly higher predicted probabilities compared to Māori and European/Other children, particularly at the low (ventiles 1-6) and medium (ventiles 7-14) risk levels. After applying constraints to ensure fairness (disparate impact and equal opportunity for both false positive and false negative rates), the predicted probabilities for Pacific children become more aligned with those of the other groups, particularly in the low and medium risk categories. This suggests that the fairness-aware adjustments have considerably reduced the disparities previously observed in the unconstrained model.

However, based on this analysis, the application of fairness constraints does not appear to have had a negative effect on Pacific children. Although Pacific children were combined with NZ European and Others, the constrained logistic regression model effectively reduced disparities in predicted probabilities across risk levels. The alignment of predicted probabilities with other ethnic groups, particularly in the low and medium risk categories (ventiles 1-14), suggests that the model is still able to capture important distinctions for Pacific children. Therefore, the fairness constraints seem to have enhanced, rather than compromised, equitable outcomes for this group.

Table 6.19: Predictive performance comparison of existing decision-making process, baseline logistic regression, and constrained logistic regression for Pacific ethnic group (n=1,806).

Metrics	Existing Decision-Making	Logistic Regression	Constrained Logistic Regression
AUC	*	0.6626	0.6605
AUC 95% CI	*	[0.6377,0.6875]	[0.6355,0.6854]
Intake	0.66	0.57	0.68
Accuracy	0.59	0.62	0.62
PPV	0.62	0.66	0.64
NPV	0.55	0.56	0.59
TPR	0.73	0.67	0.77
FPR	0.58	0.44	0.56

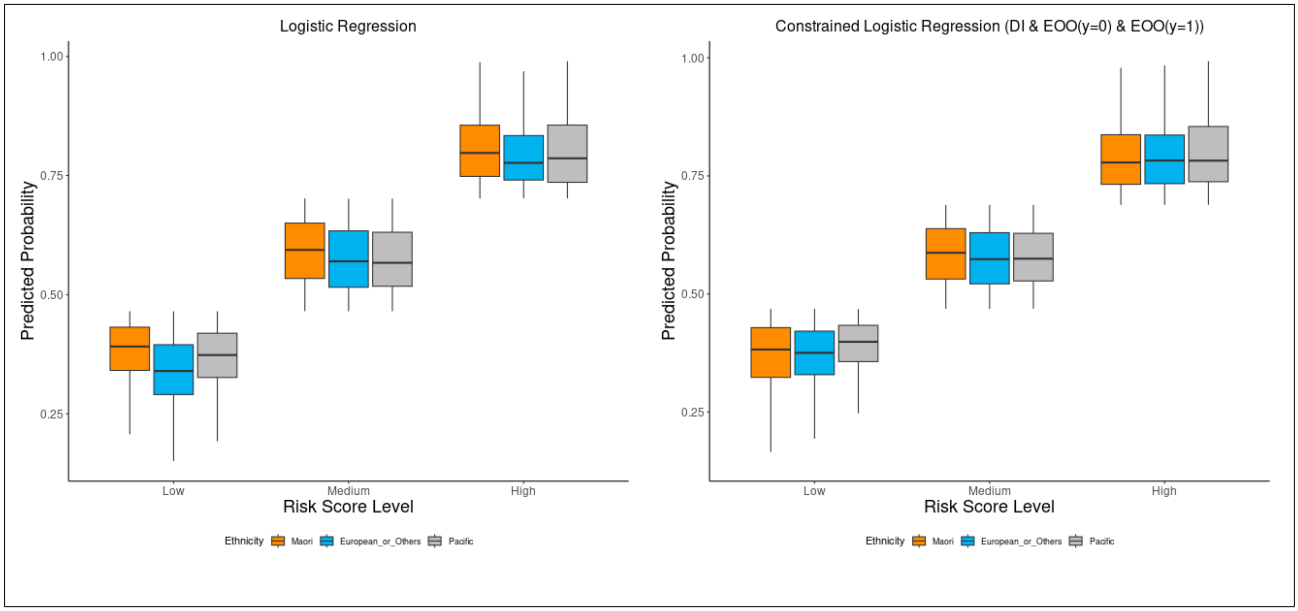

Figure 6.19: Distribution of predicted probabilities across risk score levels for the baseline logistic regression model and constrained logistic regression models. The boxplots represent the predicted probabilities for Māori (orange), Pacific (gray), and NZ European and other children (blue) across low (ventiles 1-6), medium (ventiles 7-14), and high (ventiles 15-20) risk score levels.

6.6.1.5 External Validation on the Sample Cohort 2018

Following the development and evaluation of the constrained logistic regression model on the Sample Cohort 2017, it was essential to validate its performance on an independent dataset to assess its robustness. For this purpose, the Sample Cohort 2018 was used to evaluate the model's ability to maintain fairness and predictive accuracy across different time periods. This validation step is crucial to ensure that the improvements in fairness, particularly regarding Māori children and other groups, are consistent and not limited to the original training dataset.

Table 6.20 presents the results of applying the constrained logistic regression model, incorporating both disparate impact and equalized odds constraints, to the 2018 cohort. It highlights key accuracy metrics and provides an analysis of shifts in fairness outcomes between Māori and Non-Māori ethnic

Table 6.20: Predictive performance metrics of constrained logistic regression with both disparate impact and equalized odds constraints [DI & EOO(y=0) & EOO(y=1)] on internal testing data (30% of the Sample Cohort 2017) and on external testing data (Sample Cohort 2018).

Measure	Cohort 2018 (n=53,997)	Cohort 2017 (n=16,233)
AUC	0.7027	0.7117
AUC 95% CI	[0.6983,0.7071]	[0.7030,0.7122]
Intake	0.6605	0.6438
Accuracy	0.6535	0.6682
PPV	0.6655	0.6930
NPV	0.6302	0.6233
TPR	0.7778	0.7688
F1	0.7173	0.7289
Disparate Impact (DI)	0.8372	0.8489
Equality of Opportunity for Negative Outcome [EOO(y=0)]	0.8036	0.8324
Equality of Opportunity for Positive Outcome [EOO(y=1)]	0.9224	0.9335

groups. The focus is on disparate impact and equalized odds, examining both false positive rates EOO(y=0) and true positive rates EOO(y=1).

The comparison of performance metrics between the external testing on the Sample Cohort 2018 and internal testing on 30% of randomly selected observations from the Sample Cohort 2017 reveals a consistent pattern of slightly better performance in the internal cohort across most measures. The AUC for the 2017 cohort (0.7117) is about 1% higher than for the 2018 cohort (0.7027), indicating slightly stronger discriminatory ability in the internal data.

This trend is further supported by higher accuracy (0.6682 vs. 0.6535) and a notably higher PPV in the 2017 cohort (0.6930), suggesting the model was more precise in identifying true positives in the internal cohort. The equality of opportunity metrics for both positive and negative outcomes also favor the 2017 cohort, with values of 0.9335 for EOO(y=1) and 0.8324 for EOO(y=0), reflecting a more balanced identification of positive and negative cases.

Despite these differences, the external validation on the 2018 cohort demonstrates reasonable robustness, with the model maintaining its sensitivity (TPR = 0.7778) and a slightly higher negative predictive value (NPV = 0.6302).

Importantly, the fairness assessment using the 80% rule continues to be satisfied across both cohorts, with statistical parity at 0.8372 for the 2018 cohort and 0.8489 for the 2017 cohort, ensuring equitable treatment across groups.

The small changes in performance between cohorts can be attributed to differences in the underlying population and potential temporal effects; however, these variations are minimal and are not expected to undermine the model's overall fairness and effectiveness, supporting its ability to generalize effectively.

The observed decrease of 1% in AUC when scoring children from 2018 is both expected and reasonable. Natural variations in data distribution, random variability, or shifts in patterns and population characteristics between the two years are likely contributors (Guo et al., 2021). Specifically, it is very likely that the four-year time span and the impact of the COVID-19 pandemic, including associated lockdowns in NZ, disrupted routine interactions between child protective services and the public, leading to a decline in recorded cases (Guardian, 2020; Robson, 2020; Tamariki, 2020). Such disruptions could potentially alter the underlying data structure, making it inherently more challenging for predictive models to maintain their accuracy. This highlights the importance of regular model updates to address shifts in data trends. Without periodic retraining, models may struggle to remain effective in the face of evolving socio-economic conditions and external events like the pandemic.

During this research, an additional analysis was conducted to evaluate the impact of COVID-19 on records of interaction with child protective services based on the care and protection-related event, which is detailed in Chapter 8, "Miscellaneous Topics".

Additionally, changes in policies during this period could have influenced the patterns of reported cases, further contributing to the observed reduction in AUC. While these policy changes are beyond the scope of this study, they underscore the complexities of maintaining model performance across different time periods and shifting socio-environmental contexts (Guo et al., 2021). However, given that the difference in AUC is only 1%, it is likely within the margin of variability one might expect when applying predictive models across different cohorts. This stability in AUC underscores the model's ability to maintain predictive accuracy even when confronted with data that may exhibit subtle changes over time, minimizing concerns about potential overfitting to the training data.

6.6.2 Summary of Findings

The application of the proposed in-processing fairness-aware machine learning approach in this work, which integrates fairness constraints, revealed that the constrained logistic regression model with both disparate impact and equalized odds constraints provides the optimal balance between fairness and predictive accuracy among the models evaluated.

The key findings from this approach are as follows:

- Improvement in error rate balance through reduction in the FPR and intake rates for Māori children, effectively mitigating the overestimation of risk observed in the baseline logistic regression model.
- Enhancement of TPR for Non-Māori children, though this improvement came with a trade-off, as it increased the FPR for this group.
- Slight improvement in calibration, addressing age and gender disparities present in the baseline logistic regression model.
- Preservation of accuracy equity between Māori and Non-Māori children, with consistent AUC values across both groups.
- A trade-off between maintaining calibration between Māori and Non-Māori children and improving fairness in terms of error rate balance or equalized odds.
- Outperformance of the existing decision-making process, with higher TPR and F1 scores, indicating better identification of children with actual care and protection concerns.
- Improvement in overall accuracy and precision, despite a higher FPR compared to the existing decision-making process.
- Improvements in TPR and overall accuracy for Pacific children, even though they were combined with NZ European and Others, outperforming both the existing decision-making process and the baseline logistic regression model.
- Reduction in disparities in predicted probabilities across risk levels for ethnic groups, maintaining equitable outcomes without compromising predictive accuracy.
- Demonstration of the model's potential to enhance child welfare interventions by promoting more accurate decision-making across different ethnic groups, especially for Māori and Pacific children.

In conclusion, the constrained logistic regression model, with integrated fairness constraints for disparate impact and equalized odds, represents a robust approach to addressing fairness concerns in predictive modeling. It successfully reduced disparities in error rates for Māori children while maintaining performance across ethnic groups, including Pacific children. Additionally, it outperformed existing decision-making processes, suggesting that fairness-aware predictive models have the potential to enhance child welfare interventions as assistance tools by promoting more accurate decision-making. However, ongoing validation and comparison with real-world outcomes remain critical to ensuring that these models continue to offer tangible benefits in improving the accuracy and fairness of child protection systems. The evolving nature of socio-economic conditions, along with shifts in data trends, necessitates continuous model updates and retraining to maintain their effectiveness and fairness in real-world applications.

Chapter 7

Discussion

7.1 Introduction

The persistent over-representation of Māori children in the child welfare system has long been a critical issue in NZ (Keddell & Davie, 2018; Keddell & Hyslop, 2019). This imbalance, as confirmed by our own analysis, shows that Māori children are disproportionately represented across key child welfare interactions. For instance, in the the Sample Cohort 2017 utilized for training predictive models in this thesis, Māori children comprised 60% of records for past Section 15 notifications, 65% of past intakes, and 65% of past substantiated findings of maltreatment (Table 6.8), despite representing a smaller proportion of the general population based on NZ official statistics. The systemic over-representation of Māori children in the child welfare system poses a significant risk of introducing bias into predictive models, as these models are trained on historical data that reflect existing inequalities, potentially perpetuating these disparities (Barocas & Selbst, 2016; Calders & Žliobaitė, 2013).

This phenomenon was evident in our analysis, where the models developed in this thesis consistently referred Māori children for intake at significantly higher rates, approximately 77% to 78%, compared to children from other ethnic groups. This consistent over-identification suggests a potential bias within the models, as they systematically flagged a larger proportion of Māori children as requiring further intervention. These findings align with previous research conducted in NZ, which has similarly highlighted disparities in model predictions affecting Māori children (Rea & Erasmus, 2017; Wilson et al., 2015). Consequently, predictive risk models are at risk of intensifying the over-representation of Māori, and their use in decision-making could perpetuate a cycle of bias that may lead to further disadvantage or unfair treatment of Māori (Rea & Erasmus, 2017). The majority of NZ and international studies on predictive risk modeling in the child welfare sector have opted to exclude *race* or *ethnicity*

as predictors, with the intention of avoiding direct bias (see Section 3.4.3). While this approach may appear to be a straightforward solution to prevent the reinforcement of racial stereotypes, it does not fully eliminate the potential for bias. Variables that are strongly correlated with ethnicity can still introduce bias into the model, a phenomenon known as the *redlining effect* (Lum & Johndrow, 2016; Vaithianathan, Kulick, et al., 2019; Žliobaitė, 2017). In the context of the over-representation of Māori children within the child welfare system, features reflecting historical interactions with welfare services may indirectly perpetuate bias, even when ethnicity is not explicitly included. This underscores the need for fairness-aware machine learning techniques that can proactively identify and mitigate such hidden biases, rather than relying solely on the exclusion of sensitive attributes.

Our study explicitly incorporated ethnicity as a predictor to examine its impact on model predictions and to investigate whether these disparities could be mitigated through the application of fairness-aware machine learning approaches. To facilitate this analysis, we developed a model with a robust level of predictive accuracy, enabling a meaningful evaluation of fairness-aware methodologies. Additionally, we hypothesized that refining the outcome variable definition and integrating additional data could enhance the models' predictive performance. This hypothesis was confirmed, as these factors contributed to a slight improvement in the models' predictive power (see Sections 6.3 and 6.4).

Extending the outcome variable (*estimated care and protection concern*) time frame from two to four years, thereby capturing the occurrence of long-term events, resulted in improvements in the model's predictive accuracy and calibration (Section 6.3). Furthermore, incorporating socio-economic indicators from Benefit Dynamics data, parental criminal history from Sentencing and Remand data, mental health and substance abuse records from the Programme for the Integration of Mental data (PRIMHD), and family structure data from the Stats NZ Census, each a recognized risk factor for child maltreatment, resulted in slight enhancement in performance metrics such as AUC. Despite these improvements, the findings suggest that while data linkage can moderately enhance predictive accuracy, it does not fully address the underlying fairness issues related to intake rates and error rates. Moreover, although these modifications improved the models' predictive capacity and partially reduced fairness concerns, they were insufficient in eliminating the disparities observed between Māori children and children from other ethnic groups (Section 6.4).

Addressing the challenge of developing a predictive risk model that is both more accurate and less discriminatory, we applied fairness-aware machine learning techniques to systematically identify effective strategies for mitigating biases. Although calibration and accuracy equity were generally maintained across predictive models trained on the final set of 250 predictors, achieving error rate balance,

particularly for Māori children, remained a significant challenge, as outlined in Section 6.5. To address this, an in-processing fairness method was employed, incorporating constraints during model training to improve error rate balance, also referred to as equalized odds, without compromising overall predictive performance.

The constrained logistic regression model, incorporating constraints based on disparate impact and equalized odds, effectively reduced the FPR for Māori children while increasing the TPR for Non-Māori children. This resulted in a more balanced distribution of error rates between the two groups (see Tables 6.15 and 6.16). However, this improvement came with trade-offs. Specifically, the fairness-aware approach reduced the TPR for Māori children by approximately 7% compared to the standard logistic regression model. Despite this decline, the reduction in TPR was relatively modest in contrast to the significant improvements in FPR and intake rate, which decreased by 11% and 8%, respectively (Table 6.15). Nevertheless, the TPR remained higher than that observed in existing decision-making processes. (Figure 6.18). Conversely, an increase in TPR for children from other ethnic groups led to a corresponding rise in FPR and intake rate. These findings underscore the intricate balance between fairness and predictive accuracy in the child welfare context, indicating that while fairness-aware approaches can mitigate certain biases, their broader implications for decision-making and potential unintended consequences for various demographic groups must be carefully considered.

This thesis offers a detailed examination of racial disparities, particularly between Māori children and children from other ethnic groups, in the predictions generated by risk models within the context of NZ's child welfare system. Utilizing a comprehensive dataset constructed from linked administrative and Census data via the Stats NZ IDI database (Chapter 4), this research provides insights on the complex dynamics of predictive inequity. While the primary aim was not to validate these models' efficacy within the NZ child welfare system, the empirical results presented in Chapter 6, combined with the review of relevant international and NZ literature in Chapter 3, highlight the models' potential as decision-support tools, provided that ethical and technical challenges are carefully addressed.

This chapter discusses the findings and implications of the research, focusing on the challenges and potential solutions for integrating fairness-aware predictive models into child welfare decision-making. It also synthesizes the key contributions, acknowledges the study's limitations, and outlines recommendations for future research. This sets the stage for advancing the development of robust, equitable, and ethically sound predictive models that can better support child welfare decision-making and promote fair outcomes across diverse demographic groups.

7.2 Key Findings

In the next subsections, we present the essential findings derived from the empirical analysis of predictive risk models developed in this thesis within the NZ child welfare context. The models were trained on 70% of randomly selected observations from the Sample Cohort 2017 (n=54,411) and internally tested on 30% of the remaining observations. The results are derived from this 30% of the observations. However, the size of samples utilized in the outcome variable time-span analysis (Section 6.2) varies as the sample is adjusted based on the age of children. See Section 5.2.2 for more details on the sample cohort and sample construction process.

7.2.1 Optimal Outcome Variable Selection

Our analysis revealed that extending the outcome variable (*estimated care and protection concern*) time frame from two years to four years resulted in better predictive accuracy and model calibration. The LASSO logistic regression model with *estimated care and protection concern* within four years as the outcome variable achieved the highest AUC (0.7213) and overall accuracy (0.6743), demonstrating greater performance in identifying high-risk children compared to the Two-Year and Three-Year models (Table 6.1). This improvement highlights the importance of longer observation periods for capturing more complex and long-term risk patterns, which are crucial for effective child welfare interventions.

7.2.2 Impact of Comprehensive Data Linkage

With respect to the integration of a diverse range of predictors from CYF data, Children's Action Plan records, Personal Details data, Benefit Dynamics data, Sentencing and Remand data, PRIMHD, and data from 2018 Census (5L linkage) led to an improvement in the performance of baseline logistic regression models. Specifically, incorporating these additional variables increased the model's overall accuracy by about about 1% compared to models using only CYF data, Children's Action Plan records, and Personal Details Data (1L linkage).

While this improvement may seem small, a 1% improvement in AUC in the child welfare context can be considered substantial as it enhances the model's ability to accurately distinguish between at-risk and non-risk children, which is critical in high-stakes decisions. This small gain can lead to better resource allocation, ensuring that interventions target those most in need, and help reduce both false negatives and false positives. In large populations, even minor improvements in predictive accuracy can positively affect the outcomes for many children, making the model more effective in protecting

vulnerable individuals (Vaithianathan et al., 2013).

This incremental improvement highlights the effectiveness of including a wide array of risk factors, such as parental mental health and substance abuse, criminal history, socio-economic status, and family structure, to provide a more comprehensive assessment of child welfare risks. Furthermore, these comprehensive data linkages contributed to reduced disparities across ethnic groups, with the LASSO logistic regression model showing better calibration and more consistent performance. This finding aligns with previous studies conducted in the U.S., where LASSO was selected for its robust predictive capabilities, particularly in addressing high-risk populations. These studies highlighted that LASSO not only offered high overall prediction accuracy but also achieved comparable performance across diverse racial groups, including African-American children and those from other backgrounds (Centre for Social Data Analytics, n.d. Putnam-Hornstein et al., 2022; Vaithianathan, Kulick, et al., 2019).

7.2.3 Advanced Machine Learning Models

The advanced machine learning models developed in this thesis, including support vector machine, random forest, and XGBoost, demonstrated higher AUC values compared to baseline logistic regression models, see Table 6.11. While XGBoost achieved the best performance with an AUC of 0.73 and a TPR of 0.80, it did not substantially improve fairness metrics such as disparate impact or equalized odds . In contrast, logistic regression models, particularly LASSO, offered comparable fairness outcomes while being more interpretable and easier to implement (Table 6.14). This makes logistic regression a more suitable choice in the child welfare context, where transparency and trust are critical for stakeholder acceptance and ethical decision-making.

7.2.4 Effectiveness of Fairness-Aware Machine Learning Approaches

In terms of the constrained logistic regression model, incorporating fairness constraints for both disparate impact and equalized odds effectively reduced disparities, particularly by balancing error rates between Māori and children from other ethnic groups. The model achieved an AUC of 0.71 while significantly enhancing key fairness metrics. Disparate impact increased from 0.62 to 0.85, equality of opportunity for positive outcomes EOO(y=1) rose from 0.76 to 0.93, and for negative outcomes EOO(y=0) improved from 0.51 to 0.83, all meeting the 80% fairness threshold, indicating a substantial reduction in bias (Table 6.16).

Additionally, the model maintained consistent AUC values for both Māori and Non-Māori children,

balancing predictive accuracy and fairness in terms of statistical parity and equalized odds. However, achieving fairness in error rates came at the expense of calibration, highlighting the trade-off between these objectives in predictive modeling when considering fairness constraints (see Section 6.6.1.2).

7.2.5 Model Evaluation Against Existing Decision-Making

Compared to existing child welfare decision-making practices, the constrained logistic regression model demonstrated higher accuracy (0.67 vs. 0.6), higher TPR (0.77 vs. 0.62), and improved F1 scores (0.73 vs. 0.65), highlighting its potential to more effectively identify children at risk. However, the model's increased FPR (0.47 vs. 0.42) points towards careful implementation and continuous monitoring to mitigate potential unintended consequences.

For Māori children, the model showed a higher TPR (0.8 vs. 0.7) and PPV (0.73 vs. 0.54), with a modest rise in FPR (0.52 vs. 0.43), but overall better accuracy (0.68 vs. 0.61). Similarly, for Non-Māori children, the model improved TPR (0.74 vs. 0.62) and F1 score (0.69 vs. 0.61), achieving a balanced performance with reduced false positives while maintaining high accuracy and fairness across groups (see Section 6.6.1.3). These results suggest that the constrained logistic regression model is effective in classifying children at risk while meeting fairness objectives across different demographic groups.

7.3 Contributions

This thesis contributes to the field of predictive analytics in child welfare by developing a novel dataset, proposing a fairness-aware machine learning approach, introducing a structured evaluation framework for fairness assessment, and offering comparative insights into model selection. Unlike previous studies that often focus on a limited set of predictors, this research integrates diverse administrative data sources, offering a broader and more nuanced assessment of child welfare risk factors.

- **Development of a novel integrated dataset:** This research constructed a dataset by linking multiple administrative sources within the StatsNZ IDI. These sources include child welfare (CYF data, Children's Action Plan data), social welfare (Benefit Dynamics data), criminal justice (Sentencing and Remand data), health (PRIMHD), and demographic information from the 2018 Census. The resulting dataset provides a robust foundation for building predictive models that capture a wider range of risk factors. It also enables future studies to evaluate the predictive power of these combined variables and supports the refinement of risk models in the NZ child

welfare context.

- **Development of a comprehensive evaluation framework:** A key contribution of this thesis is the introduction of a multi-metric evaluation framework designed to assess both the predictive performance and fairness of child welfare risk models. Given the ethical sensitivities and the potential for algorithmic harm in this context, the framework enables a balanced and rigorous evaluation using the following components:
  - **Predictive Accuracy:** Standard classification metrics such as AUC, accuracy, precision (PPV), recall (TPR), and F1-score were used to evaluate overall model performance.
  - **Calibration:** Assessed by comparing predicted probabilities with observed outcomes across subgroups, particularly Māori and Non-Māori children, to ensure predictions align with actual risk.
  - **Accuracy Equity:** Evaluated by comparing AUC and other metrics across ethnic groups to determine whether the model performs consistently for different populations.
  - **Statistical Parity (Disparate Impact):** Calculated as the ratio of selection (intake) rates across ethnic groups, to assess whether the model leads to disproportionate intervention rates.
  - **Equalized Odds:** Measured by assessing whether true positive and false positive rates are balanced across demographic groups.
  - **Trade-off Analysis:** Considered the interaction between fairness and performance, especially the consequences of applying fairness constraints on accuracy and calibration.

  This framework offers a robust tool for future researchers and practitioners seeking to evaluate and enhance fairness in predictive modeling, and may be applicable beyond the NZ context.
- **Implementation of a fairness-aware in-processing methodology:** This thesis advances fairness-aware machine learning by developing and applying a constrained logistic regression model. This in-processing technique introduces constraints based on disparate impact and equalized odds during model training. The resulting model reduced FPR for Māori children and improved the balance of error rates between Māori and Non-Māori groups, providing a practical strategy for reducing algorithmic bias in child welfare decision-support tools.

- **Comparative insight into LASSO versus constrained models:** This research offers practical guidance on model selection in fairness-sensitive applications. LASSO logistic regression demonstrated strong performance in terms of calibration and accuracy equity, likely due to its ability to reduce overfitting and eliminate noisy predictors. However, it does not directly address fairness. Its performance with respect to equalized odds and disparate impact depends on the structure of the data and underlying sources of bias. In contrast, fairness-aware constrained models proved more effective in reducing disparities in outcomes. While LASSO may be preferable when interpretability or predictive accuracy is the primary goal, constrained models are more appropriate when addressing systemic biases is the priority, particularly in high-stakes domains such as child welfare.

- **Emphasis on continuous evaluation and contextual adaptation:** The findings highlight the need for ongoing model monitoring, stakeholder engagement, particularly with Māori communities and adaptation to the specific social and ethical context of Aotearoa NZ. Although the focus of this thesis is on the NZ child welfare system, the findings regarding model performance, fairness trade-offs, and evaluation methodology may be applicable to other jurisdictions and domains where group-level disparities are of concern.

