IMPERIAL COLLEGE LONDON
SCHOOL OF PUBLIC HEALTH

Modelling Childhood Risk Factors and Time-Based Patterns for Respiratory Infections with Deep Learning and Life Course Data

Helen Coupland

Supervised by: Samir Bhatt, Juliette Unwin, Seth Flaxman & Swapnil Mishra

PhD submitted on March 20th, 2024.

Statement of Originality

I declare that all the work contained within this thesis is my own except where otherwise specified.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Correspondence

Helen Louise Coupland, MRes
MRC Centre for Global Infectious Disease Analysis, Imperial College, London, UK
h.coupland@imperial.ac.uk

Financial Support

This work was funded by a Wellcome Trust PhD Studentship.

Supervisors

Samir Bhatt
MRC Centre for Global Infectious Disease Analysis, Imperial College, London, UK
Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark
samir.bhatt@sund.ku.dk

H. Juliette T. Unwin
School of Mathematics, University of Bristol, Bristol, UK
MRC Centre for Global Infectious Disease Analysis, Imperial College, London, UK
juliette.unwin@bristol.ac.uk

Seth Flaxman
Department of Computer Science, University of Oxford, Oxford, UK
seth.flaxman@cs.ox.ac.uk

Swapnil Mishra
Saw Swee Hock School of Public Health, National University of Singapore, Singapore
s.mishra@nus.edu.sg

Acknowledgements

First, I would like to thank my primary supervisor Samir Bhatt for his faith in me and his guidance of my path to becoming an independent researcher. Much of this work would not have been half as comprehensible without the help of Ettie Unwin who kept me on track, offered continuous support and pushed me to improve. Thank you to Seth Flaxman for suggesting interesting directions in which to take the project and also for inspiring me with numerous ideas. I want to thank Swapnil Mishra for his advice and encouragement. I am grateful for the help of all my supervisors and have benefited from their vast amount of knowledge and insight.

I would like to thank Naja Hulvej Rod, from the University of Copenhagen, for her wealth of expertise in epidemiological modelling, advising me on study design and making my visits to Denmark enjoyable. I also would like to acknowledge Lise Kristine Højsgaard Schmidt and Anna Rubach Birkmose for helping me with countless technical problems and forgiving me for the many times my large runs crashed the server. Thank you to Kathe Jensen for supporting and coordinating my visits to the team in Copenhagen.

I am indebted to Julie Middleton for helping me with both pastoral care and endless patience when I was filling out expense forms. I would like to thank my original PhD cohort teammates Andra, Will, Katharine, Ollie, and Jonathon for all the origami, fun times, and support throughout the lockdowns. I would like to especially thank Neil for his dedication to helping me sift through the documentation of terrible Python packages and ironing out my code intricacies. Without him my Python code would not be where it is today. I would also like to thank Alex, Jacob, Johannes, KC, Lingfa and Mark for their advice, support and hilarious jokes that have made the last months of my PhD really lovely. I would also like to thank Adrian for giving me guidance about explainability methods.

I'd like to thank all my friends Oliver, Kate, Poppy, Iwona, Kwasi, Alex, Ruth, Conni, Chris, Hector, Evelina, Poonam, Nora, Lily, Charlie, Phil, Anisha, Andrew, Aran, Dave, Connah, Anand, Zoi, Emily, Sasha and Josh for keeping me going and being amazing supports. Thank you especially to those, I am sure they know who they are, who believed in me and helped me find myself again. I would like to thank Kyan, without whom this thesis could have been finished six months ago, but I would not have a documentary in a film festival.

This thesis would be nothing without my amazing family who have gone above and beyond to support me. My sister Claire who has done endless work sessions with me and is my absolute rock. My sister Lucy who is unbelievably kind and has incredible levels of emotional awareness. My mum and my dad, who are welcoming, supportive, and inspiring. As well as my beloved grandparents, Geoff, Hazel, Magaret, and Mike. This is dedicated to you all x.

Abstract

Respiratory tract infections (RTIs), including pneumonia and bronchitis, are common illnesses that significantly impact global health, posing substantial challenges to healthcare systems. This research addresses the need for a comprehensive understanding of RTI dynamics and the underlying patterns which link an individual's early-life exposures, such as medical conditions and adverse childhood experiences (ACEs), with an increased susceptibility to RTIs. Traditional approaches cannot reflect the complexity of longitudinal data, struggling to capture the dynamic interplay of numerous factors influencing health outcomes. In contrast, state-of-the-art machine learning (ML) models have achieved excellent performance in pattern recognition tasks. A simulation study was carried out that employs synthetic longitudinal data, to assess the applicability of several ML models, including ResNet and the Signature method, within life course research. This simulation provides a controlled environment to rigorously test model performance. The ML models demonstrate efficacy in capturing intricate temporal patterns, prompting an investigation of the DANLIFE cohort; a longitudinal data set comprised of Danish registry data. Suboptimal performance in predicting RTIs led to an exploration of other health outcomes, finding that ResNet excelled in predicting diabetes and neurological conditions, revealing the model's capacity to discern patterns and dependencies in the data. Explainability methods, such as permutation feature importance and SHAP, were used to identify potential risk factors influencing health outcomes, finding a link between incidents of parental alcohol abuse and diabetes. Explainability methods enhance transparency, allowing researchers to interpret and trust the relationships uncovered by ML models. In summary, this study not only leverages innovative ML models but redefines how life course analysis can be approached. Through methodological innovations, it opens avenues for more precise, interpretable, and impactful investigations into the complex dynamics of health trajectories, contributing to the evolution of epidemiological research.

Abbreviations

ACE	Adverse childhood experience
AI	Artificial intelligence
ALSPACE	Avon Longitudinal Study of Parents and Children
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
AUROC	Area under the received operating characteristic curve
BPTT	Back-propagation through time
CNN	Convolutional neural network
COPD	Chronic obstructive pulmonary disease
CPR	Civic Personal Registration
DAG	Direct Acyclic Graph
DANLIFE	DANish LIFE course
DL	Deep learning
DTW	Dynamic time warping
EHR	Electronic health register
FCN	Fully convolutional network
FN	False negative
FP	False positive
FPR	False positive rate
GBTM	Group-based trajectory model
GCM	Growth curve model
GHE	Global Health Estimate
GP	Gaussian process
HES	Hospital Episode Statistics
HMM	Hidden Markov model
ICD	International Classification of Diseases
ICE	Individual conditional expectation
LCP	Life course pattern
LCC	Life course concept
LIME	Local interpretable model-agnostic explanations
LR	Logistic regression
LRTI	Lower respiratory tract infection
LSTM	Long short-term memory
LTA	Latent transition analysis
MA	Moving average
ML	Machine learning

MLSTM-FCN	Multivariate long short term memory fully convolutional network
MTS	Multivariate time series
NLP	Natural language processing
NPI	Non-pharmaceutical intervention
PAF	Population attributable fraction
PCA	Principal component analysis
PDP	Partial dependence plot
PFI	Permutation feature importance
PSI	Pneumonia severity index
ResNet	Residual Network
RNN	Recurrent neural network
ROC	Received operating characteristic curve
RSV	Respiratory syncytial virus
RTI	Respiratory tract infection
SARS	Severe acute respiratory syndrome
SHAP	Shapley additive explanations
SVD	Singular value decomposition
TN	True negative
TP	True positive
TPE	Tree-structured Parzen Estimator
TPR	True positive rate
TSC	Time series classification
URTI	Upper respiratory tract infection
WHO	World Health Organisation
XAI	Explainable artificial intelligence
XEM	Explainable-by-design ensemble method

List of Contents

1 Introduction & Background **14**

1.1 Respiratory Tract Infections . . . 14
1.1.1 Overview of RTIs . . . 14
1.1.2 Global Burden of RTIs . . . 15
1.1.3 Treatment and Prevention of RTIs . . . 17
1.2 Significance of Risk Factors in Health Outcomes . . . 17
1.2.1 Known Risk Factors for RTIs . . . 17
1.2.2 Overview of Risk Factor Studies . . . 18
1.3 Introduction to Life Course Epidemiology . . . 20
1.3.1 The Life Course Approach . . . 20
1.3.2 Personalised Healthcare and its Potential in RTI Management . . . 23
1.3.3 Life Course Analysis in Practice . . . 24
1.4 Description of the DANLIFE Cohort . . . 26
1.5 Machine Learning and Predictive Modelling for Healthcare . . . 29
1.5.1 Traditional Machine Learning Models for Time Series Data . . . 29
1.5.2 Deep Learning Approaches . . . 31
1.5.3 Model Interpretability in Healthcare and Epidemiology . . . 32
1.6 Integrating Life Course Epidemiology and Machine Learning for RTI Research . . . 33
1.7 Research Questions & PhD Objectives . . . 34
1.7.1 Summary of Key Themes . . . 34
1.7.2 Research Questions & Objectives . . . 35
1.7.3 Research Structure and Chapter Overview . . . 36

2 Data Simulation and Deep Learning Model Evaluation **37**

2.1 Introduction . . . 37
2.1.1 Life Course Patterns . . . 38
2.1.2 Deep Learning Models . . . 39
2.1.3 Chapter Outline . . . 49
2.2 Methodology for Data Simulation . . . 50
2.2.1 Directed Acyclic Graphs and Causality . . . 50
2.2.2 Input Data Formulation . . . 51
2.2.3 Life Course Pattern Simulation . . . 54
2.3 Evaluation Metrics and Experimental Setup . . . 57
2.3.1 Model Training Pipeline . . . 57
2.3.2 Metrics to Measure Model Performance . . . 59
2.3.3 Loss Function . . . 60

2.3.4 Model Training Procedure & Hyperparameter Optimisation . . . 61
2.3.5 Experimental Design . . . 62
2.4 Results . . . 63
2.4.1 Impact of Noise Levels on Model Performance . . . 63
2.4.2 Model Performance Across Life Course Groups . . . 65
2.4.3 Computational Efficiency of Machine Learning Models . . . 67
2.5 Interpretation . . . 69
2.5.1 Model Comparison across Life Course Patterns . . . 70
2.5.2 Balancing Performance and Computational Cost . . . 72
2.6 Limitations . . . 72
2.7 Conclusions . . . 73

3 The Signature Method **75**
3.1 Introduction . . . 75
3.1.1 Background of the Signature Method . . . 75
3.1.2 Challenges of the Sparsity of Exposure Event Data . . . 75
3.1.3 The Signature Method . . . 78
3.1.4 Research Questions of the Chapter . . . 81
3.2 Data Preparation and Experimental Setup . . . 81
3.2.1 Data Preparation . . . 81
3.2.2 Data Transformation Techniques . . . 82
3.2.3 Experiment Setup for Signature Method Evaluation . . . 83
3.3 Evaluating the Signature Method's Performance . . . 84
3.3.1 Impact of Increasing Truncation Level on Model Performance . . . 84
3.3.2 Signature Method Performance with Noise . . . 85
3.3.3 Performance of Signature Augmentation Methods on LCPs . . . 88
3.3.4 Influence of Interpolation on Signature Performance . . . 91
3.3.5 Model Factors Effecting Computational Burden . . . 91
3.4 Discussion . . . 93
3.4.1 Limitations of the Signature Method . . . 96
3.5 Conclusions . . . 97

4 Predictive Modelling of RTIs using the DANLIFE Cohort **99**
4.1 Unlocking Insights in Longitudinal Health Data . . . 99
4.1.1 Existing Cohort Studies . . . 100
4.1.2 Research Outline . . . 102
4.2 Data Description, Preparation and Model Training . . . 102
4.2.1 Data Description . . . 102
4.2.2 ResNet Model Training Procedure . . . 109
4.2.3 Signature Model Setup . . . 111
4.3 Experiment Design . . . 111
4.4 Results . . . 112
4.4.1 Data Summary . . . 112
4.4.2 ResNet Performance . . . 115
4.4.3 Signature Performance . . . 117
4.4.4 Computational Efficiency . . . 117
4.5 Discussion . . . 119

4.5.1 Understanding the Results and their Implications . . . 119
4.5.2 Limitations & Future Work . . . 121
4.6 Conclusions . . . 124

5 Explainability Approaches for Deep Learning Methods 125
5.1 Introduction . . . 125
5.1.1 Overview of Explainability Methods . . . 126
5.1.2 Permutation Feature Importance . . . 128
5.1.3 SHAP . . . 129
5.1.4 Research Questions and Objectives . . . 130
5.2 Methods . . . 130
5.2.1 Permutation Feature Importance . . . 131
5.2.2 SHAP Values for Feature Attribution . . . 131
5.2.3 Estimation of Risk Factor Scales . . . 131
5.2.4 Experimental Design . . . 132
5.3 Comprehensive Comparison of Interpretability Methods . . . 132
5.3.1 Permutation Feature Importance . . . 132
5.3.2 SHAP . . . 134
5.3.3 Risk Factor Estimates . . . 149
5.4 Discussion . . . 149
5.4.1 Risk Ratio Analysis . . . 152
5.4.2 Limitations & Future Work . . . 153
5.5 Conclusions . . . 154

6 Conclusions, Limitations and Future Research 155
6.1 Contribution to Life Course Epidemiology and RTI Research . . . 156
6.2 Limitations and Scope for Future Research . . . 158
6.3 Final Remarks and Key Takeaways . . . 159

7 References 161

8 Appendices 182
8.1 Simulated Data Distributions . . . 182
8.2 GHE Causes . . . 187
8.3 The Signature Method Window Transformations . . . 192
8.4 SHAP Values Without Scale Limit . . . 192

List of Figures

1.1 Upper and lower respiratory tract infections . . . 16
1.2 Life course approach diagram . . . 22
1.3 DANLIFE dataset . . . 28

2.1 RNN and LSTM model architectures. . . . 42
2.2 LSTM cell structure. . . . 43
2.3 FCN and LSTM-FCN model architectures. . . . 46
2.4 ResNet model structure. . . . 47
2.5 InceptionTime model architecture . . . 48
2.6 Inception module structure. . . . 49
2.7 Example of a directed acyclic graph (DAG) . . . 51
2.8 DAG used for input data simulation. . . . 52
2.9 Model comparison for multiple noise levels. . . . 64
2.10 Computation time. . . . 68

3.1 Diagram demonstrating MTS data sparsity. . . . 76
3.2 The Signature method. . . . 80
3.3 Predictive performance of Signature models for different truncation levels. . . . 86
3.4 Calibration performance of Signature models for different truncation levels. . . . 87
3.5 Predictive performance of Signature models for different noise levels. . . . 89
3.6 Predictive performance of Signature models for different interpolation levels. . . . 92
3.7 Training time for different truncation levels. . . . 93
3.8 Time taken to compute signature features for different interpolation levels. . . . 94

4.1 Selection criteria for the study cohort. . . . 103
4.2 Flat optimisation surface. . . . 116

5.1 Permutation feature importance with respiratory infections as the outcome. . . . 133
5.2 Permutation feature importance with diabetes mellitus as the outcome. . . . 135
5.3 Permutation feature importance with neurological conditions as the outcome. . . . 136
5.4 Permutation feature importance for time steps. . . . 137
5.5 Individual level SHAP values for simulated data set Period2. . . . 139
5.6 Average SHAP values for simulated data set Period2. . . . 140
5.7 Individual level SHAP values for simulated data set Timing1. . . . 141
5.8 Average SHAP values for simulated data set Timing1. . . . 142
5.9 SHAP values for RTI outcome. . . . 143
5.10 SHAP standard deviation values for RTI outcome. . . . 144
5.11 SHAP values for diabetes outcome. . . . 145

5.12 **SHAP standard deviation values for diabetes outcome.** 146
5.13 **SHAP values for neurological conditions outcome.** 147
5.14 **SHAP standard deviation values for neurological conditions outcome.** 148

8.1 **Sibling illness exposure events within the population.** 183
8.2 **Distribution of sibling illness exposure events across the life course.** 183
8.3 **Poverty exposure events within the population.** 184
8.4 **Total number of instances of poverty according to education and social class.** 184
8.5 **Parental alcohol abuse exposure events within the population.** 185
8.6 **Total number of instances of parental alcohol abuse according to their sex and ethnicity.** 185
8.7 **Mild illness exposure events within the population.** 186
8.8 **Distribution of mild illness events across the life span.** 186
8.9 **SHAP values for RTI outcome without scale limit.** 193
8.10 **SHAP values for diabetes outcome without scale limit.** 194
8.11 **SHAP values for neurological conditions outcome without scale limit.** 195

List of Tables

2.1 Summary of the five time invariant variables in the simulated input data. 53
2.2 Summary of the four time dependent variables in the input data. 55
2.3 The definitions for each LCP within each LCC. 57
2.4 Relative model performance within the four LCCss; period, repeats, order, and timing. . 66

3.1 Relative model performance within four LCPs; period, repeats, order and timing. 90

4.1 Individual characteristics definitions. 104
4.2 ACE variable definitions 105
4.3 The WHO GHE categories. 107
4.4 Number of ACE exposures experienced. 113
4.5 Distribution of income in population. 113
4.6 Incidence rate of hospitalisation exposure events. 114
4.7 Incidence rates of each disease outcome. 114
4.8 ResNet performance on the DANLIFE cohort for multiple outcomes. 115
4.9 Signature performance on the DANLIFE cohort for multiple outcomes. 118
4.10 Time performance on the DANLIFE cohort for the RTI outcome. 119

5.1 Risk ratio estimates for diabetes mellitus outcome. 149

8.1 ICD-10 codes associated with GHE categories. 187

Chapter 1

Introduction & Background

1.1 Respiratory Tract Infections

1.1.1 Overview of RTIs

Respiratory tract infections (RTIs) cover all infections of the respiratory tract. The respiratory tract is divided into the upper respiratory tract, made of the nose, sinuses, mouth, pharynx and larynx, and the lower respiratory tract comprising the lungs, trachea, and bronchial tubes. RTIs are classified according to the area of the respiratory tract infected or by the causative pathogen. Multiple pathogens exist that can infect the respiratory tract, and each may result in several types of RTI according to the symptoms and location of infection (Figure 1.1). The most common pathogens that cause RTIs are influenza, respiratory syncytial virus (RSV), coronaviruses and *Streptococcus pneumoniae* [Sahadulla, 2018]. Upper RTIs (URTIs) tend to be less severe than lower RTIs (LRTIs). URTIs include the common cold, tonsillitis, laryngitis, and sinusitis. Symptoms of URTIs can include coughing, sore throat, nasal congestion, and headaches. LRTIs include pneumonia, tuberculosis, and bronchitis. There is greater variation in the symptoms of LRTIs than URTIs depending on the severity of the infection. Individuals infected with LRTIs may experience fever, coughing, chest pain and difficulty breathing.

Most cases of RTIs are mild, self-limiting and do not require treatment. However, complications may occur, leading to respiratory failure or acute respiratory distress syndrome that may result in hospitalisation, chronic disease, or death [Hoogendoorn et al., 2011]. Those that recover from the infection and possible complications may go on to develop chronic conditions, such as long COVID, which affects those who have had COVID-19, as well as asthma and chronic obstructive pulmonary disease (COPD). COPD represents the third greatest cause of premature death in Denmark [Bønnelykke et al., 2015; Hoogendoorn et al., 2011; Mahase, 2020]. Multiple or severe RTIs that occur early in life can result in decreased lung function, leaving individuals more susceptible to respiratory or cardiovascular diseases in the future [Dharmage et al., 2009; Dratva et al., 2016; Galobardes et al., 2008].

There are several modes of transmission of RTIs that are dependent on the causative pathogen; via direct contact with an infected individual, contact with contaminated surfaces or inhalation of infected respiratory droplets in the air. Most RTIs are highly contagious and have high basic reproduction (R_0) numbers, which represent the average number of secondary infections caused by one infected individual within an entirely susceptible population. For example, R_0 of the influenza virus is estimated to be in the range of 1 to 21 [Leung, 2021]. Consequently, RTIs can rapidly propagate within populations, posing challenges for containment and leading to widespread outbreaks and global pandemics. Given the endemic nature and high prevalence of

multiple RTIs, these infections frequently undergo pathogenic mutations, making it difficult to develop long-lasting vaccines and treatments.

1.1.2 Global Burden of RTIs

The global impact of RTIs has been vividly demonstrated by the COVID-19 pandemic, which has underscored the vulnerability of societies worldwide to these diseases. Since its emergence in December 2019, COVID-19 has led to the reporting of 7 million confirmed deaths globally [data.who.int]. Europe and the Americas had the greatest reported attack rates [data.who.int]. In Denmark, the focus of this study, nearly 3 million COVID-19 cases and around 30,000 intensive care unit admissions were recorded before May 2022 when data collection ended [sta]. Initially, with no existing vaccines, mitigation measures such as social distancing and restricting population mixing were implemented to curb new infections. This resulted in the governments of over 100 countries introducing lockdown measures, including the closure of schools, restaurants and often “stay at home" orders. The nationwide lockdown imposed in Denmark in March 2020 aimed to reduce mortality rates and alleviate pressure on hospitals but came at a great economic cost [Byrialsen and Raza, 2021]. Furthermore, these measures severely impacted the mental and physical well-being of the wider population.

Even prior to the COVID-19 pandemic, RTIs were the second leading cause of child mortality worldwide, accounting for an estimated 14% of all deaths in children under the age of five in 2019 [Institute for Health Metrics and Evaluation (IHME), 2019]. While the burden of RTIs is mostly concentrated in low- to middle-income countries, they remain a major concern in high-income countries as well. In Denmark, for instance, approximately 468,000 new cases of RTIs are reported each year in children under the age of five, with 29% of hospital admissions for children aged 1-4 years attributable to RTIs [sta; Institute for Health Metrics and Evaluation (IHME), 2019]. The prevalence of chronic respiratory diseases is also on the rise, with an estimated 545 million individuals affected worldwide in 2017. This represents a 39.8% increase since 1990, with Europe and Australia reporting the highest number of cases [Soriano et al., 2020]. Whilst the burden is highest in LMICs, the highest number of chronic cases is concentrated in Europe and North America.

The emergence of SARS-COV-2, the causal agent responsible for COVID-19, is not the only novel respiratory pathogen to emerge and cause a large-scale outbreak. In 1918, the Spanish influenza, a variant of influenza type A H1N1 virus, infected nearly one-third of the world’s population and resulted in approximately 4 million deaths, marking one of history’s deadliest pandemics [Taubenberger, 2006]. Both SARS-COV-2 and Spanish influenza are zoonotic diseases, which are thought to have originated in animals before spreading to humans [Andersen et al., 2020; Anderson and Tong, 2010; Reid et al., 1999, 2001]. This is a recurring pattern with RTIs; many originally emerge as novel pathogens from zoonotic transmission and there are several large waves through the population until they eventually become endemic when the population has obtained some level of immunity. Other notable examples include the H5N1 influenza outbreak in 1997, the severe acute respiratory syndrome (SARS) outbreak in 2003 and the H1N1 influenza pandemic in 2009, which are believed to have originated in poultry, civets, and pigs, respectively. While the majority of RTIs are endemic, there is an increasing risk of new zoonotic spillover events giving rise to future respiratory virus pandemics [Marani et al., 2021].

RTIs represent a considerable burden on healthcare systems, both in terms of their economic and social impacts. The high prevalence of RTIs in the general population means that there are many emergency and outpatient visits arising due to RTIs, which account for a high proportion of the costs of health expenditure [Meier et al., 2020]. Progression to severe illness in vulnerable populations leads to hospitalisation and expensive stays in ICU facilities and the use of ventilators. RTIs are also related to increasing antibiotic resistance due to viral RTIs being misclassified and overuse of antibiotics [Guitor and Wright, 2018]. This has a knock-on effect that treatment of antibiotic resistant strains of RTIs is more costly and treatments for other conditions

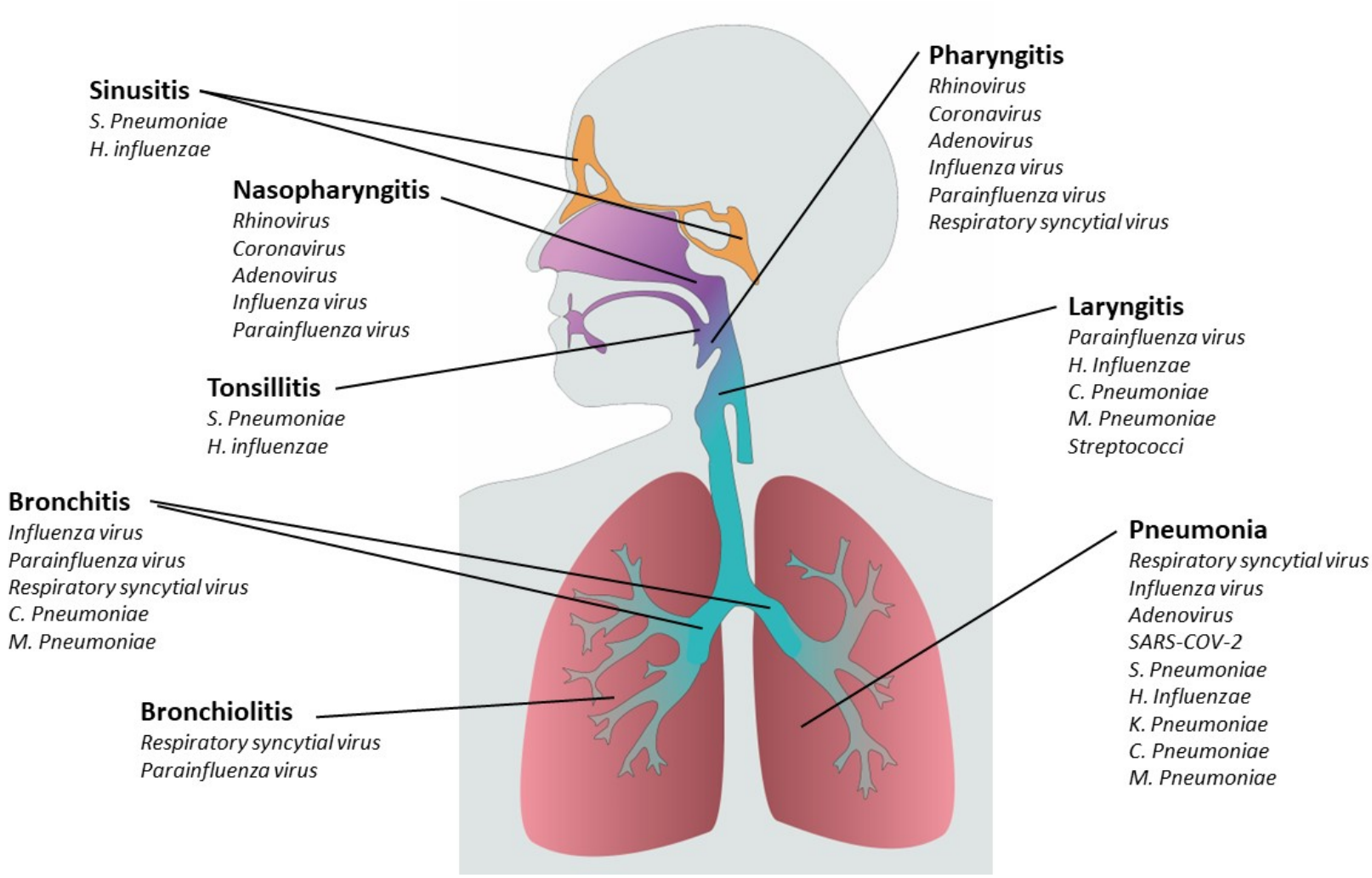


Figure 1.1: **Upper and lower respiratory tract infections** A diagram showing the main RTIs alongside the possible responsible pathogens and the location of infection. Figure created in Microsoft Powerpoint and based upon work by Calderaro et al. [Calderaro et al., 2022].

are less effective. The burden of RTIs is concentrated in socioeconomically disadvantaged populations which leads to health disparities in the population [Cortes-Ramirez et al., 2021]. They also heavily impact other vulnerable populations, such as children, the elderly and immunocompromised individuals [Pattemore and Jennings, 2008]. Additionally, RTIs contribute to a sizeable proportion of time taken off work and education due to sickness, causing a loss of productivity and academic performance [Fendrick et al., 2003]. In order to combat the effect of RTIs in the population on society, healthcare systems invest in large scale vaccination programs and public health campaigns to promote preventative measures. It is in the best interest of healthcare systems to have knowledge of which individuals are most susceptible to RTIs to enable early detection and effective vaccination schemes.

1.1.3 Treatment and Prevention of RTIs

Treatment and prevention strategies for RTIs have evolved to encompass a range of supportive and pharmaceutical interventions. While supportive care, including rest and ventilator support, remains fundamental in managing RTIs, targeted pharmacological approaches are employed in specific cases. In the early stages of the COVID-19 pandemic, when knowledge about the novel virus was limited, the primary focus was on implementing non-pharmaceutical interventions (NPIs) until vaccines could be developed [Flaxman et al., 2020]. NPIs aimed to minimize social contact and included widespread measures such as school and workplace closures, mask-wearing mandates, and stay-at-home orders, implemented across many countries. There was also a focus on promoting good hand hygiene practices and covering the mouth and nose when coughing to reduce respiratory droplet transmission. During the Spanish Flu outbreak, similar NPIs were used to control the transmission rate such as school closures and bans on large gatherings [Tomes, 2010].

RTIs pose a challenge for vaccine development due to their rapid mutation rate resulting in antigenic shift [Harding and Heaton, 2018; Kaushal et al., 2020; Kim et al., 2018]. This mutational variability complicates vaccine formulation, as multiple and evolving viral variants need to be targeted. For instance, the efficacy of the annual seasonal influenza vaccine varies substantially from year to year due to the prevalence of specific viral strains and their resistance patterns. In 2014, the effectiveness of the seasonal influenza vaccine dropped to 19% due to variant strain predominance [CDC]. Nevertheless, vaccines have been successfully developed for several respiratory pathogens. Additional vaccines against *Streptococcus pneumoniae* and SARS-CoV-2 have been widely deployed, with ongoing efforts to develop a vaccine against RSV currently underway [Anderson et al., 2017; Mazur et al., 2023].

Understanding the risk factors associated with RTIs plays a critical role in shaping effective health policies to target the most influential factors affecting population health, prioritising at-risk individuals in vaccination programmes, or shielding schemes, and pre-empting the impact of novel RTIs on the population. Identifying high-risk populations who are more susceptible to severe outcomes and implementing strategies can effectively reduce the burden of RTIs and protect the most vulnerable members of society. This is why it is imperative to gain an understanding of how certain life events acts as risk factors and lead to some individuals having an increased risk of severe RTI-related illness.

1.2 Significance of Risk Factors in Health Outcomes

1.2.1 Known Risk Factors for RTIs

Risk factors impact an individual's susceptibility to developing a disease or the severity of the outcome. They are classified into two categories: innate or external. Innate risk factors encompass genetic information and individual characteristics such as age and ethnicity. Examples of external risk factors are environmental

influences, such as air pollution levels, and being exposed to other life events, such as traumatic events or socioeconomic circumstances. One of the most significant risk factors for RTIs is age: children and the elderly are at a significantly greater risk of severe RTI-related illness [Simoes, 2003]. Many comorbidities impact the efficiency of the immune system and therefore increase an individual's likelihood of acquiring an RTI. Examples of comorbidities that have been linked to increased RTI risk include diabetes, obesity, hypertension, and HIV [Berbudi et al., 2020; Brown et al., 2017; Gao et al., 2021]. Air pollution is detrimental to lung health meaning that working or living in areas with high levels of air pollution also acts as a risk factor [Domingo and Rovira, 2020]. Similarly, smoking actively or passively is harmful to lung performance and increases one's risk of severe respiratory illness.

Lung development continues during childhood, therefore risk factors occurring within this time period can have lifelong respiratory health consequences [Grant et al., 2020]. In fact, many of the significant RTI risk factors that have been identified in studies are concentrated within childhood. These include living in overcrowded housing, low socioeconomic status, malnutrition, and poor maternal education [Brown et al., 2017; Gao et al., 2021; Haerskjold et al., 2015; Sommer et al., 2011]. Adverse childhood experiences (ACEs) are defined as potentially traumatic exposure events occurring during childhood, such as parental separation, living in poverty, or the death of a family member [Hughes et al., 2017]. The occurrence of stressful or traumatic events has been linked to an increased susceptibility to RTIs [Pedersen et al., 2010]. ACEs have also been linked to a higher risk of mental health disorders, cardiovascular diseases, diabetes, and other chronic conditions in adulthood. Certain ACEs, such as family poverty in childhood, have been found to increase the likelihood of developing asthma-like illness [Béatrice et al., 2012]. Therefore, it is believed that there may be several ACEs that are RTI risk factors.

Identifying previously unknown risk factors in childhood and understanding the interactions between different risk factors within this period could prove critical for targeting intervention strategies within public health schemes. Knowledge of the dynamic temporal interactions between risk factors and whether there are specific developmental periods during which they have greater impact can aid policymakers in intervention design. Interventions targeting specific risk factors could modify or mitigate the likelihood and severity of future RTIs. Further, this information can highlight the contribution of non-medical interventions, such as reducing pollution, addressing poverty, or introducing children's therapy services, in improving population health.

1.2.2 Overview of Risk Factor Studies

When assessing the benefits of potential interventions, policy makers will perform detailed risk assessments to generate insights into the causal links between an exposure and the outcome. These assessments generally involve a comprehensive literature review with particular focus on epidemiological studies of risk factors allowing the policy makers to estimate the impact of different interventions on the prevalence of the outcome in the population. The epidemiological studies that explore the associations between specific risk factors and the likelihood of a particular outcome or condition will generally take the form of observational cohort studies. Observational studies include cross-sectional studies, which examine the relationship between an outcome and risk factors at a set point in time, and case-control studies which compare individuals with and without the outcome to try to identify potential risk factors. In these studies, the strength and direction of association is usually measured using the odds ratio. The odds ratio is calculated as the odds in the exposed group divided by the odds in the unexposed group where the odds within a group are the number of individuals in that group who experienced an outcome divided by the number of individuals who did not [Bland and Altman, 2000]. It is also worth noting that the analyses of these types of study will usually present two versions of each measure of association, with one being raw and the other having been adjusted to account for a range of possible confounding factors that are assumed to be on the causal pathway. The most common form of

study design for examining exposures accumulated over time is a cohort study which can either prospectively or retrospectively follow a group of individuals over time to examine which individuals developed the outcome. The metrics of association used in this type of study are usually risk ratios and hazard ratios. The risk of an outcome within a particular cohort is defined as the proportion of individuals in the group that experienced that outcome during the follow up period. The relative risk, otherwise known as the risk ratio, is then calculated as the risk in the exposed group divided by the risk in the unexposed group [McNutt et al., 2003]. Similarly, the hazard ratio is the ratio of hazards between the two groups with the hazards representing the likelihood of an individual experiencing an event at a certain timepoint given they have survived and not experienced the outcome up until that point [Spruance et al., 2004]. Analysis that examines the hazard rates is known as survival analysis which models time-to-event outcomes, such as the time until disease occurrence or death. The most common approaches used within survival analysis are Cox proportional hazards models and Kaplan-Meier survival curves. The Kaplan-Meier survival curves estimate the survival function while the Cox proportional hazards model estimates the hazard ratio. The model allows for censored data and time-dependent covariates; however, this model assumes the risk factors act independently and also depends upon the assumption that the hazards are proportional, in other words that the hazard ratio is constant over time. There are other types of regression analysis used to understand the impact of exposures on the outcome, including linear, logistic or Poisson regression according on the type of outcome. Coefficients are assigned to each variable or exposure that can be used to gauge the magnitude of their impact.

In Denmark, policymakers are guided by Sygdomsbyrden i Danmark (The Burden of Disease in Denmark) which examines the burden associated with nine assorted risk factors based upon Danish Registry data [Flachs et al., 2015]. These nine risk factors and their effect on morbidity and mortality are measured using questionnaires that were given to a small subset of the population and then scaled up to represent population-level health impacts. The small cohort size limits the generalisability of the results, i.e. the extent to which these results will relate to other groups or the wider population. Surveys are expensive to conduct and therefore, it would not be cost-effective to conduct this sort of risk assessment at a greater scale to better represent the entire Danish population, especially if it were to be regularly updated to account for changes in lifestyle over time. Increasing the scale of the survey may additionally introduce considerable bias into the analysis, arising due to how participants are selected for the study and self-reporting and recall bias of individuals [Rottoli et al., 2020]. Selection bias is common in studies where participants are self-selected or all originate from the same sub-population, resulting in a lack of diversity in the cohort in terms of age, socioeconomic status, ethnicity, and pre-existing health conditions [Ellenberg, 1994]. Within Sygdomsbyrden i Danmark, the etiological fraction, otherwise known as the population attributable fraction (PAF), is used as a measure of the proportion of risk caused by a specific risk factor at the population level [Mansournia and Altman, 2018]. The calculation of the PAF depends upon the following assumptions: the risk factors act independently; a causal relationship exists between the risk factor and the disease outcome; and the effect of the risk factor is homogeneous across the entire population. The last assumption is particularly restrictive as in reality the risk represented by a certain exposure is likely to differ according to sex, age, and other individual characteristics. The calculations are also conducted at a population-level so they do not reflect how risk factors impact individuals disparately given their unique background and other exposures.

Measuring the level of impact represented by each risk factor is instrumental in developing models that can predict the risk of a certain individual developing the outcome. They can take multiple forms ranging from simple risk scoring systems, to regression models to machine learning (ML) models. Risk scoring systems assign weights to each risk factor according to their risk ratios or domain knowledge and then sum all the weights of the risk factors to which the individual has been exposed to obtain a final risk score [Aoki et al., 2016; Jones and De Cossart, 1999; Paes et al., 2009; Venmans et al., 2008]. The obvious drawback of this approach is that the different exposures and their timings can interact to magnify or mitigate the effect of

others, so these risk scores are likely to be inaccurate representations of the true underlying risk. In fact, using this approach has led to the overestimation of risk within certain population subsets [De Filippis et al., 2017]. Regression models assume that each risk factor acts as an independent variable so for each individual the risk can be calculated using the relevant covariates [Kleinman and Norton, 2009]. Again, this approach is only able to represent inter variable interactions in overly simplistic forms and the notion of temporal interactions is often neglected. These studies estimate the impact of a risk factor at the population-level, in some ways showing the magnitude of the effect for the average person. Whilst it can be useful having knowledge of how different exposures can affect the population as a whole, or indeed different subsets of the population, these measures of association can mask how both timing and an individual's unique life experiences may shape the level of risk an exposure represents to them. It is hard to assess the risk that these exposures present when multiple known risk factors have occurred due to their interactions, making it difficult to assess the true health burdens represented by different exposures. There has been a recent surge in ML methods being used within risk prediction due to advances in computation and the fact that ML models allow for a much greater level of complexity in the interactions between the risk factors [Dahiwade et al., 2019; Dritsas and Trigka, 2023]. However, the black-box nature of these models mean that it is more complicated to identify which risk factors have the greatest influence on the outcome. Recent developments have enabled some light to be shed on the inner working of these models and therefore, these models can be both highly accurate predictive models as well as providing insights on key risk factors. The ML models take many different forms with each allowing for a deeper understanding of the variable interactions or a reduced computational burden. They have the potential to calculate individual level risk scores directly from an individual's longitudinal history while also accounting for the timing and interactions of exposure events which has the potential to generate more reliable risk estimates.

1.3 Introduction to Life Course Epidemiology

1.3.1 The Life Course Approach

The life course approach is a theoretical framework that recognises the cumulative and interactive effects of social, environmental, and biological factors on an individual's health throughout their lifetime [H. et al., 2003]. These can include medical diagnoses, education level, environmental influences, socioeconomic status, and ACEs [Kalmakis and Chandler, 2014]. It emphasises the importance of understanding individuals' unique life trajectories, identifying critical periods of vulnerability and resilience, and the potential for interventions at various stages to promote health and well-being across the entire life span. This approach considers the timing, duration, and interactions of a range of factors and their long-lasting impact on future health [Ben-Shlomo and Kuh, 2002; Jones et al., 2019; Kuh et al., 2003]. The life course approach emphasises the accumulation of adverse exposure events over time and their potential causal pathways to future health outcomes, illustrated in Figure 1.2. For example, in this schematic representation, the co-occurrence of poverty and the loss of a parent during childhood may contribute to a hospital admission in adolescence.

Using the life course framework within the study of risk factors and development of risk prediction models has the potential to alleviate many of the problems experienced when using the traditional methods. Primarily, the life course approach acknowledges the importance of an individual's unique history in generating their disease risk, particularly how the same risk factor can have markedly different impacts on disease risk of two different individuals who have dissimilar backgrounds. This approach enables more accurate predictions of individual level disease risk as well as estimates of how significant risk factors are in giving rise to increased disease risk within sub-populations. To look at risk analysis through the life course lens, one must no longer consider risk factors in isolation and instead consider their timing as well as previous exposures when calculating their

associations. This allows for more realistic models that better reflect health pathways on an individual level that lead to increased RTI risk.

The core concept of life course theory is that pathways exist between multiple exposures to impact the likelihood or severity of given health outcomes [Jones et al., 2019]. In other words, there are patterns present within the structure of the data that connect a combination of several demographic factors, environmental exposures, and medical events with an increased risk of RTI-related illness. Such patterns are thought to be associated with the underlying biological and societal processes that connect life experiences to one's current health status [Halfon et al., 2014]. Several potential significant patterns of exposures have been proposed as conceptual models, each backed by experimental evidence. Moreover, it is generally believed that whilst the true underlying pathways are likely to be more complex than those described here, they will probably be a combination of these hypothesised significant exposure patterns.

Developmentally, there are certain periods within the life course during which occurrences of exposures have a greater impact on future health. For instance, when the lungs are still developing during childhood, it is known that passive smoking is more detrimental to lung health and consequently future respiratory health than during adulthood [Grant et al., 2020]. In fact, it has been established that early life experiences can have a long-lasting impact on the development of the immune system and its ability to fight infection. Specifically, several ACEs, such as abuse, neglect or low socioeconomic status can have significant health consequences and severe repercussions for respiratory well-being [Mocelin et al., 2022]. There is a paucity of research that examines both the influential risk factors as well as during which periods they have the most harmful effect. The World Health Organization put forward multiple life course conceptualisations with the aim of classifying potential life course patterns that may be present in data [World Health Organization and International Longevity Centre-UK, 2000]. Within this work, the idea of a critical period was introduced. Critical periods are characterised by the occurrence of an exposure directly giving rise to the health outcome with certainty [Ben-Shlomo and Kuh, 2002]. For example, low socioeconomic status in childhood irreversibly increases the likelihood of poor cardiovascular health [Poulton et al., 2002]. A similar concept is the sensitive period during which an individual's risk of poor health outcomes is increased by a greater margin if the exposure event happens within this period. Unlike critical periods, the occurrence of the exposure within a sensitive period is not enough to guarantee the outcome happening but it does increase the likelihood. Ben-Shlomo et al. explain this distinction as these two types of periods having "qualitatively different exposure-time interactions" [Ben-Shlomo and Kuh, 2002].

Another key paradigm within life course analysis is the concept that the occurrence of consecutive adverse experiences may substantially increase the risk of negative health outcomes. This epidemiological model is known as the accumulation-of-risk model [Kuh et al., 2003]. Cumulative exposures to stress, environmental factors, and other exposures across the life course can lead to the "weathering" effect, which accelerates biological aging and increases disease risk [Raffington et al., 2023]. Marginalised populations experiencing chronic stress and discrimination may be particularly susceptible to the weathering effect, leading to premature health decline. An example of this concept in action is that exposure to chronic poverty during childhood has been found to increase reported asthma-like attacks [Béatrice et al., 2012]. The duration of poverty exposure was found to be a significant predictor of poor health as well as whether children were chronically or transiently experiencing poverty. These results suggest that two individuals who both experienced living below the poverty threshold for exactly two years of their respective childhoods can have different risks of disease if one person experienced those years consecutively and the other did not. Logically, this makes sense as the individual experiencing the two years of poverty consecutively will likely have had to deal with long-term consequences of poverty such as the cumulative effects of lower nutrition, housing conditions and lack of access to social and educational opportunities. There are many other instances where this effect can be seen, for

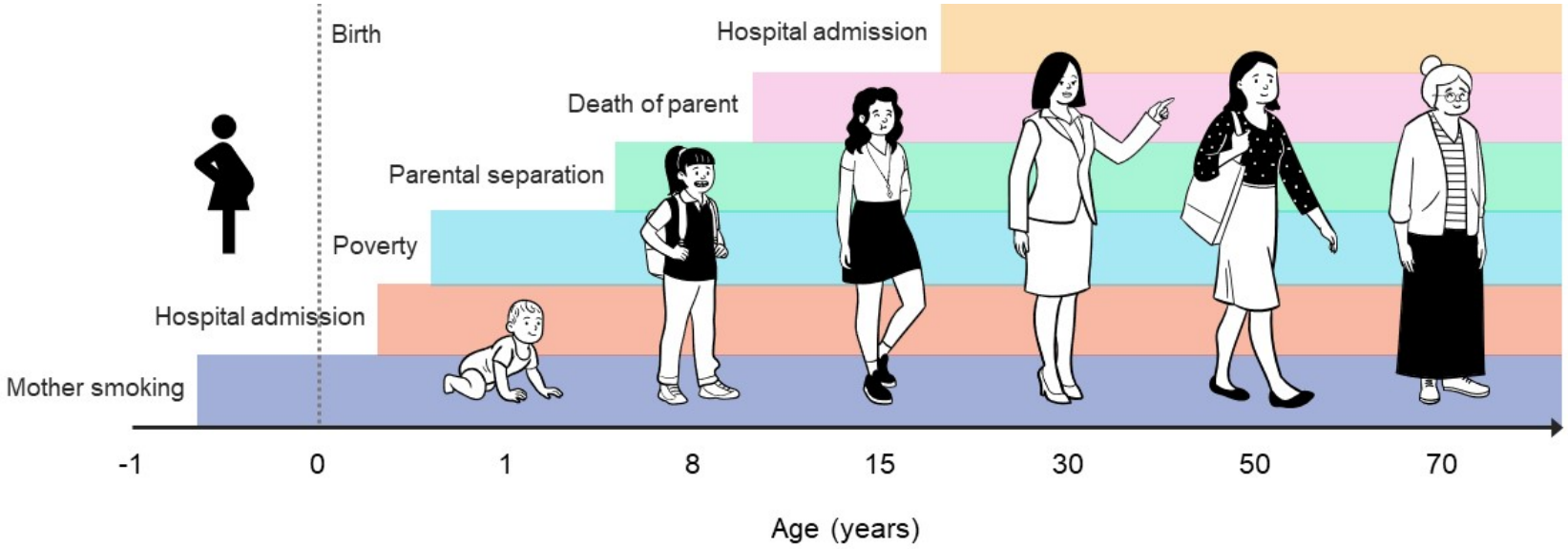


Figure 1.2: **Life course approach diagram** A schematic representation of the principles of life course theory. The x-axis follows the life course of a woman from pre-conception until old age. Multiple exposure events (for example, hospital admission, poverty etc.) are represented by coloured bars. The start points of the bars indicate when the exposure event occurred and the length of bar demonstrates the duration of the impact on health outcomes. Figure created in Microsoft Powerpoint.

example, the impact of parental drug abuse on their children's disease risk [Conners et al., 2003].

The life course concepts view the health trajectory of each individual as something that is nudged up and down by separate exposure, with the magnitude of this nudge dependent on the timing of the exposure and the previous exposures experienced. Consequently, the timing of exposure events, particularly in relation to other exposures, is important to consider within the life course framework. Multiple different adverse experiences happening within a short time frame can magnify their impact on the risk of negative health outcomes. For example, individuals exposed to both asbestos and smoking have a greater jump in risk for lung cancer than individuals with only one exposure [Ngamwong et al., 2015]. This principle applies within other settings, such as how an individual who experiences becoming unemployed and has a heart attack within the same year will likely have longer-term impacted health than if these events occurred several years apart [Dupre et al., 2012]. In a similar vein, the temporal ordering of exposures has a substantial impact on the health trajectory [Mishra et al., 2010]. Which event comes first is believed to be a factor in producing an individual's relative risk of negative health outcomes [Wahrendorf and Chandola, 2016]. For instance, an individual's likelihood of a heart attack is greater if they have experienced depression before [May et al., 2017]. All these epidemiological life course concepts cover distinct aspects of how exposure events are linked to health outcomes and provide a framework for how to approach discovering influential risk factors and how they interact with one's life history.

1.3.2 Personalised Healthcare and its Potential in RTI Management

Personalised healthcare refers to an approach that focuses on tailoring medical care and interventions to the specific needs and characteristics of each individual [Chawla and Davis, 2013]. It involves utilising information about an individual's medical history, genetic makeup, lifestyle, and environmental factors to customise their medical care and optimise health outcomes. Such knowledge can be used to guide treatment decisions by identifying the most effective treatment with fewest side effects and avoiding treatments that may work best in most cases but not for this individual. It also improves diagnoses by basing them on a patient's unique history, speeding up time to diagnosis and reducing misdiagnoses. This patient-centered approach aims to provide more precise and effective healthcare interventions while considering the individual's preferences and values. As advancements in technology and data collection continue to progress, personalised healthcare is emerging as a main area of interest in healthcare research and practice [Ross et al., 2014].

There exists significant synergy between the outlooks of personalised healthcare and life course epidemiology. Both approaches recognise the need to tailor medical interventions and healthcare strategies to the individual, utilising personal data to identify life course patterns and genetic factors influencing specific health needs. Personalised healthcare considers the inherent heterogeneity within the population that creates diverse trajectories and aims to provide tailored medical care that considers individual differences in disease risk, susceptibility, and response to treatments. The importance of critical time windows is emphasised in life course epidemiology and personalised healthcare helps identify these critical time windows for specific individuals, enabling the timely implementation of preventive measures or targeted interventions to optimise health outcomes. Personalised healthcare allows for early identification of individuals at higher risk of developing certain health conditions based on their life course patterns and genetic profiles. Early intervention strategies can then be implemented to prevent or mitigate the development of diseases, improving long-term health trajectories. Life course epidemiology recognises that individuals respond differently to treatments based on their unique characteristics and health histories and personalised healthcare tailors treatment plans to each individual's specific needs, optimising treatment efficacy while minimising potential side effects. Identifying high-risk groups and tailoring interventions can lead to more efficient allocation of healthcare resources. By integrating personalised healthcare within the framework of life course epidemiology, researchers and health-

care providers can gain a more comprehensive understanding of the dynamic interplay between individual life trajectories to inform public health policy, develop more effective and targeted interventions and improve health outcomes.

Personalised healthcare has the potential to improve RTI treatment, prevention, and outcomes. The ability to identify individuals with the greatest risk of developing an RTI allows for early identification and treatment of susceptible individuals, reducing the strain on the healthcare system. Furthermore, knowledge of individuals at increased risk of more severe illness enables them to be prioritised early for hospital beds and ventilators, reducing costs, and improving health outcomes due to timely treatment. Designing vaccination schemes that prioritise high-risk individuals can help to protect the most vulnerable people in the population against RTIs.

1.3.3 Life Course Analysis in Practice

Multiple analytical approaches have been leveraged to conduct life course analysis. The methodological approaches that have gained the most popularity in life course analysis to date are trajectory analysis and sequence analysis [Aisenbrey and Fasang, 2010; Bengtsson et al., 2021]. The primary focus of life course studies is to identify the most influential exposures or developmental periods for given health outcomes as well as which individuals have the greatest risk. Conducting life course research relies upon a large longitudinal data set that contains repeated measures of multiple exposures over an extended time frame. Consequently, the modelling methods must employ some form of data aggregation method or be able to appropriately model sparse data. The data is sparse by nature because of the rarity of health outcomes in addition to various exposures that occur at low levels in the population. In this section, the strengths and weaknesses of the traditional life course models are examined as well as approaches to maximise the usability of the big longitudinal data sets.

Data Aggregation and Event Counting

Most of the modelling methodologies used within life course analysis require the large longitudinal data sets to be pre-processed. There are a number of ways in which this can be done depending on the model structure. It is common for the longitudinal data to be aggregated to reduce its dimensionality, particularly in the temporal dimension. This can range from the exposures being grouped into counts per year to the time dimension being removed entirely to just have total counts across the entire time period. In some instances, the timeline is divided into a small number of periods that are thought to align with the most influential developmental periods, for example early-, mid- and late- childhood. This approach can be used alone or in conjunction with feature reduction. Again, there are multiple procedures used to achieve this, such as grouping similar features into broader categories and taking the sum or the average at each time point. These methods are not forms of feature selection - these techniques represent a change in the formulation or definition of the features themselves. Another form of data aggregation is to group the life courses of multiple individuals based upon their past exposures. This approach then focuses on identifying disparities between the groups' exposures and their health trajectories [Hutchison, 2019]. Each of these aggregation methods can be used alongside feature selection methods, such as principal component analysis (PCA), to further reduce the dimensionality of the data and thereby reduce computational complexity and improve model performance. Feature selection methods seek to capture the most significant information in the data whilst reducing the amount of noise in the input data.

The disadvantage of employing models that rely upon data aggregation within the analysis is that some of the rich longitudinal information in the data is discarded in these procedures. The act of grouping different exposures within a single feature essentially equates the severity of both types of exposure as well as their

interactions. Additionally, creating categories involves more assumptions about the significance of different features which limits the model's ability to find novel relationships between the variables. Aggregation of the time dimension means that subtle temporal interactions between exposures may be masked. The most detrimental procedure is the clustering of individuals into groups as the data can no longer be used to get personalised individual level insights. Consequently, the analysis is finding population averages based upon averages within subgroups leading to generic risk predictions. Therefore, it is of interest to employ modelling methodologies that enable minimal data aggregation and pre-processing in order to take full advantage of the rich longitudinal data.

Trajectory Models and their Application in Life Course Analysis

Several modelling methodologies have been used within life course analysis; a standard framework is known as trajectory modelling. This method aims to cluster individuals in the population according to the type of exposures experienced and draw conclusions about differences in their resulting health trajectories [Nguena Nguefack et al., 2020]. These clusters are called trajectory groups and are sub-populations with similar life courses that are expected to have similarities in their future disease outcome. In theory, this method can help to understand variability in large data sets and see if there are general patterns followed by the majority of individuals. Once the trajectory groups have been created then the group indicator can be used as a feature within further analysis. There are multiple types of trajectory model and they have been applied within a range of fields spanning social sciences, psychology, and public health [Eggleston et al., 2004; Nagin and Odgers, 2010; Pierce et al., 2021; Walsh et al., 2020]. Growth curve models (GCMs) aim to examine how trajectories change over time and to estimate an average trajectory that represents the general trend within the entire population [Cillessen and Borch, 2006]. This approach is a big simplification of the underlying dynamics as we know that certain sub-populations that experience specific exposures, such as lower socioeconomic status or pre-existing medical conditions, will have vastly dissimilar health trajectories than the average individual in the population. Therefore, a more popular approach is group-based trajectory modelling (GBTM) which aims to split the population into groups with similar developmental patterns of exposures [Nagin et al., 2018]. The different trajectory groups are assigned so that they best explain the heterogeneity in the population whilst also providing intuitive groupings of individuals. For example, the intuitive grouping used by Rod et al. is five trajectory groups according to the number and types of ACEs experienced across the life course [Rod et al., 2021]. The trajectory model does not assume any fixed form, making it possible to capture a range of patterns within the population. The groups are fixed across the life course as the model uses the temporal changes in variables to assign trajectory groups, reflecting temporal dynamics in the groupings. The assigned trajectory groups can be used to compare how the group members differ in terms of their individual characteristics, past exposures, and disease outcome. This assessment helps to guide policymakers in intervention design as it may highlight specific characteristics or exposures that have the greatest influence on health trajectories. There are other approaches, such as latent transition analysis (LTA), that allow for individuals to switch groups at some point in the life course. However, these models were designed for clinical analysis with a short time frame and categorical variables making them inappropriate for risk analysis [Nguena Nguefack et al., 2020].

The main limitation of trajectory models, particularly GBTMs, is that the choice of the number of trajectory groups is subjective and can have a significant impact on the results. In addition, it may not be possible to produce an intuitive understanding of how the trajectory groups differ and which exposures are most influential in generating the different outcomes. This limits the utility of the results as it is not possible to deduce which sub-populations relate to the most at-risk group. Furthermore, when the number of variables in the model increases, it becomes increasingly difficult to interpret the resulting trajectory groups. The model assumes that each trajectory group is distinct, and it is not possible to move between them, meaning that it is possible

for individual exposures to alter your health trajectory irreversibly. Furthermore, it may not be possible to capture all the sources of heterogeneity in the model, and pathways that result in changes to health outcomes are complex and may be rare and not affect enough individuals in the cohort to become a separate group. Another problem with this modelling approach is that by grouping the population in this way, individual level patterns can be missed and some of the rich information in the longitudinal data is lost. The GBTM assumes that trajectories are linear and so the outcome changes in a very predictable way that may not be able to reflect the complexities of life course patterns.

Sequence Analysis and its Benefits for Understanding Trajectories

Recently, another approach called sequence analysis which addresses some of the limitations of trajectory modelling has become popular within life course analysis [Aisenbrey and Fasang, 2010]. This method originated in the field of bioinformatics and is used to analyse the differences between two segments of genetic code and quantify the magnitude of the distance between them [Can, 2014]. Sequence analysis approaches have now been applied within many other fields, including epidemiology, political sciences, and geophysics [Blanchard, 2011; Gosselin et al., 2018; Kanasewich, 1981]. When applied to longitudinal data, the sequence analysis method assumes that the life course of an individual can be represented as a sequence of distinct states that they move through with age. These states can symbolise risk factor exposure, medical events, or differing income levels. The sequences of all individuals in the population are compared according to some measure of distance, such as dynamic time warping (DTW) [Brzinsky-Fay and Kohler, 2009]. A distance matrix is created that contains the alignment scores between all possible pairs of sequences, from this clusters can be created that group individuals with similar developmental trajectories. Similarly to GBTMs, these clusters can be compared to find the significant life course patterns that define them and their resulting differences in relation to the health outcome. Unlike GBTM, the use of clustering after comparing sequences means that the important events can be identified at the individual level, it also places a greater importance on the specific timing of events. This means that it is possible to find sensitive periods during which changes in state have the biggest impact on future health.

One big assumption within the sequence analysis framework is that events must be strictly ordered and must happen linearly [Aisenbrey and Fasang, 2010]. However, it is thought that life course transitions may be considerably more complicated than this and sequence analysis does not account for this level of complexity. Whilst this approach allows for long sequences, it performs poorly with sparse data which is often the format of medical records as hospitalisations are rare [Bai and Kremer, 2012]. The choice of distance measure has a significant impact on the results and may lead to erroneous conclusions [Halpin, 2012]. Sequence models are based upon the assumption that each state is independent of previous states, but the nature of life course theory means that childhood exposures will naturally continue to impact health trajectories into adulthood. Ultimately, sequence analysis is another method of creating clusters of individuals. However, this approach does not enable individual level insights to be gained into the intricate interactions between exposures and their timing on the disease outcome.

1.4 Description of the DANLIFE Cohort

The DANish LIFE course (DANLIFE) cohort has been compiled from multiple Danish registers to create a longitudinal data set of 2.2 million individuals followed for an average of 16.8 life years [Bengtsson et al., 2019]. This data set was designed to allow in-depth investigation of health outcomes through a life course lens, particularly the way in which ACEs influence health in adulthood. With information on medical diagnoses, socioeconomic factors, and ACEs, this cohort offers valuable opportunities to examine the influence of these

factors on individuals' health trajectories and outcomes. In Denmark, each person who enters the country, either through birth or immigration, is given a Civic Personal Registration (CPR) number. This number acts as a personal identification number and is recorded each time the person encounters public authorities or administrative centres, including hospital visits, employment records and education. Individual level linkage between multiple registers is therefore very straightforward. The registers used to create the DANLIFE cohort are the Danish Civil Registration System, the Danish Medical Birth Registry, the Integrated Database for Labour Market Research, the Register of Support for Children and Adolescents, the Income Statistics Register, the Population Education Register, the Danish National Prescription Registry, the Danish Psychiatric Central Research Register and the Danish National Patient Register [Bengtsson et al., 2019]. Figure 1.3 shows which variables are obtained from each register. The Danish National Patient Register and Danish Psychiatric Central Research Register provide detailed records of hospital admissions related to somatic or psychiatric illness. The CPR number allows for data linkage across the registers and, by using the family identifier provided in the birth register, individuals can be connected to their parents and siblings to obtain measures of ACEs. These ACEs encompass several factors including psychiatric or somatic illness within the family, parental substance abuse, death within the family and long-term parental unemployment. This comprehensive approach enables the examination of inter-generational health patterns and the exploration of familial risk factors.

All individuals born in Denmark between 1st January 1980 and 31st December 2018 are automatically included in the DANLIFE cohort. Therefore, the oldest individual in the cohort is currently 38 years old and, as the data set is updated periodically, the size of the cohort and the maximum lifespan of individuals continue to increase. Only individuals who have died or emigrated to another country prior to their 16th birthday are excluded from the cohort to prevent missingness in the data set. Therefore, aside from country of birth and emigration, the cohort is unselected which minimises selection bias and improves generalisability of findings. This is a significant improvement compared to other cohorts that have been used for life course analysis which rely upon a selection procedure. For instance, the Avon Longitudinal Study of Parents and Children (ALSPAC) is the largest life course cohort in the UK and the cohort was obtained by enrolling pregnant mothers in 1991 and 1992 in the Avon area [Jones et al., 2019]. Consequently, the selection procedure introduces substantial bias in that the population in the Avon area is largely white and the choice of pregnant women may also give rise to further bias. The response rate to the first questionnaire was 79% [Golding, 2004]. Another advantage of the DANLIFE cohort is that it is constructed solely from register data, eliminating recall bias regarding the timing or nature of different exposures. Furthermore, there is no loss to follow-up except in cases of death or emigration, which means there is no need to account for this source of bias. The DANLIFE cohort is also significantly larger than most cohorts meaning that any analysis will have greater statistical power, better represent the entire population, and record more instances of rare events. The use of record-based data alone represents a true reflection of what can be learned about population health without relying on additional variables collected through costly and time expensive questionnaires. This approach not only keeps costs low whilst studying an entire population but also facilitates the implementation of resulting policy changes or health indicators derived from this data. Increasingly, record collection of this type is expanding and there is a growing need for suitable methods to analyse these rich longitudinal cohorts effectively.

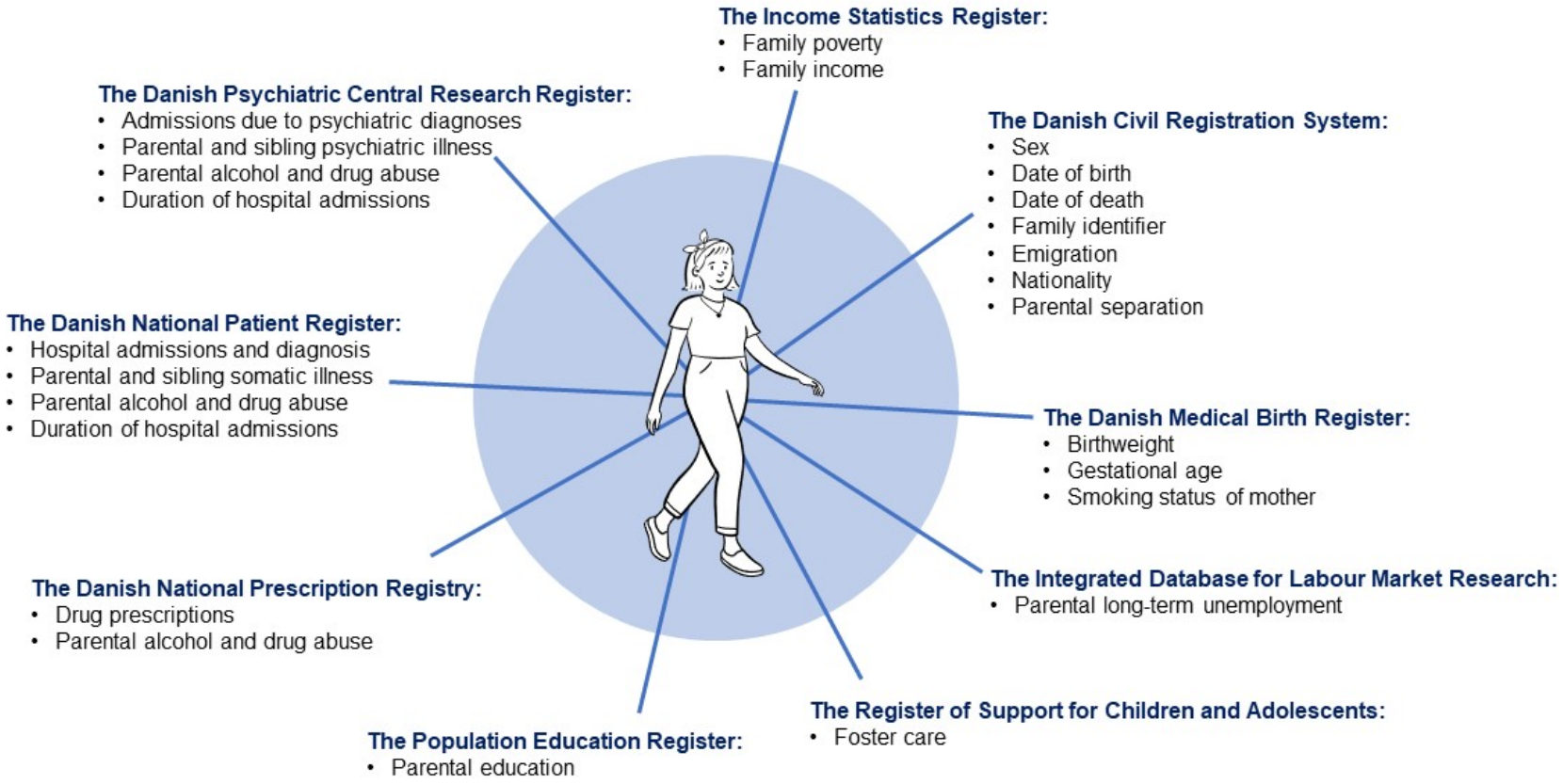


Figure 1.3: **DANLIFE dataset** The Danish registers and individual exposures that comprise the DANLIFE data set. Created in Microsoft Powerpoint and based upon a figure in [Bengtsson et al., 2019].

1.5 Machine Learning and Predictive Modelling for Healthcare

ML is a subset of artificial intelligence (AI) that describes machines which can perform human tasks, such as understanding language and object detection. The definition of ML is when systems can automatically learn and improve from experience without being explicitly programmed [Alzubi et al., 2018]. ML approaches are becoming increasingly popular across multiple fields and applications due to their flexible and powerful nature. ML models are able to identify and replicate hidden patterns in the data without requiring the interactions between variables to be pre-specified. ML is the broader category of models in which deep learning (DL) falls. DL models refer specifically to one type of ML model known as neural networks with multiple layers. Designed to replicate the functioning of the human brain, DL models are known to be the most capable of undertaking complex tasks but requiring the most data and prone to overfitting. The longitudinal data required in life course analysis can also be referred to as multivariate time series (MTS) data. Each individual has their own multivariate time series that follows the values of multiple features over time. This distinction is important as the majority of the methods mentioned in this section were designed for MTS data and have not been applied within a life course setting. In this section, a selection of the main ML methodologies used for MTS prediction tasks are described along with their current applications within healthcare settings.

1.5.1 Traditional Machine Learning Models for Time Series Data

ARIMA Models

Traditionally, MTS data has been analysed using autoregressive integrated moving average (ARIMA) models. They have been applied across fields such as stock price prediction in economics and predicting breast cancer in epidemiology [Adebiyi et al., 2014; Lin et al., 2021]. This class of models combines moving average and autoregressive models [Box et al., 1970]. Autoregressive (AR) models of order p assume that current values in the time series can be represented as a linear combination of the p proceeding values and some forecast error at that time point [Shumway and Stoffer, 2017]. The moving average (MA) model assumes that the current value in the sequence is made up of a linear combination of q previous forecasting errors [Shumway and Stoffer, 2017]. The term integrated refers to the process by which a time series is made stationary, i.e. independent of the time at which the measurement was taken. Stationarity is achieved by subtracting a previous value in the series from the current one d times until the mean and the variance of the series becomes constant over time. The ARIMA model combines the three methods so that the current value in the time series depends upon a linear combination of its previous values and forecasting errors once the data has been made stationary in time. Once fitted to the data using maximum likelihood estimation and consideration of the choice of p, d and q, this model will take in the time series data and predict future values. The ARIMA model was designed for univariate time series data but has been extended for MTS although this comes at a high computational cost [Siami-Namini et al., 2018]. There are several variations, such as the Box and Jenkins forecasting method which allow the ARIMA models to model non-stationary time series [Box et al., 1970].

The ARIMA model depends upon the assumption that the next step in the sequence can be represented as a linear function of the previous observations, which means that interdependencies between variables and nonlinear relationships are not able to be captured by this model. Life course trajectories can be complex and arise from interactions between different exposure events, so these linear relationships are unlikely to be able to capture these patterns. Another notable assumption within the ARIMA framework is that the data must be stationary over time. In life course data this assumption will likely be violated as when an individual ages certain exposures will become more or less frequent. The choice of the number of influential previous time points (p) limits this model's ability to reflect the impact of early-life events on disease risk later on. In addition,

this model is extremely sensitive to outliers which are very common within a large population. The reasoning behind certain model predictions can be difficult to interpret which is crucial within healthcare applications as practitioners must be able to justify medical decisions.

Hidden Markov Models

Hidden Markov models (HMMs) are commonly used to model time series data which is produced by an unobserved process. HMMs were first developed for computational biology and bioinformatics to analyse sequences [Eddy, 1996; MacDonald and Zucchini, 1997]. In gene prediction, the observed data is the nucleotide sequences, and the unobserved process is the functions of the DNA sequence. The models have also been applied within other contexts, such as predicting the stock market and speech recognition [Gupta and Dhingra, 2012; Tokuda et al., 2013]. The model is formed of two related sequences; one observed sequence corresponding to the data and an unobserved or hidden sequence that generates the observed sequence [Rabiner and Juang, 1986]. Within life course analysis, the hidden sequence would be the unobserved biological and societal pathways and patterns, possibly representing underlying conditions, that results in negative health outcomes. Each sequence is a series of possible states that are moved between sequentially, for instance the price of a given stock or states in disease progression. For the hidden sequence, the likelihood of moving from one state to another is governed by a matrix of transition probabilities that obey the Markov property. The Markov property states that the probability of moving to the next state depends solely on the current state and the proceeding states have no impact. The hidden sequence drives the values of the observed sequence, with the current observed state being emitted according to the current state of the hidden sequence according to a matrix of emission probabilities. Furthermore, the transition and emission probabilities are assumed to be time-independent meaning that they are the same for all points in the sequence. This assumption means that it is straightforward to conduct inference using the Baum-Welch algorithm [Rabiner and Juang, 1986]. With this algorithm, it is possible to fit the model to the observed data, estimate the transition and emission probabilities and calculate the most likely hidden sequence. It is also possible to use the fitted probabilities to forecast and predict future values. HMMs have been extended to make them more suitable for longitudinal data, by allowing intermittent observations and for the time series to be continuous [Maruotti, 2011; Zhou et al., 2020]. Coupled HMMs allow for more complex and nonlinear relationships between states to exist [Lehrmann et al., 2014]. Significantly, the mixed Markov model combines HMMs with generalised linear models and incorporates mixed effects to relax the assumption of conditionally independent observations [Maruotti, 2011].

HMMs increase in size and complexity with the addition of additional states and therefore, can quickly become prohibitively computationally expensive [Rabiner and Juang, 1986]. The other limitation of this modelling method is that they depend upon the Markov assumption being satisfied. In life course analysis, the underlying processes may be complex, time-dependent and non-Markovian rendering this model unsuitable in this context. As a result of the Markovian assumption, HMMs have a limited ability to model long-term dependencies, such as the relationship between early-life experiences and health consequences in adulthood. It is also well-known within life course research that one's health is directly related to age in the case of many diseases which breaks the assumption that transition probabilities are independent of time. It may be the case that the hidden state cannot be represented by a finite number of states and this number of states must be carefully selected by the user. Most model formulations do not allow for external covariates, i.e. additional observed states that are known to influence the outcome of interest, such as ACEs or other medical diagnoses which means that these models must discard a significant amount of information present in the raw longitudinal data. Finally, as with the ARIMA model, it can be complex to untangle why HMMs made certain decisions, limiting its utility in clinical settings.

Gaussian Processes

Gaussian Processes (GPs) are a powerful and flexible probabilistic modelling approach used in various fields, including time series analysis [Roberts et al., 2013]. Unlike ARIMA models, it is possible for nonlinear relationships to exist between variables and the approach is non-parametric. Additionally, unlike HMMs, GPs allow for more flexible, complex, and nonlinear representations of the time series. GPs were originally designed for time series forecasting and have also been used in in the fields of astronomy, bioinformatics, and public health [Futoma et al., 2017; McDowell et al., 2018; Rajpaul et al., 2015]. The GP modelling framework assumes that the output takes the form of a GP; that is there exists a function mapping the input variables to the observations that follows a multivariate Gaussian distribution and the output is equal to a sequence of the function evaluated at different time points. The Multivariate Gaussian distribution is defined according to a mean vector and covariance matrix. The covariance matrix is constructed using a covariance function, otherwise known as a kernel, which is chosen by the user to represent the correlation between any two input points based upon domain knowledge. In order to fit a GP model, first a prior distribution over the functions must be decided upon that is based upon knowledge of the data. Then the Bayesian approach is applied to obtain a posterior distribution of the functional form. GPs are remarkably effective for this type of analysis as the output includes uncertainty intervals in addition to point predictions.

The GP model has high computational cost, and this grows cubically with the number of observations. Consequently, data sets that have a large number of individuals, time points or features can result in extremely large computational requirements. Another limitation of GPs is that the model performance relies heavily upon the choice of the covariate functions. Often it is not known which kernel will be most suitable for the data and therefore, extensive experimentation is required. The data itself is required to be Gaussian for the modelling assumptions to hold and, therefore, if the outcome takes the form of counts or binary data then the GP model is not suitable. There exist a number of model extensions to resolve this problem, however they further increase the computational burden. Finally, the GP framework assumes stationarity of the data, which means that the statistical properties of the data remain constant over time. In life course analysis, trajectories can exhibit non-stationary behaviour due to age-dependent exposure rates and consequently, the stationarity assumption may not hold and significantly affect the model's accuracy.

1.5.2 Deep Learning Approaches

DL architectures offer an alternative approach to traditional MTS modelling techniques and are particularly advantageous when dealing with large-scale sequential data with complex patterns and dependencies. DL models are characterised by an increased flexibility, fewer model assumptions and an ability to learn the influential patterns directly from the data itself. These features allow for more intricate patterns present within the data to be identified. These models have become increasingly popular due to their ability to handle complex time-dependent relationships, capture seasonality and trends, and provide accurate forecasts. Due to the rise in the amount of MTS data, several DL models that were originally designed to be applied to language modelling or image processing tasks have been adjusted for application to MTS data in forecasting or classification tasks.

Briefly, there are two main categories of DL model that are most appropriate for time series data, these are recurrent neural networks (RNNs) and convolutional neural networks (CNNs). RNNs are designed to model long-term temporal dependencies in data and were originally designed for natural language processing (NLP) [Hermans and Schrauwen, 2013]. RNNs have been utilised within a range of fields, including NLP and traffic forecasting [Li et al., 2017; Mikolov et al., 2011]. They have also been applied within healthcare for early detection of heart failure onset, providing treatment recommendations, predicting clinical outcomes and diagnoses [Choi et al., 2017; Esteban et al., 2016; Ma et al., 2017; Wang et al., 2018]. Many such

clinical studies are limited to electronic health register (EHR) data and have few measures of life factors. RNNs model a time series as a sequence of events that depends on the previous states according to a series of hidden states that pass forward information. However, this model suffers from the vanishing gradient problem which limits the model's ability to retain information from early in the time series. Consequently, RNNs have been extended to better capture long range dependencies with the addition of long short term memory (LSTM) [Sherstinsky, 2018]. This architecture involves the addition of gates that control which information from the features is stored into memory, which information is forgotten at each time step and which information is outputted. The ability of this model to reflect early-life exposures in its disease outcome predictions suggests it is appropriate for life course analysis.