7.4 Implications

The findings of this thesis have significant implications for integrating fairness-aware predictive models into child welfare decision-making frameworks. The constrained logistic regression model developed and evaluated demonstrates the potential to mitigate biases and promote more equitable outcomes across demographic groups, particularly for Māori children, who have historically faced disproportionate representation in both child protection services and risk predictions. By enhancing the model's capacity to guide intake decisions more accurately, it can serve as a valuable tool for social workers, supporting them in making more informed and balanced decisions. This approach not only helps optimize resource allocation to those most in need but also addresses ethical concerns surrounding the fairness of predictive analytics in sensitive domains like child welfare. However, when compared to existing decision-making practices, the model tends to produce a 5% higher rate of false positives, incorrectly flagging cases for intake, despite achieving a high true positive rate and successfully identifying a large proportion of genuine cases requiring intervention. This discrepancy necessitates careful consideration, as the increased number of false positives could lead to unnecessary investigations, placing undue strain on resources and potentially causing harm to families.

Therefore, it is crucial for stakeholders to be actively involved in defining appropriate risk thresholds that align with their operational priorities and ethical considerations. For example, stakeholders could adopt a screening threshold similar to the Allegheny Family Screening Tool by focusing on the top 25% of risk scores (ventiles 16-20), thereby adjusting the model's sensitivity to better align with operational goals and ethical standards (Chouldechova et al., 2018).

Another key insight from this analysis is the inherent trade-off between calibration and error rate balance when developing fairness-aware models. Adjusting the model to ensure error rates are balanced across groups can lead to a loss in calibration, where the predicted probabilities no longer reflect the true likelihood of outcomes equally well for all groups. This trade-off underscores the importance of stakeholder involvement in defining and prioritizing the aspects of fairness that align with their operational goals and ethical standards. Stakeholders must decide whether to prioritize reducing false positives or maintaining calibration, as these choices will impact the practical utility and ethical implications of the model in different ways.

This research also identifies significant structural barriers to effective data integration, particularly within the NZ context. While this study utilized the IDI to develop the predictive model, child welfare agencies would require data-sharing agreements to access and incorporate information from other agencies. Establishing such agreements, while ensuring robust privacy protections, is crucial for enabling the development and application of more comprehensive predictive models. Although only a slight improvement in model performance was observed with the inclusion of additional data, this underscores the value of even incremental contributions from external sources. It highlights the need for stronger collaboration between agencies to address the barriers currently limiting effective data sharing in NZ. These barriers not only hinder timely and comprehensive data integration but also restrict the potential for more substantial improvements in predictive accuracy. Addressing these challenges through appropriate data-sharing agreements and privacy protections is essential for developing more robust and equitable predictive models in the long term.

Moreover, the research highlights the structural barriers to effective data integration, such as privacy and data-sharing constraints, particularly within the NZ context. Addressing these challenges is crucial for advancing predictive analytics in child welfare. Establishing robust data-sharing agreements and safeguarding privacy will be key to overcoming these obstacles and enabling the development and application of more comprehensive predictive models. Despite the slight improvement observed in model performance, the benefits of incorporating data from other organizations, even incrementally, highlight the importance of fostering stronger collaborations between agencies and addressing

the structural barriers that currently inhibit effective data sharing in NZ. These barriers can limit the timely and comprehensive integration of data, thereby restricting the potential for more substantial improvements in predictive accuracy. Ensuring appropriate data-sharing agreements and safeguarding privacy are crucial steps in overcoming these hurdles, which can lead to more robust and equitable predictive models in the long term. Nevertheless, the adaptable nature of the proposed approach in this thesis, which can accommodate various protected variables, risk thresholds for classification, and fairness thresholds (denoted as δ), suggests its broad applicability (see Section 5.4.2.4). This adaptability also makes it relevant for other domains where maintaining fairness in terms of error rates and ensuring predictive accuracy are critical, such as criminal justice and healthcare. By applying these models in different contexts, stakeholders can enhance decision-making processes, making them more fair and effective across a variety of sectors.

7.5 Limitations

While this research offers valuable insights into the development of fairness-aware predictive risk models in the child welfare context, our study encountered relevant limitations that, however, provide opportunities for future exploration. These limitations are categorized and outlined in the following sections, and are mainly focused on refining predictive models to support fair and effective decision-making in child welfare.

7.5.1 Data Quality and Availability

Our predictive models rely heavily on the quality and completeness of the administrative data sourced from the Stats NZ IDI database. Despite the integration of diverse data sources, the accuracy and comprehensiveness of these datasets is not exempt of variations. Missing or inaccurate data, particularly in variables extracted from 2018 Census data, could lead to biases in the model's predictions. Moreover, administrative data may not capture all relevant aspects of a child's situation, such as qualitative information about family dynamics, community support, or individual resilience, which are difficult to quantify but essential for accurate risk assessment.

Another limitation in this research is related to data availability and the temporal scope of the dataset used. The research aimed to use data that reflects the period after the transition of the NZ child welfare system from CYF to Oranga Tamariki in April 2017, with the goal of building more contextually relevant predictive models. Consequently, the study was constrained to using the 2017 cohort for

training and the 2018 cohort for testing, as these were the most recent and suitable datasets available when the research commenced in 2021. This restriction was necessary to accommodate the four-year longitudinal analysis of the outcome variable, capturing a complete period post-transition (see Sections 5.2.1 and 5.2.2).

This limitation may affect the robustness of the findings, as more recent data reflecting changes in the child welfare system were not accessible. Additionally, the longitudinal analysis of outcomes may be influenced by external factors such as the COVID-19 pandemic, which had significant social and economic impacts during this period, potentially affecting the validity of the findings (Guardian, 2020; Robson, 2020; Tamariki, 2020).

7.5.2 Trade-offs Between Error Rate Balance and Calibration

While this thesis employed multiple fairness metrics, including equality of opportunity, calibration, accuracy equity, and equalized odds, achieving a balance between these metrics turns out to be a compromised solution. The trade-off between calibration and error rate balance was clearly identified. Models optimized to equalize error rates across different demographic groups, such as false positives and false negatives, often experience a decline in calibration. In simple words, while the model becomes fairer in terms of error rates, the predicted probabilities may no longer accurately reflect the observed outcomes within each group. The decision to prioritize either calibration or error rate balance depends on the preferences of the stakeholders involved. If the primary concern is avoiding disproportionate impacts on any group due to incorrect predictions, balancing error rates may be emphasized. However, if stakeholders are more focused on ensuring that predicted risk scores closely match true outcomes, calibration may be prioritized.

7.5.3 Limited Scope of Implementation

Although this thesis presents a robust framework for fairness-aware predictive modeling, it has not been tested in real-world child welfare settings. The proposed framework and models are anticipatory in nature, suggesting potential benefits should these tools be integrated into the child welfare system in the future. Ongoing validation and practical testing are needed to fully understand the models' real-world effectiveness and impact.

Without real-world testing, the effectiveness and potential challenges of implementing these models remain uncertain. There may be unanticipated issues related to how the models interact with existing decision-making processes or how stakeholders respond to them. Additionally, the lack of practical

application could mean that the anticipated fairness improvements may not fully materialize in practice (Rudin & Ustun, 2019). Therefore, the absence of real-world validation presents a risk that the models could either overestimate or underestimate their potential to improve outcomes in the child welfare system.

7.6 Future Research Directions

Building on the findings of this thesis, the following areas present opportunities for future exploration:

1. **Incorporating New and Diverse Data Sources:** While the integration of additional variables from various government agencies has enhanced our model's performance, the observed similarity in performance across multiple models suggests a phenomenon known as *model saturation*. This phenomenon indicates that the predictive capacity of the current dataset has reached a limit, and no model can significantly outperform others (Li, 2006). To overcome model saturation, future research should focus on incorporating new, more diverse data sources or employing advanced data engineering techniques, potentially involving the integration of qualitative data, exploration of alternative data linkages, or the use of advanced feature engineering strategies to extract additional information from existing data.

2. **Balancing Fairness and Calibration:** Balancing equalized odds with calibration in predictive models remains a significant challenge. In this work, adjustments made to satisfy equalized odds across ethnic groups resulted in a decrease in calibration between the groups, where predicted probabilities no longer aligned with observed outcomes within each group (see Section 6.6.1.2). Future research should explore additional fairness metrics and methodologies to better navigate these complexities, providing stakeholders with clear guidance on aligning model design with their specific values and objectives.

3. **Understanding Stakeholder Perceptions of Fairness:** Although fairness notions like statistical parity and equalized odds may be suitable for the NZ child welfare context, the prioritization of these concepts should reflect the values and principles of relevant stakeholders. Conducting a study similar to that of Cheng et al. (2021), which examines stakeholders' perceptions of algorithmic fairness, would be valuable in the NZ context. Such a study would help identify the fairness principles most important in child welfare decision-making, ensuring that the chosen model aligns with the ethical standards and expectations of the community it serves.

4. **Exploring Post-Processing Fairness Correction Methods:** Future research should also investigate post-processing fairness correction methods, such as the one proposed by Purdy and Glass (2023). This approach involves adjusting classification thresholds based on protected attributes, such as demographic group membership, to create a multi-tiered scoring system that supports more precise and context-specific decision-making. By customizing thresholds according to the specific needs of different demographic groups, this strategy can help achieve fairer outcomes. Further studies are needed to refine these thresholds and assess the effectiveness of such methods in the NZ child welfare context, ensuring they are both practically implementable and ethically aligned with the goals of reducing disparities.

5. **Developing Frameworks for Regular Model Updates:** As socio-economic conditions and the factors influencing child welfare concerns evolve, predictive models must be periodically retrained on the most current data to remain effective and relevant. Future research should focus on developing frameworks for regular model updates, incorporating a data refreshment mechanism to ensure that the models adapt to emerging patterns and trends in child welfare. This approach should also include the re-evaluation of fairness constraints to ensure that the models continue to mitigate biases effectively. Regular retraining will not only maintain predictive accuracy but also ensure that fairness metrics are consistently upheld, thereby enhancing the model's long-term utility and ethical viability.

6. **Conducting Longitudinal Studies:** To fully understand the impact of predictive risk models on child welfare outcomes, longitudinal studies are needed to assess their long-term effectiveness in real-world settings. Evaluating the performance and fairness of these models over time will provide valuable insights into their strengths and limitations. Such studies should also examine the models' effects on different demographic groups to identify any unintended consequences and inform future model refinements. Understanding the longitudinal impact of these models will be crucial for their responsible deployment in child welfare systems.

7. **Addressing Data Bias in Administrative Records:** Administrative data, while extensive, often reflect institutional practices and operational priorities rather than a complete or representative view of the population. This can introduce data bias, particularly underrepresentation or misclassification of certain ethnic or socio-economic groups which may in turn affect model accuracy and fairness. Future research should focus on identifying, quantifying, and mitigating these biases through statistical auditing, bias detection techniques, or debiasing algorithms. Greater transparency and critical assessment of the sources and structure of administrative data will be

key to building more trustworthy and equitable predictive models in the child welfare domain.

By addressing these areas for future exploration, this thesis sets a foundation for ongoing research aimed at developing more robust, equitable, and ethical predictive models in child welfare. Tackling these challenges will be essential for the responsible and effective integration of predictive models into decision-making processes, ensuring they serve as valuable tools for supporting vulnerable children and their families while upholding the highest ethical standards in NZ.

7.7 Conclusion

This thesis has thoroughly developed a framework for fairness-aware predictive modeling within the NZ child welfare context, addressing significant concerns regarding biases in predictive risk models. By utilizing a research dataset from the Stats NZ IDI and applying machine learning techniques, the study has highlighted both, the potential and limitations of these models in mitigating ethnic disparities, particularly for Māori children. Our research introduces an in-processing approach to incorporate fairness constraints during the logistic regression learning phase, specifically focusing on disparate impact and equalized odds. Our approach demonstrated to a great extent, measurable improvements in fairness metrics, such as disparate impact, equality of opportunity, and error rate balance (also referred to as equalized odds), while maintaining reasonable predictive accuracy. Our results extensively suggest that fairness-aware models can effectively reduce ethnic disparities in error rates.

This research, on the other hand, also highlights the trade-offs between fairness and other key metrics, such as calibration. While the fairness-aware approach improved error rate balance, it resulted in a decrease in calibration between groups, raising important considerations for the real-world implementation of these models.

More importantly, our findings underscore the need for ongoing research and stakeholder engagement to refine these models and ensure their ethical and effective use in supporting vulnerable children. This work sets the stage for further advancements in the field, advocating for a nuanced, data-driven approach to developing robust, equitable, and contextually appropriate predictive tools in child welfare.

Chapter 8

Miscellaneous Topics: Additional Analysis

8.1 Introduction

This chapter presents two independent analyses that stem from the primary findings of this research, each with distinct objectives:

- To investigate the impact of the COVID-19 pandemic on child protective service records, particularly records related to care and protection events, and assess how these disruptions may have influenced the definition of the outcome variable (*estimated care and protection concern*) used in this thesis.
- To explore the application of Clustering Analysis (CA) as a methodological approach to enhance predictive risk modeling in the child welfare context, examining whether unsupervised learning techniques can improve model understanding and segmentation.

These analyses were identified as important for advancing our understanding of how both external disruptions and methodological innovations may influence model development and performance in the Aotearoa New Zealand child welfare context. They are presented as two stand-alone articles within this chapter.

The main objective of this thesis has been to examine potential unfairness associated with the development of predictive risk models and the application of fairness-aware machine learning in the NZ child welfare context. To support this investigation, it was crucial to develop a model with a sufficient level of predictive power, aiming for an AUC of at least 0.70, in order to meaningfully assess and

explore fairness-aware methodologies.

Due to confidentiality constraints, we were unable to access the data used in the NZ government-commissioned project reported in (Rea & Erasmus, 2017). Consequently, we created a research dataset by extracting information from administrative data available in the Stats NZ IDI. Baseline logistic regression models were trained using predictors closely aligned with those in (Rea & Erasmus, 2017). Similarly, the outcome variable was defined to predict the probability that one or more events listed in Table 5.1 would occur within two years for each child in our sample, referred to as *estimated care and protection concern* within two years. However, baseline logistic regression models trained on a sample of 55,287 children notified in 2019 produced an average AUC of 0.68, notably lower than the 0.75 AUC reported by Rea and Erasmus (2017). Given that the logistic regression model in the (Rea & Erasmus, 2017) study was based on data from children reported in 2014, prior to the transition of the NZ child welfare system from CYF to Oranga Tamariki in April 2017, this decrease in AUC was expected due to natural variations in data distribution, shifts in population characteristics, and changes in welfare practices. Furthermore, the selection of a 2019 sample and the two-year outcome definition included the COVID-19 pandemic period. The pandemic, particularly during lockdowns, may have disrupted routine interactions with child protective services, impacting administrative data records.

This performance gap, along with the potential effects of these contextual changes, prompted further analysis. In Section 6.2, we examined the outcome variable's time frame to account for possible shifts in child welfare practices following the reforms. We also explored enhancing the model's predictive capability through data linkage, developing a novel research dataset that incorporated additional predictors from various government agencies and the Stats NZ Census (Section 6.3). Moreover, the impact of the COVID-19 pandemic on child protective service records, particularly those relevant to the outcome variable (*estimated care and protection concern*), was investigated, as detailed in Section 8.2. Finally, in an early and independent analysis, we explored the potential of utilizing Clustering Analysis (CA) methods as a preliminary step toward improving the performance of predictive risk models for use in the intake decision-making process of child protection agencies (Section 8.3). This paper is also available at: https://arxiv.org/abs/2308.04060.

8.2 Impact of COVID-19 on Care and Protection Event Records

8.2.1 Introduction

The COVID-19 pandemic has had a significant impact on child maltreatment reports in NZ (Guardian, 2020; Robson, 2020; Tamariki, 2020). According to Guardian (2020), during the first phase of lockdown, reports of family violence to police dropped, and reports of concern to Oranga Tamariki, the country's welfare agency for children, fell by around 40%. In response to the COVID-19 pandemic, NZ promptly implemented an elimination strategy as its course of action. This approach encompassed a set of strong measures, including highly effective border controls, contact tracing, quarantine measures, and intense physical separation (lockdown) (M. G. Baker et al., 2020). While NZ's rapid response prevented it from experiencing the ravages of Delta as other countries did during the pandemic (M. G. Baker et al., 2020; Jefferies et al., 2020), the nation suffered through four phases of lockdown.

Following the first nationwide lockdown as part of the government's response to COVID-19, the NZ child welfare agency established weekly operational reporting on key statistics to maintain oversight of their operating system during the pandemic. A report by Tamariki (2020) provides a snapshot of those key statistics, including reports of concern, notifier groups, notifications that required further action, entries to care, and referrals to youth justice, and FGCs spanning the period of early March 2020 to late June 2020. According to Tamariki (2020), there was a decrease across all these key statistics when compared to the same time last year. Although this trend may appear as good news, child welfare experts viewed it differently. To them, fewer reports indicated that abuse and neglect of children had become less visible, not less prevalent (Robson, 2020).

Throughout the pandemic, many jurisdictions worldwide identified increased risks of abuse and neglect of children (Rapp et al., 2021; Rebbe et al., 2021). Considering that most abuse against children is perpetrated by family members who reside at home with the child (Bartlett et al., 2017; Hurren et al., 2018), stay-at-home restrictions, disruption of services during lockdowns, limited access to schools, and other risk factors (e.g. household stress, financial strains and unemployment) raised concerns regarding possible maltreatment of children (S. M. Brown et al., 2020; Bullinger et al., 2021; C. Katz et al., 2021; Lawson et al., 2020; S. J. Lee et al., 2021).

Despite this, various international studies have revealed a reduction in reports of abuse and neglect during the pandemic (S. M. Brown et al., 2022; I. Katz et al., 2021; McTier & Soraghan, 2022; Rebbe et al., 2021). For instance, using data from the Colorado Child Abuse and Neglect hotline and the

Colorado child welfare system, S. M. Brown et al. (2022) found a significant decrease in the number of referrals during the pandemic, particularly in the education and health reporter categories in Colorado, USA. However, it was found that the response time of the child welfare system to the referrals did not change significantly during this time.

Similarly, I. Katz et al. (2021) examined the impact of the pandemic on child maltreatment reports in Australia, Brazil, Canada, Colombia, Germany, Israel, and South Africa. This study used data from child protection services, child abuse hotlines, and official reports to compare the number of child maltreatment reports and the response time before and during the pandemic. The results of this study were consistent with (S. M. Brown et al., 2022) across all seven countries.

McTier and Soraghan (2022), on the other hand investigated the use of administrative data to understand the impact of the COVID-19 pandemic on child maltreatment in Scotland. The authors of the study found that administrative data on child protection referrals, assessments, and interventions were useful in understanding the impact of the pandemic on child maltreatment in Scotland. Based on the analysis of these data, they found that while the number of child protection referrals decreased, the number of assessments and interventions increased. The authors also found that certain populations, including families living in poverty and single-parent households, were at a higher risk of child maltreatment during the pandemic.

The findings from these studies raised concerns that the decline in reports of concern during the COVID-19 pandemic may have resulted in fewer intake decisions, investigations, and documented substantiated maltreatment cases. Consequently, if predictive models were trained on reports from this pandemic period, where the outcome variable was defined within that same time frame, there is a significant risk of misclassifying high-risk and low-risk children. This misclassification could occur because the models may not have been properly trained to capture true outcomes, given the disruptions in reporting and intervention processes during the pandemic.

In our analysis in Section 8.3, which included a dataset of 82,338 notifications involving 55,287 unique children and young people reported in 2019 under Section 15 of the *Oranga Tamariki Act 1989*, the outcome variable was defined to predict the likelihood of one or more events listed in Table 5.1 occurring within the next two years for each report of concern. Since the two-year follow-up period overlapped with the pandemic (2020-2021), and the AUC value of 0.68 was obtained from training a LASSO logistic regression model on predictors largely based on (Rea & Erasmus, 2017), we grew

concerned that the pandemic may have affected the records of interactions between NZ's child protective services and the public. As a result, this disruption likely contributed to the relatively poor predictive accuracy of our initial model, as it was trained on data that may not have fully captured the true outcomes due to pandemic-related anomalies.

The primary aim of this analysis was to utilize the Child, Youth, and Family data accessed through StatsNZ's IDI to examine the impact of COVID-19 on the records of interactions between NZ's child protective services and the public across different phases of the pandemic. This analysis focused on understanding how specific events during the pandemic influenced the outcome variables used in our initial analysis in Section 8.3 and throughout this thesis. More specifically, it aimed to determine whether statistically significant differences existed in these records during, before, and after the lockdown phases in NZ. This analysis contributes to the literature and our work in two key areas. First, it offers longitudinal data from January 2020 to April 2022, covering various phases of the pandemic, which is valuable for understanding the effects of lockdowns on administrative child welfare records. Second, it provides critical insights into how the pandemic may have impacted key events and adverse outcomes, an important consideration when selecting a sample cohort to train predictive risk models for the NZ population.

8.2.2 Method

8.2.2.1 Data Sources

We utilized records from five datasets from the CYF data (see Section 4.3.1). These datasets include comprehensive records of intakes, investigations, risk and safety assessments, and substantiated findings of maltreatment involving children reported to NZ's child welfare agency. The intake datasets capture details of each intake event, which occurs when an individual or agency reports concerns about a child being harmed, ill-treated, abused, neglected, or deprived to a child welfare agency or the police.

Following a notification or report of concern, an investigation or child family assessment is initiated when the initial assessment outcome indicates Further Action Required (FAR). Safety and risk screenings are performed for all care and protection cases that proceed to this stage. The safety and risk datasets contain the social worker's evaluation of the child's risk and safety, along with the rationale for subsequent actions. During the investigation phase, a social worker is assigned to conduct core assessments, which may result in substantiated findings of maltreatment, or a determination

that no further action is required (NFAR). In cases where additional steps are recommended, the social worker may suggest holding Family Group Conferences (FGCs) or Family Whānau Agreements (FWAs). The investigation datasets contain the full details of these assessments and actions.

Agency findings related to abuse are stored in the abuse datasets. Social workers use an abuse finding event table to record their determinations regarding whether a child has been abused. If maltreatment is identified, or if the outcome of the investigation phase is FAR, a full assessment is completed. Figure 2.1 provides a visual representation of this process.

8.2.2.2 Variables

The purpose of this analysis is to examine the child protection indicators outlined in Table 8.1 to assess how reports of child protection concerns changed between January 2020 and April 2022. This period encompasses the weeks leading up to the first lockdown and a few weeks following the final lockdown phase. Using records from the intake, safety and risk screening, investigation, and abuse finding administrative datasets, nine key variables were developed for this analysis. Specifically, we analyzed the weekly data on the number of reports of care and protection concerns received by NZ's child welfare agency, focusing on reports generated by schools, police, and primary health organizations. Additionally, the number of unique risk and safety assessments, as well as the investigations conducted, were examined. To further understand the impact of the pandemic, particularly the lockdown phases, we analyzed differences in the number of unique children receiving intake decisions, those receiving Family Group Conference (FGC) or Family Whānau Agreement (FWA) recommendations from social workers, and the number of children with substantiated maltreatment findings. This analysis aimed to assess how these variations in child protection records could influence the outcome variables used in predictive risk modeling in this thesis.

8.2.2.3 Data Analysis

For this analysis, and in alignment with the sample cohort construction described in Section 5.2.2 of this thesis, only notifications made under Section 15 of the *Oranga Tamariki Act 1989* were considered. These notifications, referred to as care and protection notifications, involve reports of concerns related to the safety and well-being of children. By focusing exclusively on these notifications, the analysis ensures consistency with the cohort criteria used throughout the thesis.

Throughout the analytical approaches in this paper, key time frames during the COVID-19 pandemic

Table 8.1: Dataset indicators used in this study.

Theme	Indicator	Tables in IDI
Reports of concern	-Number of unique care and protection notifications received. -Number of unique care and protection notifications made by police. -Number of unique care and protection notifications made by school. -Number of unique care and protection notifications made by primary health organizations (e.g., hospital).	Intake event Intake details
Assessments	-Number of unique risk and safety assessments conducted. -Number of unique Investigations carried out.	Safety and Risk Screen event Investigation event
Decision Outcomes	-Number of children with an intake decision made during the initial assessment phase. -Number of children with substantiated findings of maltreatment including emotional, physical, sexual and neglect. -Number of children with FGC or FWA recommendation from social workers.	Intake details Abuse event Investigation details

are referenced. These include the first lockdown (23-Mar-20 to 13-May-20), second lockdown (12-Aug-20 to 30-Aug-20), third lockdown (28-Feb-21 to 6-Mar-21), fourth lockdown (18-Aug-21 to 02-Dec-21), and the subsequent reopening phases: first reopening (14-May-20 to 11-Aug-20), second reopening (31-Aug-20 to 27-Feb-21), and third reopening (07-Mar-21 to 17-Aug-21).

The data analysis is organized into three primary components: descriptive analysis, statistical tests, and change point analysis. The descriptive analysis visualizes the data using time series charts, which illustrate patterns of notifications received, assessment processes undertaken, and related outcomes during the pandemic. Due to fluctuations in week-to-week data and the complexity of interpreting these charts, statistical tests were employed to determine the presence of statistically significant variations. Lastly, changepoint analysis was conducted to identify notable shifts across 2019, 2020, and 2021, taking into account school holidays. This comprehensive, multi-faceted approach provides a detailed exploration of the dynamics of child welfare indicators within the specified pandemic time frames.

8.2.3 Results

8.2.3.1 Descriptive Analysis

The initial analysis involved visualizing the data using time series charts, which illustrated how NZ's child welfare agency received notifications of care and protection concerns (reports of concern), conducted assessments, and documented the resulting outcomes during the pandemic period. These charts provided an overview of the fluctuations in reporting and response activities over time.

Notifications can be made through various channels, including calls, emails, in-person visits, fax, or automated records, such as Police Family Violence referrals. These notifications originate from a broad range of notifiers, including family members, community members, healthcare providers, schools, and legal entities (Oranga Tamariki, 2023d). The visual representation of these patterns allowed for a clear understanding of how the volume and nature of notifications, assessments, and outcomes fluctuated during different phases of the pandemic.