CNNs originated with applications within computer vision and image processing tasks, such as image restoration, image annotation and segmentation [Jia et al., 2020; Radenovic et al., 2019; Zhang et al., 2017]. They are also suitable for many other contexts and have been used for stock prediction and mechanical fault detection [Huang et al., 2022; Selvin et al., 2017]. CNNs can be applied to MTS data as well as image data. For MTS data, the two dimensions of the model that relate to the x and y axis of an image are chosen as time and the number of features. CNNs are composed of two types of layers; convolutional layers and pooling layers [Zhao et al., 2017]. Convolutional layers apply a kernel across the data that is able to extract significant temporal and inter-variable patterns. Pooling layers act as a form of dimensionality reduction and reduce the size of the previous layer by applying a transformation such as taking the maximum or average across a selection of cells. The sequences of convolutional layers can detect small scale patterns in the data and then that allows following layers to find higher level relationships between the input variables. The final layer of a CNN is the fully connected layer that reduces the layer into the final output. There are many variations of CNNs that have been proposed for specific applications, including ResNet, Inception and AlexNet [He et al., 2015; Krizhevsky et al., 2012; Szegedy et al., 2017]. There are also models that are formed of a combination of CNNs and RNNs to leverage the benefits of each model architecture.

Life course trajectories can exhibit complex and non-linear temporal patterns, influenced by a range of factors across different life stages. Deep learning models, with their ability to capture intricate non-linear relationships, are better equipped to model and represent these complex patterns than traditional trajectory and sequence analysis methods. DL models can automatically learn meaningful features and representations from the raw data, eliminating the need for manual feature engineering, which can be challenging and time-consuming. Life course data can be high-dimensional, containing numerous variables and observations over time. Deep learning models can effectively handle large-scale, high-dimensional data, which may be challenging for traditional ML approaches. However, training DL models can be computationally demanding and require a significant amount of data. Proper hyperparameter tuning and architecture design are essential to achieve optimal performance in specific time series analysis tasks. Life course trajectories often involve long-term dependencies, where early-life events can have lasting effects on future health outcomes which DL models are designed to be able to capture. Deep learning models can be efficiently scaled using GPUs and distributed computing, making them suitable for large-scale life course studies with significant computational demands. With recent variations allowing for interpretability and causal inference [Chattopadhyay et al., 2019; Choi et al., 2016; Sha and Wang, 2017], it is now possible to pick out the most influential exposures in an individual's risk of developing an RTI. Therefore, the most significant events from childhood adverse experiences, environmental exposures and medical history can be identified.

1.5.3 Model Interpretability in Healthcare and Epidemiology

DL models are known for being black boxes, with a large number of parameters compared to other methods and an opacity as to how the input variables are used to predict the output. Model interpretability, also known

as explainability, is a term that describes the extent to which it is possible to understand the inner mechanisms of the model including how much influence each feature has on the estimated outcome and how changes to the input will change the output [Linardatos et al., 2020]. Recently, there have been many developments within this area to introduce explainability to DL models, such as the development of explainable AI (XAI) frameworks [Saluja et al., 2021]. Interpretability is especially important within healthcare applications because it is important for medical decisions based upon model predictions to be transparent and justifiable [Jin et al., 2022]. For instance, it is imperative to be able to point to the reasons behind healthcare decisions and which aspects of a patient's medical records, for example, flagged them as being high risk for a particular disease. Understanding the model's decision-making process enables there to be trust in the predictions and for the models to align with ethical guidelines required by healthcare practitioners. Knowledge of the most prominent features that contributed towards the model's classification enables clinicians to make clinical decisions and also for policymakers to design interventions based around these key features. XAI also allows researchers to learnt more about disease mechanisms and hypothesise about preventative measures. The incorporation of interpretability helps to validate the model and to identify any biases or faults that the model has learnt and eradicate them. Within life course epidemiology it is possible to use XAI methods to gain insight into the most influential risk factors and time periods to allow the DL models to work in a comparable way to a logistic regression model in risk analysis.

1.6 Integrating Life Course Epidemiology and Machine Learning for RTI Research

DL techniques overcome many of the limitations of traditional modelling methodologies used as part of life course analysis, such as trajectory and sequence models. In comparison with alternative ML methods used for MTS data, they appear to be more suitable and have a greater potential for life course research. DL models have the potential to enhance the understanding and analysis of complex life course trajectories and their impact on health and well-being. DL models are designed with Big Data in mind and therefore, unlike other frameworks, have the capacity to handle multifaceted longitudinal data with many individuals, features and time points. These large data sets represent a prohibitively large computational burden for many other modelling methodologies. It is worth noting that these datasets lead to a smaller but still considerable computational requirement for DL. Furthermore, DL models can learn complex and nonlinear patterns in data and are totally non-parameteric, meaning that the model is extremely flexible and can adapt to find these underlying features. This is particularly relevant for life course analysis as it is thought that the biological and social pathways and mechanisms that link exposures with health outcomes are multifaceted and difficult to define explicitly. Unlike many other modelling methodologies, the performance of DL models does not rely upon the data conforming to some assumption of form, for instance stationarity or having a Gaussian distribution. Given that relationships between the explanatory variables and the outcome are learnt directly from the data itself instead of specified by the user, this means that the model can learn the key features and relationships from the raw data, giving it a better capacity to uncover life course patterns present. Consequently, the model predictions are likely to be more accurate using the same data and more personalised. In addition, DL models are more suited to modelling long-term dependencies present in the data, allowing for a strong influence of childhood exposures on future health. This matches what has been found through experimentation in the literature and means these model representations will be more realistic. They are better adapted to modelling imbalanced data, which is often the case within health data as outcomes are often rare, helping to reduce modelling bias. Recently, there have been many adaptations and extensions of DL models and modelling techniques to allow for greater interpretability which is essential for models to be applied to assist with clinical decisions and intervention design. Furthermore, the ability to understand which exposures are most

influential in increasing the risk of disease outcomes, provides invaluable insight into which risk factors are most associated with specific health outcomes. Overall, DL models have an untapped potential for life course epidemiology and risk factor analysis. Together, these approaches provide a powerful toolkit for unravelling the complexity of longitudinal data, guiding public health strategies, and advancing personalised healthcare interventions to improve population health outcomes.

There have been a select few studies that have attempted to integrate ML methods with life course research [Billari et al., 2006; Hoffmann, 2023]. Clark et al. applied support vector machines to predict individual's self-reported health status [Clark et al., 2021]. Other studies have predominantly focused on creating a realistic life cohort using ML and mapped individuals' paths through significant life events, such as marriage and having children [Deb and Chakrabarty, 2020; Muratova et al., 2021]. Cheng et al. used an LSTM-based model to predict childhood obesity evaluating the longitudinal EHR data requirements required to obtain good predictive performance [Cheng et al., 2022]. Rieckmann et al. investigated whether causal links between life exposures and health outcomes can be unearthed using a simple neural network model [Rieckmann et al., 2022]. The method used puts forward a potential causal graph to explain the patterns within the data and fits it using a single layer neural network. The limitation of this work is that the model fits the data very poorly and temporal relationships are not considered. In addition, the choice of causal graph impacts the results greatly and means the flexible nature of ML models is overlooked. There have been no studies using DL in applications to life course analysis. Therefore, there is a lot of scope within this area to create a new framework with which life course analysis should be conducted using DL methods.

1.7 Research Questions & PhD Objectives

1.7.1 Summary of Key Themes

RTIs pose a substantial global health burden, and with an increasing number of novel viruses causing global scale pandemics, it is more important than ever to understand the complex pathways that link exposure events and a greater susceptibility to RTIs in some individuals. Previous studies of risk factors for RTIs within the life course analysis lens do not use models that are capable of learning the complex relationships that exist between the life course exposures and health outcomes. The non-flexible nature of these models both limits the ability to use the full information captured within the rich longitudinal data and obscures the true nature of the pathways linking an individual's unique life history and their future health. Moreover, risk factors are assumed to be conditionally independent in the traditional methods despite a large body of evidence demonstrating the existence of inter-variable interactions that magnify the impact on disease risk. Current life course approaches tend to aggregate the data and lose the ability to make individual level predictions and discard the temporal or multivariate aspects of the data. Both of these aspects mean that meaningful analysis of the significant risk factors requires further analysis. With an increasing number of healthcare providers in a growing number of countries collecting longitudinal data of this nature, it is vital that methods are created that can handle its large size and can extract its full potential. Recent advancements in XAI for DL models have made it possible to use these models to gain new insights into the hidden pathways linking exposures and outcomes. In addition, models that are able to learn the complex landscape of risk factors that brings about instances of poor health will also produce far more accurate estimates of risk ratios and therefore, act as a better guide to policymakers in designing intervention strategies.

The majority of existing studies are not based upon data sets that have sufficient statistical power to conduct life course research at a population-level. This is due to many studies utilising data from cohorts with a small number of participants, participants representing little heterogeneity, too few follow up points and only a small number of exposures represented. In order to conduct meaningful life course research, it is imperative

that the data includes the most influential socioeconomic, environmental, behavioural and lifestyle exposures. Furthermore, these variables must be recorded granularly over a long time frame to allow the model to be able to detect these intricate patterns. The nature of the DANLIFE cohort allows for an in-depth exploration of health trajectories and has the statistical power to draw important conclusions. As the number of similar data sets grows then the demand for a new suitable methodological framework for this data increases. In order to address these gaps, the life course perspective will be updated by applying innovative DL methods to a large cohort. With the aim of disentangling the complex web of what causes disease, this thesis explores the intricate interplay between life course patterns, early-life experiences, and subsequent RTI occurrences.

1.7.2 Research Questions & Objectives

There are three main research questions which will be addressed within this thesis. Underneath are the specific research objectives that will aid me in answering these questions.

1. **How well do different DL models perform in capturing and predicting life course patterns of health outcomes in simulated data that mimic the characteristics of the DANLIFE cohort?**
   Answering this question gives a strong indication of which DL models are most suitable for life course analysis
   - Simulate data that replicates the characteristics of the DANLIFE cohort, including life course patterns of health outcomes and risk factors, to create a representative benchmark dataset.
   - Embed different patterns between the input data and the output to mimic life course concepts based upon research into the pathways between exposures and outcomes.
   - Apply various DL models, to the simulated benchmark data to assess their performance in capturing life course patterns and predicting health outcomes.
   - Compare and analyse the model performance with multiple metrics.
2. **What are the significant life course patterns of exposures, including ACEs, socioeconomic and demographic factors, which influence the occurrence of RTIs and other health outcomes in the DANLIFE cohort?**
   - Utilise the best performing DL model to fit to the DANLIFE cohort data.
   - Identify influential exposures within the DANLIFE cohort using explainability methods.
   - Investigate the associations between different exposures and the occurrence of RTIs to better understand potential life course patterns present in the data.
3. **What are the most influential features and risk factors associated with the occurrence of RTIs and other health outcomes, and how do they differ for individuals with and without previous RTI diagnoses?**
   - Conduct feature importance analysis on the best-performing model to determine the most influential risk factors associated with the occurrence of RTIs and other health outcomes.
   - Investigate how the importance of risk factors varies between individuals with and without previous diagnoses by performing subgroup analysis based on the presence or absence of previous health conditions.
   - Calculate risk ratios and assess the magnitude of impacts of different risk factors on the likelihood of RTIs and other health conditions.

Addressing these questions advances the field of life course epidemiology by providing a new framework to measure life course associations whilst using the data to its full potential. This allows for researchers to gain a new understanding of how different exposures interact with each other over the life course and the significance of their timing in generating the risk of an RTI for an individual. The DL framework that will be developed in this thesis acts as a guide for those wishing to conduct life course research and this work sets out a way to assess the most influential risk factors.

1.7.3 Research Structure and Chapter Overview

Throughout this thesis, the primary aim is to unravel the complex interactions between early-life experiences and RTI outcomes. The integration of life course epidemiology principles and DL models provides comprehensive insights into the dynamics of RTIs, their risk factors and new opportunities for improving public health and well-being. Chapter 2 presents the innovative data simulation process, mimicking the DANLIFE cohort, and simulating life course concepts derived from the literature. Several DL models are tested on the simulated data to evaluate their performance and ability to identify underlying life course patterns. Chapter 3 investigates the Signature Method which is an alternative ML method designed to address data sparsity and dimensionality reduction challenges. The model is assessed on simulated data and compared to DL methods in order to understand its utility within life course analysis. Chapter 4 applies the best-performing DL model to the DANLIFE cohort data, investigating the link between ACEs, medical history and RTI occurrences. The Chapter dissects potential reasons for discrepancies in model performance, shedding light on the model's strengths and limitations. In Chapter 5, the notion of model explainability is explored in depth and several advanced approaches are applied to the fitted DL model's predictions. Feature importance analysis is performed to identify crucial risk factors influencing RTI occurrence, particularly in the context of ACEs during childhood. Conditional feature importance analysis is explored for different subgroups. Risk ratios of specific risk factors are estimated to provide valuable insights into health disparities and disease risk, empowering targeted interventions. The concluding Chapter summarises the thesis's key contributions, emphasising the integration of life course epidemiology and DL to advance risk prediction models and personalised healthcare strategies. The implications of the research findings for public health interventions are discussed, and future research directions are suggested, encouraging interdisciplinary collaboration and methodological advancements.

Chapter 2

Data Simulation and Deep Learning Model Evaluation

2.1 Introduction

There is a paucity of studies that investigate the application of deep learning (DL) models within life course analysis tasks. There are numerous varieties of DL models, each possessing their own strengths and weaknesses that vary according to the application. Therefore, in order to use DL models to conduct life course research, there is a need for an assessment of multiple DL approaches to determine which models have the capacity to detect intricate life course patterns. In order to achieve this goal, an experiment has been designed to examine and compare the suitability of several DL models for analysing longitudinal life course data using simulated data designed to mimic fundamental life course concepts. Data simulation is a key component of this analysis as it offers the opportunity to control and manipulate the relationships that exist between the exposures and the outcome. By doing so, it is possible to assess the robustness and performance of DL models under controlled conditions. This analysis acts as the foundation of the thesis, providing an initial understanding of the capabilities of different DL models in multiple realistic scenarios. The performance of the DL models across the range of simulated life course patterns will also function as a guide for deciphering possible life course patterns present in the DANLIFE cohort data, furthering the understanding of how disease risk accumulates over a lifetime.

Many of the DL models within this analysis were initially designed for tasks within the fields of natural language processing (NLP) and computer vision, and have had to be adjusted for applications to multivariate time series (MTS) data. There are several key differences between MTS data and the longitudinal cohort data used within life course analysis. Longitudinal cohort data compiled from records will contain multiple features that each correspond to the occurrence of a distinct type of exposure event, for example hospitalisation with a certain medical condition or a child being placed in foster care. Within the data, these features will often be formulated as a vector containing binary or integer values with their length corresponding to the number of distinct time periods. The values of the vector each represent either whether an exposure event occurred or the total number of exposure events that occurred within the associated time window. In the case of the DANLIFE cohort, the registers are updated yearly and so the time step is also a year. A yearly time step is adapted in this analysis in order for the data to be comparable and improve consistency. This differs from standard MTS data as it usually contains continuous values at a greater number of time steps. DL models tend to perform better when modelling continuous values as the categorisation or binary nature of exposure variables

introduces further noise and uncertainty. An additional difference is that the cohort data is often very sparse since most medical conditions and adverse childhood experiences (ACEs) only happen very rarely within the population and perhaps once or twice per lifetime. As a consequence of these data discrepancies, any studies that have compared these DL models on MTS data are unlikely to be informative in understanding how the models perform on life course data. As a result, the analysis contained within this Chapter acts as a basis the real-world application of DL models in healthcare, providing insights into their potential utility for life course analysis and therefore, public health strategies.

2.1.1 Life Course Patterns

The main crux of the research contained in this thesis is the exploration of the complex pathways that link the occurrence of multiple exposure events, taking place across a person's lifespan, with future health outcomes. Exposure events encompass a wide range of factors, including socioeconomic status, early life experiences, health behaviours, and various risk factors. From a life course perspective, it is crucial to consider the timings and interactions between exposure events to accurately assess an individual's susceptibility to disease. These interactions can amplify or mitigate the impact of specific risk factors, so it is important to examine the interplay between several types of exposures. Within the life course approach, an individual's health is thought of the outcome of the accumulation of social, environmental, physical, and psychological exposures within their life history [Ben-Shlomo and Kuh, 2002; H. et al., 2003; Jones et al., 2019; Kalmakis and Chandler, 2014; Kuh et al., 2003]. The impact of a given exposure event with be heterogeneously distributed across the population, according to each individual's life history. Therefore, an accurate understanding of the scale of the risk posed by each exposure at the population level can only be obtained by accounting for each individual's unique life history and aggregating the predictions. In this way it is possible to improve the effectiveness of public health interventions. The specific patterns of exposure events that interact to increase the risk of future poor health outcomes are referred to as life course patterns. The life course patterns arise from the underlying societal and biological processes that generate disease which can be triggered by certain exposure events [Jones et al., 2019]. Knowledge of the life course patterns that exist which increase the risk of RTIs can lead to more accurate predictions of individual and population health trajectories as well as an improved awareness of how risk accumulates to give rise to disease.

In Chapter 1 Section 1.3.1, the four key life course concepts that form the pillars of life course theory are described. In order to assess the aptitude of multiple DL models for identifying life course patterns, multiple data sets will be simulated that are each designed to imitate a particular pattern based upon a life course concept. There are four broad categorisations that will be explored in this analysis; named periods, repeats, timing, and order. The periods life course pattern reflects the importance of key developmental time periods within the life course and the lifelong influence of early life experiences on an individual's health in adulthood. The simulated data sets that fall within this group are formulated to emulate critical and sensitive periods, during which the negative effect of certain exposures on future health is magnified. The repeats life course patterns are designed to resemble characteristics of the accumulation-of-risk model in which multiple exposures happening consecutively amplifies the negative health consequences. The timing life course pattern symbolises the importance of the relative timing of different exposures in generating increased risk of disease outcomes. This reflects how the size of the time frame in which the exposures occur can compound their negative effects on health. Finally, the order life course pattern encompasses the significance of the temporal ordering of two distinct exposure events. Each of these life course patterns characterises a life course concept, allowing for a detailed exploration of the strengths and weaknesses of each DL model according to the underlying relationships within the data.

2.1.2 Deep Learning Models

There are several reasons why DL models are likely to be well-suited to big longitudinal life course cohorts, such as the DANLIFE data set. Firstly, the flexible nature of DL methods means that inter-variable relationships do not need to be specified a-priori as DL models have the capacity to learn these connections directly from the data. Furthermore, unlike other traditional MTS approaches, there are fewer constraints on the format of the data necessary to apply the model. With recent variations allowing for interpretability and causal inference [Chattopadhyay et al., 2019; Choi et al., 2016; Sha and Wang, 2017], it is now possible to pick out the most influential exposures in an individual's risk of developing a respiratory tract infection (RTI). Therefore, the most significant time periods and exposure events can be identified. In this analysis, four DL models will be investigated and compared to a baseline logistic regression (LR) model. The LR model will act as a baseline model from which to examine any improvement in performance against. The choice was made not to compare model performance to a trajectory or sequence analysis model as those models generally are used to create clusters of individuals in the population and are not designed to act as predictive models. The questions commonly answered using traditional life course modelling methodologies focus on identifying the characteristics of the subgroups that may be related to any observed difference in the prevalence of the outcome. In contrast, regression models are frequently used within risk analysis, particularly to generate individual-level risk predictions. Therefore, it was decided that the performance of the DL models would be compared to an LR model. Within this section, the architectures of each of the four DL models (LSTMAttention, MLSTM-FCN, ResNet and InceptionTime) and the LR model are described.

The experimental design frames the learning task as a time series classification (TSC) problem. This means that the goal is to organise the cohort into two classes according to whether each individual is hospitalised with an RTI-related illness within the follow up period. These predictions shall be made using the longitudinal data that records the exposures each individual experienced during childhood. The primary reason behind setting up the problem as a binary TSC problem as opposed to multi-class TSC or time series regression is that it is a good first step in assessing the ability of models to conduct the analysis. Alternative options for the outcome are to use counts of hospitalisations within the follow up or to devise a measure of severity from the data. Using counts instead of a binary indicator function for hospitalisation adds little benefit to the analysis and ultimately, as a hospitalisation already implies a severe level of illness, the use of counts does not provide clinicians using the risk prediction model with any useful additional information. Designing an outcome that is representative of severity requires intimate knowledge of data collection methods and the Danish healthcare system. For instance, whilst the medical events within the DANLIFE data are recorded as being either emergency, inpatient or outpatient care, this categorisation has different implications within different regions of the country. Regions without convenient access to GPs often experience individuals presenting to emergency care when they are less acutely ill than other regions. Similarly, the number of days hospitalised with RTI-related illness during the follow up is not a useful metric of severity due to the way in patients are transferred between departments within the hospital. For instance, when an individual is transferred they are recorded in the system as a separate admission and it is common for individuals within outpatient care to have a case opened to investigate long term symptoms which is recorded as an extended hospital stay but in reality does not equate to days spent in hospital. Therefore, framing this problem as a binary TSC is the best first approach to assess the use of DL methods for life course analysis for a simplified setup with room for further analysis into the best formulation of these problems.

Standardised notation for the task is introduced to facilitate easier comparison of the assorted modelling frameworks. The matrix containing the MTS data is denoted by $X \in \mathbb{R}^{N \times M \times T}$, where N is the number of individuals, M is the number of variables and T is the number of time points at which the variables are measured. The M variables include exposures such as air pollution levels, socioeconomic status, and medical

events. Similarly, the output variables are given by $\mathbf{y} \in \mathbb{R}^N$ denoting the binary vector that contains the disease status of each individual. Altogether, the TSC problem can be written as

$$\mathbf{y} = f(X) + \varepsilon$$

where ε is the noise in the data and $f(\cdot)$ is the underlying function that generates the outcome from the MTS data that we seek to learn. Adopting terminology from DL, the matrix X will be referred to as the model input and the outcome $\mathbf{y}$ as the model output.

Logistic Regression

The LR model acts as a baseline model with which to compare the DL architectures. As outlined in Chapter 1, the LR model can be used within risk analysis, it has interpretable model coefficients which can be used to quantify the significance of each variable and identify risk factors. This model has no explicit temporal dependence and only allows for linear relationships to exist between the features. In this formulation of the model, the model is provided with multiple variables which are each the value of a given feature at a single time point. Therefore, the value of a feature at different time points are treated independently. There are alternative ways in which the LR model could be designed to better reflect temporal relationships, including lagging the variables and conducting piecewise regression. In piecewise regression the time frame is split into multiple subsections and the model receives variables that are calculated as the feature average across each shorter time period. This approach is essentially a data aggregation method, and therefore, has the same limitations described in Chapter 1, namely that there is a loss of some of the rich information contained within the original longitudinal data. Lagging adds new variables into the analysis that have been generated by moving the current features forwards or backwards in time by a given number of steps. Within the model formulation, it does not make sense to introduce further variables that would add additional noise. Therefore, the decision was made to not use these additional methods with the LR model framework.

The data is prepared for the LR model by flattening the input matrix so that it becomes $\tilde{X} \in \mathbb{R}^{N \times MT}$ where N is the number of individuals in the population, M is the number of variables and T is the number of time steps. As a result, the value of the same variable at two different time points are treated as two independent variables. The LR model equation is given by [Seber and Lee, 2003]

$$P(y_i = 1) = \frac{\exp(WX_i + \beta)}{1 + \exp(WX_i + \beta)} \quad \text{for} \quad i \in (0, N) \tag{2.1}$$

with $W \in \mathbb{R}^{MT}$ and $\beta \in \mathbb{R}$ both denoting the trainable weights of the model. It is possible to introduce interaction terms between multiple covariates, this was examined in additional experiments, and it was show that adding second order interactions increased computational burden and worsened model performance. Therefore, interaction terms have been omitted from the model.

Long Short-Term Memory & Attention

Long short-term memory (LSTM) models with attention (LSTMAttention), otherwise known as transformer models, are a popular relative of the recurrent neural network (RNN) model. RNNs originated in applications to natural language processing (NLP) tasks, for example, predicting the next word in an unfinished sentence or translating text from one language to another. They are designed specifically for sequential data and can learn patterns related to temporal dependencies between variables. RNNs have found application in various fields, including speech recognition, prediction of stock returns and traffic forecasting [Graves, 2012; Li et al., 2017; Rather et al., 2015]. They have also been applied within healthcare settings; to diagnosis prediction, forecasting influenza trends and early detection of health failure onset [Choi et al., 2017; Lipton et al., 2016;

Liu et al., 2018; Ma et al., 2017]. The following descriptions of the structures of RNNs and LSTMs are based upon the explanation given in Colah's blog [Colah's blog, 2015].

RNNs are composed of a series of cells, each of which is fed an input X_i at time $i \in \mathbb{N}$ (Figure 2.1a). Within NLP settings, the inputs take the form of words with the order being their place in a sentence. In this application, the input variables are in the format of matrices $X_i \in \mathbb{R}^{N \times M}$ which contains the values of the M variables for each member in the population of size N at a given time point $i \in [0, T]$. The beauty of the RNN architecture is the introduction of a hidden state which is passed between cells and contains information about previous time steps, enabling the model to capture sequential information and learn from past observations. At each time step, a cell is provided with and processes both the value of the input X_i and hidden state h_{i-1} at that time step and it emits an output $y_i \in \mathbb{R}^N$ and a new hidden state h_i. The RNN equations are as follows [Colah's blog, 2015]:

$$h_t = \tanh(W_x X_t + W_h h_{t-1} + b) \tag{2.2}$$

$$\hat{y}_t = \text{softmax}(W_y h_t) \tag{2.3}$$

where W_x, W_h and W_y are weight matrices, $b \in \mathbb{R}$ is the intercept and $\hat{y}_t$ is the model estimate of output y_t. The softmax function, given by softmax$(\cdot)$, is a function that is used to transform a real-valued input vector and transforms each element into a value in the interval $[0, 1]$ in such a way that they sum to one [Jang et al., 2017]. The formula for the softmax function is $\text{softmax}(\mathbf{x})_i = \frac{\exp^{x_i}}{\sum_{j=1}^{n} \exp^{x_j}}$ where $\mathbf{x}$ is the input vector that contains n elements corresponding to the number of classes. RNNs are trained using the backpropagation through time (BPTT) algorithm [Rumelhart et al., 1986].

RNNs suffer from a number of problems, including the vanishing/exploding gradient problem which describes how the gradients can become very small or very large due to the multiplicative nature of the weight matrices resulting in poor learning [Hochreiter, 1998]. As a result, RNNs struggle to capture long-range dependencies in the sequence and in practice the hidden state only holds information about the three proceeding states. Multiple variations and adaptions of the RNN model structure have been developed to mitigate these issues. A significant model variant which addresses these limitations is the introduction of LSTM. The LSTM architecture shares many features with RNNs but introduces gates that control the passage of information between the cells (Figure 2.2) [Hochreiter and Schmidhuber, 1997]. The LSTM architecture introduces an additional state called the cell state denoted by C_t, which performs a similar task to the hidden state, and it encodes memory that is passed between cells (Figure 2.1b) [Colah's blog, 2015]. The forget gate, f, selects information that is no longer required to be retained within the memory. The input gate, i, is responsible for selecting which elements to extract from the input data and hidden state to add to the current cell state. The final gate is the output gate, o, and it controls the new hidden state h_t which is equivalent to the response variable y_t. The equations are given by [Colah's blog, 2015]

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.5}$$

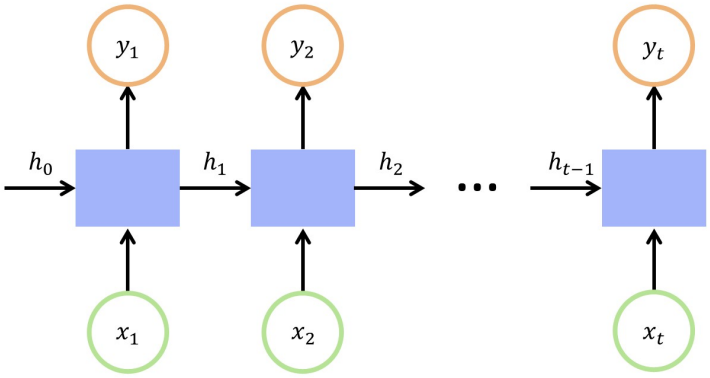
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.7}$$

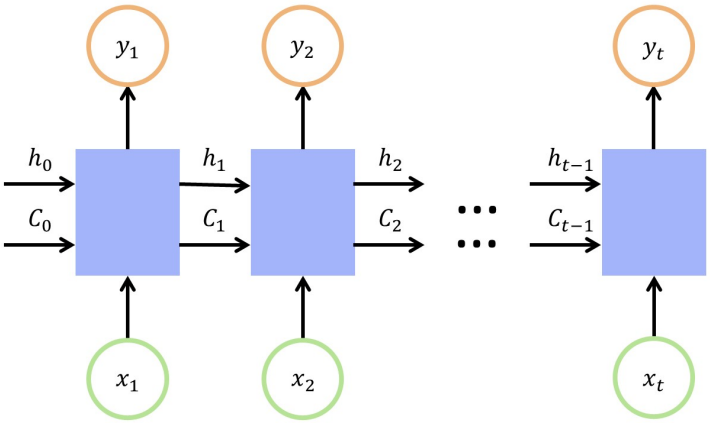
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.8}$$

$$h_t = o_t * \tanh(C_t) \tag{2.9}$$

Several architectural variations of the LSTM architecture have been developed, such as the bidirectional LSTM which allows future inputs to influence past outputs and peephole LSTMs where all cells are able to observe



(a) **RNN architecture.**



(b) **LSTM architecture.**

Figure 2.1: **RNN and LSTM model architectures.** The structure of unrolled RNN (top) and LSTM (bottom) models based upon figures in Colah's Blog [Colah's blog, 2015]. For time step i, the input variable is denoted by x_i, outputs are denoted by y_i, the hidden state by h_i and the cell state by C_i.

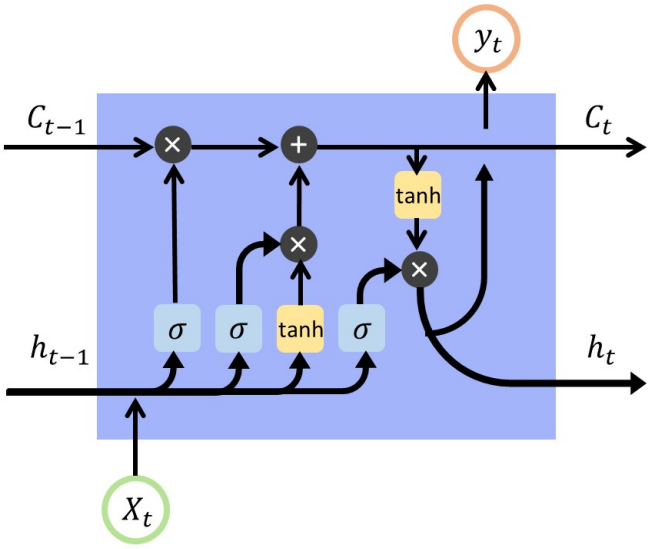


Figure 2.2: **LSTM cell structure.** The inner workings of an LSTM cell based upon figures in Colah's Blog [Colah's blog, 2015]. At each time point t, the input, x_t, and the previous hidden state h_{t-1} are concatenated and passed through sigmoid (σ) and tanh functions. The previous cell state, C_{t-1} is updated to obtain the new cell state, C_t, and the new hidden state h_t which is equal to output y_t.

the cell state [Graves, 2012]. However, LSTMs may not fully capture extremely long-term dependencies and they are not able to give higher importance to more significant developmental periods. This property is particularly important for life course applications when the early years of life are thought to have the greatest impact on health in adulthood. Another adaptation of the LSTM architecture is the addition of attention mechanisms. Attention mechanisms assign weights to the features and time steps to allow the model to focus on the most influential sections of the input sequence [Vaswani et al., 2017].

The attention mechanism allows the model to dynamically focus on different time points within the same time series. It achieves this by computing a set of weights that determine which time steps and features within the input sequence require the greatest level of priority [Vaswani et al., 2017]. The calculation is performed each time step, making it dynamic and able to reflect the different temporal dependencies thought to be important within life course epidemiology. The network uses an alignment model to assess the concordance between the input and output sequences to produce a context vector that represents the model weighting. The context vector, c_i, takes the form

$$c_i = \sum_{j=1}^{T} \alpha_{ij} \cdot X_{ij}$$

where α_{ij} is the amount of attention that the ith input should pay to the jth input and X_{ij} is the MTS for the jth input. RNN models are sometimes formulated within an encoder-decoder framework. The encoder-decoder structure is formed of two RNNs, one RNN encodes the input sequence step by step and passes the final hidden state to another RNN which decodes the information to generate a new output sequence. For example, in translation tasks the input sentence in one language is fed into the encoder to generate a hidden representation and then the decoder outputs the sequence of words that form the translated sentence. The benefit of the encoder-decoder structure within forecasting is that the model can learn patterns from the entire sequence before producing the output.

There are multiple configurations of the attention mechanism. Luong et al. introduced the concepts of global and local attention, with global attention considering the entire input sequence when computing weights and local attention only focusing on a small window of time steps around the current time step [Luong et al., 2015].

One common form of attention is known as self-attention or scaled dot-product attention [Cheng et al., 2016]. This attention mechanism is particularly useful when certain time points contribute more significantly to the prediction or when scaling is needed to control the magnitude of the attention scores. For a given time point, the attention scores are computed by taking the dot product between the current time point and all other time points. The attention scores, α_{ij} are computed as

$$\alpha_{ij} = \text{softmax}\left(\frac{X_{ij} \cdot X_{ij}^{\top}}{\sqrt{M}}\right)$$

where X_{ij} is the MTS for the jth input and M is the number of features. The softmax function is used to normalise the results between 0 and 1 and it is given by

$$\text{softmax}(x) = \frac{\exp(x)}{\sum_{k=1}^{T_x} \exp(x)}$$

Another construction of the attention mechanism is additive attention in which the alignment model takes the form of a trainable feedforward neural network with a single hidden layer [Bahdanau et al., 2014]. In this case, the attention scores are calculated by applying the alignment model to the concatenation of the current time point and all other time points. The resulting scores are then passed through a softmax function to obtain the final attention weights. In scenarios where the contribution of each time point to the prediction is non-linear or context-dependent, additive attention allows the model to adapt and learn these patterns. A final attention construction of note is multi-head attention that simultaneously applies the attention mechanism in parallel across multiple heads that each capture different relationships within the data. Each head has its set of learnable parameters, allowing it to specialise in certain types of relationships. The outputs from different heads are usually combined through concatenation to form the final attention scores.

The LSTMAttention model used in this analysis is based upon the Transformer model first proposed in "Attention is all you need" by Vaswani et al. [Vaswani et al., 2017]. The Transformer model architecture was originally designed for NLP tasks and adapted for TSC applications [Zerveas et al., 2020]. The time series adaptation of the model is composed only of the encoder part of the original model, which is formed of six identical blocks, each of which comprises the same two sub-layers. The first sub-layer is the multi-head self-attention layer and the second is a feed-forward network. Using self-attention layers enables faster computation and improves the model's ability to learn long-range temporal dependencies [Zerveas et al., 2020]. The feed-forward neural network takes the form of a pointwise, fully connected layer and takes the form

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \tag{2.10}$$

where x is the input data and W_i and b_i are the weights and biases of the model. ReLU stands for the rectified linear unit function that takes the form $ReLU(x) = max(0, x)$. Each of the six blocks within the model has its own trainable weights and biases within the feed-forward network. Around each of the sub-layers is a residual connection that allows that layer to be skipped if its inclusion in the model does not improve training. This attribute of the model allows for the vanishing gradient problem to be avoided, of which a more in-depth description is provided in the description of ResNet later in this section. Following each sub-layer there is a batch normalisation layer which improves training by normalising the batches to simply the learning procedure for the model. Following the six model blocks, there is a linear layer then a softmax layer to output the class predictions, $\hat{y}$:

$$\hat{y} = \text{softmax}(W_o\bar{z} + b_o) \tag{2.11}$$

where $z_t \in \mathbb{R}^d$ are the final representation vectors and $W_o\mathbb{R}^{(d \cdot w)}$ and $b_o \in \mathbb{R}$ are the weights and biases of the linear output layer. Here d is the dimension of the vector space into which the input data is projected prior to inputting it into the model. The linear projection takes the form

$$u_t = W_p x_t + b_p \tag{2.12}$$

where $W_p \in \mathbb{R}^{d \times m}$, $b_p \in \mathbb{R}^d$ are learnable parameters and $u_t \in \mathbb{R}^d$, $t = 0, \ldots, w$ are the model input vectors. Then positional encodings, $W_{pos} \in \mathbb{R}^{w \times d}$, are added to the projected vectors $U \in \mathbb{R}^{w \times d} = [u_1, \ldots, u_w]$: $U' = U + W_{pos}$. This process helps to retain temporal information.

Multivariate Long Short Term Memory Fully Convolutional Networks

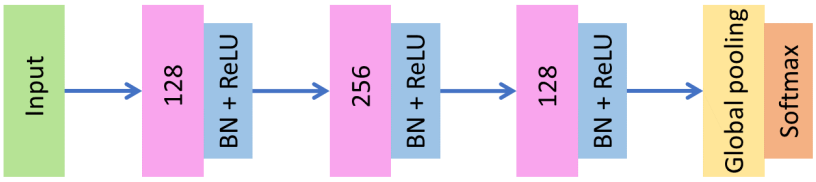
Multivariate long short term memory fully convolutional networks (MLSTM-FCNs) are a type of neural network architecture that combine fully convolutional networks (FCNs) and LSTMs to extract both temporal and inter-variable structural information from the input data [Karim et al., 2017]. This hybrid architecture is designed for processing sequential data, such as time series or sequential sensor data, and has found applications in various domains, including speech emotion recognition, earthquake detection, and in healthcare for arrhythmia detection [Bai and Tahmasebi, 2021; Park et al., 2023; Zhao Yu Zheng Zixing Zhang Haishuai Wang et al., 2018]. As described in Section 2.1.2, LSTMs were originally designed for applications within NLP and have been adapted for MTS data. The LSTM model can learn complex long-range temporal patterns within sequential data. FCNs were designed for computer vision tasks and exhibit excellent performance at spatial feature extraction [Wang et al., 2016b]. The LSTM-FCN architecture feeds the input data through both LSTM and FCN models in parallel to better leverage temporal and inter-variable information in the data [Colah's blog, 2015].

FCNs are a family of model architectures that encompasses convolutional neural networks (CNNs) which are constructed in the exact same way except CNNs do not include fully connected layers [Wang et al., 2016b]. The model is composed of multiple convolutional layers which slide a one-dimensional kernel across the surface of the data with some stride length that computes some function of the data contained within its window according to its trainable weights. The data is padded either side with zeros so that the dimension of the input as the same as the output. This model structure is formed of three convolutional layers with 128, 256 and 128 filters and kernel sizes of 8, 5 and 3, respectively (Figure 2.3a). Convolutional filters slide across the input sequence to capture local patterns and create feature maps. Between each layer is a layer of batch normalisation and a ReLU activation function. Batch normalisation normalises the batches and is used to improve computation speed and improve generalisation. The activation function is used to introduce nonlinearity into the model and produces the output of the layer. The entire convolution block takes the form

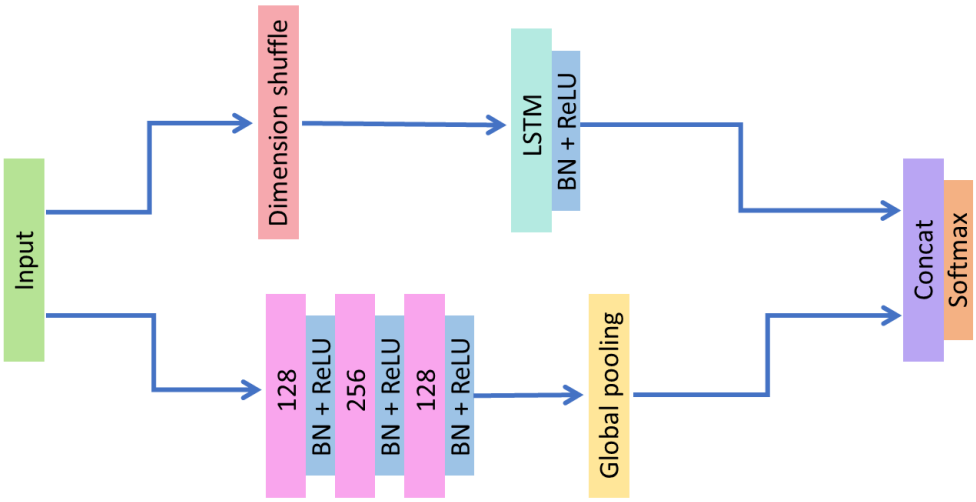
$$z = \text{ReLU}(BN(\mathbf{W} \otimes \mathbf{x} + \mathbf{b})) \tag{2.13}$$

where x is the layer input, y is the layer output and W and b are the weights and biases of the convolution. Then the output of the FCN is put through a global average pooling layer. The global average pooling layer is similar to a convolutional layer where a window is moved across the data but at each point the maximum of the elements in the window is taken to make up the output.

The LSTM-FCN network applies both LSTM and FCN models to the input data in parallel and then concatenates the result (Figure 2.3b). The equations that comprise the LSTM block have the same construction as given in the Section 2.1.2. The only difference is that the input variables go through a dimension shuffle layer before being passed to the LSTM block. The aim of this is to pass the LSTM an MTS with a single time



(a) **FCN model structure.** Based upon a figure in [Wang et al., 2016b].



(b) **LSTM-FCN model structure.** Based upon the figure in [Karim et al., 2017].

Figure 2.3: **FCN and LSTM-FCN model architectures.** The structure of the FCN (top) and LSTM-FCN (bottom) models.

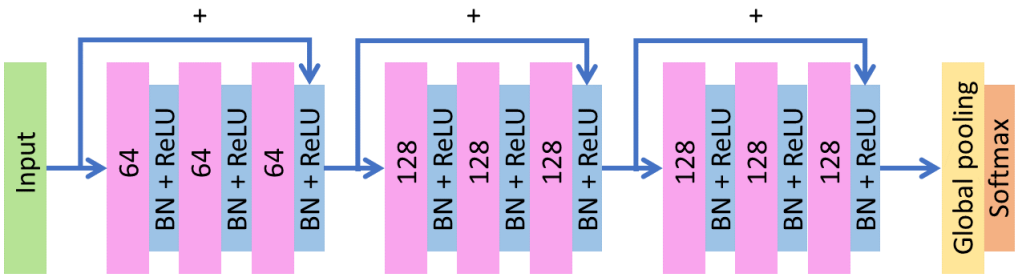


Figure 2.4: **ResNet model structure.** Based upon figures shown in [Wang et al., 2016b].

step by transposing the temporal dimension of the input data. This greatly improves the training speed. The output from both models is concatenated then it is passed through a softmax layer which gives the classification labels. LSTM-FCNs can be computationally intensive, especially when dealing with long sequences or high-resolution data. Proper hyperparameter tuning and regularisation techniques are essential to prevent overfitting.

Residual Network

Residual Network (ResNet) is a deep neural network architecture that has had a significant impact on the fields of computer vision and DL. It was introduced by He et al. in the paper “Deep Residual Learning for Image Recognition" in 2015 and has since become a fundamental building block for many state-of-the-art CNNs [He et al., 2015]. ResNet architectures have achieved state-of-the-art performance on various computer vision tasks, including image classification in traffic monitoring, medical image processing and facial recognition [Jung et al., 2017; Lu et al., 2018; Xu et al., 2023]. The central idea behind ResNet is the use of residual connections, otherwise known as skip or shortcut connections. These connections allow the network to skip over layers if the information learnt in that block does not improve model performance. Including these connections allows deep architectures to avoid the vanishing gradient phenomenon as unnecessary layers are bypassed [Szegedy et al., 2017]. Residual connections take the output from the previous step and pass over one or more layers to add this output to that of a later cell [He et al., 2015]. As a result, if the layers within the connection do not improve performance, then the gradient from the previous output can be better propagated through the model. Otherwise, the residual block will have a bigger impact than the previous output and will dominate the sum. Therefore, the residual block can be summarised by the following equations:

$$z_1 = \text{ReLU}(BN(W_1 \otimes x + b_1)) \tag{2.14}$$

$$z_2 = \text{ReLU}(BN(W_2 \otimes z_1 + b_2)) \tag{2.15}$$

$$z_3 = \text{ReLU}(BN(W_3 \otimes z_2 + b_3)) \tag{2.16}$$

$$y = \text{ReLU}(z_3 + x) \tag{2.17}$$

where z_i denote the separate convolutional layers that make up a block, x is the output from the previous block, y is the output from the current block and W_i and b_i denote the model weights and biases.

Whilst there are many variants of the ResNet model architecture, in this analysis the formulation chosen is the one described by Wang et al. [Wang et al., 2016b] (Figure 2.4). In comparison with other ResNet architectures, this configuration has fewer layers and trainable parameters which limits the model’s ability to learn especially complex parameters. However, the choice of the basic ResNet formulation allows these results to showcase the model’s capabilities at a lower bound for its performance. ResNet is composed of three model blocks, each

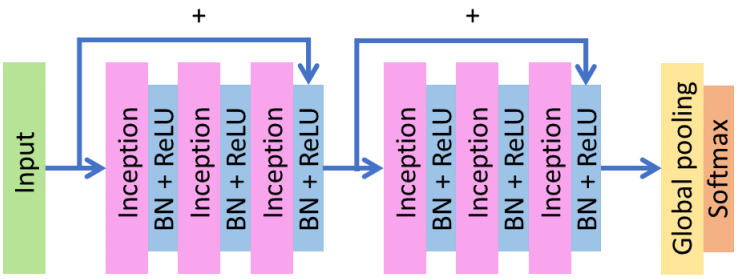


Figure 2.5: **InceptionTime model architecture** Based upon a diagram in [Fawaz et al., 2020].

of which is spanned by a residual connection. Within each block there are three convolutional layers each separated by a layer of batch normalisation and a ReLU activation function. In the first of the three blocks, each convolutional layer has 64 one-dimensional filters with kernel length of 7, 5 and 3 for the first, second and third layers, respectively. Using different filter lengths enables temporal information to be extracted at differing degrees of granularity. The type of convolutions used in the ResNet architecture can be described as bottleneck layers which use one-dimensional filters to reduce computational complexity whilst retaining the ability to learn rich features. The remaining two blocks are identical to the first except that each convolutional layer has 128 filters instead of 64. After the three blocks, there is a global average pooling layer followed by softmax function to obtain the final classification labels. The use of the global average pooling layer means the model has fewer trainable parameters and is less likely to overfit the data and also less computationally expensive.

InceptionTime

InceptionTime is a DL architecture designed specifically for TSC tasks by He et al. [He et al., 2015]. It is an adaptation of the Inception model which was designed for image classification [Szegedy et al., 2014]. InceptionTime has gained popularity for achieving excellent performance when applied to tasks that involve diverse temporal patterns in time series. The InceptionTime model and its adaptations have achieved very good performance, and this model has significantly outperformed other DL architecture based upon CNNs in several applications including detecting cardiovascular abnormalities and classifying ECGs and classifying brain activity [Borra et al., 2020; Crocker and Costall, 2021; Kastrati et al., 2021]. Here, an overview of the model structure is provided, for a more in-depth explanation of the model construction please refer to the original InceptionTime paper upon which this description is based [Fawaz et al., 2020].

InceptionTime is formed of multiple parallel convolutions designed to capture both fine-grained and long-range temporal dependencies from the MTS data [Fawaz et al., 2020]. The structure is composed of six inception modules that are arranged into two identical blocks (Figure 2.5). Each block has a residual connection across the top to improve gradient flow during training and minimise overfitting [He et al., 2015]. After the two blocks, there is a global average pooling layer followed by a fully connected softmax layer to predict class labels. Global average pooling reduces the spatial dimensions of the feature maps to produce a single vector for classification. Between each of the modules is a batch normalisation and ReLU activation function. When working with InceptionTime, it is important that the hyper-parameters are fine-tuned and data is carefully prepared with appropriate augmentations to improve model performance. Data augmentation techniques, such as time warping, scaling, and rotation, can be applied to increase the robustness of InceptionTime to variations in the input MTS data.

The inception modules are composed of multiple parallel convolutions with different filter lengths that are concatenated to form the output (Figure 2.6). Firstly, the data passes through a bottleneck layer which reduces

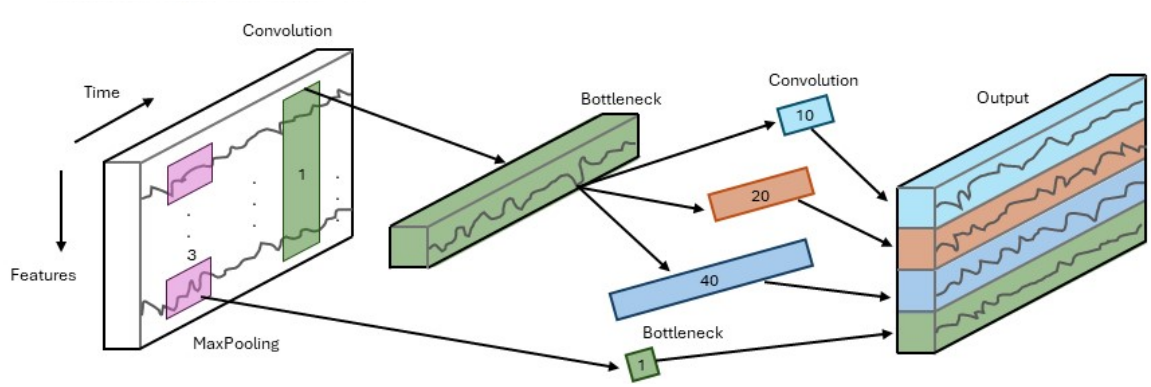


Figure 2.6: **Inception module structure.** Image based upon [Fawaz et al., 2020]. The input multivariate time series is processed to form the output multivariate time series that reflects temporal patterns in the data.

the dimensionality of the data and, as a result, the computational cost. The other benefit of the bottleneck layer is that is reduces the number of model parameters, decreasing the likelihood of the model overfitting the data. The bottleneck layer is a convolutional layer with kernel size 1 and filter length 32. Following the bottleneck layer there are three parallel convolutional layers with kernel sizes 10, 20 and 40 with filter size 32. The convolutional layers are designed to be able to pick up on short-, mid- and long- term temporal dependencies in the data, meaning that InceptionTime is appropriate for sparse temporal data [Szegedy et al., 2014]. Depthwise separable convolutions are used which factorise the standard convolution into a depthwise convolution followed by a pointwise convolution, reducing the computational cost while preserving model performance. There is a further additional layer that runs in parallel to the bottleneck and convolutions that consists of a max pooling layer with window size 4 which takes the maximum value within each window then the output is passed through a bottleneck layer. The results from each convolutional layer as well as this additional layer are concatenated to produce the output of the inception module. Maxpooling helps to make the model less sensitive to perturbations in data [Fawaz et al., 2020].

2.1.3 Chapter Outline

The main goal of this Chapter is to assess the suitability of various DL architectures for applications within life course epidemiology. Specifically, the Chapter aims to address the first research question of this thesis: *How well do different DL models perform in capturing and predicting life course patterns of health outcomes in simulated data that mimic the characteristics of the DANLIFE cohort?* The Chapter has two main objectives: first, to simulate data that encapsulates the key characteristics of the DANLIFE cohort, and second, to rigorously test the performance of various DL models on this simulated data.

In the pursuit of these objectives, a simulation study is undertaken to identify which DL architectures are best suited for life course analysis by examining their ability to learn simulated patterns of exposure events. Four state-of-the-art DL models adapted for MTS data are compared against a regression model. Plausible epidemiological mechanisms or life course patterns are generated to mirror the hypothesised patterns relevant to life course theory. Multiple simulated data sets, each containing patterns representing archetypal life course concepts, are used for an in-depth exploration of scenarios in which the models perform best. This use of synthetic data allows for a robust, controlled, and systematic comparison across key exposure patterns foundational to life course theory. The performance of the models across various patterns is examined, with an understanding that good performance is necessary but not sufficient for suitability in life course research. The best performing models demonstrate the capability to identify underlying patterns in the data, a crucial next step assessing the extent to which DL model are capable of conducting a detailed examination of the

pathways linking exposure events and health outcomes. For risk prediction models related to RTIs, it is vital to ensure that DL models applied to longitudinal data can distinguish true underlying pathways from artifacts or spurious patterns.

This research is of particular importance as life course data differs from MTS data, being less granular with multiple binary features representing exposure events. Moreover, the medical data is very sparse both in terms of the number of individuals who experience specific exposures and the frequency of these exposures in time. As a result, there are many challenges when applying traditional modelling methodology to this data and this has often meant that much of the richness in the data has been discarded in previous studies. Consequently, this work is the first step in developing a framework with which to analyse this data using DL approaches. This acts as a building block upon which a comprehensive evaluation of the risk factor interactions of RTIs can be performed, leading to a model able to give high accuracy individual-level predictions of RTI risk. Given that the processes linking different exposure events and individual biology will likely be very complex and involve both temporal and inter-variable dependencies, it is important to find which models are capable of learning these patterns. As studies of the patterns that exist in data are very limited, it is not known exactly how these processes look and how the multiple processes will interact with each other.

This Chapter is organised into distinct sections, beginning with a detailed exploration of the data simulation process and the methodologies used to create synthetic life course data. It then delves into the training procedures for DL models, showcases the results and offers a comprehensive discussion of findings and their implications for health strategies related to RTI control. Ultimately, the Chapter acts as a bridge between data simulation and real world applications, demonstrating the potential of DL techniques to advance life course analysis.

2.2 Methodology for Data Simulation

This section describes the process used to create simulated data that embodies the key characteristics of longitudinal life course data. A distinctive feature of this study lies in the incorporation of patterns within the simulated cohort that align with multiple life course concepts. This approach aims to provide insights into the capacity of DL models to learn these patterns, thereby assessing their potential for life course analysis. The first step in the process is the generation of input data, denoted by X, with a Direct Acyclic Graph (DAG). DAGs are graphical representations representing all the causative relationships within the data, accounting for uncertainties inherent in the associations. Subsequently, the outcome vector, y, is synthesised to reflect RTI-related hospitalisation during the follow up period. This is achieved by examining whether the MTS for each individual in the input data contains a specific pattern of exposures associated with one of the life course concepts. The procedure is iterated for multiple life course patterns, allowing a comprehensive examination of how the performance of various DL models depends upon the unique relationships within each life course concept. This approach provides a nuanced understanding of the models' proficiency in capturing specific life course patterns, thereby contributing valuable insights into their potential utility for life course analysis.

2.2.1 Directed Acyclic Graphs and Causality

A DAG was designed to allow longitudinal cohort data with realistic underlying causal patterns to be simulated. A DAG is composed of several nodes, each of which corresponds to a particular exposure or variable. The edges which connect the nodes indicate the existence and magnitudes of any causal associations between the variables. For example, Figure 2.7 depicts an example of a DAG showing plausible processes between different exposures that generate risk of acquiring RTIs. In the context of DAGs, a relationship between two variables is causal if the presence of the causal variable directly or indirectly gives rise to the presence of

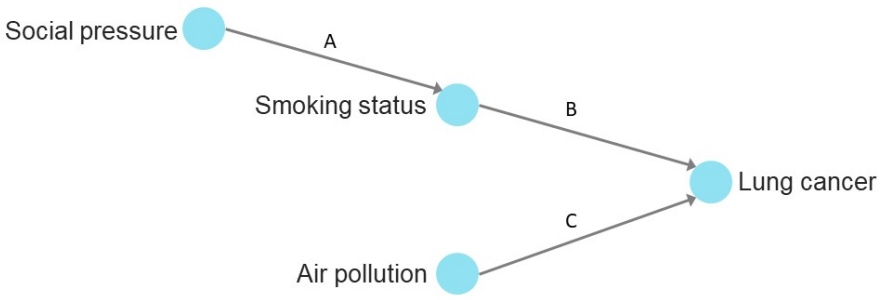


Figure 2.7: **Example of a directed acyclic graph (DAG)** DAG with four variables, each linked with causal relationships.

another variable, known as the effect. In Figure 2.7, there are multiple causal mechanisms, peer pressure increases the likelihood of a person smoking by magnitude A and a positive smoking status increases the risk of lung cancer by magnitude B. In the system, living in an area with high levels of air pollution also increases the chance of developing lung cancer by magnitude C. The DANLIFE cohort will contain multiple causal relationships between different variables and the outcome. The figure shows two types of causal relationship, and many others are possible. For example, smoking status acts as a mediator of the causal relationship between social pressure and lung cancer. Additionally, lung cancer is a collider variable as it is affected by both smoking status and air pollution. Other possible relationships are confounding variables that impact multiple variables and mean that they are correlated without being causative to each other. It is possible for causal loops to exist in the data and for there to be bidirectional relationships. Furthermore, causative relationships can either be deterministic, where one exposure directly gives rise to the outcome, or probabilistic, where an exposure only increases the likelihood of the outcome. It is well known that causal relationships and mechanisms underpin disease development and progression, however it is difficult to identify these complex relationships and to establish causality.

2.2.2 Input Data Formulation

The input data generation process involved constructing a representative DAG featuring multiple variables, including demographic, socioeconomic, and medical factors (Figure 2.8). From this structured DAG, the simulation process was employed to generate a time series dataset for each individual in the population with the number of exposures experienced in their life course being related to their unique characteristics. The DAG simulation process was written and executed in the program R using package `simcausal` version '0.5.4'. The dataset is designed to mirror a population of $N = 100,000$ individuals over a time span of $T = 40$ years. While this cohort size is notably smaller than the 2.2 million individuals in DANLIFE, it is still sufficiently large for the models to effectively capture representations. Additionally, the framework of this analysis would restrict the individuals in the DANLIFE cohort to those that reached age 16 so the cohort size becomes considerably smaller [Bengtsson et al., 2019]. The simulated cohort assumes that all individuals enter the cohort at birth, with no missing data throughout the entire 40-year duration. There is also the implicit assumption that the cohort either encompasses the first 40 years of life for all individuals, regardless of birth year, or includes only individuals born in a specific year This simplifying assumption was made to streamline analysis, as the primary experiment objective is to assess the DL models' capacity to capture significant patterns and this objective is unaffected by this particular assumption.

The DANLIFE cohort is comprised of multiple types of variables; time-independent individual-level characteristics, such as sex and ethnicity, and time-dependent exposures, such as hospital admissions and ACEs

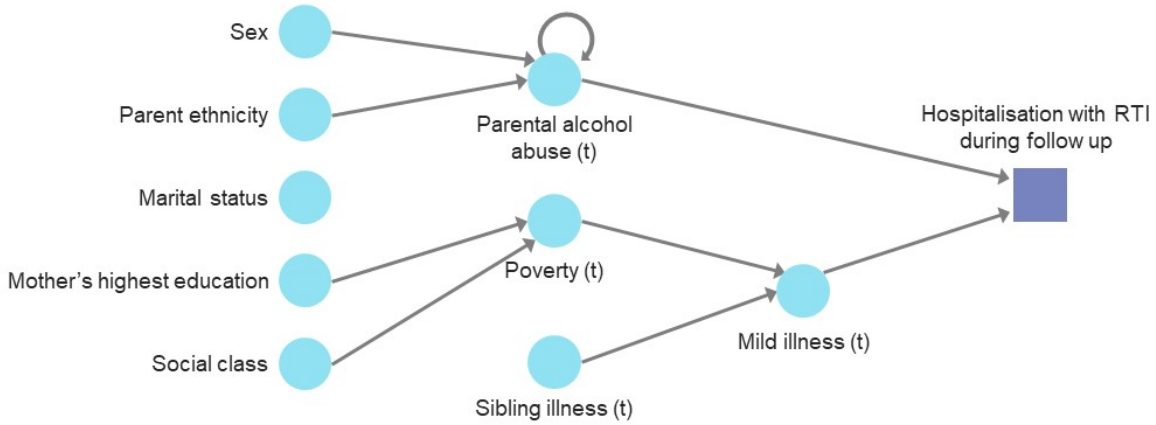


Figure 2.8: **DAG used for input data simulation.** Formulation of the DAG between the nine input variables and the outcome. The marker of (t) indicates that the variable is time dependent.

(Section 1.4, Figure 1.3). Similarly, the simulated data is composed of nine variables; five variables that represent individual-level characteristics and four variables that represent time-dependent occurrences of different exposures. In order to ensure the simulated data is realistic, the population is divided into groups according to the individual characteristics and these groups dictate the frequencies of the different types of exposures that occur. The individual-level characteristics are designed to replicate the cohort composition in the ALSPAC study by Johnson et al. [Johnson et al., 2019]. The ALSPAC study examined the association between multiple characteristics and hospitalisation rates taken from the NHS Hospital Episode Statistics (HES) [Lawlor et al., 2019]. Of the numerous characteristics recorded in the ALSPAC study, five were chosen to be represented in the simulated cohort (sex, ethnicity, marital status of mother, education of mother and social class) and randomly assigned them to cohort members according to the proportions found in the study (Table 2.1). The different characteristics were assumed to have the same distributions as given in the ALSPAC study, with the one exception of the ethnicity variable which was altered to reflect a more diverse population.

Each individual-level characteristic can fall within one of two possible categories, meaning that they take the form of binary variables that are constant over the life course for each individual with their value denoting whether an individual falls into the reference category or not. The time-independent variables are formulated as time series with a constant value. Note that the definition of ethnicity created by Johnson et al. is incredibly broad and appears to confer the perceived race of an individual instead of their ethnicity. The construction of this variable is likely to mask patterns that exist for the large number of different racial groups that are all grouped within the other category [Johnson et al., 2019]. The use of race as a variable in research is highly controversial as race is a social construct, however, in some instances, it can be beneficial to include this variable as people are still treated differently by society, and in particular, within healthcare settings based upon their perceived race and using this variable can help to identify evidence of bias or discrimination.

The remaining four variables measured within the simulated cohort data represent time-dependent occurrences of exposures, such as ACEs and hospitalisation with other illnesses (Table 2.2). Names have been assigned to these variables to encapsulate the typical exposures that they represent, however, the design of the simulated data means that these variables are representative of many other exposure interactions. In this instance, these variables act as measures of parental alcohol abuse, poverty, sibling illness and mild illness and are referred to as such. The exposure events take value 1 in every year that an exposure occurred and 0 otherwise, for example, if an individual experienced poverty for the first 3 years of their life, then the time series for poverty for the first five years of their life would look like $1\,1\,1\,0\,0$. These time-dependent exposure

Table 2.1: **Summary of the five time invariant variables in the simulated input data.** Definitions of the five time invariant variables in input data X are taken from study [Johnson et al., 2019]. The reference values are expressed as a 0 and the other values are expressed as a 1. The proportion of the individuals that are in the reference category is given along with a reference for the value. Marital status and mother's highest education are measured at time of birth.

Variable	Definition	Value	Percent of Cohort (%)	Source
x_1	Sex	Female (reference) Male	50 50	Assumption
x_2	Parent ethnicity	White (reference) Other	90 10	Assumption
x_3	Marital status	Married (reference) Single	75 25	[Johnson et al., 2019]
x_4	Mother's highest education	Degree/A level (reference) O level/CSE/Vocational	35 65	[Johnson et al., 2019]
x_5	Social class	Non-manual occupation (reference) Manual occupation	81 19	[Johnson et al., 2019]

variables are probabilistically simulated using the risk ratios for hospitalisation calculated by Johnson et al. for each of the individual characteristics to produce plausible patterns of exposures in the population [Johnson et al., 2019]. Essentially these exposures all match by eye the temporal distribution of hospitalisations but can similarly be thought of as representing ACE occurrence or other causative exposures. The use of risk ratios to concentrate adverse experiences within certain high-risk subgroups of the population ensures that the simulated data reflects reality. As a consequence, a small number of individuals will experience a much greater number of exposures than the average individual. The time-dependent variables also represent the sparsity of the DANLIFE cohort (Appendix 8.1).

Following the generation of the simulated cohort data, all $M = 9$ variables are concatenated to form the input matrix X. Multiple experiments were undertaken to assess how the exact formulation of X impacted model performance at learning the relationship between X and y. It was found that by far the most influential factor affecting model performance was the simulated relationship between X and y and that the composition of X has limited impact on model performance. Therefore, the input data has been designed to mirror typical patterns in exposures and hospital admissions which are thought to be similar to the DANLIFE cohort and allow for a realistic investigation of life course pattern detection. It is worth noting that the variables within the simulated data are not the same as those within the DANLIFE cohort and the proportions of individual-level characteristics will differ from those in the DANLIFE cohort. There will also be differences in the hospitalisation distributions due to differences between the healthcare systems in the UK and Denmark. However, these experiments demonstrate that it is more important that the simulated input data is representative of typical causative relationships between variables across the life course than that it exactly mirrors the DANLIFE cohort to allow for an assessment of the utility of DL models for life course. Moreover, the simplicity of the simulated cohort data enables a better understanding of the underlying process that may be impacting model performance that would be difficult to uncover within the full DANLIFE cohort.

2.2.3 Life Course Pattern Simulation

In order to create scenarios that represent plausible life course trajectories, multiple outcomes were simulated to represent several of the key inter-variable dependencies found within life course analysis that make up the key life course concepts discussed in Section 1.3.1. These outcomes were split into four groups: periods, repeats, timing, and order. Each group contains multiple simulated scenarios that fall into that category to allow an in-depth exploration of the capability of DL models to learn each life course pattern. Each simulated output will be referred to as a life course pattern (LCP) that falls within one of the four life course concepts (LCCs). The output is binary $y \in [0, 1]^N$ and is designed to correspond to whether an individual is hospitalised with RTI as the primary diagnoses when they are between ages 40 and 45. The outcome y is simulated from input data X according to multiple rules that each fall within one of the life course groups. The life course groups are designed to cover a range of possible temporal exposure patterns:

- **Period:** The rules for generating the outcome that fall within the period life course group represent critical periods, sensitive periods and developmental periods as described in [Ben-Shlomo and Kuh, 2002]. These are time periods during the life course when risk factors will have a greater impact on health trajectories. Rules that simulate critical period patterns generate a positive outcome if no events occur within the first given number of years. In contrast, sensitive period rules give rise to positive outcomes if the number of events that occur in the first 10 years of life surpasses a specific threshold. Developmental periods are represented as a weighted sum of the exposures that occur in different life

Table 2.2: **Summary of the four time dependent variables in the input data.** Definitions of the four time invariant variables in input data X, descriptions of their construction and the formulas used to simulate them, where t denotes the age of an individual and $\mathbb{1}_A$ is an indicator function dependent of the class of characteristic A which is equal to zero when A is the reference group and one otherwise.

Variable	Definition	Description	Formula
x_6	Sibling illness	Age-dependent designed to match hospital admissions by age (t) in [Johnson et al., 2019]	$x_6(t) \sim \text{Bernoulli}\{1/5 \log(t+3) - 0.015\}$
x_7	Poverty	Base probability of occurrence adjusted by the likelihood ratios for class and education level [Johnson et al., 2019]	$x_7(t) \sim \text{Bernoulli}\{0.04(0.28\mathbb{1}_{\text{class}} + 1)(0.22\mathbb{1}_{\text{education}} + 1)\}$
x_8	Parental alcohol abuse	Base sex-dependent probability of occurrence multiplied by the likelihood ratio of ethnicity [Johnson et al., 2019] and if an event occurred in the previous year	$x_8(t) \sim \text{Bernoulli}\{0.01(3 + \mathbb{1}_{\text{sex}})(0.17\mathbb{1}_{\text{ethnicity}} + 1) + \text{plogis}(-5 + x_8(t-1))\}$
x_9	Mild illness	Depends on the number of x_6 and x_7 events that occurred within the previous 3 years	$x_9(t) \sim \text{Bernoulli}\{\text{plogis}(-2 + (\sum_{k=0}^{3} x_6(t-k) > 2) + (\sum_{k=0}^{3} x_7(t-k) > 2))\}$

stages and if this sum is greater than a given threshold then there is a positive outcome.

$$\sum_{i=1}^{4} x_9(i) + \frac{1}{2}\sum_{i=5}^{12} x_9(i) + \frac{1}{3}\sum_{i=13}^{17} x_9(i) + \frac{1}{5}\sum_{i=18}^{25} x_9(i) + \frac{1}{10}\sum_{i=26}^{40} x_9(i) \tag{2.18}$$

- **Repeats:** The rules within the repeats group give a positive outcome if multiple events occur consecutively over a given number of years.
- **Order:** The order group encompasses scenarios where the first given number of x_8 exposure events occurring prior to the first instance of the x_9 exposure event gives rise to the outcome.
- **Timing:** The timing of exposure events, particularly their proximity to other exposures, is represented by the rules that fall within the timing group. Positive outcomes are generated when multiple x_8 exposure events occur at similar times to x_9 exposure events.

The specific formulation of the rules that fall within each LCC group are presented in Table 2.3 along with the percentage of the resulting outcomes that were positive, meaning the outcome event occurred. The outcome is designed to symbolise which individuals experience an RTI-related hospitalisation within the following 5 years after the study period. Medical outcomes, such as this, tend to be relatively rare within the population. In 2018, approximately 0.7% individuals in Denmark aged 35-44 were hospitalised with diseases of respiratory organs [sta]. Therefore, over a five-year period the expected proportion of positive outcomes would be around 3.5% assuming that no individuals are hospitalised twice, and the rate is the same each year. In order to reproduce the imbalanced nature of medical events within the simulated data, rules with greater than 5 percent of positive outcomes were excluded. Notice that x_3 is not used within the definitions for the time-dependent variables (x_6, x_7, x_8, x_9) or the equations for y, making it a nuisance variable. As each outcome is generated from the same set of input data X, the proportion of positive outcomes is different for each output. Therefore, the model performance is not directly comparable between the different simulated outputs, however, it is possible to examine the relative performance of the DL models on the same output. To minimise complexity and understand how the models perform in simple representative scenarios, the rules initially only encompass one or two of the temporal variables. In reality, it is likely that the underlying biological and societal mechanisms that give rise to negative health outcomes are considerably more complex than the relationships within this analysis. However, the LCPs are the most simplistic formulations of the hypothesised pathways that form the LCCs and consequently act as building blocks that can be used in conjunction to form more complex and realistic inter-dependencies between the variables. DL models that perform well at identifying the different rules that make up the LCCs will have the best chance of identifying more complex patterns when they exist.

Real-world complexities are reflected in the simulated data by incorporating randomness within the generation of the outcome. This process makes the synthetic data more realistic as not all individuals who experience the same set of exposures will have the same health outcomes. For example, not all individuals who smoke develop lung cancer and there are individuals with lung cancer who have never smoked. Initially the data is split into two groups: the training and testing sets. The rationale behind dividing the cohort into two sets is that this provides a fair comparison of the different modelling methods. The training of the model is entirely conducted with the training set, then the model performance is tested on the test set. The test set acts as the “unseen" data which assesses how well the model generalises to new data as it is possible to obtain excellent performance on the training set due to the flexible nature of the DL models but this can arise entirely due to overfitting and not be representative of model performance on new data. Noise was added to the deterministically generated outputs y by randomly selecting 10 percent of the individuals with a positive outcome and switching their output to being negative. Then the same number of individuals with a negative outcome

Table 2.3: **The definitions for each LCP within each LCC.** Each row gives information about a singular LCP including the LCC it falls into, an identifying number, the proportion of the resulting outcome that is positive and its definition. If the rule is met for an individual, then the output is positive, and that individual has the outcome of being hospitalised with an RTI. The variables x_8 and x_9 are defined in Table 2.2. The ID allows for each pattern to be uniquely refered to, for example Period3 or Timing2.

Group	ID	Positive (%)	Rule (Positive Criteria)
Period	1	1.9%	x_9 has no events within the first 30 years
	2	3.1%	x_9 has at least four events within the first 10 years
	3	0.6%	x_9 has at least five events within the first 10 years
	4	1.3%	Sum of weighted x_9 events is greater than four
Repeats	1	6.0%	x_9 has a repeat of three events in a row anywhere
	2	0.7%	x_9 has a repeat of four events in a row anywhere
	3	0.6%	x_9 has two repeats of two events in a row anywhere
Order	1	4.9%	The first two events of x_8 occur before the first event of x_9
	2	1.2%	The first three events of x_8 occur before the first event of x_9
	3	0.2%	The first four events of x_8 occur before the first event of x_9
Timing	1	1.9%	Any two x_8 events occur within the same year as any x_9 event
	2	5.4%	Any three x_8 events occur within two years of any x_9 events
	3	1.3%	Any four x_8 events occur within two years of any x_9 events
	4	3.5%	Any four x_8 events occur within four years of any x_9 events

are changed to having positive outcomes to keep the same proportion of positive outcomes in the population. The noise is added only to the training set and the test set is left untouched meaning that theoretically it is possible for a model to have perfect performance on the test set if it has truly learnt the underlying patterns and has not overfit to the noise.

2.3 Evaluation Metrics and Experimental Setup

This section lays out the way in which the DL models are trained, including the process of hyperparameter optimisation, the justification behind the selection of the evaluation metrics and model formulation. The section concludes with a description of the experiments that are carried out within this analysis to provide a robust understanding of the capacity of multiple DL models to identify life course patterns.

2.3.1 Model Training Pipeline

The process of training a ML model is made of several steps that are common to all model architectures. Several of the steps can be tailored so that the training process is better suited to the format of the data. Each of the choices made in this analysis and the justifications are explained in later sections, however, this pipeline is a broad overview of the entire ML model training pipeline and is as general as possible. The steps involved in model training are as follows:

1. **Data Preparation:** The data must be loaded from the source; in the case of MTS classification the input data is formulated as $X \in \mathbb{R}^{N \times M \times T}$ and the output as $y \in \mathbb{R}^N$. Pre-processing data maximises the performance of DL models. Data is normalised or standardised to ensure that the features all have similar scales so the larger features do not dominate training meaning that smaller scale features may be overlooked by the model. Another form of pre-processing is augmentation which is a process that

transforms the data to have desirable properties, for example by taking the logarithm. Finally, the data is divided into train and test sets. The model is trained using the train set and then its performance is tested on the unseen test set to assess model generalisability.

2. **Model Selection and Initialisation:** The DL model is selected, for example ResNet or MLSTM-FCN, and the model weights are initialised randomly. Each model has a different selection of hyperparameters that govern the model structure and allow it to be adjusted to suit the specific data. These include, for example, the number of layers, the kernel size and the dropout rate. Dropout is a regularisation technique that randomly ignores a given percentage of the model cells when training to prevent overfitting. Another regularisation method is weight decay, otherwise known as l1 or l2 regularisation which adds the sum of the absolute value, or the square of the weights weighted by a factor to the loss function which restricts the size of the model parameters.

3. **Model Compilation:** The model is compiled with an optimiser, loss function and suitable evaluation metrics. The optimiser is responsible for updating the model weights, for example Adam and SGD. Model training is conducted to minimise or maximise the loss function so careful choice is vital, examples include cross-entropy and mean squared error. The loss function is very specific to the data type and the nature of the problem, for instance if this is a classification or regression task. Model performance is measured using evaluation metrics that compare the model predictions with the ground truth, it is important that these are chosen based upon the specific goals of the study.