Official statistics and data show that the majority of care and protection concerns are reported by schools, police, and primary healthcare organizations over the years. Therefore, our initial investigation focused on assessing the impact of the pandemic on care and protection notifications, particularly those originating from these key sources.

The time series for the weekly number of unique care and protection notifications from police, schools, and primary health organizations, as well as notifications from all sources, is presented in a single chart (Figure 8.1). Vertical red lines indicate the first day of each lockdown, while green lines mark the first day of the corresponding reopening phases. Since the majority of notifications come from schools, and notification volumes tend to decline during school holidays, the start and end dates of these holidays are marked with vertical blue dashed lines in the charts. This provides a clearer interpretation of any fluctuations in notification patterns, helping to distinguish changes influenced by the academic calendar from other factors.

As shown in Figure 8.1, the lowest numbers of care and protection notifications are observed during the first lockdown period. This is followed by an increase during the initial phase of reopening, with another decline coinciding with school holidays. This pattern highlights a consistent trend throughout the pandemic: a drop in notifications at the onset of each lockdown, followed by a period of stabilization and subsequent increases as the reopening phases began. Moreover, a consistent pattern of significant declines in care and protection notifications is evident when assessing the blue dashed lines, which represent summer and school term breaks. This observation aligns with long-established

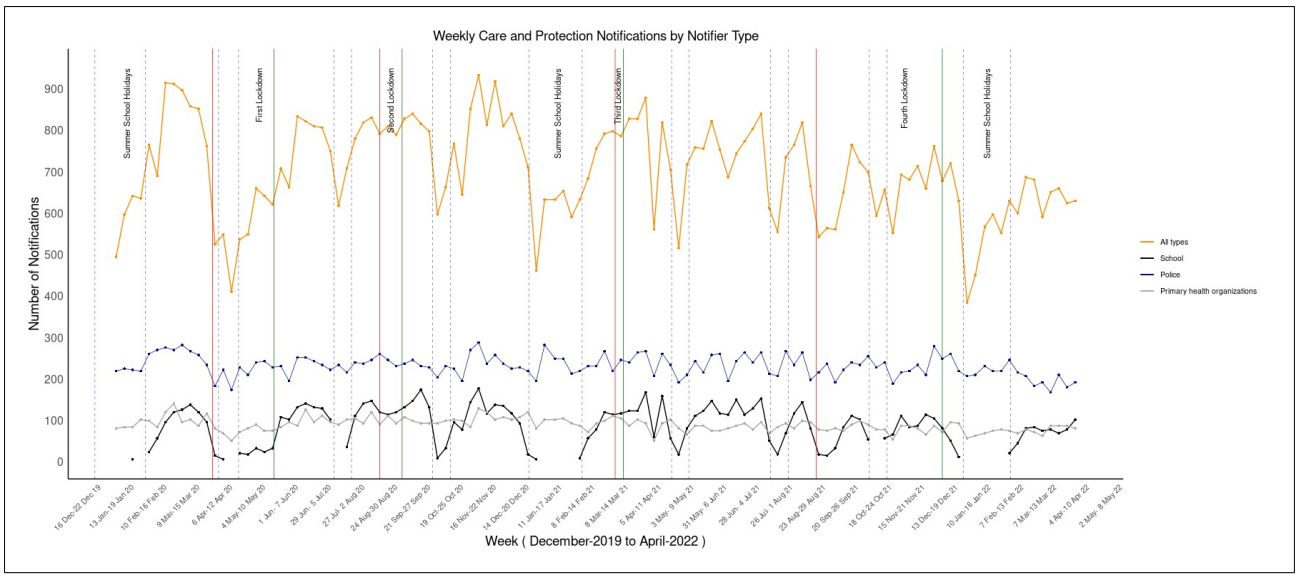

Figure 8.1: Weekly count of unique care and protection notifications received from various sources, including schools, police, primary health organizations, and all sources combined.

Note: The discontinuity in the time series for school notifications is a result of confidentiality requirements mandated by Stats NZ, leading to the suppression of counts during certain weeks. These suppressed data points indicate exceptionally low notification numbers for those specific weeks.

seasonal fluctuations, particularly those associated with school holiday periods (S. M. Brown et al., 2022; Petrowski et al., 2021). These declines emphasize the critical role schools play as notifiers in child protection systems, where the absence of regular school interactions during holidays can lead to reduced reporting, underscoring the influence of both lockdown restrictions and school schedules on the reporting of care and protection concerns.

To understand potential changes in the assessments conducted in relation to shifts in the number of notifications during different phases of the COVID-19 pandemic, three key indicators were examined. One indicator focuses on reports of concern (number of unique notifications for care and protection), while the other two are related to assessments (the number of unique risk and safety assessments and the number of unique investigations). Figure 8.2 provides a time series representation of these indicators, allowing for a comparative analysis of how each was impacted by the pandemic and its associated lockdown and reopening phases. From Figure 8.2, the impact of the decline in the number of care and protection notifications is not immediately observable in the same week but becomes noticeable in the following weeks. The reason for this delayed effect might be linked to the specific time frames associated with the assessments and investigations conducted in the child protection process (Oranga Tamariki, 2023h). This delayed manifestation emphasizes the complexity and temporal intricacies of the child protection system. It suggests that the consequences of changes in notification patterns take time to unfold, highlighting the sustained nature of child protection efforts over time.

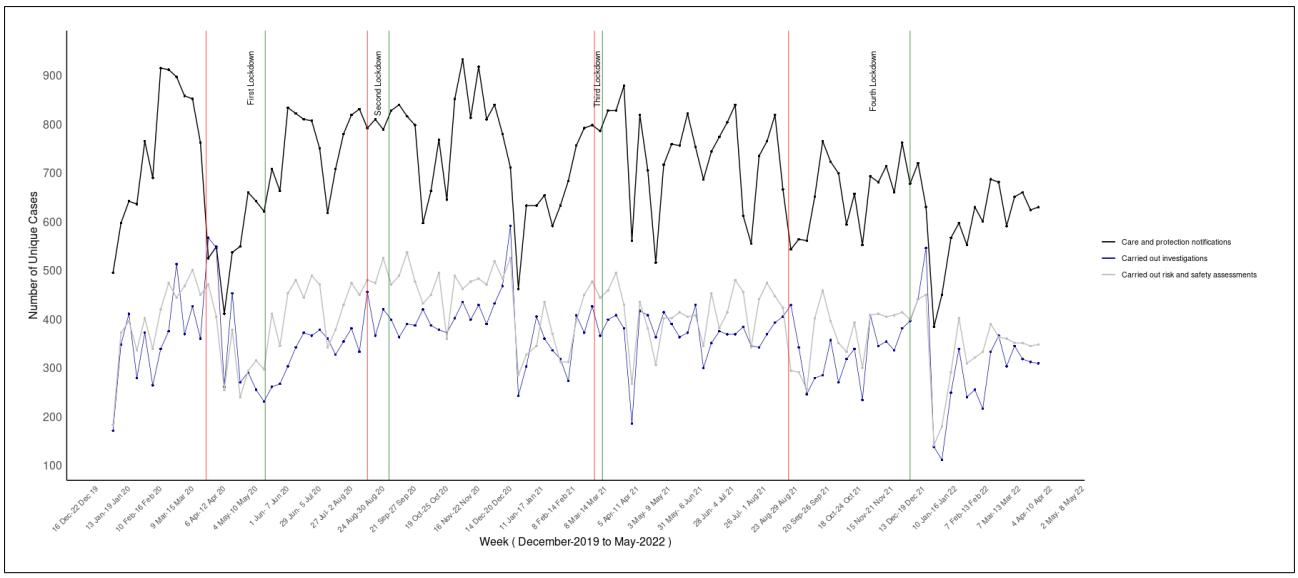

Figure 8.2: Weekly count of care and protection notifications, safety and risk screening, and investigations conducted from January 2020 to April 2022.

There are concerns that a decline in care and protection notifications may result in fewer intake decisions being made, investigations being conducted, and subsequently fewer FGC or FWA recommendations, as well as fewer substantiated maltreatment findings being recorded. To investigate this, three indicators related to events from Table 5.1 were considered. These events, referred to as care and protection-related events, have served as essential factors in defining the outcome variable used in this thesis (Section 5.2.1).

Figure 8.3 shows a consistent pattern that mirrors Figure 8.1: a decline in the number of children at the onset of each lockdown, followed by stabilization and increases during the reopening phases. A child is assigned an intake outcome when they are either referred for investigation or directed to other agencies for essential support services (Figure 8.1). This determination is made promptly during the initial assessment following notification. Consequently, a decrease in the number of notifications can directly impact the count of children with an intake outcome.

However, when examining children for whom social workers recommend a Family Group Conference (FGC) or Family Whānau Agreement, the time series does not display a significant difference between lockdown periods and reopening phases. Therefore, further analysis is required to investigate these changes in more detail.

An abuse finding event records a social worker's assessment of whether a child has experienced abuse. Since a child may have multiple abuse finding records, either due to multiple notifications requiring investigation or because they experienced more than one type of abuse (e.g., physical and emotional abuse) within the same period, this analysis considered the number of children with at

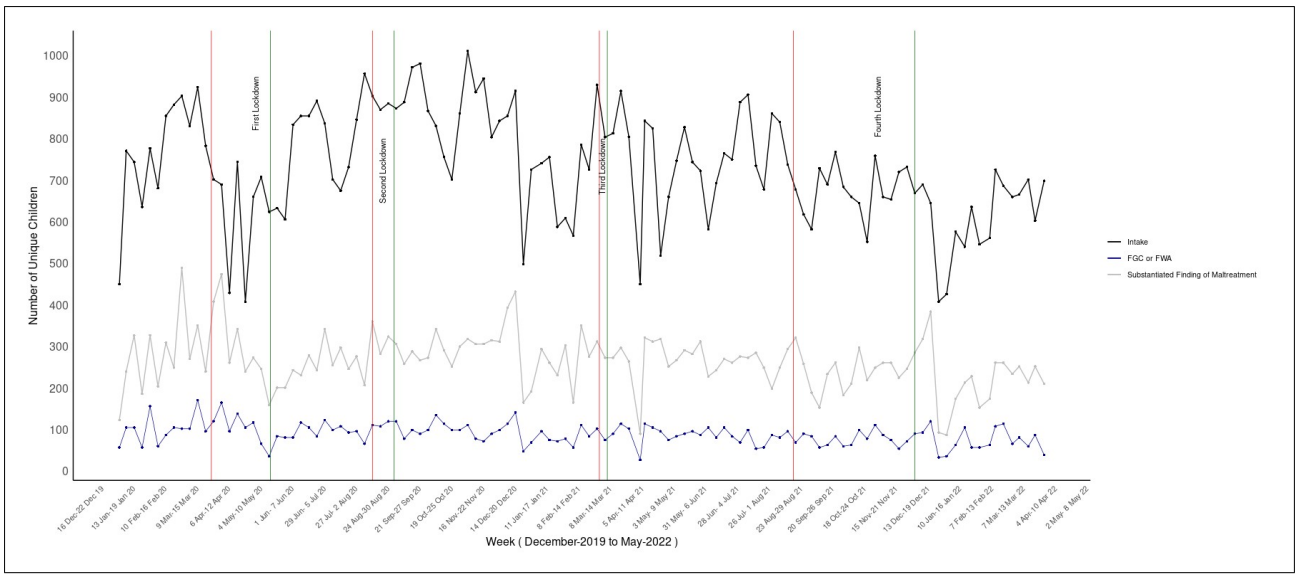

Figure 8.3: Weekly count of unique children with adverse outcomes including intake, FGC or FWA recommendations, and substantiated findings of maltreatment within January 2020 and April 2022.

least one type of maltreatment.

As depicted in Figure 8.3, the time series representing substantiated findings of maltreatment does not exhibit a consistent pattern throughout the lockdown periods. A decrease in the number of children with substantiated findings was observed only during the later weeks of the initial lockdown phase. This could be attributed to the fact that substantiated findings are documented following investigations, which in some cases can take up to 10 days (Oranga Tamariki, 2023h).

The time series charts offer a descriptive analysis, highlighting distinct and varied patterns across different phases of the COVID-19 pandemic in NZ. However, confidently interpreting these charts is challenging due to fluctuations in the week-to-week data. Therefore, statistical tests were employed to determine whether these variations were statistically significant.

8.2.3.2 Statistical Tests

To determine whether there were statistically significant variations in the variables under consideration between the lockdown periods and re-opening phases, data from the four lockdown phases (27 weeks) and three re-opening phases (61 weeks) were grouped and compared. The statistical significance of observed trends and patterns in the time series was assessed using either the Wilcoxon Rank-Sum test or the Welch two-sample t-test, depending on whether the assumptions of normality and homogeneity of variance were met.

Table 8.2: Median or mean weekly number of child protection indicators during lockdown and reopening periods with statistical comparisons using Wilcoxon rank sum test and Welch t-test.

	Lockdown Periods (n=27)	Re-opening Periods (n=61)	Test	P-value
Report of Concern				
From All sources	M=657	M=765	W=1249	0.0001203***
Schools	M<70	M<120	W=1152	0.002996***
Police	M=228	M=234	W=981.5	0.1541
Primary Health Organizations	M=81	M=96	W=1268	0.0000583***
Assessment				
Risk and Safety	M=396	M=441	W=1161	0.002293***
Investigations	μ=348.33	μ=371.46	t=1.2411	0.2226
Outcome				
Intake	M=690	M=807	W=1196	0.000762***
FGC or FWA	μ=88.67	μ=91.57	t=0.46078	0.6476
Findings of Maltreatment	M=258	M=276	W=960.5	0.2168

Note: Weekly median of notifications received from schools suppressed for data confidentiality requirements imposed by Stats NZ.

The normality and homogeneity of variance were tested using the Shapiro-Wilk (Shapiro & Wilk, 1965) and Levene statistical tests, respectively (Schultz, 1985). The Welch t-test was used for analyzing variations in the number of investigations conducted and the number of children for whom FGC or FWA were recommended. The Welch t-test is a more reliable adaptation of the t-test, especially when the two samples have unequal variances or different sample sizes (Welch, 1947).

The Wilcoxon Rank-Sum test was deemed suitable for the remaining variables. This non-parametric alternative to the independent two-sample t-test was chosen due to deviations from normality in several cases, as the Wilcoxon test remains valid for data from any distribution and is less sensitive to outliers compared to the two-sample t-test (Wilcoxon, 1992).

The Wilcoxon Rank-Sum test assesses differences in median values (M) between two groups, while the Welch t-test examines differences in mean values (μ). Therefore, Table 8.2, representing the results, includes mean values for indicators analyzed using the Welch t-test and median values for those subjected to the Wilcoxon Rank-Sum test.

The analysis illustrated in Table 8.2 reveals a substantial difference in the weekly number of unique care and protection notifications between lockdown periods (M=657) and re-opening periods (M=765), and this difference is statistically significant (W=1249, p<0.05). Examining notifications by notifier

type, notifications originating from schools and primary health organizations are significantly lower during lockdown periods than re-opening periods ($p<0.05$) compared to notifications from the police.

As previously noted, the decrease in the number of notifications correlates with intake decisions (Section 8.2.3.1). This pattern is consistently supported by statistical tests, which show that the number of unique children with an intake outcome was significantly lower during lockdown periods (M = 690) compared to re-opening periods (M = 807), W = 1196, $p < 0.05$. Importantly, there was no significant difference in the number of investigations conducted or the number of unique children for whom a social worker recommended a FGC or FWA. Similarly, no significant difference was found in the number of unique children with substantiated findings of maltreatment between the combined lockdown periods and the re-opening phases ($p > 0.05$). These findings align with the initial assumptions drawn from the time series data analysis.

Prior studies conducted by S. M. Brown et al. (2022) and McTier and Soraghan (2022) highlight that school closures hinder educational staff from effectively monitoring and observing their students. This limitation has the potential to reduce the frequency of care and protection notifications recorded in administrative data. Our findings align with these studies, as demonstrated in Figure 8.1, where consistent and sudden declines in care and protection notifications are evident during all school holidays. To ensure a more inclusive analysis and address this issue, we applied changepoint analysis, taking into account the start and end dates of school holidays.

8.2.3.3 Changepoint Analysis

Changepoint analysis was conducted to identify significant differences between selected weeks outside of school holiday periods in 2019, 2020, and 2021. The weeks were selected using data from the NZ Ministry of Education archive on school terms (Ministry of Education, 2023). While several U.S. studies have applied changepoint analysis to administrative data to examine shifts in child protection referrals during the COVID-19 pandemic (S. M. Brown et al., 2022; Nunez et al., 2023), similar research has not yet been conducted in NZ. This study addresses that gap by utilizing administrative data to explore shifts in key indicators, as outlined in Table 8.1.

Detecting changepoints involves identifying the point at which a sequence of observations experiences a significant change in its statistical properties (Killick & Eckley, 2014). The detection of a single changepoint can be treated as a hypothesis test, using a likelihood-based approach. This method, originally proposed by Hinkley (1970), was implemented using the *changepoint* package in

R to evaluate whether a significant shift in the mean occurred before and during the two major lockdowns in NZ (Killick et al., 2016).

The "at most one change" methodology was applied to each time series, testing the hypothesis that a single significant shift in mean weekly counts was likely attributable to COVID-19 restrictions. It was anticipated that the null hypothesis, indicating no changepoint during the specified period would be rejected. The analysis specifically aimed to detect statistically significant changes in the mean weekly counts of reports of concern, assessments conducted, and outcomes following the onset of the lockdown periods.

To minimize the impact of seasonal fluctuations caused by school closures, week groupings for 2019, 2020, and 2021 were selected from periods that did not overlap with school breaks in NZ. The changepoint analysis focused on two key lockdown periods: the "First Lockdown" and the "Fourth Lockdown". The analysis was subsequently divided into two parts, each corresponding to one of these lockdown phases.

Data for the first lockdown period were grouped into six full weeks prior to the arrival of COVID-19 in NZ (early February to mid-March, referred to as “pre-First Lockdown weeks”) and four weeks during the onset of COVID-19 and the resulting stay-at-home restrictions (mid-March to mid-April, referred to as “First Lockdown weeks”). For baseline comparison, similar 6-week and 4-week groupings were established for the year 2019. With respect to NZ fourth lockdown period, data were grouped into four full weeks prior to the lockdown (late July to mid-August 2021, referred to as “pre-Last Lockdown weeks”) and six weeks that included the lockdown phase (mid-August to mid-December 2021, referred to as “Last Lockdown weeks”). Comparable 4-week and 6-week groupings were defined for 2019 and 2020 as well. These groupings were selected to avoid periods influenced by school closures and seasonal variations, which are known to affect the number of child protection concerns (S. M. Brown et al., 2022). Table 8.3 and 8.4 presents the mean weekly counts for selected weeks before and after the lockdown phases in NZ, along with corresponding weeks from previous years.

Table 8.3: Weekly means of selected indicators before and during NZ's first lockdown for selected weeks, with the weekly mean of notifications received from schools during lockdown suppressed for confidentiality.

	Pre-First Lockdown Weeks		**First Lockdown Weeks**	
	Feb 05–Mar 18, 2019 (6 Weeks)	**Feb 05–Mar 17, 2020 (6 Weeks)**	**Mar 19–Apr 15, 2019 (4 Weeks)**	**Mar 18–Apr 14, 2020 (4 Weeks)**
Total Notifications	5,154	5,124	3,498	2,244
Total R & S Assessments	2,580	2,643	2,046	1,584
Total Investigations	2,082	2,580	1,656	2,043
	Weekly Mean (N)	**Weekly Mean (N)**	**Weekly Mean (N)**	**Weekly Mean (N)**
Reports of Concern				
All types	859	854	873	561
School	112	111	150	<50
Police	270	270	242	201
Health	106	105	133	80
Assessments				
Risk & Safety	429	440	510	396
Investigations	346	381	413	434
Outcome				
Intake	910	847	962	650
FGC/FWA	107	105	140	119
Maltreatment Findings	292	312	397	346

Note: Data confidentiality requirements imposed by Stats NZ. Additionally, all weekly mean values in the table are rounded down to the nearest whole number.

Table 8.4: Weekly means of selected indicators before and during NZ's fourth lockdown for selected weeks, with the weekly mean of notifications received from schools during lockdown suppressed for confidentiality.

	Pre-Last Lockdown			Last Lockdown		
	Jul 23–Aug 19, 2019 (4 wks)	**Jul 22–Aug 18, 2020 (4 wks)**	**Jul 23–Aug 19, 2021 (4 wks)**	**Aug 20–Sep 30, 2019 (6 wks)**	**Aug 19–Sep 29, 2020 (6 wks)**	**Aug 20–Sep 30, 2021 (6 wks)**
Total Notifications	3,312	3,219	2,982	5,064	4,881	3,804
Total R & S Assessments	1,770	1,836	1,782	2,853	2,976	2,091
Total Investigations	1,572	1,524	1,509	2,265	2,328	1,938
	Weekly Mean (N)	**Weekly Mean (N)**	**Weekly Mean (N)**	**Weekly Mean (N)**	**Weekly Mean (N)**	**Weekly Mean (N)**
Reports of Concern						
All types	828	804	744	844	814	634
School	123	129	103	153	137	61
Police	257	246	240	258	236	223
Health	113	102	91	104	99	83
Assessment						
Risk & Safety	442	459	446	475	495	348.5
Investigations	392	381	377	378	388	323
Outcome						
Intake	789	858	780	964	911	677
FGC or FWA	96	93	82	95	103	74
Maltreatment Findings	292	272	247	279	288	235

Note: Data confidentiality requirements imposed by Stats NZ. Additionally, all weekly mean values in the table are rounded down to the nearest whole number.

First Lockdown Analysis

During the first six selected weeks of 2019, total care and protection notifications amounted to approximately 5,154, compared to 5,124 in the same period of 2020, showing a high level of consistency (Table 8.3). However, following the introduction of COVID-19 restrictions, notifications dropped significantly from 3,498 in mid-March to mid-April 2019 to 2,244 in the same period of 2020, representing a substantial 36% decrease. This drop was most pronounced among notifications from schools and primary health organizations. While risk and safety assessments decreased by around 23%, the number of investigations completed increased by 19%.

The results of the changepoint analysis for the first lockdown are depicted in Figure 8.4, where the points of change are marked by vertical dashed lines. The analysis rejected the null hypothesis, which assumed no significant changepoint in the mean weekly care and protection notification counts between weeks 7 and 10. Instead, a statistically significant changepoint was detected at week 7, coinciding with the start of NZ's first lockdown in 2020. This shift in the mean for 2020 is represented by the red horizontal line segments in Figure 8.4. For comparison, the changepoint analysis for the same period in 2019 (weeks 1 to 10) is also included. However, no changepoint was identified at week 7 in 2019, reinforcing the conclusion that the changepoint in 2020 is likely related to the first COVID-19 lockdown restrictions in NZ.

Furthermore, the analysis identified a statistically significant decline in mean weekly intake decisions and risk and safety assessments at week 9 in 2020. This decrease, similar to the drop in notifications, aligns with NZ's stay-at-home restrictions during weeks 7 to 10. However, no significant changepoint was detected in the weekly counts of investigations, children for whom social workers recommended a Family Group Conference FGC or FWA, or in substantiated findings of maltreatment during the first lockdown period and the selected weeks.

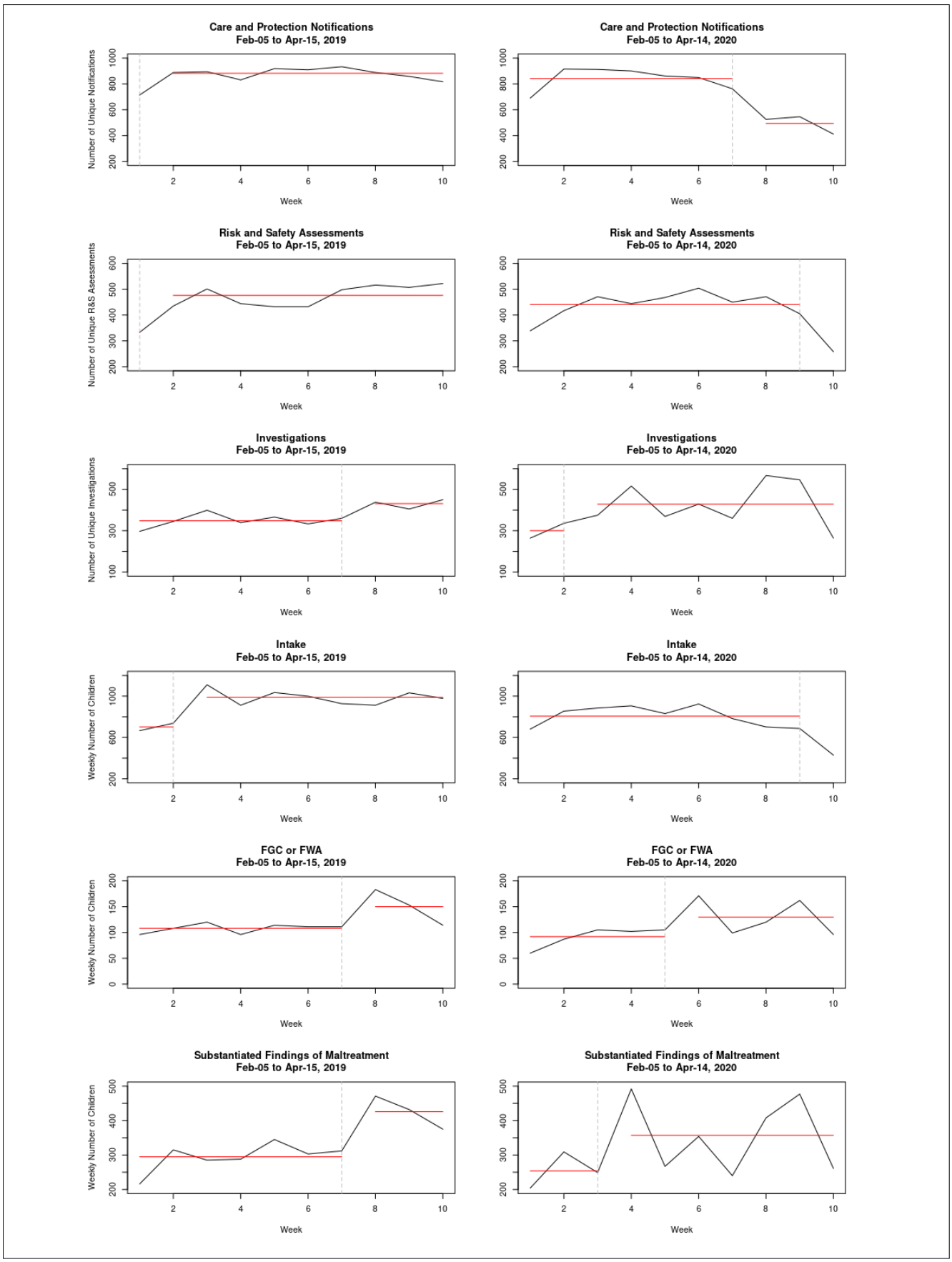

Figure 8.4: Identified changepoints in weekly care and protection, Risk and Safety assessments, investigations, Intake decisions, Recommendations for FGC or FWA as well as substantiated findings of maltreatment for 2019 versus 2020.