4. **Hyperparameter Tuning:** In addition to the model-specific hyperparameters, further hyperparameters that govern the training process need to be selected. These include the learning rate that determines the step size in optimisation, the batch size that is the number of data points used within each iteration and the number of epochs which is the number of complete passes through the data set. All the hyperparameters are tuned using k-fold cross-validation. The optimisation procedure is formed of multiple trials that each test the model performance for a different set of hyperparameters which are selected from a feature space using a sampler. The sampler may be random, systematic (gridsearch) or choose parameters based upon previous trial results, such as the Tree-structured Parzen Estimator (TPE). The training data set is split into k folds and within each trial the model is trained and validated k times, each time a different fold is used as the validation set and the remaining data as the train set. The validation set acts as a test set to assess model performance without using the unseen test set. This helps ensure robustness of parameter choice and generalisability of the results. The model performance, measured with a suitable evaluation metric, is averaged across the k runs and the final hyperparameter values are chosen as the values from the trial that maximised this value.

5. **Training Loop:** For a fixed number of epochs, the data is iterated using the following steps:

    - **Forward Pass:** A batch of training data is fed through the model and predictions are computed.
    - **Loss Calculation:** The loss is computed between the predicted and true values according to the loss function.
    - **Backpropagation:** The gradients of the loss with respect to the model parameters are computed.
    - **Gradient Descent:** The model weights are updated using the optimiser and computed gradients.

    After each epoch model performance is calculated on the validation set. Callbacks can be used within the model to adjust training, improve performance, and prevent overfitting to the train set. For example, early stopping is a callback that stops model training early if the validation performance does not improve for a set number of epochs. Another example is learning rate scheduling which decreases the learning rate when the model performance plateaus.

6. **Inference and Evaluation:** After the hyperparameters have been tuned, the model is trained again one final time on the entire training set using these hyperparameter values. Then model performance is calculated on the test set to assess generalisability.

2.3.2 Metrics to Measure Model Performance

Evaluation metrics are measures of model performance and provide insights into the strengths and weaknesses or models. Once the model has been trained using the training data, model performance is measured on the test set and indicates how well the model generalises. This experiment has been framed as a binary TSC problem which means that each sample can be classified into four groups: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). True positives and negatives are correctly classified positive and negative outputs, and false positives and negatives are incorrectly classified negative and positive outputs. There are multiple scores by which the performance of these binary classification models can be measured. The most common one is accuracy which is defined as the fraction of samples that were correctly classified [Huang and Ling, 2005]

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

For many data sets accuracy is a good measure of performance, however in the case of imbalanced classes when one class makes up the vast majority of the output which is common with medical outcomes, accuracy is no longer informative. For example, if 95 percent of the output is in class 0 and the remaining 5 percent of the output is in class 1 and the model only predicts all individuals are in class 0 then the accuracy will be equal to 0.95 suggesting good performance. However, whilst medical outcomes are rare, the health impacts are such that it is crucial that such cases are able to be identified by the model and the metric truly reflects this.

There are multiple other metrics that provide different insights into how the model is performing that are less influenced by class imbalance. Precision and recall are two metrics that penalise false positives and false negatives, respectively. They are defined as [Goutte and Gaussier, 2005]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

These metrics both prioritise correctly identifying true positives and either penalise missed positive cases or incorrectly identifying non-instances as positive cases. These metrics are both far more suitable for medical data where really the focus is on correctly identifying individuals with a high risk of certain conditions. They are combined into the F1 score metric which seeks to prioritise having high recall and precision scores, i.e. keeping true positives high and reducing both false positives and false negatives and it is defined as [Zijdenbos et al., 1994]

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

All the evaluation metrics discussed so far range from 0 to 1, where 0 indicates poor performance and 1 indicates good performance.

Another useful metric is the area under the received operating characteristic curve (AUROC) [Huang and Ling, 2005]. The received operating characteristic curve (ROC) is the line between the true positive rate, otherwise known as sensitivity, and the false positive rate. The true positive rate (TPR) and false positive rate (FPR) are

calculated as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The ROC curve is then created by plotting the TPR against the FPR for different classification threshold, i.e. the output probability at which the model will start classifying cases as class 1. Then the AUROC is defined as the integral of the ROC curve, representing the area underneath it. The AUROC score values range between 0.5 and 1, where 0.5 is poor performance and the model predictions are no better than random guesses, and an AUROC of 1 indicates perfect performance. Higher values of the AUROC score indicate a higher TPR and lower FPR so this metric prioritises reducing false positives whilst also increasing true negatives. The AUROC is a widely used metric in binary classification problems, however it is somewhat unsuitable for imbalanced data problems as it weights both classes equally.

A metric that is common in binary classification tasks is the Brier score [Brier, 1950]. It is defined as the mean squared differences between the predicted outcomes and the actual outcomes.

$$\text{Brier score} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2$$

where $\hat{y}_n$ is the predicted class for individual n and y_n is their true class and the number of samples is N. The Brier score is a useful complementary metric because it takes into account the calibration of the model as well as the reliability of the predictions. The term calibration describes the level of concordance between the model predictions and the actual observed outcomes. A model that is well-calibrated has predicted probabilities that align with the true occurrence of an event within that group [Van Calster et al., 2019]. However, unlike the F1 score or AUROC, the cost of misclassifying of positive and negative cases are equal.

AUPRC is the area under the precision-recall curve which summarises a precision-recall curve as the sum of precisions for each threshold weighted by the difference in recall as follows [Boyd et al., 2013]:

$$AUPRC = \sum_{n}(R_n - R_{n-1})P_n \tag{2.19}$$

where R_n and P_n are the precision and recall for the nth threshold. This metric directly symbolises the trade-off between precision and recall in one value that describes model performance. This metric is appropriate for quantifying model performance on data sets that are heavily imbalanced as it focuses on correctly classifying the minority positive class.

2.3.3 Loss Function

The choice of loss function heavily depends on the structure of the data. For instances of binary outcomes and a classification task, as is the case in this analysis, the loss function chosen is typically cross-entropy loss. The definition of cross-entropy is [Zhang and Sabuncu, 2018]

$$CE(p, y) = \begin{cases} -\log(p) & y = 1 \\ -\log(1-p) & y = 0 \end{cases} \tag{2.20}$$

where p denotes the probability that an observation is in a given class and y is an output. However, the simulated data was designed to reflect the significant class imbalance frequently present within healthcare data. Consequently, minimising the loss function will prioritise minimising misclassified negative cases (false positives) and it is likely that the model will not be able to identify any positive cases. This problem is often

addressed by adding weights to the cross-entropy loss so that the proportions within each class dictate the importance of correct classification. The weighted cross entropy is given as

$$WCE(p, y) = \begin{cases} -\alpha \log(p) & y = 1 \\ -(1-\alpha)\log(1-p) & y = 0 \end{cases} \tag{2.21}$$

where α is the weight given to the positive classes, which is often selected as the inverse of class frequency. However, another unaddressed difficulty of severe class imbalance is that the positive instances are rare and therefore much harder to classify. A focusing parameter, γ, is introduced to allow the loss function to reflect the increased difficulty of identifying the positive outcomes and therefore, improve the fitting to the data. This forms the focal loss which is defined as [Lin et al., 2017]

$$FL(p, y) = \begin{cases} -\alpha(1-p)^{\gamma} \log(p) & y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p) & y = 0 \end{cases} \tag{2.22}$$

2.3.4 Model Training Procedure & Hyperparameter Optimisation

There are 14 different simulated data sets which the DL models are fitted to for two noise levels which gives a total of 28 data sets. Each of the data sets are fed through the five models (LR, InceptionTime, ResNet, MLSTM-FCN and LSTMAttention) and the output evaluation metrics are used to compare model performance. The model training procedure was the same for each model except for LR, which is described separately below. The input data did not require pre-processing or augmentation as it is simulated to be binary meaning all variables are automatically on the same scale. Each data set was loaded and then split into train and test sets with an 80:20 ratio. Fitting the DL models is conducted using the `tsai` package version 0.2.22 in python which also gives the specific model formulations [Oguiza, 2022]. Below the model-specific hyperparameters and their ranges are given:

- **ResNet:**
  - Number of filters in first block: 32, 64 or 96
  - Dropout rate in fully connected layer: between 0.1 and 0.5
  - Kernel size for each convolutional layers: three numbers taken from 1, 3, 5, 7 and 9
- **InceptionTime:**
  - Number of filters in convolutional layers: 32, 64 or 96
  - Dropout rate in fully connected layer: between 0.1 and 0.5
  - Dropout rate in convolutional layer: between 0.1 and 0.5
  - Kernel size for first convolutional layer in each block: 20, 40 or 60
- **MLSTM-FCN:**
  - Kernel size for each convolutional layers: three numbers taken from 1, 3, 5, 7 and 9
  - Number of convolutional layers: three numbers taken from 32, 64, 96 and 128
  - Number of features in the hidden state: 60, 80, 100 or 120
  - Number of RNN layers: 1, 2 or 3
  - Dropout rate in fully connected layer: between 0.1 and 0.5

  - Dropout rate in cells: between 0.1 and 0.5
  - Dropout rate in RNN layers: between 0.1 and 0.5
- **LSTMAttention:**
  - Number of parallel attention heads in self-attention: 8, 16 or 32
  - Dimension of the feedforward network model: 64, 128 or 256
  - Number of sub-layers in the encoder: 2, 3 or 4
  - Number of features in the hidden state: 32, 64 or 128
  - Number of RNN layers: 1, 2 or 3
  - Dropout rate in fully connected layer: between 0.1 and 0.5
  - Dropout rate in encoder: between 0.1 and 0.5
  - Dropout rate in RNN layers: between 0.1 and 0.5

The models are each compiled with Adam optimiser, the loss function is the focal loss, and the performance is measured by multiple evaluation metrics that are AUROC, AUPRC and the Brier score. The metric that governed hyperparameter selection was the AUROC. The ResNet and InceptionTime models had the activation function changed to LeakyReLU from ReLU to improve stability [Xu et al., 2020]. Two callbacks were used in training which were early stopping with patience level 2 and learning rate scheduling. As a result, the learning rate does not need to be optimised as it is reduced by a factor of 2e-10 when learning plateaus and so it is set to an initial value of 1e-3. The batch size and the α and γ parameters in the focal loss were also tuned in the hyperparameter optimisation for all DL models with possible values:

- **Batch size:** 32, 64 or 96
- **Alpha:** between 0 and 0.5
- **Gamma:** between 1.01 and 5

Three-fold cross validation is conducted to maximise model performance, it is coded using the optuna package [Akiba et al., 2019]. This was conducted for 100 trials of hyperparameters so as to minimise the computational requirements. Given the computational limitations it was not possible to explore the high dimensional search space of some of the models with this number of trials and therefore, the random sampler was selected. The TPE sampler was not used because after experimentation it was found that it frequently got stuck in local minima. The number of epochs was set to 10. Once hyperparameters had been selected, the model is trained once more using these values with 90 percent of the train data used to train the model and the remaining 10 percent used to validate the model. Then the model is applied to the test set to compute predictions and calculate the evaluation metrics.

The process for the LR model slightly differs in that it does not have hyperparameters that need to be selected. Therefore, the data is loaded, split into train and test sets, and then flattened so it can be input into the model before the model is applied. In order to reduce overfitting L1 regularisation is used which regulates the size of the weights by adding the sum of their values to the loss function [Ng, 2004]. In addition, the LR model incorporates class balancing to counteract the heavily imbalanced classes.

2.3.5 Experimental Design

Firstly, the impact of the noise level on the performance each of the DL models and the LR baseline model will be explored. The models are fitted to all the simulated data sets with zero and ten percent noise levels. Then

their performance will be compared across three metrics: AUROC, AUPRC and Brier score. The analysis will continue by exploring model performance on the individual data sets within each of the four life course concepts with ten percent noise. This will allow for an in-depth investigation of whether model performance varies according to the type of life course processes embedded in the simulated data. Finally, the time taken to train each model will be compared to enable insights into the comparative computational cost of the different models to be gained.

2.4 Results

In this section, the results of each experiment detailed in the preceding section are described in order. The implications of these findings within RTI and life course research will be discussed within the following section.

2.4.1 Impact of Noise Levels on Model Performance

The first part of this analysis examines how the addition of noise within the training set influences the performance of the DL models. Introducing stochasticity into the simulated data makes it more representative of realistic data. The additional noise obscures the otherwise deterministic patterns that link the exposures and the outcome, acting as a test of the model's ability to detect the true underlying relationships present in the data. The predictive performance of all models, measured by AUROC, AUPRC and the Brier score, degrades when noise is introduced into the simulated data (Figure 2.9). The AUPRC scores lie on a scale from 0 to 1 with greater values indicating better performance. Similarly, the greater AUROC values also signify improved performance, but the scale is smaller and ranges from 0.5 to 1. In contrast, lower Brier scores suggest better model calibration and values fall on a scale from 0 to 1.

When there is no noise in the simulated data, ResNet, LSTMAttention and MLSTM-FCN achieve almost perfect performance with median AUROC scores above 95 percent (Figure 2.9, top). In comparison, LR has an average AUROC score of 89.3 percent. The inclusion of noise leads to a deterioration in model performance to varying extents. InceptionTime is the model most impacted by the stochasticity in the data with the mean AUROC value falling from 99.7% to 82.5%. The model that is most resilient to noise in the data is LSTMAttention with the mean AUROC dropping only 1.1% down to 98.4% from 99.5%. Particularly, when measured with the median AUROC, the DL architectures LSTMAttention and MLSTM-FCN obtain almost perfect performance with scores above 95 percent. However, for both models, there is a significant range in performance when measured using the AUROC score across the assorted simulated data sets with certain data sets scoring below 70 percent. Differences in performance according to the life course patterns simulated are investigated in greater detail in the next section.

The AUPRC scores show a distinctly different pattern to the AUROC scores (Figure 2.9, centre). Firstly, the difference between the predictive performance of LR in comparison with the DL models is far greater when measured using the AUPRC score. This is particularly noticeable in the presence of noise where LR attains a mean AUPRC score of only 11.3 percent. Moreover, the AUPRC scores of all models are more heavily impacted by the addition of noise than the AUROC score. LSTMAttention, which has the highest mean AUROC scores of all the models, is only the third highest performing model with noise and the worst performing DL model without noise when measured using the AUPRC metric. As with the AUROC score results, the models demonstrate diverse levels of robustness to the addition of noise within the simulated data. Measured using the AUPRC score, the best performing model without noise is ResNet, however with noise the mean AUPRC score drops from 89.6% to 63.0%. In comparison, MLSTMFCN experiences the smallest decrease in performance with mean AUPRC dropping to 74.3% from 86.1%. Overall, the mean and

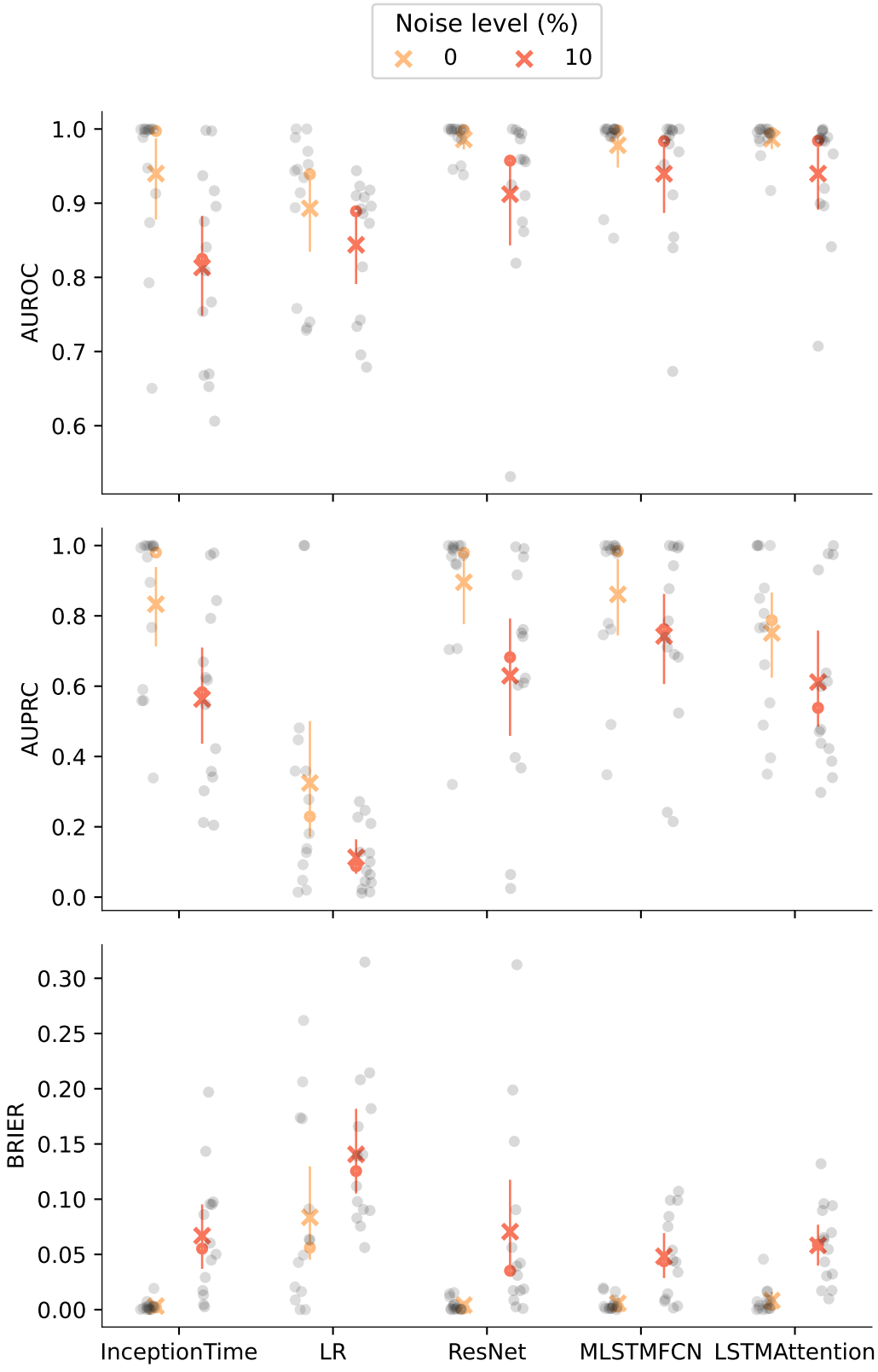


Figure 2.9: **Model comparison for multiple noise levels.** Values of evaluation metrics AUROC, AUPRC and Brier score for each outcome at two levels of noise (0 and 10 percent) for the five models (LR=Logistic Regression, MLSTM-FCN, LSTMAttention, InceptionTime and ResNet). Coloured crosses and points denote the mean and median across all outcomes, respectively. Error bars denote 95% confidence intervals obtained via bootstrapping (1000 bootstraps). Models ordered according to their mean AUROC value with no noise present. Figure taken from Coupland et al. and created by Neil Scheidwasser [Coupland et al., 2023].

average AUPRC scores tend to be lower than the corresponding AUROC scores which may be a consequence of the design of the hyperparameter selection approach.

The Brier score reflects both the predictive performance of the models alongside the calibration performance. Of all the models examined, the LR model has the greatest Brier scores indicating the poorest calibration (Figure 2.9, bottom). The Brier scores for all DL models in absence of noise are very low and signify excellent calibration. There is a noticeable increase in the Brier scores for all models when noise is introduced to the data. Although, the Brier score seems to increase broadly uniformly for all DL model architectures, unlike the associated AUPRC and AUROC scores which show more variation across models. One other feature of note is the ranges in Brier score for the different data sets, shown as grey points. LR, ResNet and InceptionTime all have larger ranges between the simulated data sets than MLSTM-FCN and MLSTMAttention.

2.4.2 Model Performance Across Life Course Groups

The performance of the DL models was evaluated for individual data sets that each fall within one of the four LCC groups (period, repeats, order, and timing), see Table 2.4. Each simulated outcome is to represent a specific type of exposure pattern that are key component of life course theory. The values of the three evaluation metrics are calculated for all models for each simulated data set all with 10 percent noise to mimic realistic real-world scenarios. The experimental design ensures a fair and robust comparison of each model's ability to identify temporal patterns, inter-variable relationships, and sequential ordering within the data. Utilising these significant patterns within the simulated data means that this analysis can highlight the strengths and weaknesses of each model alongside their suitability for life course analysis.

The period group encompasses data sets that imitate features of critical, sensitive, and developmental periods. These are time intervals in the life course during which individuals are particularly vulnerable, and the occurrence of certain exposure events increase the likelihood of negative health outcomes later in life. In this group, positive outcomes arise when the number of x9 events within a set time period surpass a given threshold. Consequently, all the underlying relationships within the period group are entirely linear. Accordingly, the DL models achieve excellent performance within this group (Table 2.4, Period). There is a substantial difference in model performance when measured by the AUROC or AUPRC scores. For instance, LR performs comparatively to the DL models when measured using the AUROC metric, however, the AUPRC scores suggest a much greater difference in performance. In fact, the LR models AUPRC scores within the periods group range between 4.2% and 22.7% showing poor predictive performance. Thie impact of class imbalance on performance is particularly apparent for the Period3 data set which has only 0.6% of outcomes that are positive. In the Period3 data set, the AUPRC score for ResNet is much lower than the corresponding AUROC value, indicating that the model is predicting essentially randomly when measured by this metric. Several models do not appear to suffer as greatly due to the imbalanced data with MLSTM-FCN attaining an AUPRC score of 87.7%. Overall, ResNet and MLSTM-FCN are the two highest performing models in this group according to the AUROC metric and MLSTM-FCN is the highest performing model according to the AUPRC metric.

The data sets that fall within the repeats group are formulated to represent how multiple exposure events can accumulate over time and their consecutive occurrence can act irreversibly to cause poor future health. When a given number of x9 events occur in succession then that individual is assigned a positive outcome which represents hospitalisation with an RTI-related illness. In this group, LR had consistently poor performance across all the data sets, particularly when measured by the AUPRC metric which ranged between 1.1% and 12.5% (Table 2.4, Repeats). The AUROC scores for LR were a little between and feel between 67.9% and 73.4% which is best described as mediocre performance. The AUPRC scores are more impacted by class imbalance, and this is especially true for the LSTMAttention and InceptionTime models. When measured

Table 2.4: **Relative model performance within the four LCCss; period, repeats, order and timing.** Each row gives information about a rule including the LCC group that it falls into, an identifying number, the proportion of the resulting outcome that is positive and its definition. If the rule is met, then the outcome is positive. For each model, the AUROC and AUPRC values as percentages are provided and the highest performing model for each data set is highlighted in bold font. Ten percent noise in the data. LR = logistic regression. Table and caption taken from Coupland et al. preprint Coupland et al. [2023].

Group		Positive	Positive Criteria	LR		InceptionTime		LSTMAttention		MLSTM-FCN		ResNet	
		%		AUC	PRC	AUC	PRC	AUC	PRC	AUC	PRC	AUC	PRC
Period	1	1.9	x_9 has no events in first 30 years	91.0	10.1	93.7	30.2	70.7	42.2	85.4	71.1	**98.7**	**96.8**
	2	3.1	x_9 has 4+ events in first 10 years	94.4	22.7	76.7	54.8	98.8	97.7	**100.0**	**99.9**	99.9	99.6
	3	0.6	x_9 has 5+ events in first 10 years	90.8	4.2	80.9	61.8	**98.9**	47.1	98.0	**87.7**	86.2	2.5
	4	1.3	Sum of weighted x_9 events is greater than 4	88.6	12.8	84.1	62.5	**98.3**	34.0	91.1	**78.6**	95.6	60.3
Repeats	1	6.0	x_9 has a repeat of 3 events in a row anywhere	73.4	12.5	91.7	84.4	**100.0**	**100.0**	**100.0**	**100.0**	53.1	6.4
	2	0.7	x_9 has a repeat of 4 events in a row anywhere	67.9	1.5	89.6	79.3	98.7	38.7	**100.0**	**99.8**	**100.0**	99.1
	3	0.6	x_9 has 2 repeats of 2 events in a row anywhere	69.6	1.1	75.4	42.2	98.5	43.8	95.2	**73.9**	**99.5**	60.9
Order	1	4.9	The first 2 events of x_8 occur before the first event of x_9	92.3	27.2	99.7	**97.3**	**99.8**	93.1	98.7	69.1	95.9	39.7
	2	1.2	The first 3 events of x_8 occur before the first event of x_9	81.4	7.5	87.5	66.9	84.1	61.3	**96.9**	21.5	92.5	**75.1**
	3	0.2	The first 4 events of x_8 occur before the first event of x_9	87.3	2.3	67.0	21.2	**90.0**	29.8	67.3	24.1	81.9	**36.8**
Timing	1	1.9	Any 2 x_8 events occur within the same year as any x_9 event	74.3	4.5	**99.8**	**97.9**	92.0	60.0	**99.8**	94.3	99.4	62.3
	2	5.4	Any 3 x_8 events occur within 2 years of any x_9 events	89.6	24.7	65.3	34.1	99.8	97.4	**99.9**	**99.3**	95.9	91.6
	3	1.3	Any 4 x_8 events occur within 2 years of any x_9 events	89.2	6.4	60.6	20.5	89.6	63.7	84.0	52.4	**91.0**	**71.6**
	4	3.5	Any 4 x_8 events occur within 4 years of any x_9 events	91.8	20.9	66.8	35.8	96.6	47.7	**99.1**	68.3	87.5	**74.1**

using the AUROC metric, the LSTMAttention and MLSTM-FCN models perform similarly. Conversely, the AUPRC scores suggest that the predictive performance of MLSTM-FCN is significantly better than that of the other models, ranging from 73.9% to 100.0%. ResNet generally performs well within this group, although there appears to be an outlier for data set Repeat1 which has an AUROC score of only 53.1% compared to scores of 100.0% and 99.5% on data sets Repeats2 and Repeats3, respectively. This indicates potential overfitting or mistuned hyperparameters.

The order group is designed to embody the phenomenon of the sequential ordering of multiple exposure events affecting an individual's risk of poor health outcomes. If one or more x_8 exposure events occur before the first instance of exposure event x_9 then a positive outcome is generated. Of all the life course concept groups, this group contains the data set with the greatest level of class imbalance, with just 0.2% of outcomes in the Order3 data set being positive. Correspondingly, the MLSTM-FCN, LSTMAttention and ResNet (excluding one outlier) models have the worst performance for this data set out of all of the other data sets in all of the other categories (Table 2.4, Order). Within the order group, the worst performing model is LR, which has particularly low AUPRC scores in this category that range between 2.3% and 27.2%. The MLSTM-FCN model performs comparatively worse in this group with low AUPRC scores ranging from 21.5% to 69.1%.

Life course studies found that the temporal separation between different types of exposure event was significant in determining future health. Within the timing category, a positive outcome is simulated when a specified number of x_8 exposure events happen within a set number of years of an x_9 exposure. This pattern, simulated within the timing group, mimics multiple kinds of exposure event occurring within a short time frame to amplify negative health effects. None of the models tested in this analysis obtain consistently high performance across all data sets in this group (Table 2.4, Timing). According to the AUPRC metric, LR is the worst performing model within this group with a range of 4.5% to 24.7% signifying poor performance. However, the AUROC evaluation metric suggests that LR performs competitively. Excluding the Timing1 data set, InceptionTime performs the worst according to the AUPRC metric with a range between 20.5% and 34.1%. No models obtain excellent performance when measured with the AUPRC metric and there is much variation between the data sets in this group.

Generally, all the DL models perform well at identifying the different life course patterns. There are considerable differences in how well the models perform according to which evaluation metric is used. The AUROC score identifies LSTMAttention as the best performing model across all data sets, whilst the AUPRC score indicates comparable performance across methods with each model having strengths and weaknesses. LR is outperformed by all the DL models, except for within the period group, indicating that this model tends to predict more false positives than the DL models. This demonstrates how the lack of temporal relationships within the model limit the complexity of the representations it can capture. Class imbalance is clearly a crucial factor in model performance and impacts the models to different extents. Whilst methods to address class imbalance, such as choice of loss function and evaluation metric, have been used within the modelling framework, the results demonstrate that these do not fully address the problem. This highlights the importance of model comparison across multiple metrics to obtain a holistic understanding of performance.

2.4.3 Computational Efficiency of Machine Learning Models

In this section, the results of the analysis on the computational efficiency of various ML models are presented. The investigation focused on the training time for each model, across the simulated data sets. A significant drawback of DL models is that they frequently represent a significant computational burden, especially when conducting hyperparameter optimisation as this involves training the model multiple times. The computational cost of different architectures is an important factor to consider when selecting a model, especially as in many healthcare data analyses settings there are computational restrictions that will prevent the use of some DL

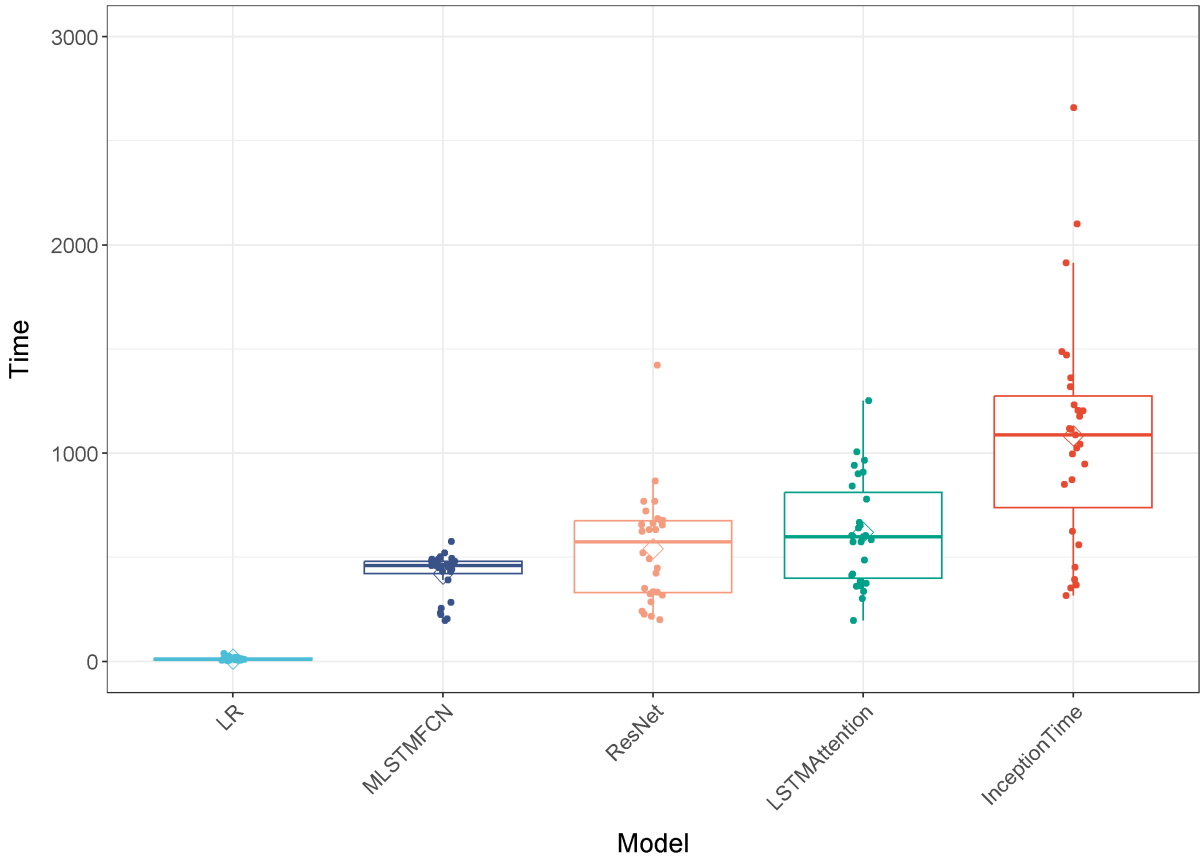


Figure 2.10: **Computation time.** Training time in seconds for the five models (LR = Logistic regression, InceptionTime, ResNet, MLSTM-FCN and LSTMAttention) with pre-selected hyperparameters for each outcome both with and without noise. Boxes indicate quantiles with the centre line depicting the median value, whiskers indicate the range of the data and diamonds show the mean value. One point has been removed from the plot which is the training time of InceptionTime for one data set with a time of 5578 seconds.

methods. The personal data are usually stored on secure servers that are not designed for DL studies and have limited memory and power. Additionally. the data sets can be very big, such as the DANLIFE cohort, which increases computation time significantly.

The training time results indicate considerable differences in the computational efficiency of the models (Figure 2.10). Notably, InceptionTime emerged as the most time-consuming model by a considerable margin. It exhibited a mean training time of approximately 1244 seconds, which was substantially longer than any other model in our evaluation. LSTMAttention, the second slowest model, also exhibited longer training times, with a mean training time of about 615 seconds. It trailed behind InceptionTime but still required slightly more computational resources compared to the other models. MLSTM-FCN and ResNet, while similar to each other in terms of training time, fell between LSTMAttention and LR in computational efficiency. They exhibited mean training times of approximately 421 seconds and 542 seconds, respectively. In contrast, LR demonstrated remarkable computational efficiency. It had the shortest training time with a mean of 12 seconds, making it the model that requires the least computational resources.

The analysis demonstrates the variation in the computational efficiency of the DL models. InceptionTime and LSTMAttention, while powerful in terms of predictive capability, demand significant computational resources. Conversely, LR stands as an efficient option for tasks with limited time constraints. The difference in computational burden becomes starker when considering hyperparameter optimisation during which the model is repeatedly trained. For example, 3-fold cross-validation with 100 trials requires training the model 300

times, meaning that slight differences in training time between models significantly impact the overall computation time. In the subsequent discussion section, the implications of these findings will be explored as well as how the computational efficiency of these models aligns with their practical utility in various real-world scenarios.

2.5 Interpretation

The examination of model performance with varying noise levels provided valuable insights into their robustness when faced with real-world, noisy data. In scenarios with no noise, all the DL models exhibited impressive performance, demonstrating their superior capability to effectively capture underlying patterns. However, when noise levels increased to ten percent, distinct patterns in model behaviour emerged. As expected, the added noise within the data masked the underlying trends and deteriorated model performance. LR remained the weakest performer, indicating its susceptibility to noise and limited ability to capture complex patterns within noisy data. This behaviour aligns is expected since linear models are not equipped to handle the intricacies of the temporal and inter-variable relationships, with this model formulated to treat adjacent time points independently. Accordingly, LR performs comparably within the period group which contains data sets that possess entirely linear points. Therefore, these results demonstrate that the LR model can learn linear relationships within the data but not more complex temporal or inter-variable connections. Life course patterns are likely to be far more intricate than simple linear relationships which suggests that LR is not suitable for life course research.

InceptionTime, although a promising model that has achieved exceptional performance in many other applications, demonstrated sensitivity to noise exhibiting a large decrease in predictive performance. A possible cause is the high level of sensitivity to hyperparameter selection displayed by InceptionTime. This finding suggests that InceptionTime may require additional augmentations or preprocessing steps to be applied to the data to mitigate the impact of noise effectively. The performance of ResNet was comparatively stable, with excellent performance except for a single potential outlier, indicating the resilience of this DL architecture in noisy scenarios. The outlier could be an instance of overfitting to the training data, which is a term that describes the vast number of parameters in these models allowing them to fit to the noise in the data and not just the key patterns. Overfitting allows DL models to appear to achieve good performance on the training set, but they generalise poorly attain low performance scores for the test set [Selvaraju et al., 2016]. MLSTM-FCN and LSTMAttention, while not entirely immune to noise, exhibited consistent and robust performance, demonstrating suitability for capturing temporal dependencies and inter-variable relationships in the data. Both of these architectures contain LSTM modules, meaning that time is explicitly accounted for which may explain their strong performance and make them well-suited for challenging longitudinal life course data. Overall, the promising performance of the DL models demonstrates they are suitable for data that is formulated as indicators of exposure event occurrences over time, even though this data differs from the continuous MTS data which these architectures were adapted for.

The two main evaluation metrics used to assess predictive performance, AUROC and AUPRC, showcased markedly different trends. Both with and without noise, the AUPRC scores generally suggested considerably worse performance than the AUROC scores. A potential explanation is that the hyperparameter selection procedure was designed to select the parameters that maximised the AUROC score. This may mean that the chosen hyperparameters are not allowing the models to reach their highest potential AUPRC values. Differences in the performance when measured by these two metrics arise since each metric assigns a different cost to various types of misclassifications. The AUROC score more heavily penalises instances of false negatives, with high AUROC scores tending to signify that a large proportion of positive outcomes were correctly classified. In contrast, the AUPRC metric seeks to minimise both false negatives and false positives

equally and takes into account the raw numbers as well as the proportions. Furthermore, the metrics also differ when there is significant class imbalance since the AUROC score can often be very high if the rare outcomes are correctly classified, whereas the AUPRC will more greatly penalise higher numbers of false positives and will be relatively lower [Davis and Goadrich, 2006]. Therefore, it is important to use multiple metrics to gain an in-depth understanding of model strengths and weaknesses [Johnson and Khoshgoftaar, 2019]. This is especially critical to consider when measuring disease risks since there is a balance between minimising missing high-risk cases whilst not incurring too many low-risk individuals being misdiagnosed. Within healthcare, there are ramifications of both possibilities with heavy costs associated with late diagnosis and poorly targeted interventions.

The AUROC metric highlighted ResNet, MLSTM-FCN and LSTMAttention as the top three performing models with only a slight degradation in performance observed with additional noise. Moreover, when using the AUROC metric, LR achieved almost comparably good performance as the DL models. In contrast, the AUPRC scores suggest that LR had far worse performance than all the DL models. InceptionTime, ResNet and MLSTM-FCN were the best performing models without noise when measured by AUPRC, and MLSTM-FCN was the best performing model with noise. The analysis also considered calibration performance which is a metric that is becoming of growing interest within the community [Van Calster et al., 2019]. All the DL models exhibited good calibration performance with little variation between models shown by the low Brier score values. However, MLSTM-FCN and LSTMAttention had more consistent performance across the different simulated scenarios. LR had the highest Brier scores both with and without noise, indicating poor calibration to the data. This goes against the common observation that DL models tend to have poor calibration in comparison with LR [Guo et al., 2016]. Moreover, it mirrors the findings of other studies which found that the use of the focal loss function improves the calibration of DL models [Mukhoti et al., 2020]. Poorly calibrated models are not suitable for healthcare applications because the underlying predicted model probabilities must align with the true risk of disease shown within the population. For example, the models assign each individual a probability between 0 and 1 which is then mapped with the softmax function to obtain a class prediction. The probabilities of a well-calibrated model are closely related to the true instances of the disease outcome. This means that of the individuals assigned a probability of 60 percent, there will be 60 percent of those who truly had the outcome and the remainder did not.

2.5.1 Model Comparison across Life Course Patterns

The remaining experiments delved into the specific strengths and weaknesses of each model according to the type of life course pattern that the outcome was classified as. The life course patterns, each mimicking a real-world scenario, encompass a range of relationships including temporal, sequential, and inter-variable dependencies. The patterns within the period group are designed to link the exposures and the outcome according to a linear relationship. The patterns within the repeats group are related to consecutive exposure events occurring and can be categorised as incorporating relative temporal relationships when the outcome arises based upon the proximately of exposures to each other but not their precise timing. The patterns contained within the order and timing groups reflect a combination of temporal and inter-variable dependencies. This means that both the relative timing of exposure events and links between the timings of different features are key in the patterns that generate the outcomes. Therefore, the model performance within each life course group provides different insights into the kinds of patterns each model is capable of learning.

LR performs comparably to the DL models within the period category as it can identify linear relationships by construction. However, it performs significantly worse than the other approaches within the other life course groups, particularly the repeats category. Consequently, these results suggest that the LR model is not capable of identifying temporal relationships between the variables and cannot recognise patterns related

to the sequential nature of the data. The LR model was formulated so that each variable value at each time point acts as an independent variable and there is no additional structure in the model to connect this variable to past or future values. Therefore, exposures that happened consecutively are treated in the same way as those which happened many years apart by the LR architecture. As a result, the LR model predictably has inferior performance across all nonlinear patterns, indicating a lack of capacity for application within life course analysis.

The examination of model performance within the period group demonstrated that all the DL models maintained consistent and robust performance across linear scenarios, even in the presence of noise. The repeats group tests the models' ability to learn temporal dependencies and propagate information from early on in the time series since the repeated events may occur within the early years. Models with good performance have successfully learnt the sequential patterns and the temporal placement of exposures relative to each other. MLSTM-FCN performed exceptionally within this life course group across both metrics. This model feeds the data through both an RNN to lift temporal patterns and an CNN to identify inter-variable relationships. LSTMAttention also achieved good performance, possibly as a direct consequence of the attention mechanisms that form this architecture which are designed to enable the model to focus on influential time periods and exposures. All other DL models had reliable performance within the repeats group displaying an ability to learn sequential patterns which is essential for application within life course research.

The two remaining life course groups give rise to positive outcomes based upon interactions between multiple variables as well as sequential and temporal patterns. When measured using the AUPRC metric, ResNet has the highest performance within the order group for two scenarios and the lowest performance for the final pattern. In general, the model performance in the order group is highly dependent on the level of class imbalance. However, ResNet shows the greatest resilience to class imbalance within this group. In addition, ResNet and MLSTM-FCN generally perform well within the timing group. InceptionTime performed poorly within the timing group, which is unexpected as InceptionTime is composed of multiple convolutional layers which are designed to allow it to identify short- and long-term temporal patterns. One possible explanation is that InceptionTime is extremely sensitive to hyperparameter choice leading to subpar performance. No models obtain excellent performance on the Timing4 pattern, despite the lower level of class imbalance, which suggests that the complex relationships giving rise to the outcome are difficult for the models to identify.

There was a negative correlation between overall model performance and higher levels of class imbalance. The data was designed to mimic the significant class imbalance that is a common property of medical outcomes. It is important that models are resilient to imbalanced data and can determine which individuals are at higher risk of hospitalisation with RTI-related illness even though there may be a low occurrence within the wider population. Model performance suffers with increased class imbalance because there are fewer examples of positive instances for the model to learn. Additionally, as the majority class dwarfs the minority class, it is possible within many modelling frameworks for the model to achieve good performance by prioritising the majority class. In order to reduce the impact of class imbalance and force the models to put greater importance on fitting the minority class, multiple methods were used including focal loss which allows for difficult and infrequent outcomes to be more heavily weighted when training as well as multiple evaluation metrics designed to consider class imbalance. Some models were better able to handle imbalanced data than others. Notably MLSTM-FCN frequently was the best performing model in scenarios with large levels of class imbalance, suggesting that this architecture is more resilient to highly imbalanced outcomes. This result may arise due to the parallel modules within this architecture that each focus on a different type of relationship present in the data.

2.5.2 Balancing Performance and Computational Cost

The analysis of computational time across different ML models reveals valuable insights into their efficiency for life course analysis. Notably, InceptionTime emerged as the most computationally intensive model, requiring significantly more time to train than the other models. This substantial computational cost is primarily attributed to its complex architecture and the number of parallel convolutions involved. Then ResNet and LSTMAttention followed, with similar computation times that were approximately twice as fast as InceptionTime on average. The fastest DL model was MLSTM-FCN, making this architecture the most efficient choice. LR stood out as the fastest model by a significant margin, this is a result of the simple model formulation, however, this model performed poorly on nonlinear patterns, so the lower computation time comes at a cost. It is important to consider the time taken to train a model in addition to its predictive performance since often there are limited computational resources available to researchers. This is especially true within healthcare because the data is identifiable and therefore, must be stored on secure servers that are generally not designed for high level programming.

These findings shed light on the models' suitability for different life course scenarios, emphasising the importance of model selection based on the specific characteristics of the data and the patterns it exhibits. Primarily, the findings underscore the importance of aligning model choice with the specific life course patterns present in the data and the trade-off between model complexity and efficiency. While LR may serve adequately in simpler scenarios, its struggles with complex temporal and sequential dependencies which makes it less suitable for intricate pattern recognition. InceptionTime, while powerful in capturing complex patterns, may be better suited for applications where computational resources are not a limiting factor. The sensitivity of InceptionTime to the choice of hyperparameters additionally indicates that it would require extensive hyperparameter tuning which is not possible given the existing computational restrictions. In contrast, ResNet, MLSTM-FCN and LSTMAttention offer a balanced combination of performance and efficiency, making them strong candidates for real-world life course analysis tasks.

2.6 Limitations

The simulation study presented in this research has certain limitations that must be considered. Firstly, the patterns of exposures investigated are intentionally designed to mimic common relationships observed within life course studies. While these patterns serve as a foundational representation, they are inherently simplistic and in real life scenarios the interactions are likely to be considerably more complex. In theory, the relationships within real life data are composed of multiple of the patterns examined in this research which form building blocks to create the more intricate interactions. However, the models' ability to learn these simple patterns does not ensure that they can capture the more intricate and nuanced relationships that may exist in real-world scenarios. The true underlying mechanisms of health outcomes may be more elaborate and nuanced, making the model performance in this study a poor or inaccurate representation. However, it is worth emphasising that these patterns were derived from a comprehensive review of the literature. This suggests that these patterns are related to existing known life course patterns in some capacity. In addition, the formulation of the patterns captures model performance for temporal and inter-variable relationships. Future work could explore more elaborate and diverse patterns to evaluate whether models which excel at learning these foundational patterns are also able to handle more complex representations.

Secondly, the simulated data and outcomes in this study represent a highly simplified version of real-life data. While the simulated outcomes in this study serve as indicators of the models' capacity to learn basic patterns, they may not fully represent the complexity of real-life data. The simulated data only includes a small number of the numerous demographic factors and ACEs encompassed by the DANLIFE cohort. Furthermore, it

assumes independence among these characteristics, disregarding correlations that often exist in real-world data, such as the link between the parental education level and their social class. Experimentation demonstrated that the precise formulation of these characteristics had limited impact on model performance and the most important factors in creating representative data are noise and class imbalance. Ultimately, model performance is primarily influenced by the relationship generating the outcome from input data rather than intricate interactions among the input variables. To address some of these limitations, future research could involve the exploration of more complex patterns, the inclusion of continuous variables or counts, and more extensive simulations encompassing a wider array of model selections and hyperparameter exploration. In essence, while the simulation study provides valuable insights into DL models' performance on fundamental life course patterns, it is essential to recognise its limitations in representing the complexity and diversity inherent in real-world data and relationships. Consequently, further research and experimentation are needed to explore these aspects in more depth.

Moreover, the simulated data assumes a binary data format, while the DANLIFE dataset incorporates variables formulated as counts, such as counts of hospitalisations or ACEs. The binary representation limits the complexity of the simulated relationships that generate the outcome and may inadequately capture the complexity inherent in real-life data. Additionally, the event-based nature of the simulated data mirrors the DANLIFE data, where exposure events are recorded annually. This differs from traditional MTS data which the DL models are best suited for, which are generally more granular and continuous. Furthermore, the binary nature of the simulated outcomes introduces challenges associated with imbalanced data, particularly when positive outcomes are infrequent. While techniques to address imbalanced data are used, including the choice of the loss function and evaluation metrics, the optimum approach for handling this imbalance in the context of DL models is not thoroughly explored in this study. Often class imbalance is addressed using resampling of the data, however when developing a risk prediction model, these methods have been shown to be detrimental to model performance [Van Den Goorbergh et al., 2022]. This constitutes the first study of DL model performance on event data and further investigation into the impact of data format and class imbalance approaches on model performance is warranted.

A constraint that needs to be acknowledged is the computational limitations. Due to constraints in computational resources, the study could only execute a limited number of trials for each model. Some models possess a substantial number of parameters, making hyperparameter optimisation computationally intensive. Conducting an exhaustive study to explore all possible combinations and relationships among variables in the simulated data was computationally unfeasible. This meant that conducting a full gridsearch for all hyperparameters was not possible, after experimentation with the optimisation method the random sampler was chosen over the TPE sampler because the TPE sampler was found to frequently get stuck in local minima. However, the limited number of trials may mean that the sampler did not explore the full parameter space. In addition, it was not possible to run hyperparameter optimisation for multiple seeds and thereby reflect the uncertainty inherent in the model predictions. Although hyperparameter optimisation was performed, it is crucial to recognise that some models may be more sensitive to hyperparameters than others, and the results may not necessarily represent the absolute best performance for each model. Therefore, this study provides a baseline understanding of model performance within the explored parameter space but may not cover all possible scenarios. In future, many more runs should be conducted to ensure there is the time to run them for multiple seeds.

2.7 Conclusions

In this simulation study, the performance of a selection of DL models was compared across a range of different scenarios and metrics to assess their suitability for conducting life course analysis. The Chapter emphasises

the importance of using data simulation to critically assess and compare the performance of multiple ML models under controlled conditions and to gain insights into their strengths and limitations. The exploration of the impact of noise on model performance underscored the importance of model robustness for real-world applications. LR and InceptionTime displayed vulnerability to noise, whereas MLSTM-FCN and LSTMAttention exhibited resilience to noise. The examination of model performance in various LCC groups revealed nuanced results. The ability of models to identify temporal, inter-variable, and timing patterns varied, emphasising the importance of selecting an appropriate model for specific life course contexts. These findings underscore the models' sensitivity to the unique characteristics of life course patterns and inform a nuanced understanding of DL model selection for life course analysis, thereby contributing to the broader field of life course epidemiology.

There are multiple benefits of using DL methods over the traditional approaches employed within life course analysis. The data requires substantially less pre-processing in order to utilise the models and therefore, the longitudinal data retains its rich information about exposure events and their timings over the life course. Consequently, more intricate, and complex dependencies between the exposures are able to be detected. Better model performance means that the models have detected the underlying patterns in the data and will make more accurate risk prediction models. The methods described here are applicable to other instances of noisy and imbalanced data. For example, non-communicable diseases, such as cancers that usually develop over an extended period of time due to the accumulation of multiple exposures. These methods could act as an early warning system to identify individuals at a greater risk of becoming ill with certain cancers.

In light of the insights gained within this simulation study, the model chosen for further analysis on the DANLIFE data is ResNet. Its comparative performance to other DL models, combined with its efficiency and adaptability, aligns well with the goals of this analysis, especially given the limited computational resources available for working with the DANLIFE data set. ResNet has demonstrated its versatility and robustness in handling various life course patterns while being relatively efficient in terms of computational resources. The best performing model overall, MLSTM-FCN, has considerably more parameters requiring optimisation, meaning that whilst each model train has comparable computation costs to ResNet, when multiplied by the required number of trials the computational burden quickly becomes prohibitively large. The fact that ResNet has fewer parameters and is amenable to hyperparameter optimisation is especially valuable when working with data that is resource-intensive like the DANLIFE cohort. Therefore, the decision to select ResNet will allow for more extensive hyperparameter optimisation to be conducted and enhance model performance.

In summary, this Chapter provides a comprehensive understanding of DL model behaviour in the presence of different data characteristics and noise levels. By leveraging these findings, researchers and practitioners can make informed decisions when selecting models for life course analysis and develop strategies to address noise challenges in real-world applications. These insights will facilitate informed decision-making, ultimately contributing to the improved application of DL techniques in life course analysis and personalised healthcare interventions. The next steps in this body of research are assessing model performance on the DANLIFE data set itself and employing explainability techniques to gain insights into what specific risk factors drive RTI susceptibility and investigate strategies for mitigation and public health interventions.

Chapter 3

The Signature Method

3.1 Introduction

3.1.1 Background of the Signature Method

The Signature method is a feature transformation and dimensionality reduction technique that can be used to reduce the size of time series input data whilst still retaining the important patterns, which decreases computational burden and allows for the incorporation of temporal effects within a regression analysis. It originated within rough path theory, being used to describe geometries and approximate continuous functions [Lyons, 2014]. However, recently it has garnered significant attention within the ML and time series modelling fields as an accurate low-dimensional representation of data that is still able to capture the key patterns that drive the output. It has been applied to a range of fields, including handwriting recognition, gas network optimisation and financial market anomaly detection [Akyildirim et al., 2022; Kreimeier et al., 2022; Yang et al., 2016]. In addition, it has been used for health applications, including early identification of sepsis onset and classification of bipolar disorder [Morrill et al., 2020b; Perez Arribas et al., 2018]. The study on classifying bipolar disorder has many similar features to the DANLIFE data set, the input data are self-reported discrete scores that represent different scales of a range of emotions, e.g. elation, anger, or sadness, which is similar to the discrete counts of hospital admissions. Wu et al. chose to use the signature in order to handle the missingness present in their data and to reduce dimensionality. In this chapter, the challenges posed by discrete and sparse data sets, like the DANLIFE cohort, are explained along with why the Signature method may be a suitable alternative to the deep learning (DL) frameworks investigated in the previous chapter to reduce computational burden. A description of the Signature method is provided along with an explanation of the properties which make it promising for life course applications. Then a simulation study will be conducted to assess its suitability for conducting life course analysis.

3.1.2 Challenges of the Sparsity of Exposure Event Data

Time series data that records the occurrence of multiple exposure events can be very sparse. Sparse data arises due to the high dimensionality of the multivariate time series (MTS) data combined with the rarity of certain exposure events. For example, being below the poverty threshold is a risk factor for respiratory tract infections (RTIs) and the proportion of the population who have ever experienced a year of poverty within their first 16 years of life is very small. Further, the number of years spent below the poverty line for most individuals is also very small. When there are multiple exposure events with this rare nature then the input data becomes very sparse. To demonstrate this, imagine a cohort of 1000 individuals followed for the first 16

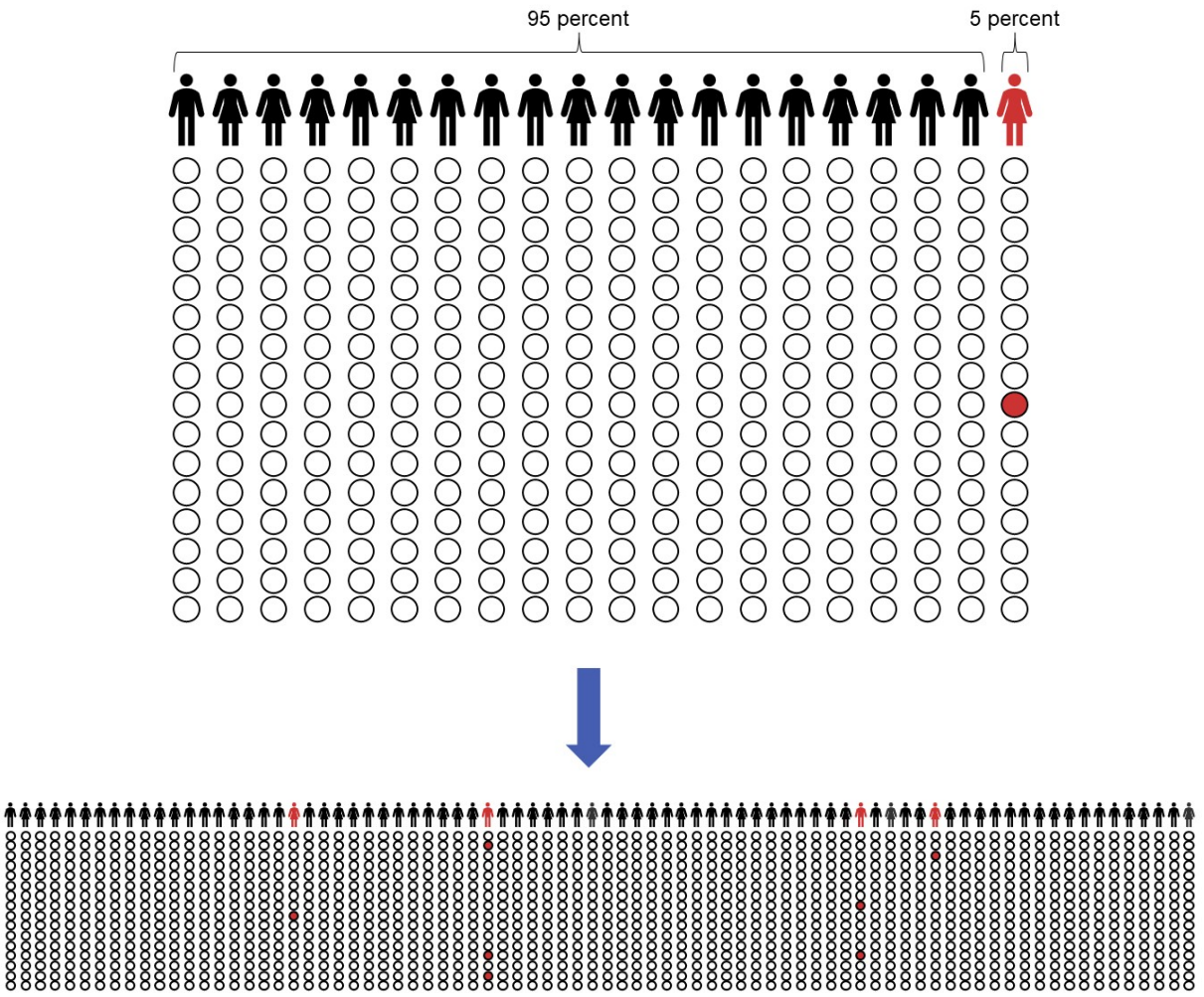


Figure 3.1: **Diagram demonstrating MTS data sparsity.** Illustration of sparse data with red individuals being those affected by a medical condition that impacts five percent of the population. Circles represent points in the time series and those filled in red represent when an exposure event occurs.

years of their life, there are records of a rare exposure affecting five percent of the population (Figure 3.1). If this exposure events occurs once only for 50 percent of people affected, twice for 30 percent and thrice for 20 percent of individuals then the input matrix is composed of $1000 * 0.95 * 16 + 1000 * 0.05 * 0.5 * 15 + 1000 * 0.05 * 0.3 * 14 + 1000 * 0.05 * 0.2 * 13 = 15200 + 375 + 210 + 130 = 15,915$ zeros compared to just $1000 * 16 - 15915 = 85$ positives. Consequently, over 99 percent of the data is zero which is a serious problem for many modelling methodologies. Due to the nature of life course data, where exposure events occur irregularly and each individual experiences a different number of total exposures, the data becomes sparse.

The reasons why traditional models struggle to model sparse input data are; most methods are overwhelmed by the vast number of covariates, the assumptions of independence start to break down, or they lack the ability to find temporal interactions within this data. As discussed in Chapter 1, traditional life course methods frequently depend upon the aggregation of data to reduce the sparsity and dimensionality. Generally, this is done by grouping the time dimension into broader time periods, combining multiple exposures into a single feature or clustering individuals according to their exposure histories. However, these population ensemble methods are not able to output individual-level risk scores that relate to that individual's unique life course. Even if only the time and feature dimensions are reduced, these aggregation methods lead to a loss of the richness of the longitudinal data. The grouping of different exposure types acts to equate the severity of each

exposure which may in actuality have differing magnitudes of impact on risk. Additionally, creating broadertime periods results in the loss of the temporal order of exposure events.

An alternative approach to traditional modelling methods is to use methodology designed for MTS data, including DL models such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs). These approaches, as discussed in Chapter 2, are powerful and capable of finding relationships within the data even when data is sparse. The drawback of DL models is that they represent a substantial computational burden. Moreover, these models may be assessing all the features at each time point unnecessarily when in reality there are only a small number of influential exposures that are relevant to the outcome. Therefore, in some cases it is preferable to use a dimensionality reduction method designed for MTS data that can extract the key patterns in the data and dramatically reduce the computational requirements both in terms of power and time taken.

Common methods for dimensionality reduction focus on reducing the number of features and time points by applying a transformation. Principle component analysis (PCA) is one prominent statistical method used to reduce the number of features or time points whilst retaining the important patterns. This method constructs new features from the old features such that the new features are linear combinations of the old features that are all independent and represent the greatest amount of variation in the data [Wold et al., 1987]. Another similar method is singular value decomposition (SVD) which decomposes the matrix of input features into multiple matrices that represent the eigenvectors and eigenvalues of the input data [Stewart, 2006]. These approaches share some of the same limitations, primarily they both result in a loss of interpretability as it is not possible to retrieve the original features. Moreover, the data is assumed to be Gaussian with linear relationships between the features. These assumptions are unlikely to hold in the case of longitudinal life course data since the inter-variable relationships can be complex and nonlinear. Both these approaches assume that the features which correspond to the greatest variance in the data are most significant, whereas in life course theory more subtle patterns may drive the outcome. Another popular approach to dimensionality reduction is wavelet transformations which convolve wavelet functions of different scales and frequencies to extract patterns present in the data at multiple granularities [Bruce et al., 2002]. One can select a subset of the wavelet functions to decompose the signal and reduce dimensionality whilst retaining the important patterns. Wavelet transformations also assume stationarity of the data which is unlikely with life course data when certain exposure events will have different frequencies at different ages. Autoencoders have been used for dimension reduction which feeds the data through an encoder to obtain a reduced form of the data before it is fed through a decoder and compared to the original data in order to fit the model. Autoencoders allow complex nonlinear relationships between variables and do not assume variance is the most important data characteristic. The disadvantage of this approach is the considerable computational burden [Wang et al., 2016a].

The Signature method has emerged as a promising approach for analysing MTS data, reducing dimensionality whilst not losing the key patterns that link the input data and the outcome [Lyons and Mcleod, 2023]. It can be used to represent and analyse complex paths and additionally resolves some of the main problems encountered when analysing MTS data. This method has not yet been investigated in relation to sparse data. The Signature method is essentially a dimensionality reduction technique that can be used in the same way as a feature transformation. The Signature method can capture sequential patterns in the data as well as temporal relationships and inter-variables interactions. This makes it appear suitable for life course analysis since it can retain temporal ordering information and is capable of learning dynamic nonlinear patterns between features. It is also able to handle missing data and differing lengths of time series for each individual. It is possible to recover the original features using the inverse signature transform which retains interpretability. The multiple advantages of the Signature method suggests that it may be a promising approach for conduct-

ing life course analysis, achieving reliable performance whilst avoiding the computational burden associated with DL approaches.

3.1.3 The Signature Method

In this section, the mathematics that underpins the Signature method is explained. The Signature method is composed of two parts; firstly, the signature of the input data is calculated and then the resulting matrix is fed through a logistic regression (LR) model. The signature can be viewed as a feature transformation which is a transformation applied to the input data prior to fitting the model. For example, within ML feature transforms are often used to reduce noise in the input data or lessen model sensitivity to scaling using normalisation. This description of the signature is based upon the work of Chevyrev et al. alongside the explanation provided by Morrill et al. [Chevyrev and Kormilitzin, 2016; Morrill et al., 2021].

Firstly, consider the MTS data associated with one individual, denoted by $X \in \mathbb{R}^{M\times T}$ where M is the number of features (referred to as the dimension) and T is the number of time points. For example, the features may be records of exposure events or ACEs the variables may be whether an exposure occurred, income or hospital admissions with specific diagnostic codes, and the number of time points is the number of years at which the variables are measured. For a ML framework, the data is a collection of samples $\{(X_i, y_i); i \in \{1,\dots,n\}, X_i \in \mathbb{R}^{M\times T}, y_i \in \mathbb{R}\}$ where X_i is the input data for an individual as above and y_i is the output variable. In the case of the DANLIFE data, y_i is a binary outcome indicating whether hospitalisation with an RTI-related illness occurred between ages 16 and 21. Within Rough Path Theory a data set X is thought of as a collection of multiple realisations $\mathbf{x}$ from a dynamical system at some time points $\{t_1,\dots,t_T\}$ [Lyons, 2014]. That is, using life course analysis as an example, there is some large system which governs how exposures encountered throughout life dictate your health status in later life of which the data is just a subset of all the possible trajectories and resulting outcomes. It is assumed that the data set is a representative subset that allows the patterns that link the exposures with given outcomes to be learnt. In other words, there is some function $f(\cdot)$ that links the input and the output in the form $y_i = f(X_i) + \varepsilon_i$ where ε_i is random noise associated with individual i. The Signature method proposes transforming X_i by augmentation and then taking its signature which gives $f(\cdot)$ the desirable property of linearity, i.e. the equation now becomes

$$y_i = f(S(\tilde{X}_i)) + \varepsilon_i \tag{3.1}$$

where $S(\cdot)$ is the signature transformation and $\tilde{X}_i$ is the augmented version of X_i which is continuous. The signature of the augmented input data, $\tilde{X}_i$, takes the form of a vector which is made up of multiple elements that are referred to as signature coefficients. The signature coefficients are the set of iterated integrals of $\tilde{X}_i$. The following formulae for the signature coefficients make use of notation introduced by Morrill et al. where $Z = \tilde{X}_i$ is used to denote the augmented input data [Morrill et al., 2021]. Moreover, and the simplification $Z_{t_j} = Z(t = t_j) = X_i(t = t_j)$ is used to denote the data for an individual i at a specific time point $t_j \in \{t_1,\dots,t_T\}$, i.e. with $j \in [1, T]$. For features $(s_1,\dots,s_k) \in \{1,\dots,M\}^k$, the corresponding signature coefficient $S(Z)^{(s_1,\dots,s_k)} \in \mathbb{R}$ is defined as

$$S(Z)^{(s_1,\dots,s_k)} = \int \cdots \int_{t_1\le u_1<\cdots<u_k\le t_T} dZ^{s_1}_{u_1} \cdots dZ^{s_k}_{u_k} \tag{3.2}$$

$$= \int_{u_1=t_1}^{t_T} \left(\int_{u_2=u_1}^{t_T} \left(\int_{u_3=u_2}^{t_T} \cdots \int_{u_k=u_{k-1}}^{t_T} dZ^{s_k}_{u_k} \cdots dZ^{s_3}_{u_3} \right) dZ^{s_2}_{u_2} \right) dZ^{s_1}_{u_1} \tag{3.3}$$

The signature coefficients make up the Signature truncated at level K according to the following equa-

tion

$$\text{Sig}^K(Z) = (\{S(Z)^{(s_1)}\}_{s_1=1}^M, \{S(Z)^{(s_1,s_2)}\}_{s_1,s_2=1}^M, \ldots, \{S(Z)^{(s_1,\ldots,s_K)}\}_{s_1,\ldots,s_K=1}^M) \in \mathbb{R}^{\frac{M^{K+1}-1}{M-1}} \tag{3.4}$$

The use of K to truncate the signature improves computation but reduces the amount of information the signature captures about the sequence. If truncation level, K, is chosen to be equal to one, then the equations simplify:

$$\begin{aligned}
\text{Sig}^1(Z) = \{S(Z)^{(s_1)}\}_{s_1=1}^M &= \left\{\int_{u_1=t_1}^{t_T} dZ_{u_1}^{s_1}\right\}_{s_1=1}^M && (3.5)\\
&= \{Z_{t_T}^{s_1} - Z_{t_1}^{s_1}\}_{s_1=1}^M && (3.6)\\
&= \{Z_{t_T}^1 - Z_{t_1}^1, \ldots, Z_{t_T}^M - Z_{t_1}^M\} && (3.7)\\
&= Z_{t_T} - Z_{t_1} \in \mathbb{R}^M && (3.8)
\end{aligned}$$

This formulation of the equation shows that when the truncation level is equal to one, the signature is equal to the change in each feature between the start point and the end point of the time series. Similarly, when K is set to two then the second order terms correspond to the Lévy area (very roughly the area above and below a line) under the curve which holds similar properties to the variance. Generally, K must be below ten to lower the computational burden so that it is feasible to run, however the signature length also depends on the number of features in the data. It is essential to select a truncation level that minimises the computational requirements whilst still capturing enough detail about the sequence to conduct meaningful inference. The signature depth can be expressed as a summation as shown in the equation:

$$\text{signature depth}_K = \sum_{k=0}^{K} M^k. \tag{3.9}$$

This demonstrates that the number of features and the truncation level both greatly affect the length of the signature. The signature does not require each individual to have the same number of temporal measurements because the length of the truncated signature is not dependent on the number of time points. Therefore, the output is the same length for all individuals regardless of the length of the input time series. Consequently, it is possible to input signatures calculated from time series of varying length into the same model and this approach naturally accounts for data recorded with different temporal intervals and for assorted sequence lengths. Additionally, notice that the signature is one-dimensional, and the single dimension captures information from both the features and their timings.

The definition of the signature requires the data to be continuous - in rough path theory terminology the input data for each individual i, X_i, is required to be a path. A path is a continuous mapping from some overarching temporal interval into $\mathbb{R}^{d\times m}$. However, in reality, time series are often measured at a discrete number of points and therefore, require augmentation and interpolation in order for them to approximate a continuous path. As set out by Morrill et al., there are a number of different transformations that are used to give the data desirable properties and make it suitable for the signature transformation [Morrill et al., 2021]. They classify transformations within four categories; augmentations, windows, transformations and re-scaling. Augmentation methods can be performed in addition to the signature transformation with the goal of introducing desirable properties to the signature, such as removing the signature invariance to temporal translation or lowering the dimensionality of the input data. Examples of augmentation methods include time augmentation, basepoint augmentation and the lead-lag transformation (Section 3.2.2). Windows are a method used frequently within other forms of time series analysis, in which a window of assorted sizes are moved across the data to capture different levels of temporal patterns (Appendix 8.3). The transform category

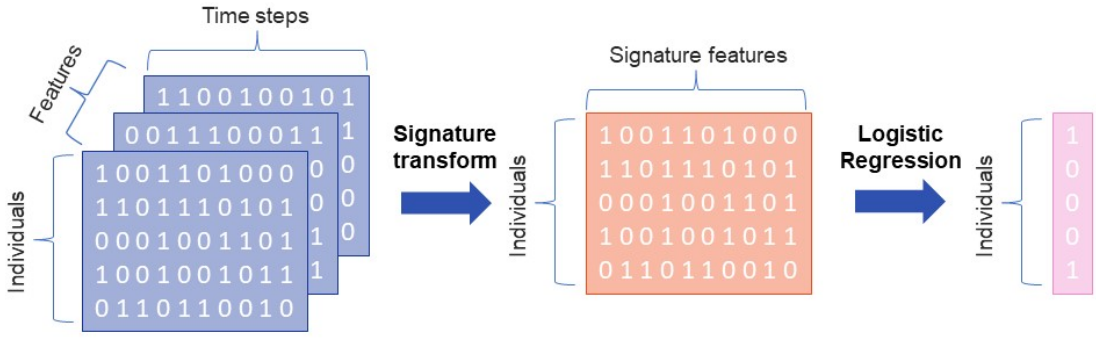


Figure 3.2: **The Signature method.** A diagram showing the steps taken in the Signature method. The input data that has been augmented is transformed using the signature transform before it is passed to a logistic regression or ML model to obtain a binary output.

of augmentations encompasses the signature and log signature transforms. Rescaling takes the same form as pre-processing the data and includes things like standardisation or normalisation of the different features to prevent ones with larger scales dominating the signature calculation. The primary challenge is making the discrete input data continuous. In practice, the input data can be made to appear continuous by using various interpolation methods; generally either linear interpolation, piece-wise linear interpolation or rectilinear interpolation [Chevyrev and Kormilitzin, 2016].

The signature has several mathematical properties that make it suitable for handling multivariate data and life course analysis. Firstly, if Z contains at least one monotone feature, i.e. a coordinate that increases between every time point, then the signature is a unique representation of Z up until translations [Chevyrev and Kormilitzin, 2016]. This property means that the signature manages to maintain interpretability as each Z corresponds to a unique signature and therefore, it is possible to use the inverse signature transform to obtain the original data from its signature. A monotonically increasing coordinate can be engineered by using the time augmentation to introduce a temporal coordinate which increases naturally. The next property of the signature is that it is unchanged if the input is reparametrised, in practice this means that if t is altered in some continuous way and remains monotonic then the signature is unaffected. The final property relevant to life course analysis is that any continuous function of Z can be represented as a linear function of the signature coefficients of Z [Chevyrev and Kormilitzin, 2016]. In other words, it is possible to assume that the function $f(\cdot)$ from equation 3.1 is linear and therefore, linear regression can be used to predict y_i. This means that the signature has a lower computational burden than ML methods and is also fully interpretable. This property also implies that for large enough K, the signature acts as an arbitrarily good representation of the input features.

The Signature method is composed of three steps (Figure 3.2); augmentation, signature transform and regression. Firstly, the raw input data is transformed to make it continuous and to add the monotone coordinate required for the uniqueness property to hold. Then the signature transformation is applied to the augmented data for each individual to obtain a matrix where the rows correspond to individuals and the columns correspond to the signature coefficients, which depends on the truncation level K and the number of features in the augmented data. It is important that K is selected carefully to ensure that the uniqueness and linearity properties hold, and the signature provides a sufficiently detailed representation of the data. The last step is using the signature coefficients as variables within a linear or logistic regression to produce estimates of y_i and to fit the model.

In the context of life course analysis, the Signature method is a promising approach with the ability to capture temporal evolution within MTS data and the capacity to identify patterns within individuals' trajectories leading to specific health outcomes. In theory, the signature can retain key features in the data whilst reducing dimensionality and lowering the computation requirement compared with ML methods. It is also particularly beneficial when dealing with complex and irregular life course trajectories, where traditional statistical methods may fall short and model assumptions may not hold.

3.1.4 Research Questions of the Chapter

The effectiveness of the Signature method in conjunction with LR for modelling longitudinal life course data remains an open question. Whilst it has shown promising results within other healthcare applications, it depends upon the input data being continuous. Whether the interpolation and augmentation methods are sufficient to make life course data appropriate for the Signature method is unknown. The nature of life course patterns that link exposures with health outcomes are complex, meaning that whilst the signature may be able to represent patterns within the variables, the deeper, hidden life course patterns may be lost using this transformation. In this chapter, the utility of the Signature method for longitudinal life course data will be examined as well as the impact of the truncation level and augmentation methods on model performance. The main research questions are

1. How does the setup of the Signature method, including truncation level and augmentation methods used, impact the predictive performance?
2. Does the Signature method accurately capture the life course patterns present in the data?
3. Is the Signature method a viable alternative to computationally expensive DL approaches?

By systematically comparing different data transformation, interpolation, and truncation techniques for signature feature extraction, this chapter aims to determine the optimal approach that enhances the interpretability and utility of signature features in the context of life course analysis.

3.2 Data Preparation and Experimental Setup

3.2.1 Data Preparation

The experiments are conducted on synthetic data that allows for a systematic and controlled comparison of Signature model performance. This is the same simulated data used in Section 2.2 of Chapter 2, which was designed to resemble the characteristics of real-life life course data. Here the key details are briefly recapitulated, and the reader is referred to the previous description for a more in-depth explanation of the pre-processing procedure and rationalisation of the methods used. Each individual $i \in N$ has associated input data X_i which takes the form $X_i = \{x_{i,1}, \ldots, x_{i,M}\}$ where each $x_{i,j} \in \mathbb{R}^T$ is a time series and output $y_i \in [0, 1]$. In this instance, the number of features is $M = 9$, the number of time points is $T = 40$ and the number of individuals is $N = 100,000$. The input data is comprised of five time invariant individual characteristics and four time dependent variables designed to mimic exposure events. Individual characteristics encompass feature that are designed to reflect multiple individual-characteristics, such as sex, parent ethnicity, marital status, mother's highest education and social class. The time dependent exposure variables mirror the patterns seen in features such as sibling illness, poverty, parental alcohol abuse and mild illness. Each variable in the input data is binary indicating whether at least one exposure event occurred in each year.