Fourth Lockdown Analysis

The total number of care and protection notifications during the first four selected weeks of 2019 was approximately 3,312, compared to 3,219 in 2020 and 2,982 in 2021 (Table 8.4). While these numbers show a gradual decline over the three years, the reductions were not statistically significant, with a roughly 3% decrease from 2019 to 2020, and a 7% decrease from 2020 to 2021. However, following the COVID-19 restrictions associated with NZ's fourth lockdown, a more substantial drop in notifications occurred. Between mid-August and late September, notifications decreased by approximately 25% compared to 2019 (from 5,064 to 3,804) and by 22% compared to 2020 (from 4,881 to 3,804). Similarly, the number of risk and safety assessments conducted, as well as investigations carried out, showed declines relative to both 2019 and 2020.

The changepoint analysis identified a statistically significant decline in the mean weekly care and protection notifications at week 3 of 2021, which occurred two weeks prior to the start of the fourth lockdown (Figure 8.5). Although this does not coincide with the official commencement of the fourth lockdown (week 5), it is evident that the mean weekly count was lower in 2021 than in 2019 and 2020. Changepoints were also identified at week 4 for risk and safety assessments, as well as for unique children receiving intake decisions. Additional changepoints were observed at week 6 during the lockdown phase for completed investigations and substantiated findings of maltreatment.

Although these results may not fully align with the changepoint analysis from the first lockdown, Figure 8.5 illustrates a consistent decline in the weekly counts of all considered indicators when compared to 2019 and 2020. To gain a clearer understanding of the changes across these years, time series analyses were conducted for the indicators associated with care and protection events, as outlined in Table 5.1. These indicators were used to define the outcome variable in this thesis. The time series comparisons for 2019, 2020, and 2021 are presented, with the weeks corresponding to the two major lockdown periods in 2020 and 2021 marked by vertical lines (Figure 8.6).

Figure 8.6 shows that the weekly count of unique children with intake decisions was generally higher in 2019 compared to the early days of the lockdown periods. While the counts in 2020 more closely align with the 2019 pattern, a more pronounced disparity is evident between 2019 and 2021. A two-sample t-test found no statistically significant difference between 2019 and 2020 ($t=1.8557$, $p=0.06638$). However, significant differences were observed between 2019 and 2021, as well as between 2020 and 2021 ($p<0.05$), with the lowest counts recorded in 2021.

Regarding FGC or FWA recommendations, although the weekly count was higher in some weeks,

it did not show a significant increase during the 2020 lockdown period. Statistical testing revealed no significant difference between 2019 and 2020 ($t=0.53206$, $p=0.5958$). However, as with intake decisions, significant differences were found between 2021 and both 2019 and 2020 ($p<0.05$).

A similar pattern was observed for substantiated findings of maltreatment. While no statistically significant difference was found between 2019 and 2020, a significant decline was identified in 2021 compared to both 2019 and 2020. Overall, the administrative data indicate that the lowest records for all key indicators were observed in 2021.

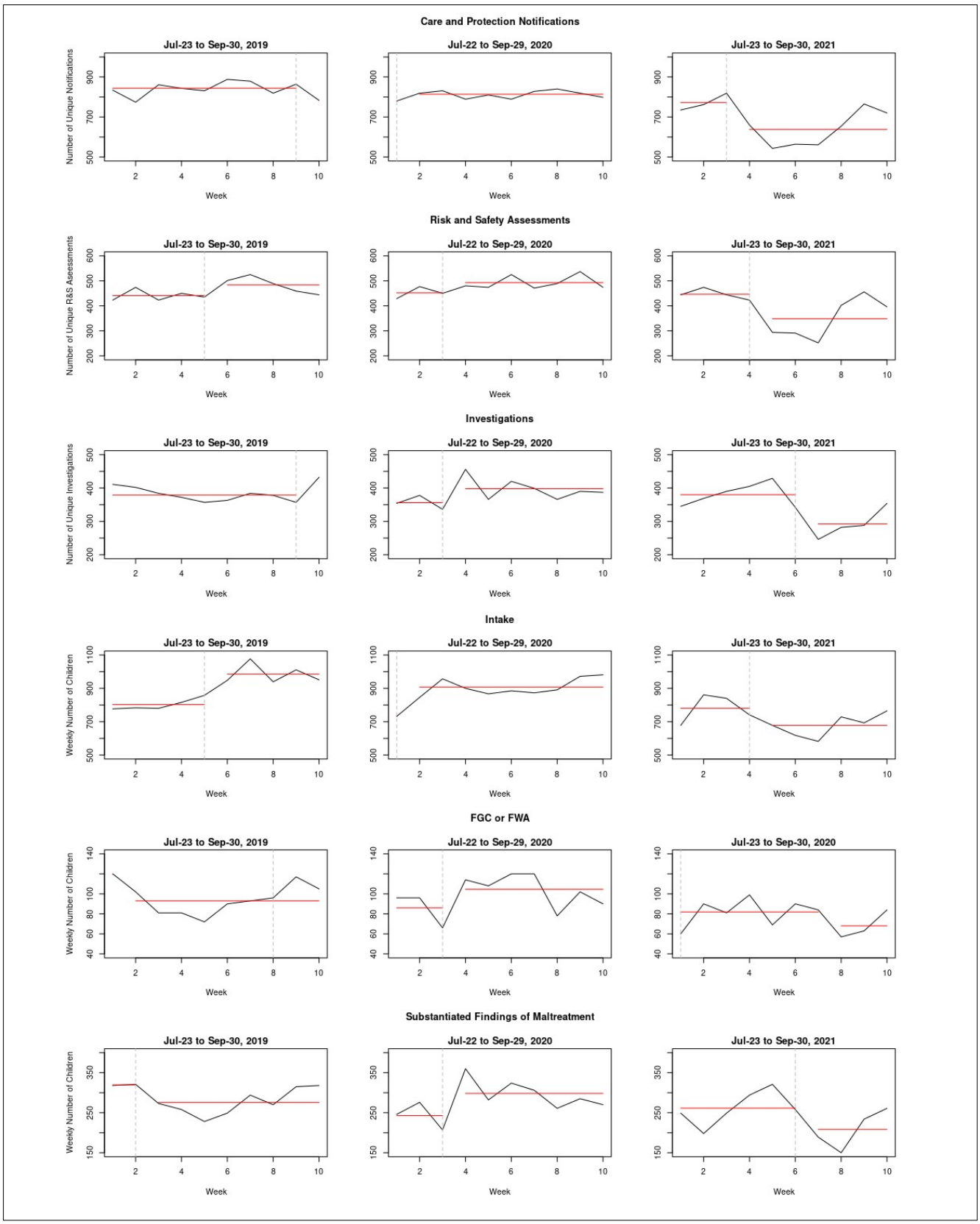

Figure 8.5: Identified changepoints in weekly care and protection notifications, risk and safety assessments, investigations, intake decisions, recommendations for FGC or FWA as well as substantiated findings of maltreatment for 2019 versus 2020 and 2021.

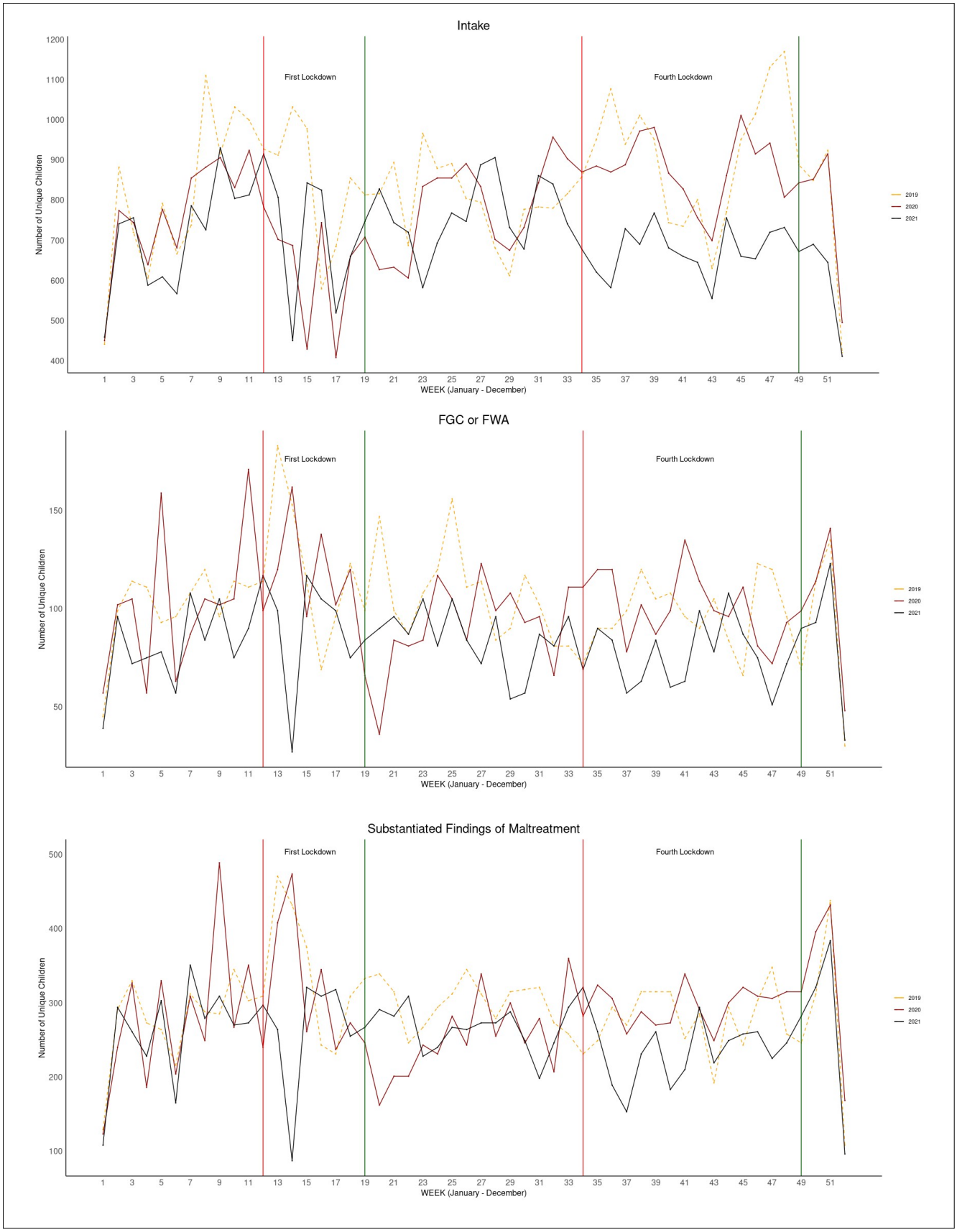

Figure 8.6: Identified changepoints in weekly care and protection notifications, risk and safety assessments, investigations, intake decisions, recommendations for FGC or FWA as well as substantiated findings of maltreatment for the years 2019, 2020, and 2021.

8.2.4 Discussion and Conclusion

This analysis explored the impact of the COVID-19 pandemic on interactions between child protective services and the public during different phases of the outbreak in NZ, using data from the Stats NZ IDI database, specifically the CYF records. The primary aim was to assess changes in intake decisions following reports of concern, FGC or FWA recommendations, and substantiated maltreatment findings. These records are crucial in defining the outcome variables used for training the predictive risk models discussed in Section 5.2.

A longitudinal approach was employed, examining data from January 2020 to April 2022 to capture shifts in care and protection notifications, assessments, and outcome records throughout the pandemic. For comparative purposes, data from January to December 2019 was also analyzed to establish pre-pandemic baselines.

The initial analysis revealed a consistent global trend; a reduction in reports of concern and notifications as well as intake decisions during lockdown phases, which align with findings from various jurisdictions (S. M. Brown et al., 2022; I. Katz et al., 2021; McTier & Soraghan, 2022). Factors such as stay-at-home restrictions, disrupted services, and limited access to schools have been identified as key contributors to this initial decrease, as reports typically received through schools and primary health organizations were significantly affected.

Child welfare experts globally have raised concerns that the decrease in reports may not reflect a true reduction in the prevalence of abuse and neglect, but rather a decline in its visibility (Robson, 2020). This discrepancy between visibility and prevalence highlights the complexities in interpreting child maltreatment data during a pandemic, as societal dynamics, reporting mechanisms, and protective factors undergo significant shifts. This issue raised concerns that a reduction in notifications could affect the definition of the adverse outcomes that child maltreatment predictive risk models aim to predict. The more nuanced impact becomes evident when considering the potential consequences of this decline in reporting on the responses of child protection agencies.

From this analysis, the impact of the decline in care and protection notifications on investigations, as visualized through time-series charts, was not immediately apparent in the same week as lockdown onsets but became noticeable in subsequent weeks. The reason for this delayed effect might be linked to the specific time frames associated with the investigations conducted in the child protection process (Oranga Tamariki, 2023h). This delayed effect may be attributed to the specific time frames required for completing assessments and investigations in the child protection process

(Oranga Tamariki, 2023h). These findings suggest that shifts in notification patterns take time to influence assessment outcomes, underscoring the sustained nature of child protection efforts. When combining the lockdown and reopening phases and applying statistical tests, no significant differences were observed in the record of investigations conducted by child protective services during the pandemic as well as substantiated findings of maltreatment. Similarly, no significant differences were found in the number of FGC or FWA recommendations by social workers as these outcomes are dependent on the results of investigations.

Changepoint analysis was employed as an additional method to examine significant shifts, taking school holidays into account to control for seasonal fluctuations caused by school closures. This approach has proven valuable in analyzing variations in child protection indicators during the COVID-19 pandemic (S. M. Brown et al., 2022; Nunez et al., 2023). Changepoint analysis enables the detection of significant changes in mean values, providing insights into the effects of key events, such as lockdowns, on care and protection notifications, assessments, and outcomes. One limitation of this analysis was the restricted number of weeks that could be included. Given the length of school terms in NZ, which range from 9 to 11 weeks, it was not feasible to include additional weeks without violating the predefined criteria. 8.2.3.3. Consequently, the analysis was conducted on carefully selected time frames to ensure consistency and reliability in detecting changepoints across the two distinct lockdown phases.

The changepoint analysis of the first COVID-19 lockdown revealed a statistically significant shift at week 7, coinciding with the commencement of the first stay at home restrictions (Figure 8.4). Additionally, a significant drop in mean weekly intake decisions and risk and safety assessments was identified at week 9 of 2020, aligning with the stay-at-home restriction period (Weeks 7 to 10). However, no significant changepoints were detected for the weekly count of investigations, recommendations for Family Group Conferences (FGC) or Family Whānau Agreements (FWA), or substantiated findings of maltreatment during this phase.

Examining the last lockdown phase, the changepoint analysis identified a statistically significant decline in the mean weekly care and protection notifications at week 3 of year 2021. This drop occurred two weeks before NZ Last COVID-19 lockdown phase. Although this does not align with the week of the fourth lockdown's onset (Week 5), it's apparent that the mean weekly count has overally decreased compared to 2019 and 2020. Changepoints were identified at week 4 for the weekly count of risk and safety assessments, as well as unique children with intake decisions. Change points for weekly count of completed investigations, unique children with FGC/FWA, and substantiated findings

of maltreatment was found at the weeks following the last lockdown phase in NZ. While these results may not be entirely consistent with the results from statistical tests and changepoint analysis applied to the first lockdown, a decline in the weekly count of all considered indicators was observed when compared to 2019 and 2020.

Several factors may explain this decline. One potential reason is the extended duration of COVID-19 restrictions in 2021, which may have affected reporting mechanisms, access to services, and the overall dynamics of child protection. Additionally, changes in child welfare policies or practices implemented by the NZ child welfare agency in 2021 could have influenced the recorded indicators. Modifications in reporting protocols, assessment criteria, or intervention strategies may have contributed to shifts in how cases were identified, assessed, and documented.

Given the potential influence of these policy changes or shifts in practice, further investigation into the NZ child welfare agency's policies during 2021 is crucial for future research. However, such an investigation is beyond the scope of this thesis.

Considering a longer time span for defining the outcome variable could be an effective approach to account for these shifts, especially when training predictive models. For instance, using a cohort of children referred in 2017 would likely be less affected by the fluctuations seen in 2021, as the longer time frame smooths out short-term disruptions. By employing a longer time frame, the model can better capture stable patterns in child protection indicators, mitigating the impact of transient changes, such as those caused by the pandemic. This approach enhances the robustness of predictive risk models, ensuring they remain reliable even in the face of significant societal or policy shifts.

8.3 Enhancing Predictive Risk Models with Clustering Methods

8.3.1 Introduction

Making decisions that affect children's lives is often a critical task and that predicting a child future risk of maltreatment at the time of notification is complicated. A predictive risk model is subject to errors and consequently may incorrectly identify as low risk some children who go on to experience abuse or neglect (FN) as well as identify as high-risk some children who do not (FP) (de Haan & Connolly, 2014). Clearly, these two types of errors may result in different harms: a FP may result in an unnecessary intervention or even the separation of a family. At the same time, a FN could lead the agency to fail to intervene when it should have . In child welfare, both classification errors are of concern. FNs can be dangerous to the child, while FPs can result in poor targeting of agency resources (Blank et al., 2015; Dare, 2013). Therefore, careful attention must be paid to the accuracy of these models.

This study is motivated by the need for further research on improving the accuracy of predictive risk models developed for potential use within the NZ child welfare system. Due to confidentiality obligations, we were unable to obtain access to the data used in the NZ government-commissioned project (Rea & Erasmus, 2017). As a result, we created our own research dataset mostly based on those described in (Rea & Erasmus, 2017). More details on the data creation process are provided in Section 8.3.2.1.

For the purpose of enhancing the accuracy of these models, we investigate the possibility of using Clustering Analysis (CA) methods as an early step in the development of predictive risk models for use within the child protection agency's intake decision-making. By using CA methods we expect children to be assigned to groups (clusters) so that children within each group are like one another. In contrast to the classification problem where each observation belongs to one of several groups, and the aim is to predict the group to which an observation belongs, CA seeks to discover the number and composition of the groups by identifying discrete, potentially hidden, groups of children. As part of our approach, we analyze the inner structure of these clusters to identify certain characteristics of children that may have a substantial impact on the error rate of these models and so be able to determine whether these errors can be reduced by training separate models for these subgroups of children.

The specific objectives are **1)** to create a research dataset including the outcome variable and predictor variables, **2)** to explore the existence of differing clusters within children reported with care

and protection concerns, **3)** to assess the performance of predictive models such as LASSO logistic regression on each subgroup of children (clusters) and, **4)** to determine whether separate models must be developed for children with specific features to attain better results.

8.3.2 Methods

8.3.2.1 Data

This study relied on a unique dataset that we constructed by record linkage between CYF data, Benefit Dynamics data, Personal details data as well as the 2018 Census records. For more details on these data please refer back to Section 4.3 of this thesis.

Starting with intake records from CYF data, our research dataset included 82,338 notifications involving 55,287 unique children and young people. For inclusion, the notification must have been made between 1 January 2019 and 31 December 2019 and under Section 15 of *Oranga Tamariki Act 1989*. Since some children are reported more than once during the year or in their lifetime, their initial care and protection notification in 2019 were considered for analysis. Children involved in these notifications were then linked to their records from Benefit Dynamics Data, the 2018 Census data and the Stats NZ demographic records (Personal Details data) to define the outcome variable and predictor variables. A more detailed explanation of the linking process is displayed in Figure 8.7.

Outcome Variable The outcome variable was defined based on events from Table 8.5. Our model intends to predict, for each report of concern, the probability that one or more of the events in Table 8.5 would take place within the next two years.

To form the outcome variable, subsequent events within the next two years were identified for the population of children and young people under analysis. The outcome variable reflected their status as to whether they have experienced at least one of the events from Table 8.5 within two years of their initial notification in 2019.

To allow for two years of records' follow up, we considered unique children and young people who were under 16 years of age at the time of notification in our analysis. The reason is that child protection datasets often contain records for children and young people under 18 years of age. As a result, the final dataset included observations involving 55,287 unique children and young people under the age of 16. From the children and young people notified in 2019, 48 percent were found to have experienced at least one of the events mentioned above within two years indicating that the data is balanced.

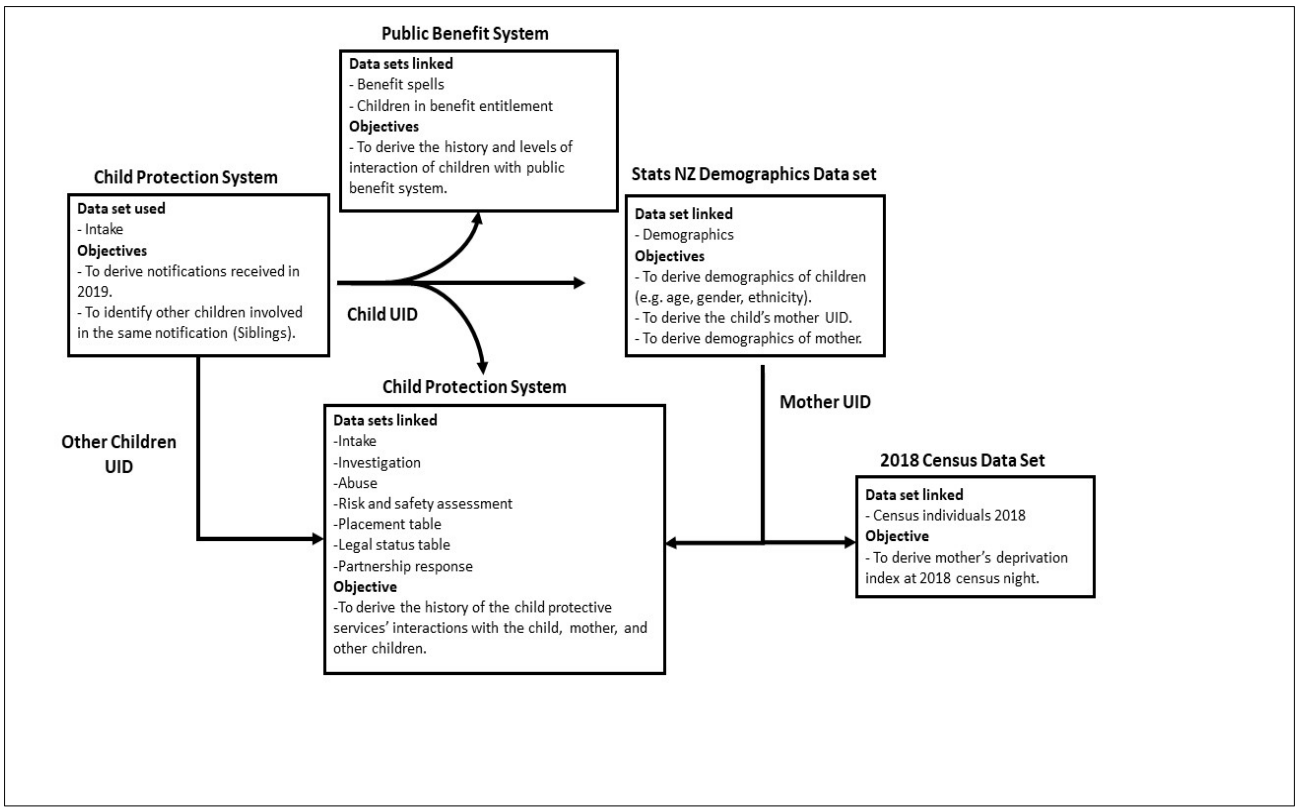

Figure 8.7: The process of data linkage and research dataset development.

Table 8.5: Care and protection-related events used to define the outcome variable (*estimated care and protection concern*).

Event 1	A substantiated finding of maltreatment including physical, sexual, emotional abuse, or neglect.
Event 2	A site social worker recommending a FGC or developing a FWA.
Event 3	The child or young person being the subject to a further notification which will be assessed as an intake.

8.3.2.2 Predictor Variables

The predictor variables were classified into four main groups such as *child* predictors, *caregiver (mother)* predictors, *family* predictors and *others*. These variables were created using the records available in the datasets we had access to through IDI and based on the risk factors of abuse and maltreatment identified in literate and academic studies, mainly the ones described in (Rea & Erasmus, 2017). These variables and their descriptions are represented in Tables C.1-C.4 of Appendix C.

8.3.2.3 Handling Missing Data

For training the predictive model for the total population, the missing values for the mother's *age* was imputed by the mean. However, the missing age for observations within each cluster was imputed

Table 8.6: Numerical predictor variables used for clustering analysis (SD stands for standard deviation).

Numerical Variable (n=55,287)	Mean	SD
Child age at the time of notification	7.20	4.68
Number of previous care and protection notifications	3.22	4.31
Number of days since last intake	609.10	917.01
Number of substantiated findings of maltreatment	1.15	2.27
Mother's age at the time of notification	33.70	7.48
Number of other children reported at the same time	1.76	1.55
Number of previous notifications for the children reported at the same time	7.46	13.16

based on the mean age of mothers within that cluster. To deal with the missing values for the categorical variables (e.g.,*gender, Ethnic group, Deprivation Index, etc.*), an additional category indicating the missing value was created (e.g. unknown). More information regarding the categories included in these variables can be found in Tables C.1-C.4 of the Appendix C.

8.3.3 Clustering Analysis

Clustering is an unsupervised machine learning technique used for grouping data into clusters based on their similarity. In CA process, observations are grouped according to their homogeneity and distinctiveness, regardless of the correlation structure of the predictors (Tryfos, 1998). In practice, it is generally challenging to determine whether clusters are an indication of phenotypic grouping, or if they are simply the result of dependency between variables. Machine learning-wise, we aim to identify reliable clusters of children reported with care and protection concerns using numerical predictor variables presented in Table 8.6 and then assess our predictive risk models trained on these clusters.

8.3.3.1 K-Means Clustering and Principal Component Analysis

The K-Means clustering algorithm is a partition-based CA approach where 'K' observations are selected as initial cluster centers. Each observation is then assigned to its nearest cluster based on its Euclidean distance to each cluster center. Afterward, all cluster averages are updated, and the process is repeated until a convergence of the criterion function has been achieved (Vora, Oza, et al., 2013). The K-Means algorithm clusters large datasets into a specified number of clusters (K) by minimizing the squared error function. Therefore, some data may be misclassified as a result of outliers (Prabhu & Anbazhagan, 2011). Principal Component Analysis (PCA) has been proven to be a continuous solution to the cluster membership indicators for K-Means clustering (Ding & He, 2004a,

2004b).

PCA is a widely used statistical approach adopted as an effective method for unsupervised dimension reduction by extracting relevant information from datasets (Joliffe & Morgan, 1992). It also facilitates the discovery of hidden relationships and improve data visualization, detection of outliers, and classification within the newly defined dimensions (Prabhu & Anbazhagan, 2011). ding2004principal showed that the continuous solutions of the discrete K-Means clustering membership indicators are the data projections on the principal directions which are the principal eigenvectors of the covariance matrix. PCA reduces the dataset to a lower dimension, while ensuring that the least information is lost, and provides a better centroid point for clustering (Zhu et al., 2019).

8.3.3.2 Results

In this study, PCA was applied before K-Means clustering was performed. As PCA generates a feature subspace that maximizes the variance along the axes, the dataset was first standardized onto a unique scale with a Mean of 0 and a Variance of 1. Scaling improves the results produced by PCA which is a requirement for the optimal performance of many machine learning algorithms. Based on the output produced from PCA, and the eigenvalue criterion (Boehmke & Greenwell, 2019), three principal components were selected. After the selection of principal components using this criterion, the components were passed for K-Means clustering.

The initial step in the application of K-Means clustering algorithm involves the identification of an optimal number of clusters (K). For this purpose, the elbow method was used. Basically, the K-Means clustering algorithm is applied multiple times with different K values, and the within cluster sum of squares is calculated and plotted for each K value (Figure 8.8). On the elbow plot, the optimal number of clusters is defined as the point beyond which there is only a minor reduction in within-cluster variability. This is visually represented as the bend of the elbow. According to our elbow plot (Figure 8.8), the optimal number of clusters (group of children and young people) should be 4. Consequently, K-Means clustering algorithm (K=4) was applied on the three selected principal components as input. Figure 8.9 Visualizes the clusters identified by applying K-Means clustering via PCA with principal components (PC1-PC3) on the axes.

After completing the CA process and assigning each child to a cluster, we divided the children into four groups (datasets) based on their cluster membership. As a next step, separate models were trained for the total population and each subgroup of children. We describe the modelling process and the results from assessing the performance of trained models in the following section.

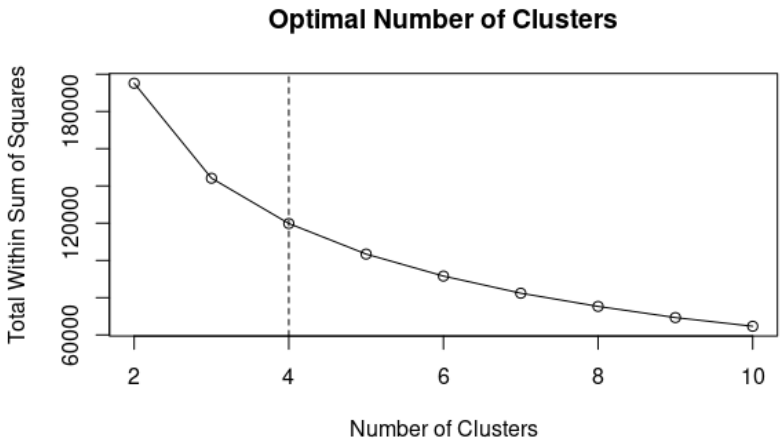

Figure 8.8: Elbow plot to select the optimal number of clusters.

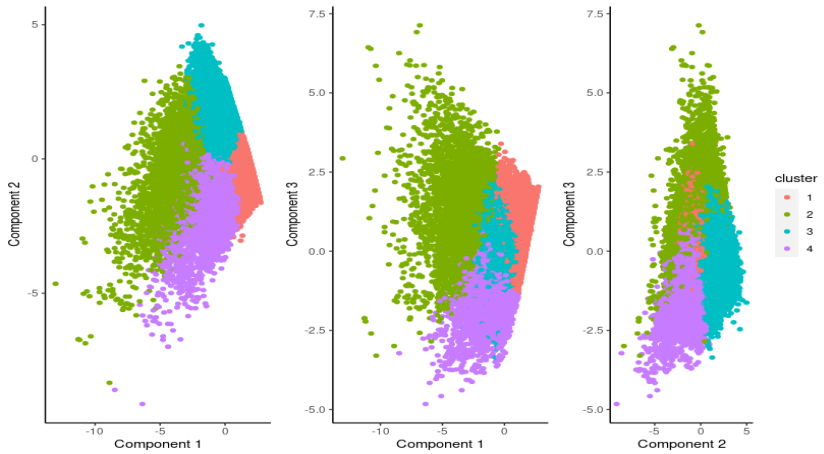

Figure 8.9: K-Means clusters (4) derived from PCA.

8.3.4 Modelling

Children and young people notified in 2019, and the datasets created based on their cluster membership were randomly split into a training dataset containing 70 percent of records and a testing dataset containing the remaining 30 percent. A training dataset was used to develop a model, and a testing dataset to assess how well the model can correctly identify children who will be subject to care and protection concerns within the next two years. As the main method of modelling in this study, we used LASSO logistic regression.

8.3.4.1 LASSO Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of an outcome variable based on predictor variables. Due to its simplicity and easy interpretation, this statistical modelling algorithm has been widely used in several fields, including biological sciences, social sciences, and machine learning (G. James et al., 2013).

Table 8.7: Performance results for models under analysis.

Model	AUC	TPR[1]	TNR[2]
Entire population	0.68	0.64	0.63
Cluster 1	0.70	0.63	0.66
Cluster 2	0.63	0.62	0.60
Cluster 3	0.63	0.57	0.61
Cluster 4	0.68	0.66	0.62

In its simple form, logistic regression utilizes weights for all predictors despite their significance and the potential for model overfitting. By contrast, the LASSO regularized form of logistic regression effectively selects only the most important predictor variables by shrinking the regression coefficients with the least important predictor variables to zero while minimizing prediction error, given the sum of the absolute value of the weights is less than a constant. Thus, it is capable of both predictor selection and regularization, which results in more easily interpretable and more accurate models (Tibshirani, 1996).

In addition to these advantages, there were other reasons why we chose LASSO as our method for training models. As a first point, although logistic regression was selected as the most appropriate model in previous NZ studies (Rea & Erasmus, 2017; Vaithianathan et al., 2013; Wilson et al., 2015), and version 1 of the AFST, (Vaithianathan et al., 2017), however LASSO regularized form of logistic regression has not been tested for the NZ population. Furthermore, in most recent US studies, LASSO was selected as the best model based on its overall performance and accuracy for the specific high-risk groups in addition to equivalent level of accuracy for black children versus non-black children (Vaithianathan, Dinh, et al., 2019; Vaithianathan, Kulick, et al., 2019). Upon this approach we will be able to compare the results in terms of accuracy with the state-of-the-art approaches in child welfare settings. The LASSO model was implemented with the R package named glmnet (Friedman et al., 2021). In the model training process, the constant – often symbolized as lambda – was optimized using 10-fold cross validation approach.

8.3.4.2 Model Performance Assessment

A number of test set classification metrics are provided in Table 8.7 for the LASSO models trained for the entire population of unique children and young people notified in 2019 and their four clusters. AUC was used to evaluate the performance of LASSO when predicting outcomes using the test datasets. The ROC curves for LASSO applied to four clusters of children are shown in Figure 8.10.

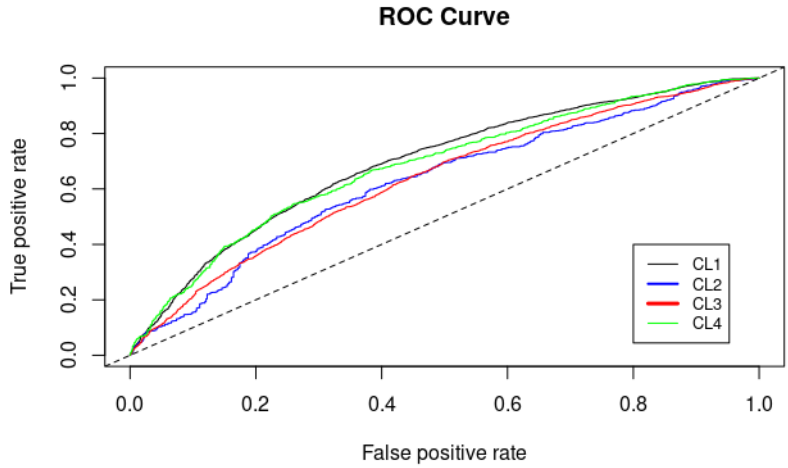

Figure 8.10: The ROC curves for LASSO Logistic Regression applied to four clusters of children.

The reason for considering AUC to measure the performance of these models is that child protection systems are often interested in developing a predictive tool that can supplement and standardize clinical decisions through a risk score or a summary statistics weighting information from the administrative data (Vaithianathan et al., 2017). A tool which allows for the use of empirically derived scores in combination with clinical judgement and other sources of data that are not instantly available to generate a screening decision. In this context, the AUC is a useful statistic for assessing the goodness of fit or prediction accuracy. There are various interpretations of AUC, but the one that is particularly useful in this context is that it can be understood as the probability that a randomly selected child that is a true positive (i.e., has had a care and protection concern) has a higher risk score than a randomly selected child that is a true negative (i.e., has not been the subject of a care and protection concern within 2 years).

8.3.4.3 Results

According to the AUC measures in Table 8.7 and the ROC curves shown in Figure 8.10, LASSO appears to perform slightly better for group of children in cluster 1 than other clusters (0.70), whereas its performance is the poorest for group of children in cluster 2 and cluster 3 (0.63). The AUC measures also indicate that the models generated for group of children in cluster 1 and cluster 4 perform very closely. True positive rate (TPR), however, is higher by 3 percent for cluster 4 (0.66), and True Negative Rate (TNR) is higher by 4 percent for cluster 1 (0.66). It is unclear why models trained on

[7] **True Positive Rate (TPR)** is the fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, TP/(TP+FN).

[8] **True Negative Rate (TNR)** is the fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, TN/(TN+FP).

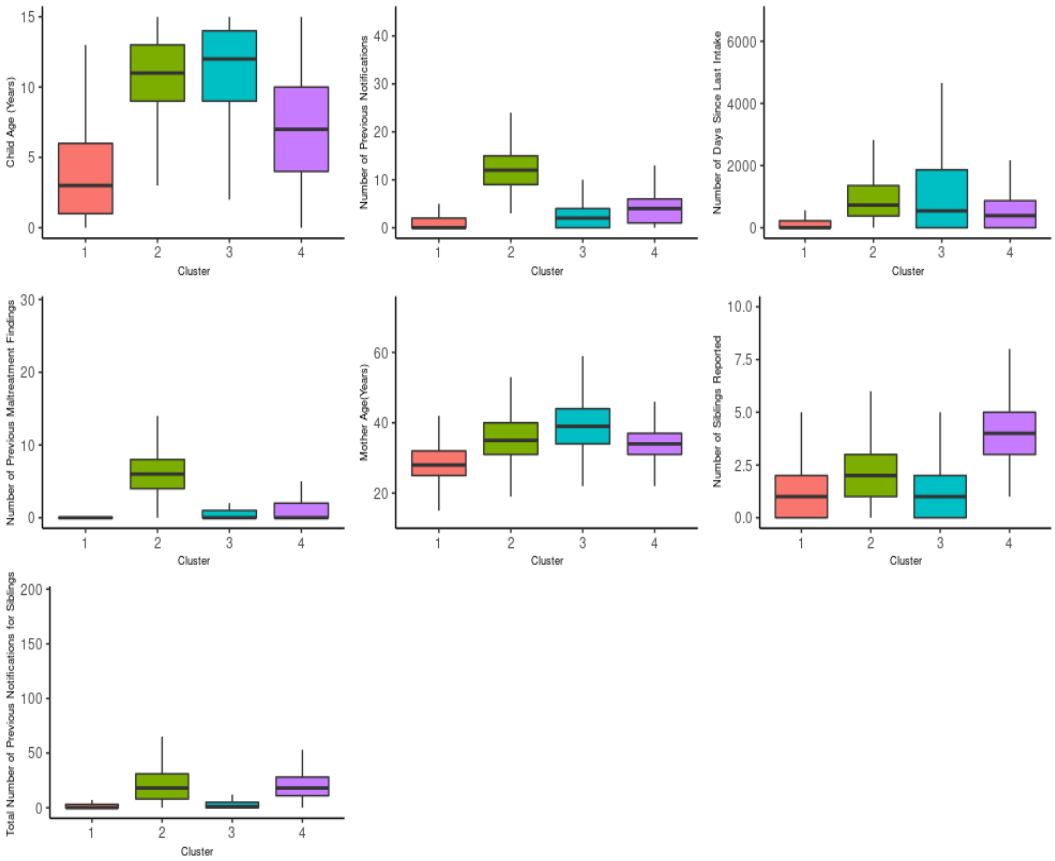

Figure 8.11: Distribution of numerical variables in four clusters.

subgroups of children in cluster 1 and 4 have slightly lower error rate. However, one way to investigate this was to examine the characteristics of children within each cluster by analyzing descriptive statistics of the variables.

The baseline characteristics for four clusters are described using standard descriptive statistics for the seven numerical variables used in the cluster analysis (Table 8.8). In addition to Table 8.8, Figure 8.11 provide a better representation of cluster characteristics based on these 7 numerical variables.

Table 8.8: Descriptive statistics for four clusters (data in the table are Means ± SD).

Variables[3]	Cluster1 (n=23,658)	Cluster2 (n=5,544)	Cluster3 (n=17,901)	Cluster4 (n=8,187)
Var1	3.44 ± 2.94	10.78 ± 2.96	11.18 ± 2.86	6.92 ± 3.85
Var2	1.11 ± 1.75	12.50 ± 5.02	2.81 ± 2.89	3.91 ± 3.03
Var3	193.90 ± 397.11	979 ± 805.02	1050 ± 1233.32	594.10 ± 665.15
Var4	0.30 ± 0.84	6.21 ± 3.38	0.70 ± 1.17	1.15 ± 1.47
Var5	28.41 ± 5.29	35.47 ± 6.12	39.47 ± 6.49	34.10 ± 4.91
Var6	1.25 ± 1.04	2.14 ± 1.56	1.24 ± 1.08	4.11 ± 1.36
Var7	2.31 ± 3.83	22.56 ± 20.85	2.96 ± 4.23	21.96 ± 17.58

Table 8.9: Performance results for models developed based on children age group.

Age group	AUC	TPR	TNR
Newborn	0.71	0.67	0.64
Newborn to Five	0.70	0.66	0.62
Newborn to Ten	0.68	0.66	0.61
Newborn to 15	0.68	0.64	0.63
Six to Ten	0.68	0.63	0.63
Eleven to Fifteen	0.67	0.60	0.64

Based on a comparison of the means within each cluster (Table 8.8), and the distribution of numerical variables illustrated by Figure 8.11, it appears that the means are lower for children in cluster 1. Specifically, the average age is lower for group of children in cluster 1 and cluster 4 with a higher AUC. As a result, there is a possibility that models will perform differently for young children as opposed to older children. To investigate this further, we categorized the children under 16 years of age notified in 2019 into different age groups and trained the Lasso models separately on each group. Several test set classification metrics are provided in Table 8.9.

Although the results did not demonstrate a significant difference in AUC for different age groups, the models did perform slightly better for children under the age of 5, and especially for new-born children, TPR and TNR appeared to be the highest (Table 8.9).

8.4 Discussion and Future Work

This study was conducted based on the data from 55,287 unique children and young people aged under 16 who were reported to the NZ child protective services in 2019. During the predictive risk

[9]Var1:= Child age at the time of notification ; Var2:=Number of previous care and protection notifications; Var3:=Number of days since last intake; Var4:= Number of substantiated findings of maltreatment; Var5:= Mother's age at the time of notification; Var6:=Number of Other children reported at the same time; Var7:= Number of previous notifications for the children reported at the same time.

modeling process using our developed research dataset, the LASSO algorithm trained on the entire population didn't perform as expected to produce a model to the desired predictive ability. While this may have been a consequence of the limited number of predictors, but we examined whether certain characteristics of children and young people were also influencing its performance. This paper presented our initial findings of our investigation on the identification of such features through CA techniques, as well as their potential effect on risk-modelling.

In a novel approach not previously considered in other studies from NZ, we utilized K-Means clustering via PCA to identify subgroups (clusters) of children based on observations of the numerical variables presented in Table 8.6. As well as being an effective method for reducing dimensions and extracting relevant data from datasets without supervision, PCA has been shown to facilitate the discovery of hidden relationships, improve data visualization, and detect outliers, resulting in more effective K-Means clustering (Ding & He, 2004a, 2004b; Prabhu & Anbazhagan, 2011). From the CA via PCA, children were divided into four groups and separate models were trained using LASSO. Results from CA and predictive modeling process revealed no significant differences in their performance (Table 8.7).

Two models, however, performed slightly better (Cluster 1 and Cluster 4) including younger children only (see Table 8.7 and Table 8.8). This is further evidence that our research dataset may not capture all potential risk factors for older children or generally predictive models are more accurate when developed for children under the age of five (Table 8.9). These assumptions are consistent with (Palusci, 2011) who examined risk factors for child maltreatment among US children. Infants and young children in the child welfare system were found to have different risk factors and are provided with different services than older children. Results from our study strongly suggest that age is a crucial factor for predicting child maltreatment and we are currently investigating this matter.

Generally, the predictive risk models trained in this study did show a slightly lower-than-standard predictive power and we could not find a significant difference in performance between the models trained for each cluster. However, our results can help identify potential factors to consider in future stages. Given that our research dataset may not capture all potential risk factors or predictors for older children, further research is required to determine whether the addition of new predictors will contribute to accuracy improvements of the models and reduction of the error rates (see Table 8.7). Consider, for example, the dataset created here included variables related to the mother (see Appendix), but similar variables could be generated for the father as well. Furthermore, it is possible to create new predictors by combining data from other organizations. Records from the Department

of Corrections available in the Stats NZ IDI database may also be extracted to verify whether a child lives with and adults who has recently been released from prison for an offence related to family violence. A further source of predictors is data provided by the Ministry of Health concerning addictions and mental health records of caregivers of children. Considering that these characteristics are identified as risk factors for child maltreatment (A. Austin, 2016; Ayers et al., 2019; Lopes et al., 2021), incorporating predictor variables that reflect these factors might enhance the accuracy of models that are developed for predicting child maltreatment.

Furthermore, our results on younger children, involving more accurate models, shed light on the importance of such age group in this analysis and can be used by child protection systems in other situations. In this work we investigate the modeling framework for potential use when notification is received by NZ child protective services and for children and young people under the age of 16. But, there are other potential situations in which a predictive risk model can be utilized, including pre or immediately after the childbirth. Moreover, using the model pre or post childbirth help to identify children who are at risk and provide high risk families with appropriate services in advance. For such models we will need to add live events data and maternity data available in Stats NZ IDI database. Maternity data helps gain information about the baby's condition at birth and mother's condition before and after giving birth to the child. Life event data is about life events relating to births, deaths, marriages, and civil unions registered in NZ.

Aside from the absence of certain predictors, there are other factors that may affect the performance of predictive risk models. Training models on biased data is one of them. Crucially, the source of any predictive model is the data on which the algorithm is trained. Any source of error will be translated into the output of the model (Barocas & Selbst, 2016). In particular, an error that changes an important factor about a child will reduce the accuracy of the model for that child (Glaberson, 2019). In the context of child welfare predictive tools, there are reasons to be concerned about the level of error present in the data being fed into the algorithm (Glaberson, 2019; Rea & Erasmus, 2017). The data used to develop such predictive models are often extracted from the child welfare agency's database systems and are linked to data collected by other government agencies. Data from government administrative systems is entered, initially, by human and therefore is subject to human error. For instance, names, addresses, or other vital information may be incorrect, information from one individual may be incorrectly linked to another, old and outdated information may persist, or information may be missing altogether (Glaberson, 2019; Rea & Erasmus, 2017). Feasibility studies and ethical reviews on the use of these models suggest that although the linkage of administrative data to

support predictive risk modeling is feasible, but the linkage is subject to error and a system for review would be needed in any implementation (Ministry of Social Development, 2014). This work, on the other hand, rely upon Stats NZ's data quality and have used identifiers created by Stats NZ for data linking.

Moreover, an important ethical concern has been whether predictive analytics methods will worsen existing racial disparities in child protection systems. In particular, past studies suggest that the presence of persistent racial bias reflected in administrative data have the potential to increase error rates and might lead to discrimination or unfairness against a certain group of people (Cuccaro-Alamin et al., 2017). This is an important factor that needs to be addressed during the process of developing predictive risk models. Specifically in the NZ child welfare system, the over-representation of Indigenous people of NZ (Māori) or other low socioeconomic status groups in the child welfare systems might be intensified by predictive risk models. If the data exaggerates risk, then its use in decision-making has the potential to feed a cycle of bias that leads to different population groups (such as Māori) being disadvantaged or discriminated against (Rea & Erasmus, 2017). Further work will be required to explore these findings and make sure the model can identify the risk as accurately as possible and does not unintentionally add to an over-representation of Māori within the NZ child welfare system.

Chapter 9

Ethical Considerations

9.1 Introduction

This chapter addresses the ethical considerations associated with our research on predictive risk modeling in the child welfare context. The Auckland University of Technology Ethics Committee (AUTEC) initially raised several concerns regarding this research. To address these concerns, we undertook a series of actions to satisfy AUTEC's requirements, culminating in full ethical approval granted on March 21, 2021.

9.2 Reconsidering the title of this research

Based on advice from an advisor at Oranga Tamariki, we changed the title of our research from "Non-discriminatory Predictive Risk Modeling in the child welfare context" to "Potential Unfairness Associated with the Development of Predictive Risk Models in the NZ Child Welfare Context." This decision was made with careful consideration of the ethical implications and to more accurately reflect the scope of our research.

This change was deemed appropriate because, although our research made significant efforts to mitigate unfairness, achieving full fairness while maintaining model accuracy is challenging. The new title better captures the investigative nature of our study, which aims to explore potential discrimination or unfairness in the process of developing predictive risk models.

Our completed research revealed that the predictive models developed had higher False Positive Rates (FPR) for Māori children compared to children from other ethnic groups. This finding underscores the importance of our title change, as it highlights the potential for discrimination and the need

for ongoing efforts to address these disparities in predictive modeling.

9.3 Reconsideration of the Treaty obligations involved in this research

In this section, we address the first condition on reconsideration of the Treaty obligations involved in this research.

The principles of Te Tiriti o Waitangi, as articulated by the Courts and the Waitangi Tribunal, provides the framework for how we needed to meet our obligations under Te Tiriti. The three principles derived from the Treaty of Waitangi, Partnership, Participation, and Protection should inform the boundary between Māori and research. Below we describe how we applied each of these principles in this research.

9.3.1 Partnership

We are committed to working collaboratively with iwi (tribe), hapū (sub-tribe), whānau (family), and Māori communities to ensure that Māori individual and collective rights are respected and protected. Partnership involves active engagement and mutual respect. To fulfill this principle, we initiated consultations with a senior advisor at the Office of the Children's Commissioner, and Dr. Valance Smith, Assistant Pro-Vice-Chancellor from AUT Māori Advancement Office. These consultations were aimed at identifying potential risks and ensuring that Māori perspectives are integral to the research process.

9.3.2 Participation

Involving Māori in the design, governance, management, implementation, and analysis of research is crucial to ensuring that their voices and perspectives are adequately represented. To adhere to this principle, We consulted with:

- Dr. Valance Smith, Assistant Pro-Vice-Chancellor from AUT Māori Advancement Office;
- Professor Tania M. Ka'ai, Senior Māori Cultural Advisor to the Pro-Vice-Chancellor and Dean;

Their involvement ensures that the research is culturally informed and that Māori participation is meaningful and impactful.

9.3.3 Protection

Actively protecting Māori and their collective rights, as well as Māori data, culture, cultural concepts, values, norms, practices, and language, was a central focus throughout the research process. Respecting the privacy and confidentiality of Māori and other communities highlighted in this research was our utmost priority.

All data used in this thesis was de-identified by Stats NZ. This process involved removing personal information such as names, dates of birth, and addresses. Identification numbers, such as IRD and NHI numbers, were encrypted and replaced with other numbers to ensure anonymity. Our access was strictly limited to the data necessary for our research project, and we did not have access to all information in the Integrated Data Infrastructure (IDI).