The output variable is designed to reflect hospitalisations due to RTI-related illness in the following five years

after the observation period. This is generated to represent the occurrence of life course patterns that are categorised into four groups each typifying a different life course concept. Each life course pattern represents a specific type of relationship or dependency between the input variables and their timing that gives rise to positive outcome values. The four life course concepts are referred to as period, repeats, order, and timing. The dependencies within these groups contain multiple types of interactions, including linear relationships, temporal dependencies, and inter-variable patterns. For each life course concept, multiple rules were used to generate the outcome from the same input data, this resulted in multiple outcomes assessed within this analysis. These rules served as a computational representation of the underlying mechanisms believed to influence health outcomes. To account for the inherent uncertainty and variability in real-life data, noise was introduced within the simulation process. Noise in this context refers to the presence of random variations or errors in the data, which can mimic the uncertainty and imperfections often observed in real-world data. The noise is only added to the training set so that the model can achieve perfect performance on the test set if it has correctly identified the underlying patterns One common challenge when conducting public health research is that the outcome is imbalanced or zero-inflated since the medical outcome is rare. This characteristic was recreated in the simulated data by selecting rules that gave rise to outcomes with fewer than six percent being positive.

3.2.2 Data Transformation Techniques

There are multiple transformations that can be applied to the data prior to the signature transform that can give the data desirable properties and improve the performance of the Signature method. These are rescaling, interpolation and augmentations. Rescaling is commonly used in ML methods to ensure the features lie on the same scale so that variables with larger values, such as year of birth, do not dominate the analysis. This can be achieved using normalisation or standardisation [Shanker et al., 1996; Sola and Sevilla, 1997]. Rescaling is not required within this analysis as the simulated data is binary.

The Signature method is defined for continuous paths of data and whilst it is possible to apply the signature transform to discrete data, the model assumptions may not hold, and model performance may suffer. There are many methods that can be used to interpolate the data and make it more closely resemble continuous data [Morrill et al., 2021]. The most common method used with the signature is piecewise linear interpolation which joins each time point in the data to the next with a linear polynomial. This avoids having to fit a global curve through all the points and can be done with higher order polynomials, although it is discontinuous at each timepoint. Once an interpolation function has been created which goes through all the time points then it is possible to sample from this function and select the new number of timepoints, with more timepoints more closely approximating a continuous path. The benefit of using the Signature method is that these additional interpolated time points do not increase the length of the signature and so do not add additional computational burden to the LR. However, there are additional computation requirements associated with calculating the signature coefficients when the input data has more time steps.

It is common for different augmentation methods to be used to improve the performance of the Signature method. Therefore, this analysis will explore the most frequently used augmentation methods to assess their relationship to model performance when applied to life course data. The augmentation methods that will be investigated are basepoint augmentation and Lead-Lag augmentation [Morrill et al., 2021]. An additional augmentation called time augmentation will be applied to all input data as it is essential. Time augmentation involves adding an additional covariate to the input data that takes the form of time points, i.e. for $X_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m}\} \in \mathbb{R}^{d \times m}$ with $\mathbf{x}_{i,} \in \mathbb{R}^d$ measured at timepoints $\{t_{i,1}, \ldots, t_{i,m}\}$ then the transformed sequence

becomes

$$\phi_t(X_i) = ((t_{i,1}, \mathbf{x}_{i,1}), \ldots, (t_{i,m}, \mathbf{x}_{i,m})). \tag{3.10}$$

The use of the time augmentation gives the signature the desirable property of uniqueness up until translation, allowing it to be used as a feature transformation. It is worth noting that the time augmentation adds an additional dimension and therefore increases the length of the signature. A further augmentation method is basepoint augmentation, which adds an origin point to the beginning of the paths. This grounds the path and removes the property of translational invariance from the signature, which means that the signature is no longer the same for two identical patterns of exposures that occur at different times. In this way the timing of the exposure events within the life course are now represented in the model. The definition of basepoint augmentation is

$$\phi_b(X_i) = (0, \mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m})). \tag{3.11}$$

Without the basepoint augmentation, two life courses will have the same signature if they have the same patterns of exposures but at different time points. In some instances, this may improve model performance as it may allow the model to better identify temporal relationships that link exposures regardless of when they occur. For instance, the repeats pattern gives positive outcomes when two exposure events occur consecutively regardless of when in the life course this happens. The second augmentation investigated is the lead-lag augmentation which is designed to introduce short- and long- term temporal dependencies into the signature. It is defined as follows

$$\phi_{LL}(X_i) = ((\mathbf{x}_{i,1}, \mathbf{x}_{i,1}), (\mathbf{x}_{i,2}, \mathbf{x}_{i,1}), (\mathbf{x}_{i,2}, \mathbf{x}_{i,2}), (\mathbf{x}_{i,3}, \mathbf{x}_{i,2}), (\mathbf{x}_{i,3}, \mathbf{x}_{i,3}), \ldots, (\mathbf{x}_{i,m}, \mathbf{x}_{i,m})) \tag{3.12}$$

This transform doubles the number of features and therefore the signature depth is increased substantially. The inclusion of this augmentation may improve the model's ability to identify inter-variable temporal dependencies. It is possible to use multiple augmentation methods before computing the signature so experiments will be conducted to see how model performance is impacted.

3.2.3 Experiment Setup for Signature Method Evaluation

The performance of the Signature method will be compared with multiple arrangements of augmentations explored as well as the truncation level. There are a total of 28 data sets which are composed of 14 individual life course patterns at two levels of noise (0 and 10 percent). The life course patterns each fall within one of the four life course concept groups: four rules in the period group, three rules in the repeats group, three rules in the order group and four rules in the timing group. The nature of the simulated data means that it does not require any rescaling or other pre-processing methods to prepare it for the models. The input data and output were split into train and test sets with an 80:20 ratio that was the same for all life course patterns. The time augmentation and any further augmentation transforms are applied to the data before the signature transform is applied. The resulting signature coefficients are fitted using LR model as described in Section 2.1.2. The LR model uses class-balancing to more heavily weight the minority positive class and improve performance and used l1-regularisation to prevent overfitting. The entirety of the train set was used to fit the LR, and performance was tested on the test set. The four Signature models explored are:

1. **Basic signature:** Time augmentation alone
2. **Basepoint signature:** Time and basepoint augmentation
3. **Lead-Lag signature:** Time and lead-lag augmentation

4. **Full signature:** Time, basepoint, and lead-lag augmentation

The four signature augmentation formulation models are applied for three different truncation levels: $K = 1$, $K = 2$ and $K = 3$. The larger the truncation level, then the more detail about the patterns in the data the signature will contain. Larger truncation levels increase the number of coefficients in the signature, which also increases the computational time and memory required. Consequently, it was not possible to examine larger truncation levels due to limited computational memory that means that these truncation levels are not able to be studied using the compute resources available. The lead-lag augmentation doubles the number of features in the input data and so the signature depth increases from $\sum_{k=0}^{K} d^k$ to $\sum_{k=0}^{K} 2^k d^k$, increasing computational time further. Another model formulation aspect that will be investigated in this analysis is the impact of interpolation on model performance. This shall be done by using piecewise linear interpolation to increase the number of time points by a factor of ten. The resulting input data sets will be compared to the raw data to assess whether the closer approximation to continuity improves model performance. This analysis was coded in python using the `signatory` package.

Within these experiments, first the impact of the choice of truncation level and augmentation formulation on the Signature method's performance across three evaluation metrics will be examined. Then the level of noise and how this degrades model performance will be assessed, followed by a detailed exploration into the individual life course patterns. The aim of this part of the analysis is to learn how the different augmentation strategies perform for the distinct life course pattern groups and whether certain methods can better capture the complex temporal dependencies and characteristics of each pattern. Finally, the effect of interpolation on predictive ability will be highlighted before the relative computational burden associated with each model formulated is compared. Similarly to Chapter 2, the predictive performance of the Signature models was examined using three performance metrics; AUROC, AUPRC and the Brier score (described in Section 2.3.2). Each metric provides different insights into model performance, especially in the case of class imbalance, and provide a holistic evaluation of model suitability for life course studies. Computation time was also recorded to compare this method with the computational costs of the DL approaches explored in Chapter 2. The computation time of the Signature method is split into two periods; the time taken to compute the signature coefficients and the time taken to train the LR model on the signature features. The impact of the interpolation and truncation levels on the two time periods was investigated.

3.3 Evaluating the Signature Method's Performance

In this section, the multifaceted factors that shape the performance of the signature approach and its suitability for life course analysis are explored.

3.3.1 Impact of Increasing Truncation Level on Model Performance

The truncation level (K) is possibly the most significant parameter within the Signature model formulation. This parameter controls the complexity of the model and the type of interactions in the data that can be reflected. At each level of truncation, K, the signature contains all the same features as the signature with truncation level $K - 1$ as well as extra features that represent higher-order interactions. Therefore, within life course applications a higher truncation level should mean the signature better reflects the temporal and inter-variable relationships present in the data. However, higher truncation levels lead to a significant computational burden and a greater likelihood of overfitting. Therefore, it is important to carefully select the truncation level. Here the performance of each formulation of the Signature model (basic, basepoint, lead-lag and full) is assessed for a range of truncation levels with no interpolation method used.

AUROC and AUPRC Performance

Increasing the truncation level consistently improved the AUROC score across all the Signature models (Figure 3.3, top). The addition of second-order features to the basic model has a smaller impact on model performance than the further addition of third-level features on average. The median AUROC jumps from 0.515 to 0.705 to 0.897 as K increases for the basic Signature model. In contrast, the full model, which incorporates more temporal dependencies by design, has similar AUROC scores for truncation levels two and three meaning that the third-level features do not have a substantial impact on model performance. The median AUROC values go from 0.510 to 0.900 to 0.914 as K increases for the full model. The models with more augmentations generally have improved performance across the different truncation metrics, with this pattern especially noticeable for $K = 2$. Moreover, there is a greater range in AUROC scores for the simpler models across the different data sets. This will be explored in greater detail in the next section of the analysis. Overall, for $K = 3$, the models all seem able to obtain high AUROC values with median values varying between 0.897 for the basic model and 0.919 for the lead-lag model.

For the AUPRC metric, the overall trend is similar to the AUROC evaluation with increased levels of the truncation parameter leading to higher median AUPRC scores for all Signature models (Figure 3.3, bottom). However, there are some key differences. Most notably, the magnitude of the AUPRC scores indicate lower model performance overall with low AUPRC values even for the highest truncation level $K = 3$. The median values at $K = 3$ ranged between 0.120 for the basic model and 0.157 for the basepoint model which corresponds to poor model performance. This means that whilst there are very few false negatives there are a high proportion of false positives leading to low precision. This discrepancy can be attributed to the nature of the AUPRC metric, which places greater emphasis on correctly classifying the negative class than the AUROC metric. Potentially this indicates that $K = 3$ is not sufficient to be able to distinguish positive and negative cases to balance precision and recall.

Brier Score Evaluation

The brier score metric evaluates the calibration and reliability of the model, with well-calibrated models having lower brier scores. It does this by assessing how the predicted model probabilities align with the true probabilities of the health outcome. This is especially important within healthcare applications as any models that are used as decision-making tools need to be trustworthy. The brier score shows a similar trend to the AUROC and AUPRC evaluation metrics in that the higher truncation levels improve model calibration since they have lower brier scores. The lowest level of truncation, $K = 1$, has the greatest range of brier score for all models (Figure 3.4). Most model formulations with truncation level $K = 3$ are well-calibrated. The median brier score for $K = 3$ ranges between 0.107 for the basic model and 0.074 for the full model.

The result across all the evaluation metrics, indicate that the highest truncation level, $K = 3$, is required to optimise model performance. The AUROC scores for the full Signature model appear to start levelling off which indicates that greater truncation levels would likely not further improve model performance. Therefore, for the remainder of this analysis the models will be compared using truncation level $K = 3$.

3.3.2 Signature Method Performance with Noise

It is important that models can perform well on data that has a realistic level of noise that better reflects the random and systemic noise that masks the underlying patterns in the data. The model performance has been investigated in the deterministic case (with no noise) and the case where there is 10 percent of noise. It is thought that these scenarios will provide a good overview of how noise impacts the model performance and the influence of other noise levels can be extrapolated from these results. The different signature augmenta-

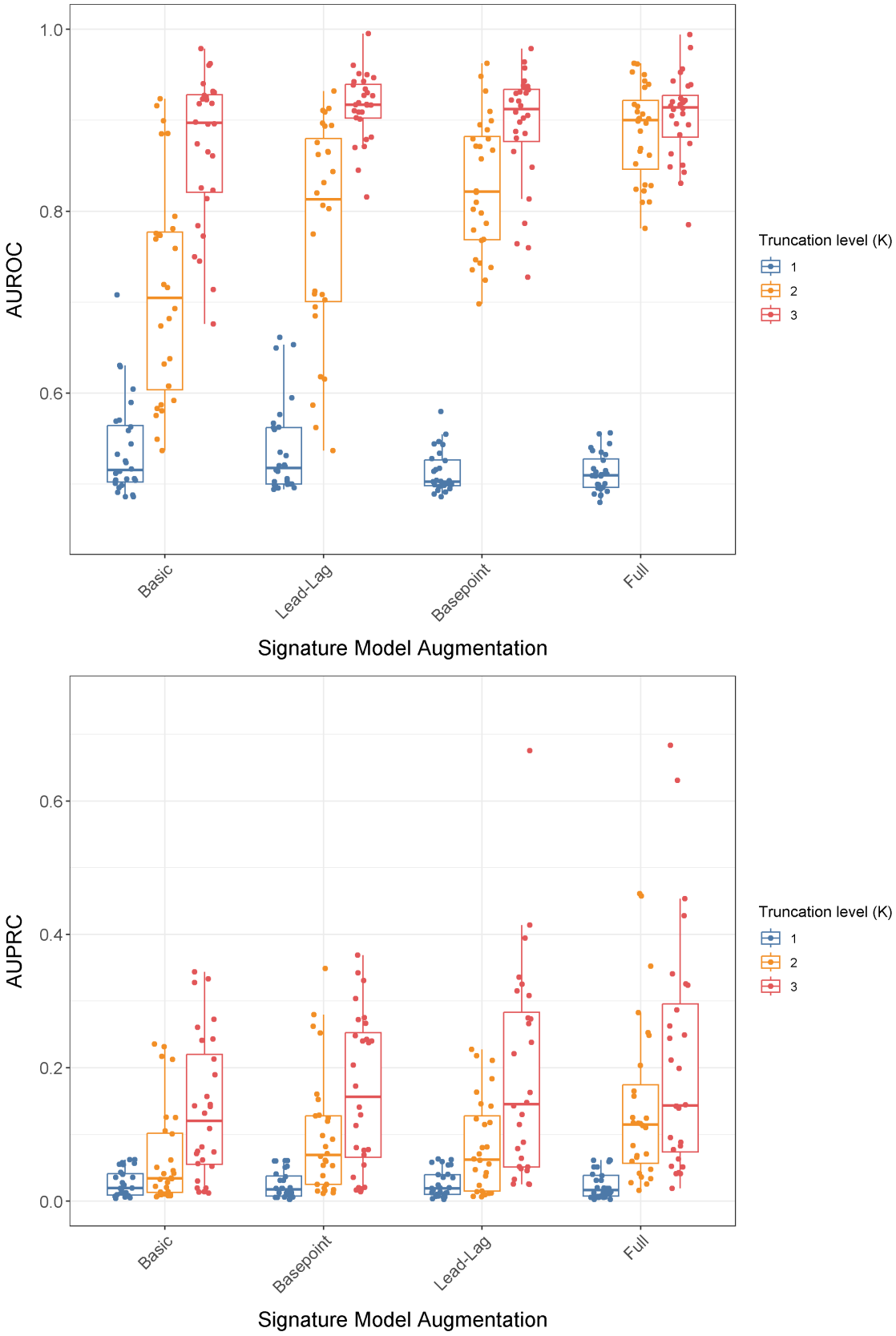


Figure 3.3: **Predictive performance of Signature models for different truncation levels.** Values of evaluation metrics AUROC (top) and AUPRC (bottom) for each outcome at both of the two levels of noise (0 and 10 percent) for the four Signature models (basic, basepoint, lead-lag and full) at three truncation levels ($K = 1, 2, 3$). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. No interpolation transform used.

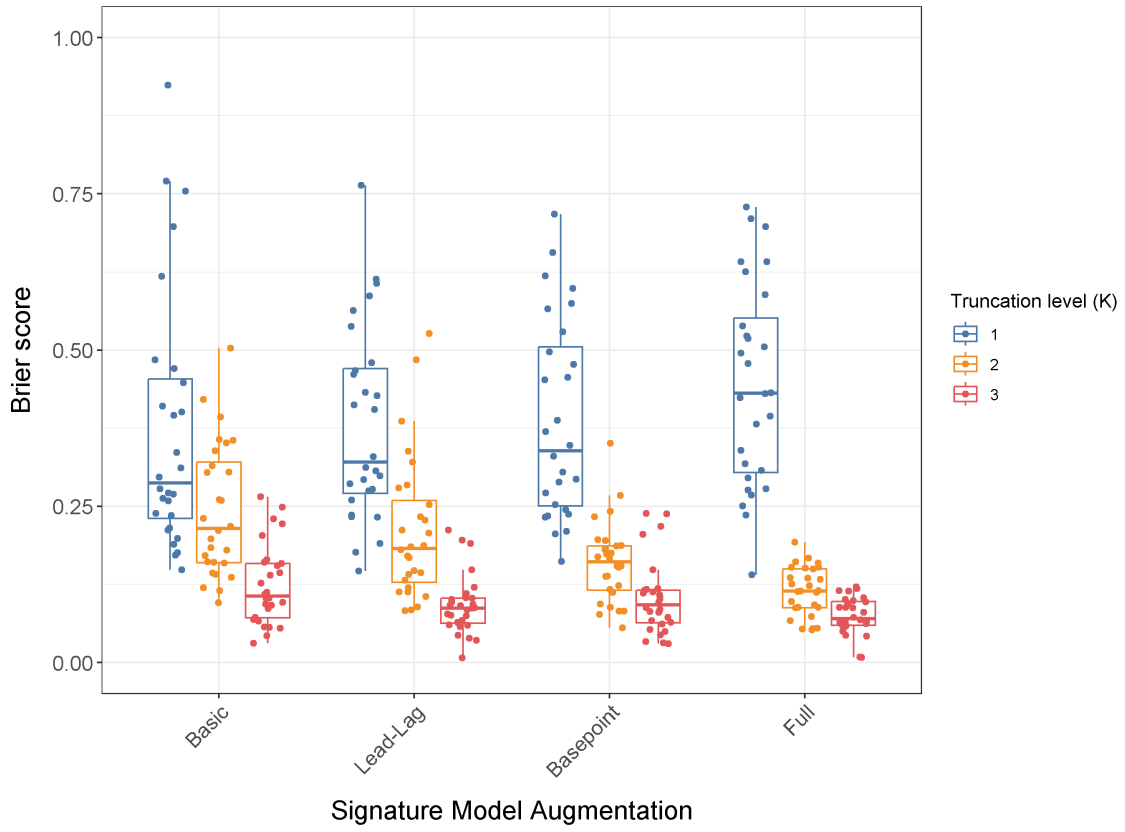


Figure 3.4: **Calibration performance of Signature models for different truncation levels.** Brier scores for each outcome at both of the two levels of noise (0 and 10 percent) for the four Signature models (basic, basepoint, lead-lag and full) at three truncation levels ($K = 1, 2, 3$). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. No interpolation transform used.

tion levels are examined and compared to the LR model which acts as a baseline model.

The addition of noise degrades the model performance of all models when measured by the AUROC metric (Figure 3.5, top). The LR and basic Signature models are most impacted by the increased noise levels with drops in median AUROC score of 0.061 and 0.055, respectively. The model that is least affected is the lead-lag Signature model which only decreases by 0.018 so this is the most robust model. The Signature models all outperform the LR model overall, however, the LR model appears to outperform the Signature models for some data sets. This will be explored in depth in the next section.

The AUPRC metric has quite a different trend to the AUROC metric (Figure 3.5, bottom). The LR model outperforms all the Signature models on average, particularly at the 10 percent noise level with the highest median AUPRC score of 0.230. This value is greater than the highest performing Signature model at 10 percent noise which is the full model with a median AUPRC score of 0.124. The Signature models degrade far more with the addition of noise when using the AUPRC metric. The model most affected by the noise is the basepoint model with a substantial drop of 0.120 in the median AUPRC score. This demonstrates that the addition of noise to the data has a big impact on model performance and therefore, with the rest of the analysis performance will be assessed for data sets with 10 percent noise to better reflect realistic scenarios.

3.3.3 Performance of Signature Augmentation Methods on LCPs

The next part of the analysis is centred around analysing the performance variations among the different signature augmentations (basic, basepoint, lead-lag and full) in comparison with LR on the individual data sets which each fall within a broader life course concept group. This analysis highlights how the choice of augmentation method impacts the quality of the signature features and their suitability for identifying distinct forms of temporal and inter-variable relationships within the data.

Within the period group, there is a clear pattern present in the relative model performance (Table 3.1, period). The LR model achieves exceptionally good performance within this group, particularly for Period1 and Period2 rules where it outperforms all the Signature model formulations and obtains almost perfect scores in both evaluation metrics, most notably with the AUPRC scores which are over 98 percent. All the patterns in the period group are linear and therefore, it is expected that LR obtains the best overall performance. All the signature-based models perform similarly to each other across the different data sets. They perform very well when measured with the AUROC metric displaying a capacity to identify linear relationships with all scores above 80 percent. However, the AUPRC metric is lower, especially for Period3 and Period4. The AUPRC metric seems to be affected by the level of class imbalance, with the most imbalanced data sets receiving the lowest AUPRC scores. Of the four Signature model formulations, the lead-lag model has the best overall performance when measured using the AUROC metric.

The first thing to notice in the repeats group is that the LR model performs very poorly and has very low AUROC and AUPRC scores across all simulated patterns (Table 3.1, repeats). LR has AUROC scores that are close to 50% which indicates random performance and that the model has completely failed to identify the significant patterns. Using the AUROC metric, in comparison to LR, the Signature models all obtain good performance with scores above 70%. Overall, the full Signature model performs best in the repeats category. The full signature is a big improvement on the basic signature with AUROC scores approximately 20% greater for the Repeats1 and Repeats2 data sets. In fact, the inclusion of the lead-lag augmentation significantly improves performance and generates the biggest increase in AUROC scores across the repeats group. The full Signature model obtains a AUPRC score of 42.8% for Repeats1 which drops to 4.1% for the other two data sets in this group.

Within the order group, the Signature models all perform well with AUROC scores above 85 percent (Table 3.1,

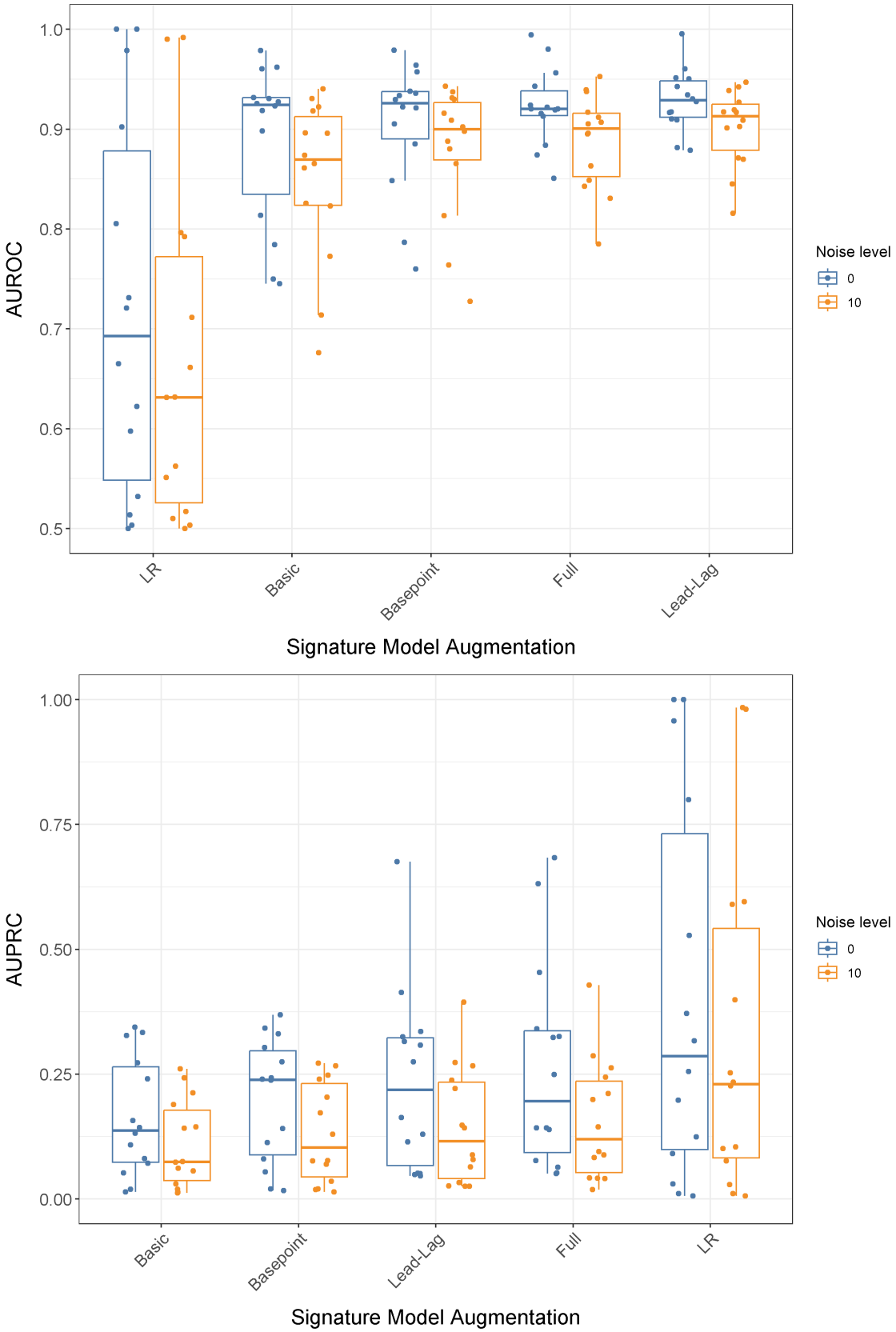


Figure 3.5: **Predictive performance of Signature models for different noise levels.** Values of evaluation metrics AUROC (top) and AUPRC (bottom) for each outcome at two levels of noise (0 and 10 percent) for the four Signature models (basic, basepoint, lead-lag and full) and logistic regression (LR). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. Truncation level $K = 3$ and no interpolation transform used.

Table 3.1: **Relative model performance within four LCPs; period, repeats, order, and timing.** Each row gives information about a rule including the LCP it falls into, an identifying number, the proportion of the resulting outcome that is positive and its definition. If the rule is met, then the outcome is positive. For each model, the AUROC and AUPRC values are provided as percentages and the highest performing model for each data set is highlighted in bold font. Ten percent noise in the data and no interpolation method used. The performance of the four Signature models (basic, basepoint, lead-lag and full) and logistic regression (LR) is assessed.

Group		Positive	Positive Criteria	LR		Basic		Basepoint		Lead-Lag		Full	
		%		AUC	PRC	AUC	PRC	AUC	PRC	AUC	PRC	AUC	PRC
Period	1	1.9	x_9 has no events in first 30 years	**99.0**	**98.1**	94.0	14.2	93.1	12.9	94.2	14.8	93.9	14.4
	2	3.1	x_9 has 4+ events in first 10 years	**99.2**	**98.4**	92.2	18.9	93.0	20.4	92.7	22.1	91.7	19.9
	3	0.6	x_9 has 5+ events in first 10 years	79.6	**59.5**	**87.4**	3.0	81.3	3.5	87.1	2.5	83.1	4.3
	4	1.3	Sum of weighted x_9 events is greater than 4	79.2	**59.0**	86.1	5.6	89.8	7.7	**90.3**	7.9	89.5	9.5
Repeats	1	6.0	x_9 has a repeat of 3 events in a row anywhere	51.7	7.6	77.3	14.5	86.5	24.0	94.7	39.4	**95.3**	**42.8**
	2	0.7	x_9 has a repeat of 4 events in a row anywhere	50.3	1.1	67.6	1.4	76.4	2.0	81.6	2.6	**84.9**	**4.1**
	3	0.6	x_9 has 2 repeats of 2 events in a row anywhere	50.0	0.6	71.4	1.2	72.8	1.4	**84.5**	2.6	84.3	**4.1**
Order	1	4.9	The first 2 events of x_8 occur before the first event of x_9	71.1	**39.9**	**91.8**	26.0	91.6	26.7	91.7	26.6	91.2	26.3
	2	1.2	The first 3 events of x_8 occur before the first event of x_9	63.2	**22.7**	89.6	7.4	88.8	6.9	**91.7**	8.8	89.6	8.3
	3	0.2	The first 4 events of x_8 occur before the first event of x_9	55.1	**10.4**	86.5	2.0	**90.2**	1.9	90.1	3.3	90.5	1.9
Timing	1	1.9	Any 2 x_8 events occur within the same year as any x_9 event	51.0	2.9	82.6	7.5	**93.7**	17.2	91.9	14.3	86.3	**21.1**
	2	5.4	Any 3 x_8 events occur within 2 years of any x_9 events	63.1	23.4	89.6	24.2	**90.9**	27.2	**90.9**	27.3	90.7	**28.6**
	3	1.3	Any 4 x_8 events occur within 2 years of any x_9 events	56.3	**10.1**	82.3	6.2	**88.0**	7.6	87.0	6.4	78.5	8.8
	4	3.5	Any 4 x_8 events occur within 4 years of any x_9 events	66.1	**25.2**	92.8	20.8	**94.3**	24.7	93.8	23.7	93.8	24.4

order). There is little variation in model performance measured by AUROC across the three different patterns, despite changes in the levels of class imbalance. However, when measured with the AUPRC metric, the LR model outperforms all the Signature models. The AUPRC scores across all the models and patterns are low with none being greater than 40% signifying poor performance. All models have especially low AUPRC scores within the Order3 pattern which is also the pattern with the highest level of class imbalance.

The Signature models outperform LR in both the AUROC and AUPRC metrics within the timing group (Table 3.1, timing). The LR model performs poorly across all patterns. The basepoint model is the best performing Signature model when measured by its AUROC score with the lowest score of all the patterns being 88%. The full model that incorporates both basepoint and lead-lag augmentations perform worse than the basepoint model with an AUROC score of 86.3% compared to 93.7%. When measured by the AUPRC metric, the full model performs best for Timing1 and Timing2, whereas the LR model performs best for Timing3 and Timing4. However, once again all the AUPRC scores are quite low with none above 30 indicating poor predictive performance.

Overall, all the Signature models seem to outperform or achieve similar performance to the LR model in all life course concept groups when assessed using the AUROC metric. Across the period and order groups, the different Signature model formulations perform similarly, whereas in the repeats group the lead-lag augmentation achieves optimal performance and in the timing group the basepoint augmentation maximises mode performance. Across all models and groups, the AUPRC score is generally very low, except for the LR model for rules Period1 and Period2. The augmentation methods and model selected depend heavily on the types of patterns that exist within the data.

3.3.4 Influence of Interpolation on Signature Performance

The signature transform is designed to be applied to continuous data and therefore, it is considered essential to transform discrete data using interpolation. However, binary input data is very different to discrete input data that has many more potential values, and it is unclear what the impact of interpolation is on model performance in the binary case. The application of piecewise linear interpolation to the input data is investigated by comparing raw data to interpolated data that has ten times as many points in the time dimension.

The AUROC metric is slightly influenced by the interpolation level to different extents for each Signature model formulation (Figure 3.6, top). The AUROC scores of the full and the basic modes increase with the use of an interpolation method by only 0.009 and 0.015, respectively. In contrast the AUROC score of the basepoint model decreases by 0.006. The AUPRC score appears to be unchanged by interpolation for all augmentation methods (Figure 3.6, bottom). The greatest change in AUPRC is witnessed for the basepoint model which increases by only 0.007. Therefore, overall, it does not seem that the interpolation has a significant and consistent impact on the model performance.

3.3.5 Model Factors Effecting Computational Burden

One of the main benefits of the Signature method as opposed to the DL models is the greatly reduced computational burden. However, the memory requirements to compute and train the Signature model with a truncation level greater than three is prohibitively large for the computational resources available. The computation time of the Signature method is split into three periods; time taken to compute the signature features, time taken to train the model on the signature features and time taken to conduct inference.

The greater the truncation level, the greater the number of signature coefficients in the model and therefore, the longer the training time to fit the LR model to the signature features (Figure 3.7). The increase in training time from truncation level one to two is the lowest for the basic model at 19.4 seconds on average whilst it is 35.8 seconds for the full Signature model. This difference is even greater when increasing the truncation level from two to three, where the basic model increases by 167.6 seconds on average compared to 1612.3 seconds for the full model. The choice of augmentation methods also drastically increases the computation time which is particularly noticeable at higher truncation levels. When $K = 3$, the full Signature model has a training time of 1691.2 seconds on average, followed by the lead-lag model with 1385.9 seconds which are

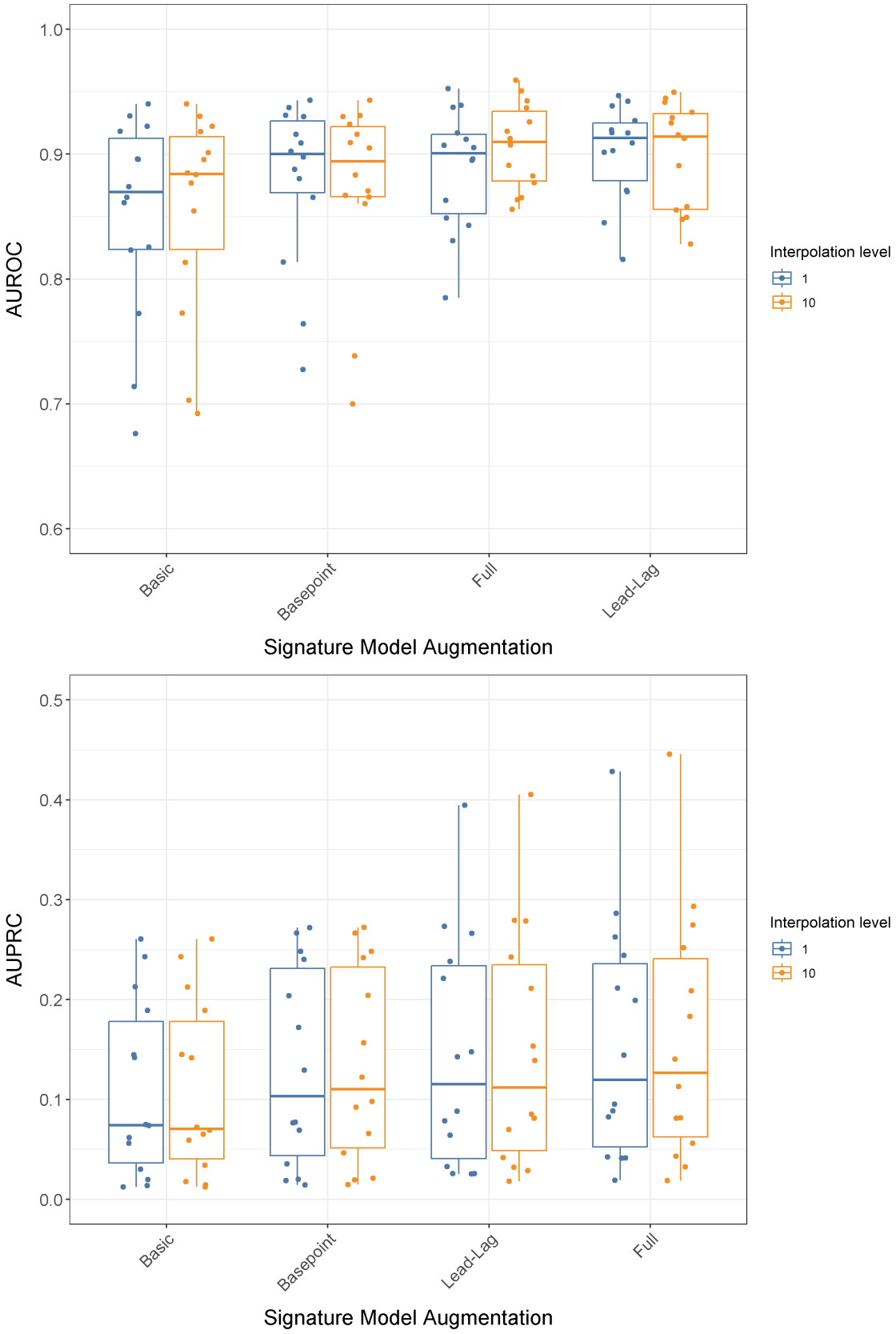


Figure 3.6: **Predictive performance of Signature models for different interpolation levels.** Values of evaluation metrics AUROC (top) and AUPRC (bottom) for each outcome at two interpolation levels (none and ten times) for the four Signature models (basic, basepoint, lead-lag and full) and logistic regression (LR). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. Truncation level $K = 3$ and 10 percent noise.