Furthermore, Stats NZ rigorously reviewed all outputs before they were released from the Data Lab. This review process ensured that the information was aggregated and grouped in a manner that made it impossible to identify individuals. These measures were essential to maintain the confidentiality and protect the privacy of all individuals and communities involved in the research.

9.4 Provision of evidence of consultation with Māori

In this section, we address the second condition from AUTEC about consultation with Māori on the use of their data, and the potential impact of this project's findings for Māori.

9.4.1 Consultation

Several organizations, advisors, and AUT academics were consulted to seek advice regarding the conditions required for obtaining full AUT Ethics approval.

Valuable insights and guidance were provided by key advisors from Oranga Tamariki, the Office of the Children's Commissioner, and the AUT Māori Advancement Office. It was recommended to contact the Office of Māori Advancement at AUT and consult with Māori academics in the field of social work to gather their perspectives on the research. Following this advice, discussions were held with relevant experts.

Based on these recommendations, the Māori Data Sovereignty Principles developed by "Te Mana Raraunga" were considered to address common issues related to the use of data about Māori (Te Mana Raraunga, 2018). Additionally, the principles of "Ngā Tikanga Paihere" were integrated into the

research framework (Stats NZ, 2020).

Furthermore, ethical commentary on the implications of predictive risk modeling for Māori, provided to the NZ Ministry of Social Development, was reviewed (Blank et al., 2015).

9.4.2 Use of Data about whānau Māori

We considered the application of the Māori Data Sovereignty Principles developed by "Te Mana Raraunga" and "Ngā Tikanga Paihere". Te Mana Raraunga provided us with the necessary framework to unpack and address common issues surrounding the use of data about Māori. Several sources provide further elaboration of these ideas, including Indigenous Data Sovereignty: Towards an Agenda (Kukutai & Taylor, 2016) and Keegan's work on Māori algorithmic sovereignty (P. T. Brown et al., 2023). We thoroughly studied these principles and integrated them into our research approach. The Te Mana Raraunga Brief offers a general overview of key Māori Data Sovereignty terms and principles (Te Mana Raraunga, 2018). These principles include:

1. Rangatiratanga (Authority)
2. Whakapapa (Relationships)
3. Whanaungatanga (Obligations)
4. Kotahitanga (collective benefit)
5. Manaakitanga (Reciprocity)
6. Kaitiakitanga (Guardianship)

Ngā Tikanga Paihere was developed in 2018 by the StatsNZ and Māui Hudson, Associate Professor at Te Pua Wānanga ki te Ao, Faculty of Māori and Indigenous Studies, University of Waikato. The framework was designed to guide the appropriate use of microdata in the IDI, with a focus on how data about Māori and other under-represented subgroups is used for research purposes. Ngā Tikanga Paihere also guides data users and researchers on how they could bring better insights to the data, by building relationships with communities from whom the data originates. The framework draws on 10 tikanga principles (te ao Māori/Māori world concepts) and aligns with the current model of the 5 Safes Framework that is used to manage safe access to integrated data at Stats NZ (Stats NZ, 2022b). The 10 tikanga principles outlined are represented in Table 9.1. The framework provided us with a clear understanding of the principles necessary to apply and follow in this research. By adhering to the guidelines of Te Mana Raraunga (TMR) and Ngā Tikanga Paihere, we ensured that

Table 9.1: Māori principles (tikanga principles) of Ngā Tikanga Paihere framework.

Tikanga	Description
Pūkenga	Knowledge and expertise
Kaitiaki	Data stewardship and governance
Whakapapa	Community relationships
Wairua	Community good
Pono	Accountability and transparency
Mauri	Data transformation and provenance
Tika	Value for all
Tapu	Sensitivity and risk
Wānanga	Organisations
Noa	Benefit and opportunity

our work with data was conducted in a culturally appropriate manner. Below, we outline the specific actions we took to fulfill these obligations, organized according to the relevant principles.

1. **Pūkenga (Knowledge and expertise)**: Appropriate expertise, skills, and experience is required for this research (Pūkenga).
    - As researchers, we possess the necessary data analysis skills and experience to conduct this study effectively.
    - We made every effort to work with the data in a culturally appropriate manner.
    - We respect the cultural values of Māori and the communities highlighted in the research.
2. **Whakapapa (Relationships)**: Suitable relationships are required to be established with whānau, hapū, iwi, Māori advisory groups, special councils, community members, expert advisors, or interested groups who are assisting the research.
    - We consulted with the AUT Māori advancement office.
    - We consulted with the Research and Evaluation Committee at Oranga Tamariki (Ministry of Social Development).
    - We consulted with the office of the Children's Commissioner.
3. **Pono (Accountability and transparency)**: We are required to be accountable and transparent to the communities in the use of data about them. Accountability means providing evidence that we are responsible to the communities impacted by our research findings. This includes demonstrating that the communities of interest or their representatives, advisors, leaders, or

advocates understand the objectives of our research.

- We were able to show evidence that the office of Māori Advancement, the office of children commissioner and the ministry of social development understand the objectives of our research.

4. **Tika (Value for all)**: The research must contribute to better outcomes for Māori and all NZers. It is essential to clearly explain the benefits this research will bring to communities, particularly those directly impacted by the data.

    - The objectives and significance of this study are thoroughly discussed in Chapter 1 of this thesis.

    - The findings of this research will help researchers and policymakers better understand the potential risk of discrimination in the government's use of algorithms in decision-making.

    - This thesis provides a detailed technical explanation of the steps taken to mitigate the risk of discrimination in algorithms, offering transparency to readers and the public.

    - If the findings are found effective for potential use by child welfare systems, they will enhance the ability of child protection staff to make more efficient and consistent decisions. This can help avoid unnecessary investigations, which are costly for the system and troublesome for families.

    - The research may significantly impact the lives of tamariki (children) and whānau (families) at risk of maltreatment by accurately identifying risk scores and preventing severe future outcomes.

5. **Wānanga (organisations)**: We required to gain the support of AUT and Oranga Tamariki in undertaking the research.

    - The School of Engineering, Computer, and Mathematical Sciences have supported the research by approving the initial research proposal.

    - Stats NZ IDI approval for use, and its approval through Ngā Tikanga Paihere.

6. **Kaitiaki (Data stewardship and governance)**: We required to make data management plans that have been developed with appropriate people. We also needede to apply careful, responsible, and ethical practices when using data.

- Data management plans were made by consulting StatsNZ and the academic supervisors at AUT.
- We applied Te Mana Raraunga (TMR) and Ngā Tikanga Paihere frameworks in this research.
- We attained AUT ethics committee and StatsNZ approval to get access to data.
- We used de-identified data to protect the privacy and confidentiality of individuals.

7. **Wairua (community good)**: We need to consider any potential harm, disadvantages, or risks to the communities of interest, particularly those made most vulnerable.
   - We addressed potential risks or negative impacts in this thesis.

   For the model to be ready for real-world use, it requires further validation, stakeholder input (particularly from Māori communities), and ongoing monitoring to ensure fairness and accuracy. Ethical and operational issues must also be addressed. This thesis will be shared with the Oranga Tamariki Research Centre and Stats NZ to support future research and discussion.

8. **Mauri (Data transformation and provenance)**: We required to clearly describe the datasets that we used, why the datasets were important to the research, and whether this research enhances or aligns with the original data collection purpose.
   - We assign a section to data description in this thesis.
   - We explained the significance of the datasets and variables used in this research.

9. **Tapu (Sensitivity and risk)**: We needed to ensure data safety, protect privacy and confidentiality, and ensure appropriate use.
   - Stats NZ provided us access to integrated data by meeting their 'Five Safes' conditions: safe people, safe projects, safe settings, safe data, and safe output.
   - We accessed research data in Stats NZ's secure virtual environment (the Data Lab).
   - We used data from the IDI, which is de-identified. This means that information such as names, dates of birth, and addresses were removed by Stats NZ. Numbers that could identify individuals, such as IRD and NHI numbers, were replaced with other numbers.
   - The data was not removed from the Data Lab; all analyses were conducted within this

secure environment.

- Stats NZ reviewed research results before their release to ensure that individuals could not be identified.

10. **Noa (Benefit and opportunity)**: We needed to demonstrate that potential risks have been balanced with benefits.

    - This research aimed to tackle potential discrimination associated with the process of developing Predictive risk models in the child welfare context. This research is exploratory, and it is only a theoretical possibility with no specific decision on whether the Predictive Risk Model developed will ever be used. The decision to demonstrate whether potential risks have been balanced with benefits depends on the stakeholders perspectives.

    - We intend to show that people of equal relevant characteristics must achieve equal outcomes by using these models regardless of their ethnicity (Equal Opportunity).

Noa (Benefit and Opportunity): We needed to demonstrate that potential risks have been balanced with benefits.

- This research aimed to address potential discrimination associated with the development of predictive risk models in the child welfare context. The research is exploratory and theoretical, with no specific decision made on whether the predictive risk model developed will be implemented. The assessment of whether potential risks have been balanced with benefits depends on the experience and perspectives of stakeholders.

- By this research, we intended to demonstrate that individuals with equal relevant characteristics should achieve equal outcomes using these models, regardless of their ethnicity, thereby promoting equal opportunity.

9.4.3 The potential impact of this project's findings for Māori

The development of a predictive risk model to assist professionals in identifying and investigating children at risk of abuse or neglect as part of a preventive early intervention strategy in NZ carries significant ethical implications. Recognizing these potential ethical risks, the NZ Ministry of Social Development commissioned two comprehensive ethical evaluations to scrutinize the development and implementation of predictive risk models. The first evaluation, conducted by Professor Tim Dare from

the University of Auckland's Philosophy Department, reviewed the ethical issues associated with implementing predictive risk models. Dare (2013) identified several potential sources of harm that could arise from the application of predictive risk models and proposed methods for mitigating these risks. His evaluation concluded that the use of predictive risk modeling could be ethically justified, provided that the recommendations outlined in his report are thoroughly addressed.

The second evaluation, carried out by Blank et al. (2015), provided a Māori ethical review of the predictive modeling using the Te Ara Tika framework (Hudson et al., 2010). This framework, grounded in tikanga Māori (Māori protocols and traditions), offers guidelines for Māori health research ethics. The review incorporated relevant ethical issues highlighted by Dare (2013) and raised by Māori commentators, adapting the Te Ara Tika framework to evaluate predictive risk modeling. This framework emphasizes four key principles derived from tikanga Māori:

1. Tika (Appropriate research, project, and design)
2. Manaakitanga (Cultural and social responsibility)
3. Whakapapa (Relationships)
4. Mana (Justice and social equity)

Both evaluations underscore the necessity of addressing ethical concerns to justify the application of predictive risk models. These reviews have been integral in shaping the ethical framework for our research, ensuring that the development and use of predictive risk modeling are conducted with rigorous ethical scrutiny and cultural sensitivity.

We reviewed these ethical evaluations and outline below some of the risks and potential benefits of using predictive risk models for Māori, as mentioned in (Blank et al., 2015). Since Māori are disproportionately represented in the high-risk group, there is an argument that a model of this kind, which helps target resources, is beneficial to Māori. Targeting based on the risk score will draw more resources towards Māori.

One of the risks highlighted by Blank et al. (2015) is hyper-vigilance. This occurs when families identified by predictive risk models have increased contact with social workers and other professionals, who might be more likely to identify maltreatment. Hyper-vigilance can be both beneficial and harmful. It is beneficial because it can lead to a reduction in harm caused by maltreatment, such as the removal of a child from a harmful environment. However, it can also be harmful if it leads to false accusations or the incorrect removal of children. Given that some children identified as high-risk do

not subsequently experience a maltreatment event, the sense of being monitored can be unpleasant for the majority of families. Since high-risk families and maltreated children are more likely to be Māori, both the potential benefits and harms of hyper-vigilance will disproportionately affect Māori. Our research aimed to explore the effects of the over-representation of Māori in the NZ child welfare system on model performance and to apply approaches to mitigate this bias stemming from predictive models. To address these risks, we have focused on developing methods that more accurately identifies true risk without inadvertently increasing the over-representation of Māori within the NZ care and protection system. We explored the statistical methods to reduce over- and under-identification errors in the model. Our efforts were directed towards creating a balanced model that minimizes these risks while enhancing fairness and accuracy in predictive risk modeling.

9.5 Conclusion and Discussion

Reviewing the Te Mana Raraunga (TMR), Ngā Tikanga Paihere, and Te Ara Tika frameworks, we conclude that we needed to develop a strong thesis of how the model is developed and to consult widely with Māori. The consultation agenda should include the rationale for predictive risk modeling, its performance, and an assessment of its relative benefits and burdens.

Appendices

Appendix A

This appendix provides supplementary material for this thesis, detailing the initial features encoded and predictor variables utilized. Tables A.1-A.7 present the initial features extracted from the data available in the Stats NZ IDI. These tables are categorized by data sources, including Child, Youth and Family, the Children's Action Plan, Personal Details, Sentencing and Remand, the Programme for the Integration of Mental Health (PRIMHD), and the 2018 Census. Additionally, Tables A.8-A.11 present the final set of predictors used for predictive modelling in this work. These predictors are organized based on levels such as child predictors, parent predictors, family predictors, and others.

Table A.1: Overview of initial features extracted from Child, Youth and Family data.

Feature Level	Domain	Description
Child	History of Interaction with CPS	• Substantiated findings of behavioral or relationship difficulties in the past. • Substantiated findings of suicide or self-harm in the past. • Substantiated findings of emotional, physical, neglect, or sexual abuse, or any type of maltreatment within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Section 15 reports of concern, intakes, investigations, risk and safety assessments, and placements within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Spells of care and protection custody or guardianship to the chief executive of Oranga Tamariki or another service provider within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Section 15 reports of concern resulting in no further action required within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • FGCs or FWAs within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Maximum and minimum number of days the child was placed in out-of-home care in the past. • Maximum and minimum number of days the child received services from other organizations (partnered response) in the past. • Number of days since last Section 15 intake. • Reports of family violence incidents to the police within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Whether the last placement phase ended within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Whether the last partnered response ended within the last 3, 6, 12, 24 months, or prior with respect to the date of notification. • Whether the child was in custody or under the guardianship of the chief executive of oranga tamariki or another service provider at the time of notification. • No history of report of concern, intake, substantiated findings of maltreatment, placement, or care and protection custody or guardianship spell. • Whether the child is already in an open social work phase at the time of notification, including investigation, risk and safety assessment, partnered response, or placement. • Whether the child was in full-time placement at the time of notification. • Last investigation outcome.

Table A.1 continued from previous page

Feature Level	Domain	Description
Parents	History of Interaction with CPS	• Whether they are known to Oranga Tamariki. • Whether there were substantiated findings of behavioral or relationship difficulties during their childhood. • Whether there were substantiated findings of suicide or self-harm during their childhood. • Whether there were substantiated findings of emotional, physical, neglect, or sexual abuse, or any type of maltreatment during their childhood. • Whether they had a history of intake by CPS as a child. • Whether they were placed full-time in out-of-home care during their childhood. • Whether they were placed due to care and protection concerns during their childhood. • Whether they were placed with the Youth Justice System (YJS) in the past. • Whether they were identified as a perpetrator of another child in the past. • Whether they perpetrated against the subject child in the past.
Family	History of Interaction with CPS	• Number of children in the same report of concern as the subject child. • Number of prior notifications for children in the same report of concern as the subject child. • Number of prior notifications for all siblings identified based on parents, including half-siblings.
Community	Others	• Notifier role type including court, family or *whānau* , health professionals, midwife or plunket, neighbours or friends, police, school or early childhood centre, unknown or others.

Table A.2: Overview of initial features extracted from Children's Action Plan data.

Feature Level	Domain	Description
Child	Demographis	• Whether the child has a disability.
	Supports and Services	• Whether the child and their family referred to children's teams in the past.
Family	Criminal History	• Whether family violence is present. • Whether drug or alcohol issue is present. • Whether there is involvement in the prison system. • Whether there is involvement in the corrections system.

Table A.3: Overview of initial features extracted from Personal Details Data.

Feature Level	Domain	Description
Child	Demographics	• Age at the time of notification. • Gender • Ethnicity
Parents	Identity	• Parents encrypted ID in StatsNZ IDI.
	Demographics	• Age at the time of notification. • Whether they were deceased at the time of notification.
Family	Identity	• Siblings encrypted ID in StatsNZ IDI.
	Familly Structure and Dynamics	• Number of siblings, including half-siblings.

Table A.4: Overview of initial features extracted from Benefit Dynamics data.

Feature Level	Domain	Description
Child	History of Interaction with Social Welfare System	• Whether included in a main public benefit at the time of notification including sole parent support, job seekers support, support living payment, and young parent support.
Parents	Family Structure and Dynamics	• Whether the main benefit caregiver is one of the parents. • Whether the main benefit caregiver is the partner of one of the parents.
Family	Family Structure and Dynamics	• Whether the relationship of the caregiver to the child is unknown.
	Socioeconomic Status	• Number of days on benefit in the past. • Number of benefit spells in the last 5 years. • Percentage on benefit in the last 5 years. • Number of food payments in the last 1, 2, 3, 4, and 5 years. • Number of power payments in the last 1, 2, 3, 4, and 5 years. • Number of clothing payments in the last 1, 2, 3, 4, and 5 years.

Table A.5: Overview of initial features extracted from Sentencing and Remand Data.

Feature Level	Domain	Description
Parents	Custody and Detention History	• Whether in custody at the time of notification. • Whether has been in custody in the last 5 years. • Whether under home or community detention in the last 5 years.
	Support and Services	• Whether completed an alcohol or drug rehabilitation program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether completed a rehabilitation program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether completed an education or training program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether completed a violence management program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether completed a sexual offense program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether completed any other program while in custody in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification.
	Record Status	• No record in sentencing and remand data.

Table A.6: Overview of initial features extracted from Programme for the Integration of Mental Health Data (PRIMHD).

Feature Level	Domain	Description
Parents	Mental Health Issues	• Whether diagnosed with mental health issues in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification.
	Substance Use Issues	• Whether diagnosed with drug use issues in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether diagnosed with alcohol use issues in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification.
	Support and Services	• Whether referred to mental health and addiction services in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification. • Whether received mental health, alcohol, and drug use support in the past 1, 2, 3, 4, 5 years, or earlier with respect to the date of notification.
	Record Status	• No record in PRIMHD data.

Table A.7: Overview of initial features extracted from the 2018 Census data.

Feature Level	Domain	Description
Child	Demographics	• Whether the child has a disability.
	Family Structure and Dynamics	• Whether the subject child is living with the mother only, father only, or both parents.
Parents	Socioeconomic Status	• Highest qualification achieved by parents.
	Family Structure and Dynamics	• Marital status of the parents. • Partnership status of the parents. • Whether either parent has a disability.
Family	Family Structure and Dynamics	• Sole parent household. • Number of people in the household. • Age group of the youngest child in the family. • Number of adult children in the family. • Number of dependent children in the family.
	Socioeconomic Status	• Deprivation index of the household.

Table A.8: Child Predictors

Variable	Coding Definition	Description	Number of Features
Age (Days)	Integer value	Age at the time of notification.	1
Age Group	Categorical: 0-Newborn One-Four Five-Eight Nine-Twelve Older Than twelve	Age group at the time of notification.	5
Gender	Categorical: -Male -Female	-	2
Ethnic Group	Categorical: -Māori -Pacific -European and Others	This variable classifies ethnic groups as follows: Māori group: Includes children who identify Māori as one of their ethnicities. Pacific group: Includes children who identify Pacific (but not Māori) as one of their ethnicities. European and Others: Includes children who do not identify Māori or Pacific as any of their ethnicities. This group encompasses NZ European, European, Asian, Middle Eastern, Latin American, African, and other ethnicities.	3
Disability Indicator	Binary: 0,1	This variable indicates whether the child has a disability.	1
Total Behavioural or Relationship Difficulty Instances	Integer value	This variable represents the total number of instances in which the child or young person had a substantiated finding of behavioural or relationship difficulties.	1

Table A.8 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Total Maltreatment Findings	Integer value	This set of variables represents the total number of substantiated findings of maltreatment for the child, including emotional abuse, neglect, or any type of maltreatment, within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	15
Total Sexual Abuse Findings	Integer value	This set of variables indicates the total number of substantiated findings of sexual abuse for the child within the last 6, 12, and 24 months, as well as prior to the date of notification.	4
Total Section 15 Reports of Concern	Integer value	This set of variables indicates the total number of Section 15 reports of concern for the child within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5
Total Section 15 Reports with Intake Outcome	Integer value	This set of variables indicates the total number of Section 15 reports of concern with an intake outcome for the child within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5
Total Child Investigations Completed	Integer value	This set of variables indicates the total number of investigations completed regarding the child within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5
Total Child Risk and Safety Assessments Completed	Integer value	This set of variables indicates the total number of risk and safety assessments completed regarding the safety of the child within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5
Total Section 15 Reports with "No Further Action Required" Outcome	Integer value	This set of variables indicates the total number of Section 15 reports of concern with a "No further action required" outcome for the child within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5

Table A.8 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Total Family Group Conferences Held	Integer value	This set of variables indicates the total number of Family Group Conferences held within the last 6, 12, and 24 months, as well as prior to the date of notification.	4
Duration of Child's Out-of-Home Care Placement	Integer value	This set of variables indicates the maximum and minimum number of days the child was placed in out-of-home care prior to the date of notification.	2
Duration of Partnered Response Support	Integer value	This set of variables indicates the maximum and minimum number of days the child and their family received support from other organizations.	2
Total Police Family Violence Reports	Integer value	This set of variables indicates the total number of police family violence reports within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5
Total Previous Custody Guardianship Spells	Integer value	This set of variables indicates the number of previous custody guardianship spells within the last 12 months, as well as prior to the date of notification.	2
Physical Abuse Substantiation Indicator	Binary: 0,1	This set of variables indicates whether the child was physically abused within the last 3, 6, 12, and 24 months, as well as prior to the date of notification.	5

Table A.8 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Outcome of Last Investigation	Categorical: -Assessment to be integrated into existing Intervention -Family Court Orders -Family Group Conference -Further Action -Family Whānau Agreement -No Further Action -Safety Assessment -Partnered Response -Refer to Service	This variable indicates the outcome of the last investigation completed regarding the child prior to the date of notification.	9
Out-of-Home Care Status at Notification	Binary: 0,1	This variable indicates whether the child was placed full-time in out-of-home care at the time of notification.	1
Child Placement History	Binary: 0,1	This set of variables indicates the child's placement history, including whether the child was placed in the last 3, 6, 12, and 24 months, or prior, and whether the placement phase ended in the last 12 months, 24 months, or earlier.	8
Recent Partnered Response Termination Status	Binary: 0,1	This set of variables indicates whether the child and their family stopped receiving assistance from other organizations within the last 3, 6, 12, or 24 months, or earlier.	5
Family Whānau Agreement History	Binary: 0,1	This set of variables indicates whether a Family Whānau Agreement (FWA) was signed within the last 3, 6, 12, 24 months, or earlier.	5
Custody or Guardianship Status at Notification	Binary: 0,1	This variable indicates Whether the child was in custody or under the guardianship of the chief executive of Oranga Tamariki or another service provider at the time of notification.	1

Table A.8 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Open Social Work Phase Status at Notification	Binary: 0,1	This variable indicates whether the child is already in an open social work phase at the time of notification. The phases include: Investigation, Risk and safety assessment, Partnered response, Placement	1
Absence of Previous Child Welfare Involvements	Binary: 0,1	This set of variables indicates whether there is any record of the following for the child: Section 15 reports of concern, Intakes, Substantiated findings of maltreatment, Placements, Custody guardianship spells	5
Past Referral to Children's Teams	Binary: 0,1	This variable indicates whether the child and their family were referred to Children's Teams in the past.	1
Child Living Arrangements	Binary: 0,1	This set of variables indicates the child's living arrangements: Living with mother only (1) or not (0), Living with father only (1) or not (0), Living with both parents (1) or not (0)	3
Inclusion in a Main Public Benefit	Binary: 0,1	This variable indicates whether the child is included in a main public benefit at the time of notification, including: Sole Parent Support, Job Seekers Support, Supported Living Payment, Young Parent Support	1

Table A.9: Parent Predictors.These predictors were encoded separately for both the mother and father of the child, unless otherwise indicated.

Variable	Coding Definition	Description	Number of Features
Age Group	Categorical: Under 20 20-25 26-35 Over 35 Unknown	This variable indicates the parents' age group at the time of notification.	10
Disability Indicator	Binary: 0,1	This variable indicates whether the parents have a known disability.	2
Highest Qualification	Categorical: -No Qualification -School Certificates -Post-School Certificate or Diploma -University Qualification -Unknown	This variable represents the highest qualification achieved by the parents, if known.	10
Marital Status	Categorical: -Never Married -Divorced -Widowed -Separated -Married -Unknown	This variable indicates the marital status of the parents, if known.	12
Partnership Status	Categorical: -Partnered -No partner -Unknown	This variable indicates the partnership status of the parents, if known.	6
Behavioural/Relationship Difficulties Indicator	Binary: 0,1	This variable indicates whether the parents were found to have behavioural or relationship difficulties as a child.	2

Table A.9 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
CPS Intake History Indicator	Binary: 0,1	This variable indicates whether the parents had a history of intake by Child Protective Services (CPS) as a child.	2
Childhood Maltreatment History Indicator	Binary: 0,1	This set of variables indicates whether the parents experienced various types of maltreatment as a child, including emotional abuse, physical abuse, sexual abuse, neglect, or any other type of maltreatment.	10
Childhood Out-of-Home Placement Indicator	Binary: 0,1	This variable indicates whether the parents experienced full-time out-of-home care placement during childhood.	2
Childhood Placement for Care and Protection Indicator	Binary: 0,1	This variable indicates whether the parents were placed out of their home due to care and protection concerns during their childhood.	2
Past Youth Justice System Placement Indicator	Binary: 0,1	This variable indicates whether the parents were placed within the youth justice system in the past.	2
Childhood Self-Harm or Suicide Indicator	Binary: 0,1	This variable indicates whether there were substantiated findings of self-harm or suicide during the parents' childhood.	2
Past Perpetration Indicators	Binary: 0,1	This set of binary variables includes: 1. Whether the parents were identified as a perpetrator of another child in the past. 2. Whether the parents perpetrated against the subject child in the past.	4
Maternal Deceased Status at Notification	Binary: 0,1	This variable indicates whether the mother was deceased at the time of notification.	1

Table A.9 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Custody and Detention Indicators	Binary: 0,1	This set of variables includes: 1. Whether the parents were in custody at the time of notification. 2. Whether the parents have been in custody in the last 5 years.	4
Home or Community Detention Indicator	Binary: 0,1	This variable indicates whether the parents were under home or community detention in the last 5 years.	2
Rehabilitation Program Completion Indicators	Binary: 0,1	These are separate binary variables indicating whether the parents completed rehabilitation programs while in custody in the past 5 years. 1. Alcohol or Drug Rehabilitation Program. 2. Mental Health Rehabilitation Program. 3. Violence Management Program Completion. 4. Other programs.	8
Father's Sexual Offense Program Completion	Binary: 0,1	This variable indicates whether the father completed a sexual offense program while in custody in the past 5 years.	1
No Custody or Detention Record	Binary: 0,1	This variable indicates whether no record was found for the parents in the sentencing and remand data.	2
Mental Health Referrals	Binary: 0,1	This variable indicates whether the parents referred to mental health and addiction services in the past 5 years.	2
Mental Health Issues Diagnosis	Binary: 0,1	This variable indicates whether the parents were diagnosed with mental health issues in the past 5 years.	2
Drug Use Issues Diagnosis	Binary: 0,1	This variable indicates whether the parents were diagnosed with drug use issues in the past 5 years.	2

Table A.9 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Alcohol Use Issues Diagnosis	Binary: 0,1	This variable indicates whether the parents were diagnosed with alcohol use issues in the past 5 years.	2
No Mental Health Issue Record	Binary: 0,1	This variable indicates whether no record was found for the parents in the Program for the Integration of Mental Health Data (PRIMHD).	2

Table A.10: Family Predictors

Variable	Coding Definition	Description	Number of Features
Co-reported Children Count	Integer value	This variable indicates the total number of children included in the same report of concern as the subject child.	1
Prior Notifications for Concurrently Reported Children	Integer value	This variable represents the total number of previous notifications involving children who are reported concurrently in the same instance.	1
Identified Sibling Count	Categorical: Zero or unknown 1-2 3-4 More than 4	This variable categorizes the total number of siblings identified for the subject child including half-siblings.	4
Total Notifications for All Siblings	Integer value	This variable represents the total number of prior notifications for all siblings identified based on parents, including half-siblings.	1
Youngest Child Age Group	Categorical: 0-Newborn 1-3 4-5 6-9 9 plus Unknown	This variable categorizes the age group of the youngest child in the family.	6
Count of Adult Children in Family	Categorical: 0 1 2 3 4 plus Unknown	This variable categorizes the total number of adult children in the family.	6

Table A.10 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Count of Dependent Children in Family	Categorical: <=2 3-4 >=5 Unknown	This variable categorizes the total number of dependent children in the family.	4
Family Violence Presence Indicator	Binary: 0,1	This variable indicates whether family violence is present.	1
Prison Involvement Indicator	Binary: 0,1	This variable indicates whether any family member is currently or has previously been involved with the prison system.	1
Corrections System Involvement Indicator	Binary: 0,1	This variable indicates whether there is involvement in the corrections system.	1
Substance Abuse Indicator	Binary: 0,1	This variable indicates whether drug and alcohol issues are present.	1
Main Benefit Caregiver as Parent's Partner Indicator	Binary: 0,1	This variable indicates whether the main benefit caregiver for the benefit spell at the time of notification is the partner of the parents.	1
Benefit Spells Count	Integer value	This variable represents the total number of benefit spells in the last 5 years.	1
Proportion of Time on Benefit in Last 5 Years	Categorical: ->80% -20-80% -Up to 20% -No time	This variable represents the proportion of time the parent or caregiver spent supported by benefits in the last 5 years. It is categorized into groups.	4
Hardship Payments Received Indicator	Binary: 0,1	This variable indicates whether the parents or main caregiver received any type of hardship payments in the last five years.	1

Table A.10 Continued from previous page.

Variable	Coding Definition	Description	Number of Features
Food Payment Requests in Last Year	Integer value	This variable indicates the total number of times parents, or the main benefit caregiver requested food payments in the last year.	1
Power Payment Requests in Last Year	Integer value	This variable indicates the total number of times parents, or the main benefit caregiver requested assistance to pay for power bill in the last year.	1
Sole Parent Household Indicator	Binary: 0,1	This variable indicates whether the child lives in a sole parent household.	1
High NZ Deprivation Index	Binary: 0,1	This set of two binary variables indicate whether the child's household has a deprivation index of 8 or higher and whether the father's household has a deprivation index of 8 or higher.	2

Table A.11: Other Predictors

Variable	Coding Definition	Description	Number of Features
Notifier's role	Categorical: -Anonymous -Court -School or Early Childhood Centre -Family or *whānau* -Health Professionals -Midwife or Plunket -Neighbours or Friends -Police -Unknown -Others	This variable is aggregated and includes the role of the notifier.	10

Appendix B

This appendix provides the distributions of observed care and protection-related events, as outlined in Table 5.1, as well as the estimated care and protection concern in the Sample Cohort 2017, used for training, and the Sample Cohort 2018, used for external validation. Table B.1 presents the distributions at different time frames of two, three, and four years. Additionally, Table B.2 outlines the distributions of outcomes within four years across different ethnic groups.

Table B.1: Distribution of observed care and protection-related events and the outcome based on these events for sample cohorts of unique children and young people across different time frames.

Time Frame	Outcome	Cohort 2017			Cohort 2018		
		All	Incidence	Incidence rate	All	Incidence	Incidence rate
2 years	Estimated care and protection concern		28,236	47%		28,227	47%
	Further notification with an intake outcome		21,897	37%		21,642	36%
	FGC or FWA recommended by a social worker	59,475	7,815	13%	59,511	7,884	13%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)		17,448	29%		17,517	29%
3 years	Estimated care and protection concern		30,558	54%		30,102	53%
	Further notification with an intake outcome		25,185	44%		24,510	43%
	FGC or FWA recommended by a social worker	57,039	9,036	16%	56,958	8,907	16%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)		19,338	34%		19,173	34%
4 years	Estimated care and protection concern		31,404	58%		30,516	57%
	Further notification with an intake outcome		26,871	50%		25,695	48%
	FGC or FWA recommended by a social worker	54,111	9,705	18%	53,997	9,393	17%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)		20,250	37%		19,689	36%

Table B.2: Distribution of observed care and protection-related events and the outcome variable based on these events within four years across Ethnic groups (Maori, Pacific, European and Others).

Ethnic Group	Outcome	Cohort 2017 (n = 54,111)		Cohort 2018 (n = 53,997)	
		Incidence	Incidence rate	Incidence	Incidence rate
Maori	Estimated care and protection concern	19,041	61%	18,348	60%
	Further notification with an intake outcome	16,677	62%	15,714	61%
	FGC or FWA recommended by a social worker	6,099	63%	5,880	63%
	Full time placement	2,736	66%	2,337	64%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)	12,462	62%	12,135	62%
Pacific	Estimated care and protection concern	3,324	11%	3,147	10%
	Further notification with an intake outcome	2,550	9%	2,391	9%
	FGC or FWA recommended by a social worker	906	9%	822	9%
	Full time placement	330	8%	261	7%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)	2,361	12%	2,196	11%
European and Others	Estimated care and protection concern	9,039	29%	9,021	30%
	Further notification with an intake outcome	8,583	30%	7,593	30%
	FGC or FWA recommended by a social worker	2,700	28%	2,688	29%
	Full time placement	1,101	26%	1,071	29%
	Substantiated findings of maltreatment (including sexual, emotional, physical and neglect)	5,427	27%	5,358	27%

Appendix C

The following appendix provides a list of the predictor variables used in the analysis discussed in Section 8.3. These variables are grouped into child predictors, caregiver predictors, family predictors, and other relevant predictors.

Table C.1: Child Predictors

Variable	Type	Coding Definition	Description
Age(Years)	Numerical	Integer number	Age at the time of notification.
Gender	Categorical	Male Female Unknown	Gender of the child or young person.
Ethnic group	Categorical	Māori Māori and Pacific Pacific European Other Unknown	Since one child can have more than one ethnicity, this variable was created based on Oranga Tamariki approach regarding prioritised ethnicity. • Māori children who identify Māori (but not Pacific) as one of their ethnicity. • Māori and Pacific children who identify both Māori and Pacific as their ethnicity. • Pacific children who identify Pacific (but not Māori) as one of their ethnicity. • New Zealand European and Other children who do not identify Māori or Pacific as one of their ethnicity.
Previous risk and safety assessment flag	Binary	1,0	This variable indicates whether the child has previously been the subject of a risk and safety assessment.
Number of previous care and protection notifications	Numerical	Integer value	This variable includes the number of previous care and protection notifications for the child.
No previous care and protection notification flag	Binary	1,0	Since the above variable is zero inflated, this binary variable was created to indicate whether the child has not been the subject of a care and protection notification in the past.
Number of days since last intake	Numerical	Integer value	Number of days since the child was the subject of a Section 15 intake where further action was required by child protection services. This is not including the current notification.

Table C.1 Continued from previous page.

Variable	Type	Coding Definition	Description
No previous intake flag	Binary	1,0	This variable indicates whether the child has not been the subject of a notification with an intake outcome in the past.
Number of previous maltreatment findings	Numerical	Integer value	This variable includes total number of previous substantiated findings of maltreatment for the child including emotional abuse, physical abuse, sexual abuse, and neglect.
No previous maltreatment finding	Binary	1,0	This variable indicates whether the child has not been the subject of a maltreatment finding in the past.
Previous custody guardianship spell flag	Binary	1,0	This variable indicates whether the child has previously had care and protection custody or guardianship to the Chief Executive of OT or another service provider. This does not include section 205 which is a temporary order and section 42 which allows a police constable, who believes its necessary to protect a child from injury or death and detain the child.
Open phase flag	Binary	1,0	This variable indicates whether the child is already in an open social work phase at the time of notification such as: • Investigation • Risk and safety assessment • Partnered response • Placement

Table C.1 Continued from previous page.

Variable	Type	Coding Definition	Description
Main public benefit inclusion flag	Binary	1,0	This variable indicate whether the child is included in a caregiver's main benefit currently or in the past. The main benefit here refers to: • Sole parents • Job seekers • Support living payment • Young parent
Level of contact with MSD and OT	Ordinal	1,2,3,4	Two binary variable were created in order to define this variable. A variable that indicated whether the child has previously had a contact with NZ public benefit system (MSD) and another one that indicated whether the child has previously been in contact with NZ child protective services (OT). For the final variable which is used in the analysis, each child is categorised into one of four groups: **Level 1**: No previous public benefit system or OT contact. **Level 2**: Previous OT contact, no previous contact with public benefit system. **Level 3**: Previous contact with public benefit system, no previous OT contact. **Level 4**: public benefit system and OT contact.

Table C.2: caregiver Predictors (Mother of the child).

Variable	Type	Coding Definition	Description
Age (Years)	Numerical	Integer value	Age of the child's mother at the time of notification.
Level of contact with child protective services	Ordinal	1,2,3,4	The level of contact with child protective services during the childhood of the caregiver (mother). **Level 1**: No involvement. **Level 2**: At least one intake as a child. **Level 3**: Finding of maltreatment. **Level 4**: Placement. **Note**: We are aware that complete history of contact with child protective services is only available for younger caregivers.
NZ deprivation index	Categorical	1 ,2, 3, 4 5, 6, 7, 8 9, 10, unknown	NZ Deprivation Index for the caregiver based on 2018 census.

Table C.3: Family Predictors.

Variable	Type	Coding Definition	Description
Number of children reported at the same time	Numerical	Integer value	The number of children involved in the notification (siblings).
Number of previous notifications for children reported	Numerical	Integer value	Total number of previous notifications for the children involved in the notification (siblings).

Table C.4: Other Predictors.

Variable	Type	Coding Definition	Description
Notifier's role	Categorical	-Anonymous -Court -Family -Health Professionals -Midwife or Plunket[1] -Neighbours or Friends -Police (FVI[2]) -Police (Other) -School or Early Childhood Centre -Unknown -Others	This variable is aggregated and includes the role of the notifier.

Bibliography

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*, 60–69.

Akalin, A. (2020). *Computational genomics with r* [Available from: https://compgenomr.github.io/book/logistic-regression-and-regularization.html]. CRC Press.

Akehurst, R. (2015). Child neglect identification: The health visitor's role. *Community Practitioner*, *88*(11), 38–43.

Allegheny County. (n.d.). https://www.alleghenycounty.us/Services/Human-Services-DHS/News-and-Events/Accomplishments-and-Innovations/Allegheny-Family-Screening-Tool

Allegheny County Analytics. (2023, September 9). *Hello baby prevention model and program supports parents of new babies*. https://www.alleghenycountyanalytics.us/2023/09/09/hello-baby-program-to-support-parents-of-new-babies-in-allegheny-county/

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*, 402–418.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Archives New Zealand. (2022, July 11). *Child welfare.* https://www.archives.govt.nz/research-guidance/research-guides/welfare/child-welfare#11-vaccination-registers

Ards, S. D., Myers Jr, S. L., Chung, C., Malkis, A., & Hagerty, B. (2003). Decomposing black-white differences in child maltreatment. *Child Maltreatment*, *8*(2), 112–121.

AsadZadehZanjani, N. (2022). *A study of administrative data representation for machine learning (Publication No. 29065054.)* [Doctoral dissertation, George Mason University]. ProQuest Dissertations and Theses Global.

Assink, M., van der Put, C. E., Meeuwsen, M. W. C. M., de Jong, N. M., Oort, F. J., Stams, G. J. J. M., & Hoeve, M. (2019). Risk factors for child sexual abuse victimization: A meta-analytic review. *Psychological Bulletin*, *145*(5), 459–489.

Austin, A. (2016). Is prior parental criminal justice involvement associated with child maltreatment? a systematic review. *Children and Youth Services Review*, *68*, 146–153.

Austin, A. E., Lesak, A. M., & Shanahan, M. E. (2020). Risk and protective factors for child maltreatment: A review. *Current Epidemiology Reports*, *7*, 334–342.

Avdibegović, E., & Brkić, M. (2020). Child neglect-causes and consequences. *Psychiatria Danubina*, *32*(Suppl. 3), 337–342.

Ayers, S., Bond, R., Webb, R., Miller, P., & Bateson, K. (2019). Perinatal mental health and risk of child maltreatment: A systematic review and meta-analysis. *Child Abuse Neglect*, *98*, 104172.

Baird, C., & Wagner, D. (2000). The relative validity of actuarial- and consensus-based risk assessment systems. *Children and Youth Services Review*, *22*(11-12), 839–871.

Baker, M., & Plessis, R. D. (2018, Jun 29). *Family welfare - welfare, work and families, 1918–1945*. https://teara.govt.nz/en/family-welfare/page-3

Baker, M. G., Kvalsvig, A., Verrall, A. J., & Wellington, N. (2020). New zealand's covid-19 elimination strategy. *Med J Aust*, *213*(5), 198–200.

Barmomanesh, S., & Miranda-Soberanis, V. (2023). Potential biased outcomes on child welfare and racial minorities in new zealand using predictive models: An initial review on mitigation approaches. *arXiv preprint arXiv:2308.00243 [stat.AP]*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732. http://dx.doi.org/10.15779/Z38BG31

Bartlett, J. D., Kotake, C., Fauth, R., & Easterbrooks, M. A. (2017). Intergenerational transmission of child abuse and neglect: Do maltreatment type, perpetrator, and substantiation status matter? *Child abuse neglect*, *63*, 84–94.

Belsky, J. (1980). Child maltreatment: An ecological integration. *American Psychologist*, *35*(4), 320.

Benesh, A. S. (2017). *Predicting child welfare future placements for foster youth: An application of statistical learning to child welfare (Publication No.10258386.)* [Doctoral dissertation, The Florida State University]. ProQuest Dissertations and Theses Global.

Bennet, P. (2012). *The white paper for vulnerable children, volume 1* (tech. rep.). Ministry of Social Development. https://orangatamariki.govt.nz/assets/Uploads/Support-for-families/childrens-teams/white-paper-for-vulnerable-children-volume-1.pdf

Bennet, P. (2014). *The white paper for vulnerable children, volume 2* (tech. rep.). Ministry of Social Development. https://orangatamariki.govt.nz/assets/Uploads/Support-for-families/childrens-teams/whitepaper-volume-2.pdf

Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13.

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

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Conference on fairness, accountability and transparency*, *81*, 149–159.

Blank, A., Cram, F., Dare, T., de Haan, I., Smith, B., & Vaithianathan, R. (2015). Ethical issues for māori in predictive risk modelling to identify new-born children who are at high risk of future maltreatment.

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with r*. Chapman; Hall/CRC.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–152.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531.

Brown, P. T., Wilson, D., West, K., Escott, K.-R., Basabas, K., Ritchie, B., Lucas, D., Taia, I., Kusabs, N., & Keegan, T. T. (2023). M\= aori algorithmic sovereignty: Idea, principles, and use. *arXiv preprint arXiv:2311.15473*.

Brown, S. M., Doom, J. R., Lechuga-Peña, S., Watamura, S. E., & Koppels, T. (2020). Stress and parenting during the global covid-19 pandemic. *Child abuse neglect*, *110*, 104699.

Brown, S. M., Orsi, R., Chen, P. C. B., Everson, C. L., & Fluke, J. (2022). The impact of the covid-19 pandemic on child protection system referrals and responses in colorado, usa. *Child maltreatment*, *27*(1), 3–11.

Bullinger, L. R., Raissian, K. M., Feely, M., & Schneider, W. J. (2021). The neglected ones: Time at home during covid-19 and child maltreatment. *Children and youth services review*, *131*, 106287.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Byrne, G. (2018). Prevalence and psychological sequelae of sexual abuse among individuals with an intellectual disability: A review of the recent literature. *Journal of Intellectual Disabilities*, *22*(3), 294–310.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *IEEE International Conference on Data Mining Workshops*, 13–18.

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. *2013 IEEE 13th international conference on data mining*, 71–80.

Calders, T., & Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. Springer.

Centre for Social Data Analytics. (2022, September). *Insights from the centre for social data analytics – september 2020*. https://csda.aut.ac.nz/__data/assets/pdf_file/0010/432928/eNews-September-2020.pdf

Centre for Social Data Analytics. (n.d.). *Douglas county decision aid*. https://csda.aut.ac.nz/research/our-projects/all-projects/Douglas-County-Decision-Aid

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & et al. (2024). *Xgboost: Extreme gradient boosting* [R package version 1.6.0.1]. https://CRAN.R-project.org/package=xgboost

Cheng, H.-F., Stapleton, L., Wang, R., Bullock, P., Chouldechova, A., Wu, Z. S. S., & Zhu, H. (2021). Soliciting stakeholders' fairness notions in child maltreatment predictive systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Conference on Fairness, Accountability and Transparency*, *81*, 134–148.

Chouldechova, A., & G'Sell, M. (2017). Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*.

Chung, G., Lanier, P., & Wong, P. Y. J. (2022). Mediating effects of parental stress on harsh parenting and parent-child relationship during coronavirus (covid-19) pandemic in singapore. *Journal of Family Violence*, *37*(5), 801–812.

Chung, G. S. K. (2021). *Risk factors of subsequent allegations of child maltreatment (Publication No. 28648614.)* [Doctoral dissertation, The University of North Carolina]. ProQuest Dissertations and Theses Global.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2019). Leveraging labeled and unlabeled data for consistent fair binary classification. *arXiv preprint arXiv:1906.05082*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Conrad-Hiebner, A., & Byram, E. (2020). The temporal impact of economic insecurity on child maltreatment: A systematic review. *Trauma, Violence, Abuse*, *21*(1), 157–178.

Coohey, C., Johnson, K., Renner, L. M., & Easton, S. D. (2013). Actuarial risk assessment in child protective services: Construction methodology and performance criteria. *Children and Youth Services Review*, *35*(1), 151–161.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2009). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cox, C. E., Kotch, J. B., & Everson, M. D. (2003). A longitudinal study of modifying influences in the relationship between domestic violence and child maltreatment. *Journal of Family Violence*, *18*, 5–17.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, *79*, 291–298.

Daley, D., Bachmann, M., Bachmann, B. A., Pedigo, C., Bui, M.-T., & Coffman, J. (2016). Risk terrain modeling predicts child maltreatment. *Child Abuse Neglect*, *62*, 29–38.

D'Andrade, A., Austin, M. J., & Benton, A. (2008). Risk and safety assessment in child welfare: Instrument comparisons. *Journal of Evidence-Based Social Work*, *5*(1-2), 31–56.

Dare, T. (2013). *Predictive risk modelling and child maltreatment: An ethical review*. University of Auckland. https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/predictive-modelling/00-predicitve-risk-modelling-and-child-maltreatment-an-ethical-review.pdf

Dare, T., & Gambrill, E. (2017). *Ethical analysis: Predictive risk models at call screening for allegheny county*. Allegheny County. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf

de Haan, I., & Connolly, M. (2014). Another pandora's box? some pros and cons of predictive risk modeling. *Children and Youth Services Review*, *47*, 86–91.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845.

DePanfilis, D., & Zuravin, S. J. (2001). Assessing risk to determine the need for services. *Children and Youth Services Review*, *23*(1), 3–20.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, *7*(4), 1–36.

Ding, C., & He, X. (2004a). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*, 29.

Ding, C., & He, X. (2004b). Principal component analysis and effective k-means clustering. *Proceedings of the 2004 SIAM International Conference on Data Mining*, 497–501.

Drake, B., & Jonson-Reid, M. (2018). *Administrative data and predictive risk modeling in public child welfare: Ethical issues relating to california*. Brown School of Social Work, Washington University. http://www.caichildlaw.org/Misc/Ethical_Review_of_Predictive_Risk_Modeling.pdf

Drake, B., Jonson-Reid, M., Ocampo, M. G., Morrison, M., & Dvalishvili, D. (2020). A practical framework for considering the use of predictive risk modeling in child welfare. *The ANNALS of the American Academy of Political and Social Science*, *692*(1), 162–181.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Eckerd. (2014, March 10). *Rapid safety feedback : Blue ribbon commission on child protection*. https://file.lacounty.gov/SDSInter/bos/supdocs/83688.pdf

Edelman, B. G., & Luca, M. (2014). *Digital discrimination: The case of airbnb.com* [Preprint number 14-054].

Elgin, D. J. (2018). Utilizing predictive modeling to enhance policy and practice through improved identification of at-risk clients: Predicting permanency for foster children. *Children and Youth Services Review*, *91*, 156–167.

English, D. J., Marshall, D. B., Brummel, S., & Orme, M. (1999). Characteristics of repeated referrals to child protective services in washington state. *Child Maltreatment*, *4*(4), 297–307.

Feldman, M. (2015). *Computational fairness: Preventing machine-learned discrimination* [Doctoral dissertation].

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Finkelhor, D. (2008). *Childhood victimization: Violence, crime, and abuse in the lives of young people*. oxford university Press.

Fish, B., Kun, J., & Lelkes, A. D. (2015). Fair boosting: A case study. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Fluke, J. D., Yuan, Y.-Y. T., & Edwards, M. (1999). Recurrence of maltreatment: An application of the national child abuse and neglect data system (ncands). *Child Abuse Neglect*, *23*(7), 633–650.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187–232.

Freisthler, B., Bruce, E., & Needell, B. (2007). Understanding the geospatial relationship of neighborhood characteristics and rates of maltreatment for black, hispanic, and white children. *Social Work*, *52*(1), 7–16.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B., & Qian, J. (2021). *Glmnet: Lasso and elastic-net regularized generalized linear models* [R package version 4.1-1]. https://CRAN.R-project.org/package=glmnet

G. D. P. Regulation. (2018). *The general data protection regulation (gdpr)*. Intersoft Consulting. https://gdpr-info.eu/

Gavighan, C., Knott, A., Maclaurin, J., Zerilli, J., & Liddicoat, J. (2019). *Government use of artificial intelligence in new zealand*. The New Zealand Law Foundation.

Gillingham, P. (2016). Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the 'black box' of machine learning. *British Journal of Social Work*, *46*(4), 1044–1058.

Glaberson, S. K. (2019). Coding over the cracks: Predictive analytics and child protection. *Fordham Urban Law Journal*, *46*, 307–345.

Goldhaber-Fiebert, J. D., & Prince, L. (2019). *Allegheny county predictive risk modeling tool implementation: Process evaluation*. Allegheny County. https://www.alleghenycountyanalytics.us/

wp - content / uploads / 2019 / 05 / Impact - Evaluation - from - 16 - ACDHS - 26 _ PredictiveRisk _ Package_050119_FINAL-6.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Guardian, T. (2020, 30 July 2023). New zealand braces for spike in child abuse reports as covid-19 lockdown eases. https://www.theguardian.com/world/2020/may/18/new-zealand-braces-for-spike-in-child-abuse-reports-as-covid-19-lockdown-eases

Guo, L. L., Pfohl, S. R., Fries, J., Posada, J., Fleming, S. L., Aftandilian, C., Shah, N., & Sung, L. (2021). Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied clinical informatics*, *12*(04), 808–815.

Gustafsson, M., Hornquist, M., & Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network: Lasso-constrained inference and biological validation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *2*(3), 254–261.

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hibbard, R. A., Desch, L. W., Abuse, C. o. C., Neglect, & Disabilities, C. o. C. W. (2007). Maltreatment of children with disabilities. *Pediatrics*, *119*(5), 1018–1025.

Hindley, N., Ramchandani, P. G., & Jones, D. P. H. (2006). Risk factors for recurrence of maltreatment: A systematic review. *Archives of Disease in Childhood*, *91*(9), 744–752.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables.

Horikawa, H., Suguimoto, S. P., Musumari, P. M., Techasrivichien, T., Ono-Kihara, M., & Kihara, M. (2016). Development of a prediction model for child maltreatment recurrence in japan: A historical cohort study using data from a child guidance center. *Child Abuse Neglect*, *59*, 55–65.

Hornby Zeller Associates. (2018). *Allegheny county predictive risk modeling tool implementation: Process evaluation*. Allegheny County. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Process-Evaluation-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-4.pdf

Hu, L., & Chen, Y. (2020). Fair classification and social welfare. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545.

Huckle, T., & Romeo, J. S. (2023). Estimating child maltreatment cases that could be alcohol-attributable in new zealand. *Addiction*, *118*(4), 669–677. https://doi.org/10.1111/add.16111

Hudson, M., Milne, M., Reynolds, P., Russell, K., & Smith, B. (2010). Te ara tika: Guidelines for māori research ethics: A framework for researchers and ethics committee members. *Auckland: Health Research Council of New Zealand*.

Hurren, E., Thompson, C., Jenkins, B., Chrzanowski, A., Allard, T., & Stewart, A. (2018). Who are the perpetrators of child maltreatment. *Criminology Research Advisory Council*, 2020–05.

James, A., McLeod, J., Hendy, S., Marks, K., Rusu, D., Nik, S., & Plank, M. J. (2019). Using family network data in child protection services. *PLoS ONE*, *14*(10), e0224554.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jefferies, S., French, N., Gilkison, C., Graham, G., Hope, V., Marshall, J., McElnay, C., McNeill, A., Muellner, P., & Paine, S. (2020). Covid-19 in new zealand and the impact of the national response: A descriptive epidemiological study. *The Lancet Public Health*, *5*(11), e612–e623.

Johndrow, J. E., & Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, *13*(1), 189–220.

Johnson, K. D., Foster, D. P., & Stine, R. A. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*.

Johnson, S. G. (2008). *The nlopt nonlinear-optimization package*. https://github.com/stevengj/nlopt

Johnston, P. (2021). Assessing risk of re-offending: Recalibration of the department of corrections' core risk assessment measure. *Practice: The New Zealand Corrections Journal*, *8*(1), 13–18.

Joliffe, I. T., & Morgan, B. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, *1*(1), 69–95.

Jolley, J. M. (2012). *Applying neural network models to predict recurrent maltreatment in child welfare cases with static and dynamic risk factors(Publication No.3542505.)* [Doctoral dissertation, The Florida State University]. ProQuest Dissertations and Theses Global.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6.

Kamiran, F., & Calders, T. (2010). Classification with no discrimination by preferential sampling. *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, 1–6.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *2010 IEEE International Conference on Data Mining*, 869–874.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50.

Katz, C., Priolo Filho, S. R., Korbin, J., Bérubé, A., Fouche, A., Haffejee, S., Kaawa-Mafigiri, D., Maguire-Jack, K., Muñoz, P., & Spilsbury, J. (2021). Child maltreatment in the time of the covid-19 pandemic: A proposed global framework on research, policy and practice. *Child Abuse Neglect*, *116*, 104824.

Katz, I., Katz, C., Andresen, S., Bérubé, A., Collin-Vezina, D., Fallon, B., Fouché, A., Haffejee, S., Masrawa, N., & Muñoz, P. (2021). Child maltreatment reports and child protection service responses during covid-19: Knowledge exchange among australia, brazil, canada, colombia, germany, israel, and south africa. *Child abuse neglect*, *116*, 105078.

Kazdin, A. E., Kraemer, H. C., Kessler, R. C., Kupfer, D. J., & Offord, D. R. (1997). Contributions of risk-factor research to developmental psychopathology. *Clinical Psychology Review*, *17*(4), 375–406.

Kearney, A. D., Wilson, E. S., Hollinshead, D. M., Poletika, M., Kestian, H. H., Stigdon, T. J., Miller, E. A., & Fluke, J. D. (2023). Child welfare triage: Use of screening threshold analysis to evaluate intake decision-making. *Children and Youth Services Review*, *144*, 106710.

Keddell, E. (2015). The ethics of predictive risk modelling in the aotearoa/new zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy*, *35*(1), 69–88.

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences*, *8*(10), 281.

Keddell, E., & Davie, G. (2018). Inequalities and child protection system contact in aotearoa new zealand: Developing a conceptual framework and research agenda. *Social Sciences*, *7*(6), 89.

Keddell, E., & Hyslop, I. (2019). Ethnic inequalities in child welfare: The role of practitioner risk perceptions. *Child Family Social Work*, *24*(4), 409–420.

Keddell, E. (2018). The vulnerable child in neoliberal contexts: The construction of children in the aotearoa new zealand child protection reforms. *Childhood*, *25*(1), 93–108.

Killick, R., & Eckley, I. A. (2014). Changepoint: An r package for changepoint analysis. *Journal of statistical software*, *58*, 1–19.

Killick, R., Haynes, K., Eckley, I., Fearnhead, P., & Lee, J. (2016). Package 'changepoint'. *R package version 0.4.-2011.-http://cran. rproject. org/web/packages/changepoint/index. html*, *109*.

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

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk. *Archives of General Psychiatry*, *54*(4), 337–343.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., & Team, R. C. (2020). Package 'caret'. *The R Journal*, *12*(3), 7–22.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kukutai, T., & Taylor, J. (2016). *Indigenous data sovereignty: Toward an agenda*. ANU press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lawson, M., Piel, M. H., & Simon, M. (2020). Child maltreatment during the covid-19 pandemic: Consequences of parental job loss on psychological and physical abuse towards children. *Child abuse neglect*, *110*, 104709.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Society*, *5*(1).

Lee, S. J., Grogan-Kaylor, A., & Berger, L. M. (2014). Parental spanking of 1-year-old children and subsequent child protective services involvement. *Child Abuse Neglect*, *38*(5), 875–883.

Lee, S. J., Ward, K. P., Lee, J. Y., & Rodriguez, C. M. (2021). Parental social isolation and child maltreatment risk during the covid-19 pandemic. *Journal of family violence*, 1–12.

Li, L. (2006). *Data complexity in machine learning and novel classification algorithms* [Doctoral dissertation, California Institute of Technology]. https://doi.org/10.7907/EW2G-9986

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

Liston-Lloyd, M., & Sun, H. (2019). *Children's teams evaluation*. Oranga Tamariki Evidence Centre. https://www.orangatamariki.govt.nz/assets/Uploads/About-us/Research/Latest-research/Childrens-Teams-evaluation/Childrens-Teams-Evaluation-Report.pdf

Littell, J. H., & Schuerman, J. R. (2002). What works best for whom? a closer look at intensive family preservation services. *Children and Youth Services Review*, *24*(9-10), 673–699.

Lopes, A. I., Leal, J., & Sani, A. I. (2021). Parental mental health problems and the risk of child maltreatment: The potential role of psychotherapy. *Societies*, *11*(3), 108.

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

Luong, B. T., Ruggieri, S., & Turini, F. (2011). K-nn as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510.

Mancuhan, K., & Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, *22*(2), 211–238.

Mancuhan, K., & Clifton, C. (2012). Discriminatory decision policy aware classification. *2012 IEEE 12th International Conference on Data Mining Workshops*, 386–393.

Marcal, K. E. (2018). The impact of housing instability on child maltreatment: A causal investigation. *Journal of Family Social Work*, *21*(4-5), 331–347.

Masten, A. S., & Wright, M. O. (1998). Cumulative risk and protection models of child maltreatment. *Journal of Aggression, Maltreatment Trauma*, *2*(1), 7–30.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

McNellan, C. R., Gibbs, D. J., Knobel, A. S., & Putnam-Hornstein, E. (2022). The evidence base for risk assessment tools used in us child protection investigations: A systematic scoping review. *Child Abuse Neglect*, *134*, 105887.

McTier, A., & Soraghan, J. (2022). The utility of administrative data in understanding the covid-19 pandemic's impact on child maltreatment: Learning from the scotland experience. *Child maltreatment*, 10775595221108661.

Meinck, F., Cluver, L. D., Boyes, M. E., & Mhlongo, E. L. (2015). Risk and protective factors for physical and sexual abuse of children and adolescents in africa: A review and implications for practice. *Trauma, Violence, Abuse*, *16*(1), 81–107.

Melz, H., Fromknecht, A. E., Masters, L. D., Richards, T., & Sun, J. (2023). Incorporating multiple data sources to assess changes in organizational capacity in child welfare systems. *Evaluation and Program Planning*, *97*, 102231.

Ministry of Education. (2023). *School terms and holidays archive*. https://www.education.govt.nz/school/school-terms-and-holiday-dates/school-terms-and-holidays-archive/#Cal2022

Ministry of Social Development. (2014). *The feasibility of using predictive risk modelling to identify new-born children who are high priority for preventive services—companion technical report*. https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/predictive-modelling/00-feasibility-study-report-technical-companion.pdf

Ministry of Social Development. (n.d.). *Historical timeline*. https://www.msd.govt.nz/about-msd-and-our-work/about-msd/history/index.html

Moody, G., Cannings-John, R., Hood, K., Kemp, A., & Robling, M. (2018). Establishing the international prevalence of self-reported child maltreatment: A systematic review by maltreatment type and gender. *BMC Public Health*, *18*(1), 1–15.

Morris, M. C., Marco, M., Maguire-Jack, K., Kouros, C. D., Bailey, B., Ruiz, E., & Im, W. (2019). Connecting child maltreatment risk with crime and neighborhood disadvantage across time and place: A bayesian spatiotemporal analysis. *Child Maltreatment*, *24*(2), 181–192.

Mortimer, S. M., Kuhn, M., Carlson, M., Mayer, Z., Fisher, A. J., Engelhardt, A., & Kapourani, A. C. (2018). *Modelmetrics: Rapid calculation of model metrics* [R package version 1.2.2.2]. https://CRAN.R-project.org/package=ModelMetrics

Mulder, T. M., Kuiper, K. C., van der Put, C. E., Stams, G.-J. J. M., & Assink, M. (2018). Risk factors for child neglect: A meta-analytic review. *Child Abuse Neglect*, *77*, 198–210.

Muthukrishnan, R., & Rohini, R. (2016). Lasso: A feature selection technique in predictive modeling for machine learning. *2016 IEEE international conference on advances in computer applications (ICACA)*, 18–20.

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*.

New Zealand Family Violence Clearinghouse. (2017). Child protection reform bill passes into law.

New Zealand Government. (2023, December 12). *State care timeline*. https://www.abuseinquiryresponse.govt.nz/rauemi-resources/state-care-timeline/

New Zealand police & Oranga Tamariki. (2021). *Child protection protocol:joint operating procedures*. https://practice.orangatamariki.govt.nz/assets/resources/Documents/child-protection-protocol-joint-operating-procedures-dec2021.pdf

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd). Springer.

Nunez, J. J., Fluke, J. D., Shusterman, G. R., & Fettig, N. B. (2023). Understanding the effects of covid-19 on child maltreatment reporting among rural versus urban communities in the united states. *International journal on child maltreatment: research, policy and practice*, *6*(2), 149–164.

Oliver, W. J., Kuhns, L. R., & Pomeranz, E. S. (2006). Family structure and child abuse. *Clinical Pediatrics*, *45*(2), 111–118.

Oranga Tamariki. (2020, September 1). *Full assessment phase*. https://practice.orangatamariki.govt.nz/our-work/assessment-and-planning/assessments/intake-and-early-assessment/core-assessment-phase/

Oranga Tamariki. (2021, July 16). *Overview of the intake decision response tool*. https://practice.orangatamariki.govt.nz/core-practice/practice-tools/intake-decision-response-tool/overview-of-the-intake-decision-response-tool/

Oranga Tamariki. (2022, March 8). *New ways of working*. https://www.orangatamariki.govt.nz/about-us/our-work/new-ways-of-working/

Oranga Tamariki. (2023a). *Factors associated with disparities experienced by māori children in the care and protection system. Wellington, New Zealand: Oranga Tamariki—Ministry for Children*. https://www.orangatamariki.govt.nz/assets/Uploads/About-us/Research/Latest-research/Factors-associated-with-disparities-experienced-by-tamariki-Maori-2022/Factors-associated-with-disparities-experienced-by-Maori-children-in-the-Care-and-Protection-System-2023.pdf

Oranga Tamariki. (2023b, September 1). *Children's teams*. https://www.orangatamariki.govt.nz/support-for-families/childrens-teams/

Oranga Tamariki. (2023c, July 30). *Considerations when responding to information received*. https://practice.orangatamariki.govt.nz/core-practice/practice-tools/intake-decision-response-tool/considerations-when-responding-to-information-received/

Oranga Tamariki. (2023d, July 19). *Contact us*. https://www.orangatamariki.govt.nz/about-us/contact-us/

Oranga Tamariki. (2023e, July 13). *Core assessment phase*. https://practice.orangatamariki.govt.nz/our-work/assessment-and-planning/assessments/intake-and-early-assessment/core-assessment-phase/

Oranga Tamariki. (2023f, July 30). *Initial assessment phase*. https://practice.orangatamariki.govt.nz/our-work/assessment-and-planning/assessments/intake-and-early-assessment/initial-assessment-phase/

Oranga Tamariki. (2023g, September 19). *Intake and early assessment*. https://practice.orangatamariki.govt.nz/our-work/assessment-and-planning/assessments/intake-and-early-assessment/

Oranga Tamariki. (2023h, July 30). *Report of concern response pathway*. https://practice.orangatamariki.govt.nz/core-practice/practice-tools/intake-decision-response-tool/report-of-concern-response-pathway/

Oranga Tamariki. (2023i, July 30). *Report of concern response timeframe*. https://practice.orangatamariki.govt.nz/core-practice/practice-tools/intake-decision-response-tool/report-of-concern-response-timeframe/

Palusci, V. J. (2011). Risk factors and services for child maltreatment among infants and young children. *Children and Youth Services Review*, *33*(8), 1374–1382.

Parker, E. M., Williams, J. R., Pecora, P. J., & Despard, D. (2022). Examining the effects of the eckerd rapid safety feedback process on the occurrence of repeat maltreatment among children involved in the child welfare system. *Child Abuse Neglect*, *133*, 105856.

Parkinson, S. (2017). *Child neglect: Key concepts and risk factors* (tech. rep. No. 1) (A report to the NSW Department of Family and Community Services Office of the Senior Practitioner). Australian Centre for Child Protection.

Parycek, P., Schmid, V., & Novak, A.-S. (2023). Artificial intelligence (ai) and automation in administrative procedures: Potentials, limitations, and framework conditions. *Journal of Knowledge Economy*, *33*(1), 1–20.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. *Proceedings of the 2009 SIAM international conference on data mining*, 581–592.

Pelton, L. H. (2015). The continuing role of material factors in child maltreatment and placement. *Child Abuse Neglect*, *41*, 30–39.

Petrowski, N., Cappa, C., Pereira, A., Mason, H., & Daban, R. A. (2021). Violence against children during covid-19: Assessing and understanding change in use of helplines. *Child Abuse & Neglect*, *116*, 104757.

Posit team. (2022). *Rstudio: Integrated development environment for r*. Posit Software, PBC. Boston, MA. http://www.posit.co/

Prabhu, P., & Anbazhagan, N. (2011). Improving the performance of k-means clustering for high dimensional data set. *International Journal on Computer Science and Engineering*, *3*(6), 2317–2322.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Purdy, J., & Glass, B. (2023). The pursuit of algorithmic fairness: On "correcting" algorithmic unfairness in a child welfare reunification success classifier. *Children and Youth Services Review*, *145*, 106777.

Putnam-Hornstein, E., Vaithianathan, R., McCroskey, J., & Webster, D. (2022). *Los angeles county risk stratification model: Methodology implementation report*. Children's Data Network. https://dcfs.lacounty.gov/wp-content/uploads/2022/08/Risk-Stratification-Methodology-Report_8.29.22.pdf

Putnam-Hornstein, E., Vaithianathan, R., Prindle, J., Cuccaro-Alamin, S., Nghiem, H., & Gupta, T. (2018, September). *Predictive risk modeling: Findings from california's proof-of-concept* [PowerPoint slides]. https://www.datanetwork.org/wp-content/uploads/PRM_CWDAB_2018.pdf

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Radovanović, S., & Ivić, M. (2021). Enabling equal opportunity in logistic regression algorithm. *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*.

Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2020). Enforcing fairness in logistic regression algorithm. *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–7.

Rapp, A., Fall, G., Radomsky, A. C., & Santarossa, S. (2021). Child maltreatment during the covid-19 pandemic: A systematic rapid review. *Pediatric Clinics*, *68*(5), 991–1009.

Rea, D., & Erasmus, R. (2017). *Report of the enhancing intake decision-making project*. Ministry of Social Development. https://orangatamariki.govt.nz/assets/Uploads/About-us/Research/Latest-research/Helping-social-workers/Enhancing-Intake-Decision-Making-executive-summary.pdf

Rebbe, R., Lyons, V. H., Webster, D., & Putnam-Hornstein, E. (2021). Domestic violence alleged in california child maltreatment reports during the covid-19 pandemic. *Journal of family violence*, 1–8.

Rittner, B. (2002). The use of risk assessment instruments in child protective services case planning and closures. *Children and Youth Services Review*, *24*(3), 189–207.

Robson, S. (2020, 30 July 2023). New zealand braces for spike in child abuse reports as covid-19 lockdown eases. https://www.rnz.co.nz/news/national/417518/reports-of-child-abuse-dropped-during-covid-19-lockdown

Rodriguez, M. Y., DePanfilis, D., & Lanier, P. (2019). Bridging the gap: Social work insights for ethical algorithmic decision-making in human services. *IBM Journal of Research and Development*, *63*(4/5), 8: 1–8: 8.

Rouland, B., Vaithianathan, R., Wilson, D., & Putnam-Hornstein, E. (2019). Ethnic disparities in childhood prevalence of maltreatment: Evidence from a new zealand birth cohort. *American Journal of Public Health*, *109*(9), 1255–1257.

Rudin, C., & Ustun, B. (2019). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *49*(5), 365–387.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(2), 9.

Schölkopf, B., Burges, C. J., & Smola, A. J. (1999). *Advances in kernel methods: Support vector learning*. MIT press.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Schönlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29.

Schultz, B. B. (1985). Levene's test for relative variation. *Systematic Zoology*, *34*(4), 449–456.

Schwartz, I. M., York, P., Nowakowski-Sims, E., & Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The broward county experience. *Children and Youth Services Review*, *81*, 309–320.

Sedlak, A. J. (2014). Risk factors for the occurrence of child abuse and neglect. In *Violence and sexual abuse at home* (pp. 149–187). Routledge.

Sen, A. (2021, November 9). *Ensemble modeling for neural networks using large datasets – simplified!* https://www.analyticsvidhya.com/blog/2021/10/ensemble-modeling-for-neural-networks-using-large-datasets-simplified/

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Stats NZ. (2018a). *Data confidentiality principles and methods report*. https://www.data.govt.nz/assets/Uploads/data-confidentiality-principles-methodology-report-oct-2018.pdf

Stats NZ. (2018b). Māori population estimates: At 30 june 2018 [Accessed: 2024-07-11]. https://www.stats.govt.nz/information-releases/maori-population-estimates-at-30-june-2018/

Stats NZ. (2020). Ngā Tikanga Paihere: A Framework Guiding Ethical and Culturally Appropriate Data Use [Retrieved from www.data.govt.nz].

Stats NZ. (2022a, August 23). *How we keep integrated data safe*. https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/

Stats NZ. (2022b, August 23). *How we keep integrated data safe*. https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/#five

Stats NZ. (2022c, August 23). *Integrated data infrastructure*. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Stats NZ. (2022d, September 28). *Introduction to the new zealand census*. https://www.stats.govt.nz/reports/introduction-to-the-new-zealand-census

Stats NZ. (2023, March 22). *Your information in the idi*. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/your-information-in-the-idi/

Stats NZ. (n.d.). *Idi ministry for children data (cap/cyf/ot)*. https://datainfoplus.stats.govt.nz/item/nz.govt.stats/3082c4e3-5c08-4137-b43b-68c23f7d571f

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an abcd for validation. *European Heart Journal*, *35*(29), 1925–1931.

Swanston, H. Y., Parkinson, P. N., Oates, R. K., O'Toole, B. I., Plunkett, A. M., & Shrimpton, S. (2002). Further abuse of sexually abused children. *Child Abuse Neglect*, *26*(2), 115–127.

Taib, M., & Messier, G. G. (2024). Efficient observation time window segmentation for administrative data machine learning. *arXiv preprint arXiv:2401.16537*.

Tamariki, O. (2020). Oranga tamariki statistics-covid-19 response snapshot. https://www.orangatamariki.govt.nz/assets/Uploads/About-us/How-we-work/COVID-19/COVID-19-Oranga-Tamariki-Statistics-June-2020.pdf

Te Mana Raraunga. (2018). Principles of māori data sovereignty. https://cdn.auckland.ac.nz/assets/psych/about/our-research/documents/TMR%2BM%C4%81ori%2BData%2BSovereignty%2BPrinciples%2BOct%2B2018.pdf

Thornberry, T. P., Matsuda, M., Greenman, S. J., Augustyn, M. B., Henry, K. L., Smith, C. A., & Ireland, T. O. (2014). Adolescent risk factors for child maltreatment. *Child Abuse Neglect*, *38*(4), 706–722.

Thurston, H., Freisthler, B., Bell, J., Tancredi, D., Romano, P. S., Miyamoto, S., & Joseph, J. G. (2017). Environmental and individual attributes associated with child maltreatment resulting in hospitalization or death. *Child Abuse Neglect*, *67*, 119–136.

Thurston, H., & Miyamoto, S. (2018). The use of model-based recursive partitioning as an analytic tool in child welfare. *Child Abuse Neglect*, *79*, 293–301.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Tryfos, P. (1998). *Methods for business analysis and forecasting: Text and cases*. John Wiley & Sons Inc.

Vaithianathan, R. (2012). *Can administrative data be used to identify children at risk of adverse outcomes?* Business School, Department of Economics, University of Auckland. https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/vulnerable-children/auckland-university-can-administrative-data-be-used-to-identify-children-at-risk-of-adverse-outcome.pdf

Vaithianathan, R., Benavides-Prado, D., Dalton, E., Chouldechova, A., & Putnam-Hornstein, E. (2021). Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine*, *42*(1), 53–60.

Vaithianathan, R., Benavides-Prado, D., & Putnam-Hornstein, E. (2020). *Implementing the hello baby prevention program in allegheny county: Methodology report*. Auckland: Centre for Social Data Analytics. https://www.alleghenycountyanalytics.us/wp-content/uploads/2020/12/Hello-Baby-Methodology-v6.pdf

Vaithianathan, R., Dinh, H., Kalisher, A., Kithulgoda, C., Kulick, E., Mayur, M., Ning, A., & Prado, D. B. (2019). *Implementing a child welfare decision aide in douglas county: Methodology report*. Auckland: Centre for Social Data Analytics. https://csda.aut.ac.nz/__data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf

Vaithianathan, R., Kulick, E., Putnam-Hornstein, E., & Benavides-Prado, D. (2019). *Allegheny family screening tool: Methodology, version 2*. Auckland: Centre for Social Data Analytics. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, *45*(3), 354–359.

Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., & Maloney, T. (2017). *Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation*. Auckland: Centre for Social Data Analytics. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V1-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL.pdf

Vaithianathan, R., Rouland, B., & Putnam-Hornstein, E. (2018). Injury and mortality among children identified as at high risk of maltreatment. *Pediatrics*, *141*(2).

Van der Put, C. E., Assink, M., & van Solinge, N. F. B. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. *Child Abuse Neglect*, *73*, 71–88.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Veale, M. (2019). *Draft ethical report concerning predictive modelling in the allegheny babies and families project*. Allegheny County. https://www.alleghenycountyanalytics.us/wp%20content/uploads/2020/09/DraftEthicsReport_MVeale.pdf

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7.

Vora, P., Oza, B., et al. (2013). A survey on k-mean clustering and particle swarm optimization. *International Journal of Science and Modern Engineering*, *1*(3), 24–26.

Walker, H. E., & Wamser-Nanney, R. (2022). Revictimization risk factors following childhood maltreatment: A literature review. *Trauma, Violence, Abuse*, 15248380221093692.

Walsh, M. C., Joyce, S., Maloney, T., & Vaithianathan, R. (2020). Exploring the protective factors of children and families identified at highest risk of adverse childhood experiences by a predictive risk model: An analysis of the growing up in new zealand cohort. *Children and Youth Services Review*, *108*, 104556.

Wandalowski, S., & Vaithianathan, R. (2023). *Northampton county: Pursuing better child welfare outcomes with a decision aid tool to support caseworkers*. Springer.

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, *34*(1-2), 28–35.

White, O. G., Hindley, N., & Jones, D. P. H. (2015). Risk factors for child maltreatment recurrence: An updated systematic review. *Medicine, Science and the Law*, *55*(4), 259–277.

White, R., Benedict, M. I., Wulff, L., & Kelley, M. (1987). Physical disabilities as risk factors for child maltreatment: A selected review. *American Journal of Orthopsychiatry*, *57*(1), 93–101.

Widom, C. S., Czaja, S. J., & Paris, J. (2009). A prospective investigation of borderline personality disorder in abused and neglected children followed up into adulthood. *Journal of Personality Disorders*, *23*(5), 433–446.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution* (pp. 196–202). Springer.

Wilson, M. L., Tumen, S., Ota, R., & Simmers, A. G. (2015). Predictive modeling: Potential application in prevention services. *American Journal of Preventive Medicine*, *48*(5), 509–519.

Wu, Y., & Wu, X. (2016). Using loglinear model for discrimination discovery and prevention. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 110–119.

Yang, M.-Y. (2015). The effect of material hardship on child protective service involvement. *Child Abuse Neglect*, *41*, 113–125.

Younas, F., & Gutman, L. M. (2022). Parental risk and protective factors in child maltreatment: A systematic review of the evidence. *Trauma, Violence, Abuse*, 15248380221134634.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, *20*(75), 1–42.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating pca and k-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv:1511.00148 [cs.CY]*.

Zliobaite, I. (2017). Fairness-aware machine learning: A perspective. *arXiv:1708.00754 [cs.AI]*.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.