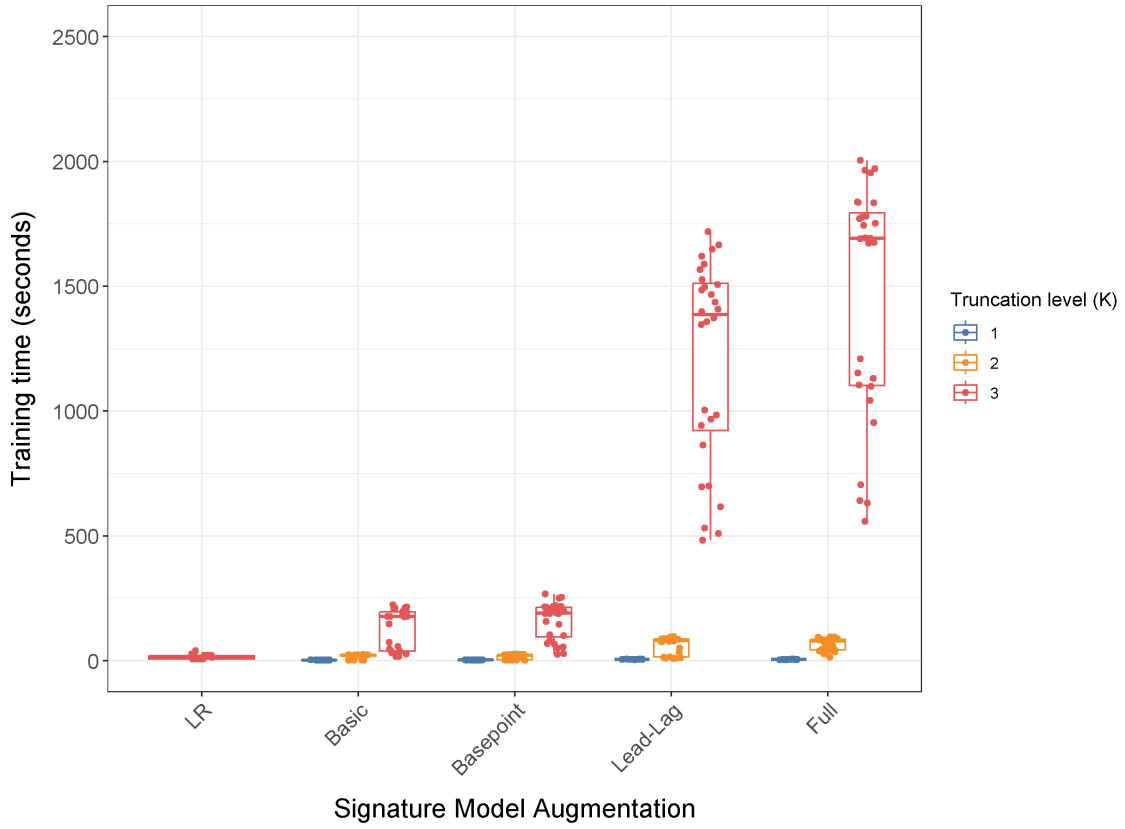


Figure 3.7: **Training time for different truncation levels.** Model training time in seconds for three values of truncation level ($K = 1, 2, 3$) for the four Signature models (basic, basepoint, lead-lag and full) and logistic regression (LR). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. No interpolation used and both levels of noise. The training time of LR is shown for comparison.

both considerably greater than the computation times of the basic and basepoint models which have times of 177.1 seconds and 190.2 seconds, respectively. It follows that the time taken to compute higher truncation levels for these models may be much smaller than the other augmentation methods, possibly shorter than $K = 3$ when incorporating the lead-lag transformation. The LR model is the fastest model with a median computation time of 12.5 seconds which is slower than the Signature models for $K = 1$.

Experiments show that the signature calculation time is minimally impacted by the truncation level. However, the augmentation methods and interpolation level have a substantial effect on the time taken when averaged across the different truncation levels. Higher interpolation factors increase the computation time (Figure 3.8). The time taken to compute the signature features increases by 133.0 seconds for the basic Signature model compared to 186.6 seconds on average for the full Signature model when interpolation is introduced. Once again the full and lead-lag model formulations give rise to substantially longer computation times.

3.4 Discussion

Higher truncation levels were found to improve model prediction and calibration performance, particularly when measured with the AUROC metric. The high AUROC scores obtained are indicative of the models' enhanced ability to capture intricate patterns within the life course data. Higher values of K effectively allow the models to incorporate higher-order interactions between variables, which can be particularly valuable for discerning temporal and inter-variable relationships that link the input data to the outcome. This trend

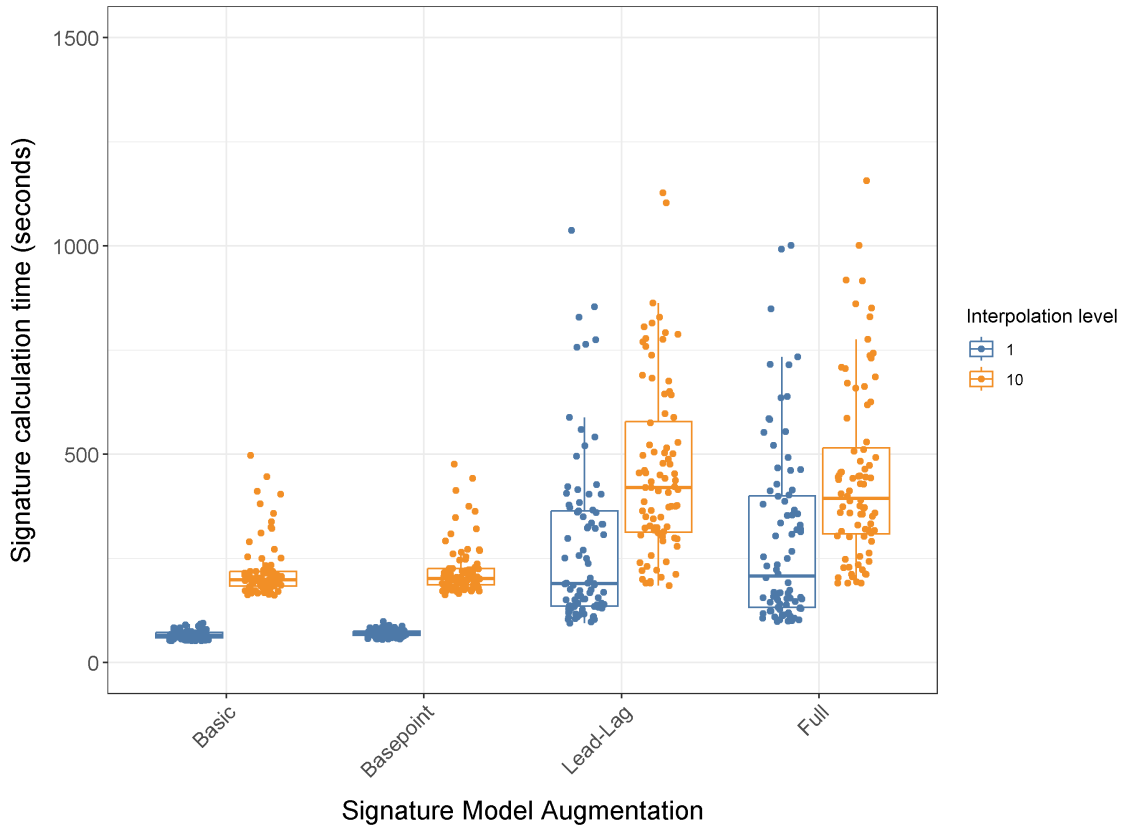


Figure 3.8: **Time taken to compute signature features for different interpolation levels.** Computation time in seconds taken to compute signature features at two interpolation levels (none and ten times) for the four Signature models (basic, basepoint, lead-lag and full). Boxes denote the 25% and 75% quantile, the midline denotes the median and whiskers indicate the range. Models ordered according to their overall mean value. For all truncation levels and for both levels of noise.

remains true when predictive performance is measured using the AUPRC metric, however, even with the highest truncation level, none of the Signature models have an AUPRC score above 30 across all the life course patterns. The reason for a model may obtain a high AUROC score and a low AUPRC score is that the two evaluation metrics assign different costs to the types of misclassification. The AUROC score prioritises correctly classifying positive instances above incorrectly classifying negative instances. In other words, the cost of a false negative is greater than the cost of a false positive. They type of scenarios in which this score is most useful are when the financial and health costs associated with missing an individual who is positive with a disease are far greater than the cost required to perhaps give extra tests or treatment to an individual without it. For example, if one designs a model to predict breast cancer risk and is used to identify high-risk individuals to undergo additional screenings and enable early detection. In the case where an individual is incorrectly given a high-risk outcome then the cost is an additional screening and in the other case of missing a high risk individual then the cost is the medical treatment of more advanced breast cancer and the impact to the individual's wellbeing and lifespan which is considerably more. Therefore, in this scenario and similar situations it is preferable to use the AUROC metric. The AUPRC metric is based upon balancing the recall and precision metrics which gives equal weight to both false positives and false negatives. Settings in which this evaluation metric is more suitable are when there are equal costs associated with both false positives and false negatives. For example, if a model is used to assign an expensive treatment to high-risk individuals and if a well individual were treated unnecessarily then this is also costly and can damage their health. When there is significant class imbalance, as is the case here to replicate real medical data, the AUROC score can overestimate model performance as it views the performances of each class misclassified equally [Saito and Rehmsmeier, 2015]. Therefore, when assessing model performance, the signature appears to lean towards reducing false negatives instead of false positives, especially in instances of high class imbalance. Therefore, the signature approach is best utilised in scenarios which seek to minimise false negatives. This may be due to the class balancing added to the LR model and warrants further investigation in future work.

As expected, the addition of noise into the life course patterns had a detrimental effect on model performance. This is a result of the deterministic mechanisms that generate the outcome values from the input being severed with the introduction of stochasticity. The performance of each model is affected to a differing extent. Notably the drop in the AUROC scores of the full and lead-lag Signature models is small in comparison to that of the LR model. When measured using AUPRC, the drop in performance is similar across the models. These effects come about as the patterns are more difficult to learn. The AUROC metric suggests that the Signature methods outperform the LR model at both levels of noise and the AUPRC suggests the opposite. This result highlights the importance of selecting the correct evaluation metric for the problem at hand.

The relative predictive performance of different augmentation formulations varied across the specific life course patterns. Notably, there was no effect of including the additional augmentations in the period or the order group. A possible reason that there were little differences in these groups is that the period group is entirely linear, so additional timing information is not required in order for the patterns to be identified. In the order group, the patterns do not depend upon the exact timing, but instead only which exposure occurred first. These results suggest that only patterns that have temporal relationships require additional augmentation methods. There is little variation in model performance measured by AUROC across the three different patterns in the order group, despite changes in the levels of class imbalance. This demonstrates that the choice of augmentation method is not significant for this type of pattern. However, the AUPRC scores in this group reflect the metrics increased sensitivity to imbalance. The basepoint augmentation is found to be essential for good performance in the timing group. This demonstrates that in this instance the property of translational invariance must be removed in order for the patterns to be able to be identified from the signature coefficients. The further addition of the lead-lag augmentation to the basepoint augmentation to create the full model appears to slightly degrade performance in the timing group. One reason for this may be that a

high truncation level leads to a vast number of signature coefficients input into the LR model which may mask some significant patterns. This demonstrates that it is not always the best cause of action to use the full model without exploring the performance of other augmentation formulations since performance can vary considerably and using the full model does not always yield the optimum result. In contrast, the lead-lag augmentation gives rise to a substantial increase in model performance in the repeats group. This is to be expected because the repeats category is based upon patterns of consecutive exposure events which is better captured by the lead-lag transform. The basepoint augmentation also increases model performance in the repeats category with the full Signature model obtaining the highest AUROC scores. The removal of translational invariance in this case may be helping to emphasise the temporal elements that give rise to positive outcomes for this life course concept.

The results demonstrate that the addition of the interpolation transformation did not improve the Signature model performance. Given that the Signature model relies upon the assumption that the data is continuous, the use of an interpolation method for discrete data is thought to be essential. However, in these experiments there was no increase in model performance using either evaluation metric when an interpolation method was used. It is possible that this is an effect of the input data being binary so even with interpolation, the data does not resemble continuous data. In terms of model implication, this result means that linear interpolation does not need to be used in scenarios of binary data which has the additional benefit of reducing overall computation time as interpolation methods were found to substantially increase the time required to compute the signature coefficients.

The higher truncation levels resulted in increased training time, particularly for the models that incorporated the lead-lag augmentation. The reason this augmentation increases the computation time considerably is that the addition of different augmentation methods increases the signature depth and consequently, the number of variables given to the LR model. The number of signature coefficients in the truncated signature is given by the formula $\sum_{k=0}^{K} M^k = \frac{M^{K+1}-1}{M-1}$ where K is the truncation level and M is the number of features after the augmentation methods have been applied. Therefore, greater values of the truncation level increase the number of signature coefficients exponentially. The number of features after augmentation also has a significant impact on the signature depth. If the basic Signature model without augmentation methods has M features, then the basepoint model has $M+1$ features, the lead-lag model has $2M$ features and the full model has $2M + 2$ features. In this analysis, $M = 10$ because there are nine features and one additional feature related to the time augmentation. It follows that the resulting signature depths are, $11, 12, 21, 23$ when $K = 1$ and $1111, 1464, 8421, 11155$ when $K = 3$. The signature depths of the models with the lead-lag augmentation are about twice as great as those without when $K = 1$, however, when the truncation level is increased to three then they become around ten times as large. As a result, if one were to try optimising model performance and keep the training time manageable then there is a balance between using more augmentation methods, increasing the number of input features, and increasing the truncation level. It may be possible here to use feature selection to reduce the number of input features and enable a higher truncation level to have been explored. Experimentation of which yields the best performance for a specific application is vital.

3.4.1 Limitations of the Signature Method

The primary limitation of the Signature method is the computational burden. Whilst it has the potential to be more computationally efficient and faster to run than the DL methods, for high truncation levels the CPU requirements can become prohibitively large. The time taken when using the lead-lag augmentation method can still be sizeable, although lower than that of hyperparameter optimisation. It was not possible to run the Signature model for a truncation level greater than three and this may have limited the model performance seen. There were also limitations in the number of interpolation levels that were investigated due to time

constraints. However, preliminary experiments demonstrated that when the input data was interpolated at a finer granularity there was no perceivable improvement in model performance and for very high numbers of interpolated points, model performance was negatively impacted.

There are multiple other data transformations that are often used when using the Signature method that were not investigated in this analysis due to time constraints. Notable examples include higher order lead-lag transforms which may allow the signature coefficients to capture long-range temporal dependencies that are present in the data. Another method that merits investigation is the use of windowing which transforms the input data by applying a sliding window in the time dimension and applying the signature locally to each window [Morrill et al., 2020a]. The windows can be sliding, expanding or hierarchical and this approach is thought to improve model performance by allowing different ranges of temporal patterns to be better captured and represented in the signature coefficients. Further experiments analysing the influence of these additional data transforms on model performance could be conducted in additional analysis and is recommended as further work.

While the Signature method offers a promising framework for representing and analysing longitudinal life course data, its combination with LR presents challenges that must be carefully considered. The signature has been used as a feature transform in multiple other architectures including ML models such as simple feedforward neural networks and more complex models such as RNNs and CNNS. Whilst the signature has the property that any function can be formulated as a linear combination of the signature coefficients, this only holds for a large truncation level. Therefore, in the absence of the full set of signature coefficients due to computational restrictions that limit the truncation level, other model structures may be better able to find the link between the signature coefficients and model outcome. This would require further investigation. Furthermore, the choice of class balancing and L1 regularisation in the logistic method were only assessed for performance improvements for a few data sets and not chosen based upon overall model performance. Therefore, researchers should explore various alternative approaches in comparison with the formulations presented in this analysis, considering the unique characteristics and requirements of their longitudinal life course data, to ensure accurate and meaningful modelling results.

A key drawback of utilising the signature approach instead of other modelling frameworks is the lack of interpretability of the modelling results. The term interpretable describes the extent to which it is possible to infer the reasons and input variables that give rise to the model's decisions. It is possible to use the built-in interpretability of the LR model and inverse signature transform (in addition to inverse augmentation methods) to explore how the LR coefficients relate to the original input MTS data. However, there is currently no existing structured framework by which to conduct this kind of analysis. In terms of the healthcare system and deployment, it is essential that model decisions can be justified, robust and the most influential inputs can be identified for patient safety and bias eradication. Therefore, the Signature method is not currently suitable for use within clinical and medical settings and further work to establish an accurate framework to infer interpretability is required.

3.5 Conclusions

This chapter contributes to the field of life course epidemiology and MTS data analysis by assessing the performance of a new methodology that has achieved excellent performance in other settings. This analysis acts as a clear guide for researchers conducting similar studies as to the most important aspects of this modelling framework and its strengths and weaknesses when compared to alternative approaches, such as LR and DL models. This knowledge allows them to make informed decisions when selecting model parameters and augmentation techniques, optimising their predictive models. Whilst the Signature method does not perform

as well as the DL approaches that were investigated in Chapter 2, it is a slightly more computationally efficient method. By avoiding the time consuming and computationally expensive procedure of hyperparameter optimisation, this model has a big advantage when working with limited computational resources. This necessitates further investigation into how the Signature method performs when applied to real data, such as the DANLIFE cohort. Therefore, the next stage of this body of research will focus on the application of both methods to the DANLIFE cohort and compare their respective predictive performances. This work contributes to developing more accurate risk prediction models for RTIs that enable key risk factors and at-risk individuals to be identified. A comprehensive understanding of risk factors enables more cost effective and powerful interventions and targeted vaccination schemes to address the population health burden of RTIs.

Chapter 4

Predictive Modelling of RTIs using the DANLIFE Cohort

4.1 Unlocking Insights in Longitudinal Health Data

In the two preceding chapters, both the Signature method and deep learning (DL) models were found to have the capacity to learn common life course patterns that were embedded in simulated data. Therefore, these models are likely to be able to learn life course patterns and obtain good performance when applied to real patient data. Once the models have learnt the patterns, they are able to be used as risk prediction models to identify individuals who have the greatest susceptibility to respiratory tract infections (RTIs) as well as investigating which exposures may act as risk factors for disease.

Within healthcare and epidemiology, big longitudinal data sets offer valuable insights into how the trajectories of health are shaped over the life course. These insights hold the key to understanding how health conditions evolve over time, what factors contribute to their progression, and how interventions can be designed to improve patient outcomes. Routine healthcare records are increasingly being collected to create comprehensive repositories of longitudinal health records, such as the DANLIFE cohort [Todd et al., 2020]. Such data, when harnessed effectively, has the power to transform the field of healthcare and offer a unique opportunity to explore the links between life course exposures and health outcomes. DL models have revolutionised the ways in which data is analysed, allowing for more complex and intricate underlying patterns to be learnt by the models leading to higher predictive accuracy. These models minimise the number of assumptions required in order to fit a model, resulting in reduced bias and greater potential to find novel connections within the data. The have been a range of applications of DL models to multivariate time series (MTS) data with excellent results but no study of how these models perform within life course analysis.

The DANLIFE cohort which follows a large cohort of individuals over a time frame spanning from birth to early adulthood, provides a crucial foundation for investigating the life course patterns and RTI risk factors that have remained elusive in traditional epidemiological studies [Bengtsson et al., 2019]. The DANLIFE cohort exhibits several unique features and strengths that make it an exceptional resource for exploring how multiple exposures generate RTI risk. The DANLIFE data set encompasses the life course trajectories of 2.2 million individuals, allowing for comprehensive investigations across diverse populations and demographics. This breadth ensures that the data set captures a multitude of health-related variables and contextual factors for each individual, from chronic conditions to adverse childhood experiences and demographic characteristics. Additionally, the DANLIFE cohort is updated yearly which makes it an ideal tool to uncover subtle temporal

dependencies and complex life course patterns that are associated with RTIs. However, the complex nature of the data set has meant that traditional epidemiological methods have encountered limitations that prevent them from fully capturing and realising the full potential of the rich data set.

In comparison to traditional cohort survey data sets, the DANLIFE cohort stands out in several key aspects. Traditional cohort studies are often limited by small cohort sizes and either a short time frame or infrequent observations. As DANLIFE includes all individuals born in Denmark after 1980, it has an extremely large cohort size followed yearly over a long period. These studies generally employ detailed questionnaires to gather information about different aspects of the lives of study participants. These questionnaires collect rich individual level information, including lower-level illnesses, individual lifestyle factors and behaviours (e.g., smoking, drinking, exercise), and living conditions. This detailed information provides a comprehensive view of multiple key factors, but the surveys are susceptible to recall bias, introducing inaccuracies due to individuals misremembering events. Moreover, the selection bias inherent in smaller study populations may limit the representativeness of the findings for the broader population. Furthermore, these studies depend upon extensive data collection efforts to conduct the surveys and detailed questionnaires. This makes interventions and analysis more resource-intensive and thereby introduces challenges in terms of scalability and efficiency. Loss to follow-up is a significant challenge in cohort studies, since participants may drop out over time, which can introduce bias into the analysis. In contrast, the DANLIFE cohort minimises bias as individuals leave the cohort only at point of death or emigration. Additionally, the DANLIFE data is automatically collected in the administrative registers and requires no further data collection. As a result, it is possible to use this growing resource to carry out passive analysis and potentially inform identification of individuals for intervention without needing individuals to come into contact with the healthcare systems. Additionally, DANLIFE boasts superior statistical power, particularly for rare medical conditions, making it a robust resource for studying health outcomes with broader implications.

In this chapter, the aim is to leverage the power of state-of-the-art ML techniques, specifically ResNet and the Signature method, to analyse the intricate patterns hidden within the DANLIFE data. The DANLIFE cohort is a unique and powerful resource that has many properties that make it ideal for exploring how exposures interact dynamically to generate disease burden at the individual level. The primary objectives include exploring the effectiveness of these advanced techniques in capturing temporal patterns, unravelling the intricate interplay between the exposure events contained in DANLIFE and their timing and evaluating their performance in predicting occurrences of RTIs at the individual level. By analysing this high-dimensional dataset with advanced DL techniques, this chapter aims to shed light on the intricate interplay between life course dynamics and the risk of RTIs, advancing the understanding of these infections and their associated risk factors. Addressing these goals, will significantly contribute to the fields of epidemiology and public health.

4.1.1 Existing Cohort Studies

Nordic national registry data, including the DANLIFE dataset, is becoming more accessible to the research community, prompting a surge in cohort studies for epidemiological research [Smith Jervelund and De Montgomery, 2020]. As the health data registries are longitudinal and updated over time, the long-term epidemiological trends may only become discernible after a number of years. This was highlighted by Schmidt et al. in 2019, who stated that the DANLIFE data set, then encompassing 40 years of data, was ripe for epidemiological research [Schmidt et al., 2019]. Traditional epidemiological approaches, such as Cox regression and the calculation of odds ratios, do not utilise the rich information contained within these vast longitudinal data sets to their full extent. Therefore, the research community has turned to DL methods, which have proven extremely successful when applied to large and complex data sets in other fields, to obtain epidemiological insights using these new data sources. Few studies have been published demonstrating the utility

of these methods on the data sets, here an overview of the questions that have been answered using these approaches and the exact DL methods and predictors used is provided.

Several studies have used ML approaches to answer epidemiological or medical questions. For instance, Chen et al. used the Swedish national registry data to predict the risk of an individual attempting or dying by suicide within the first few months after being discharged from psychiatric speciality care [Chen et al., 2020]. They used ensemble learning with multiple ML models and data containing 425 detailed predictors covering socioeconomic, medical and criminality records as well as history of criminality in the family. They obtained an AUROC of 0.88 for suicide attempts within 90 days of discharge which indicates excellent performance. A similar study examined suicide risk in Denmark with 1339 variables additionally including income and prescriptions, but not criminal records [Gradus et al., 2020]. Classification trees and random forest models were used to assess the most critical risk factors for suicide risk (AUROC of 0.77 and 0.88 respectively). However, in both cases the models did not incorporate time directly. Instead, each variable was measured as a total number of occurrences within a few broad time periods, which may have masked dynamic interactions and temporal dependencies.

Trinhammer et al. answer a similar question of whether Danish psychiatric patients are likely to commit a crime to avoid inaccurate and costly risk assessments [Trinhammer et al., 2022]. They use four ML models (logistic regression, random forests, XGBoost and LightGBM) along with their data comprising socio-economic factors, an individual's crime records and their psychiatric history to achieve an AUROC score of 0.96. They summarise the variables over time as counts and therefore provide no temporal information to the model. Similar approaches are taken in multiple other studies that use either random forest or XGBoost models without explicitly including temporal effects within the model. In some instances, high values of AUROC scores indicate that temporal information may only play a minor role in prediction, such as the prediction of whether a breast cancer patient will experience recurrence which had an AUROC score of 0.93 [Lauritzen et al., 2023]. This study uses the pathology register that contains tissue sample information and the DBCG clinical database alongside the Danish National Patient Register which are likely to be more informative than exposures alone and not require additional temporal information.

In contrast, Hartonen et al. utilise a large data set containing 2890 different variables, as well as each individual's genomic data to predict COVID-19 vaccination status in the Finish population and got an AUROC score of 0.801 suggesting that in this case timing may play a role [Hartonen et al., 2023]. Vabalas et al. used a recurrent neural network (RNN) to predict the risk of 1-year mortality for the Finish population [Vabalas et al., 2023]. They used each individual's full medical records and socioeconomic indicators, amounting to over 8000 features. The RNN framework explicitly includes an individual's age at which each exposure event occurred, thereby incorporating temporal factors into model predictions. They obtained an AUROC of 0.94 with this method which is perhaps because of incorporating temporal effects.

Another study that incorporates time into the modelling framework was conducted by Savcisens et al. who are the only other individuals who are currently seeking to combine the life course epidemiological approach with ML to understand how dynamic interactions between life exposures impact future health and well-being [Savcisens et al., 2023]. They utilise the Danish National Patient registry and the Labour Marker Registry, both of which are used in creating DANLIFE. Each exposure event along with its temporal information is transformed into multiple sentences which are then fed through **life2vec** (a transformer-based model). In this instance they demonstrate the utility of this approach when applied to predicting early mortality and differences in personality. They show a corrected F1 score of 0.44 at predicting mortality. This score is not as high as one would expect and motivates further exploration of these life course methods with ML models. This analysis demonstrates the potential of incorporating life course factors into long-term epidemiological analysis, however, they do not examine individuals from birth onwards and incorporate additional variables such as adverse child-

hood experiences (ACEs) that have been shown to impact future health substantially. Their approach also may not be fully able to capture all temporal dependencies due to the nature of the embedding. This paper was published whilst this thesis was in the final stages, so this analysis was conducted without knowledge of their methods, however, it is important to highlight this paper to demonstrate that there is much work being done in the field of life course epidemiology using ML tools.

Vartiainen et al. are the only group to have studied RTI risk using national register data along with ML methods [Vartiainen et al., 2023]. They use data from Finland and Sweden to examine the likelihood of a child being hospitalised with RSV within the first year of life given individual level characteristics and the medical histories of their parents and siblings. They use a XGBoost model which does not include temporal dependencies, although this is less relevant as it is within the first year and they also do not use ACE indicators. Once the model was fitted, they used odds ratios to assess the most significant predictors, but these will not include dynamic interactions and will likely not uncover any new insights over traditional epidemiological methods.

In summary, the increased access to Nordic register data has opened up many new opportunities to conduct insightful epidemiological research and gain insight into key predictors of disease as well as the development of improved risk prediction models. However, the current work in the field often lacks the inclusion of temporal factors to reflect the dynamic interactions and biological mechanisms that arise due to concurrent exposures. Therefore, an understanding of how influential different predictors are within different sensitive developmental periods of life are largely unknown. This work furthers the understanding of ML models used for life course.

4.1.2 Research Outline

The primary objective of this chapter is to apply state-of-the-art ML models, including ResNet and the Signature method, to the rich and extensive DANLIFE dataset, with a specific focus on predicting RTIs. By focusing on predictive accuracy and discriminative power, the chapter aims to gauge the suitability of ML techniques for handling the complexity of the DANLIFE dataset. Furthermore, the analysis aims to address the computational challenges associated with working with such a large-scale dataset and to assess the efficiency and scalability of the chosen models. This work advances the investigation into the utility of registry data within epidemiological research and explores how to maximise its potential. A detailed overview of the DANLIFE data set will be provided alongside an explanation of how each of the features has been chosen, pre-processed and designed. Then rigorous assessment of the performance of the DL and Signature models follows, with multiple performance metrics employed that reflect the models' capacities for discrimination and calibration. Computational challenges arising from the scale and security of the data are discussed, providing insights into the intricacies of working with such rich and extensive datasets. Finally, the findings are interpreted within the context of RTIs, shedding light on their implications for infectious disease epidemiology.

4.2 Data Description, Preparation and Model Training

4.2.1 Data Description

The DANLIFE cohort is a data set compiled from Danish Registry Data which allows for linkage between children and their parents and siblings and consequently the identification of multiple adversities which may impact their health trajectories. Bengtsson et al. extracted multiple variables from the register data, covering individual characteristics and ACE indicators [Bengtsson et al., 2019]. They also have extracted hospitalisation variables using the International Classification of Diseases (ICD)-10 codes to explore specific conditions

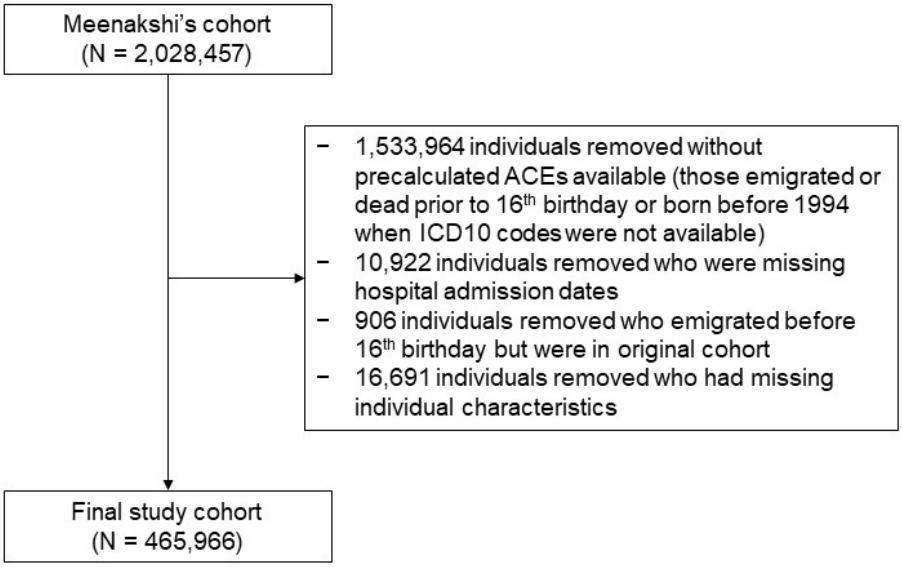


Figure 4.1: **Selection criteria for the study cohort.**

or how the number of hospitalisations are impacted by ACE occurrence within each ICD-10 Chapter [Rod et al., 2021]. The original data used is the same as Rod et al. with some changes made to how the type of diagnostic codes were categorised and how the data was pre-processed to avoid double counting. Both changes are explained in the hospitalisation section below [Rod et al., 2021]. These hospital admission variables were used and regrouped into new groups to capture illnesses within multiple disease categories.

The initial cohort and their hospitalisation data was pre-processed for me by Meenakshi Mishra based upon work carried out by Naja Hulvej Rod and Jessica Bengtsson. This initial cohort contained 2,028,457 individuals born in Denmark after 1st January 1980 and before 31st December 2018. Then the ACE confounder variables that are calculated for each individual were used, these were created by Naja Hulvej Rod, Jessica Bengtsson and Andreas Rickman. All individuals who died or emigrated prior to their 16th birthday were excluded from the cohort so that full ACE data for 16 years could be tracked for each individual. Restricting the cohort to only individuals with a full set of ACE records available until their 16th birthday left 494,493 individuals, as shown in Figure 4.1. Prior to 1994, ICD-8 codes were used so individuals born before this point were excluded for there to be consistency. Finally, the individuals which did not have a full set of individual characteristics were removed from the cohort leaving a total size of 465,966 individuals to be used in the analysis.

Individual Characteristics Data

Multiple registers had been previously compiled to obtain individual characteristics by Naja Hulvej Rod and Jessica Bengtsson, which have been described by Rod et al. [Rod et al., 2021]. Sex, age, and parental origin for each individual was extracted from the Danish Civil Registration System, parental education level from the Population Education Register and preterm birth status, birthweight and maternal age at birth from the Danish Medical Birth Register. Each of these variables, except for age, is constant across the life course and their descriptions are given in Table 4.1. In Denmark, there are 10 years of mandatory education since 2007, followed by three years of vocational and college education and any further level is university level. Low birthweight is defined as having a birthweight lower than the 10th percentile of sex-specific standard reference curves for intrauterine growth [Rod et al., 2021]. In the previous work which defined these characteristics,

Table 4.1: **Individual characteristics definitions.** The definitions of the seven individual characteristic variables as defined by Rod et al. [Rod et al., 2021]. The percentage of the analysis cohort (n=465,966) with each characteristic is provided.

Characteristic name	Definition	Categories	Percentage (%)
Sex	Sex given in the Danish Medical Birth Register	Female	51.8
		Male	48.2
Year of birth	Year of birth given in the Danish Medical Birth Register	1994	12.8
		1995	13.0
		1996	12.7
		1997	11.6
		1998	12.3
		1999	12.5
		2000	12.7
		2001	12.4
Age	Age each year based upon an individual's date of birth	0-15 years	-
Parental origin	Defined as European if both parents are of European descent and non-European otherwise	European	95.7
		Non-European	4.3
Parental education	Highest education obtained by either parent at time of birth	Low ($\leq$9 years)	12.5
		Middle (10-12 years)	49.5
		High (>12 years)	38.1
Maternal age at birth	Age of mother at time of birth	<20 years	1.8
		20-30 years	60.5
		>30 years	37.7
Preterm birth	Child born before 37 weeks of gestation	Preterm	6.2
		Not preterm	93.8
Birthweight	Child born with low birthweight for gestational age	Underweight	5.0
		Normal weight	91.7
		Overweight	3.3

being of European descent is defined as having a European, North American, Australian or New Zealand nationality. The parental origin variable is very broad and does not reflect the vast array of nationalities, racial and ethnic identities that exist. However, including this variable that acts as a proxy for race in the analysis may help to highlight the experiences of different racial groups within the healthcare system and potentially indicate biases that can be addressed. Using these individual characteristics, enables me to identify underlying biases in the medical system and highlight vulnerable groups in the population. Moreover, it is possible to examine the disparate effect of risk factors within each group.

ACE Variables

Within the input data 14 variables were designed to allow for and exploration of the impact of ACEs on a child's health trajectory and future life course. Twelve of these variables were defined and extracted from the registers by Rod et al. [Rod et al., 2021]. Rod et al. defined three categories of adversities recorded each year for the first 16 years of a child's life: material deprivation, loss of threat of loss and family dynamics. Each of these categories focuses on a different element of family life that can cause children to experience high levels of stress. The definition of each of the ACE variables is given in Table 4.2, which was taken directly from Rod et al. [Rod et al., 2021]. The ACE variables classified as material deprivation are family poverty and parental long-term unemployment which are extracted from the Income Statistics Register and the Integrated

Database for Labour Market Research, respectively. Poverty is a binary variable indicating if family income was below the poverty threshold each year and parental long-term unemployment is a count giving the total number of parents in the family that were unemployed each year. These variables are likely to be linked to important RTI risk factors such as household overcrowding and the neighbourhood the child grows up in which correlates with environmental factors, quality of school attended and street safety. Within loss or threat of loss there are the following variables: death of a parent, parental somatic illness, and sibling somatic illness. These are each defined as counts of the number of parents or siblings in the family each year who either died or were diagnosed with a severe illness. The severe somatic illnesses are defined as life threatening using the Charlson comorbidity index [D'Hoore et al., 1993]. The Danish Civil Registration System was used to identify deaths and the Danish National Patient Register was used for finding occurrences of familial somatic illness. The family dynamics group encompasses foster care placement, parental & sibling psychiatric illness, parental alcohol abuse, parental drug abuse and parental separation. Parental separation was identified using the Danish Civil Registration System and foster care with the Register of Support for Children and Adolescents. Both variables are binary indicators of whether the ACE event occurred or not each year. The Danish Psychiatric Central Research Register was used to find cases of familial psychiatric illness as well as cases of parental alcohol and drug abuse. Cases of parental substance abuse were also identified using the Danish National Patient Register and the Danish National Prescription Registry. These variables are counts of the number of parents or siblings affected each year.

DL approaches perform best with continuous data and can learn underlying patterns that are more intricate than whether a certain threshold was met. Consequently, two additional ACE variables were included in the analysis - family disposable income and parental average personal income - both of which were extracted from the Income Statistics Register. Parental average personal income is an average of the personal income of both the living parents. Family disposable income is the average of the equivalent family disposable living income of both parents which acts as an indicator of the amount of income the parents have to spend on lifestyle improvements and child rearing.

Table 4.2: **ACE variable definitions** The definitions of the twelve ACE variables taken directly from Rod et al. [Rod et al., 2021]. ICD: International Classification of Diseases

ACE	Definition
Family poverty	Family income below 50% of the median national family income in a given year
Parental long-term unemployment	A parent being unemployed for at least 12 months
Death of a parent	Death of a parent
Death of a sibling	Death of a sibling
Parental somatic illness	A parent diagnosed with one of the ICD-8 codes included in the Charlson comorbidity index38 in the period 1980–1993 or the ICD-10 codes included in the updated version of the Charlson comorbidity index39 in the period 1994–2015
Sibling somatic illness	A sibling diagnosed with one of the following somatic illnesses related to mortality in children aged 0–18 years (ICD-8/ICD-10 codes): malignant neoplasm (140–199/C00–C96); congenital anomalies of the heart and circulatory system (746–747/Q20–Q28); congenital anomalies of the nervous system (743/Q00–Q07); cerebral palsy (343–344/G80–G83); epilepsy (345/G40–G41); cardiomyopathy (425/I42–I43); congenital disorders of lipid metabolism (272/E75)
Foster care	Being placed in out-of-home care

Parental psychiatric illness	A parent's admission for at least one day with an ICD-8 or ICD-10 diagnostic code related to psychiatric illness. ICD-8 codes: psychoses (290–299 except 291); neuroses, personality disorders and other nonpsychotic mental disorders (300–309 except 303–304); and mental retardation (310-315); ICD-10 codes: mental, behavioural and neurodevelopmental disorders (F00–F99 except F10–F19)
Sibling psychiatric illness	A sibling's admission for at least one day with an ICD-8 or ICD-10 code related to psychiatric illness. ICD-8 codes: psychoses (290–299); neuroses, personality disorders and other nonpsychotic mental disorders (300–309); and mental retardation (310–315); ICD-10 codes: mental, behavioural and neurodevelopmental disorders (F00–F99)
Parental alcohol abuse	A parent diagnosed with an illness related to alcohol abuse or receiving a prescription of a drug used in treatment of alcohol addiction. ICD-8 codes: alcoholic psychosis (291); alcoholism (303); alcoholic cirrhosis of the liver (571.09); alcoholic steatosis of the liver (571.10); ICD-10 codes: alcohol psychosis and abuse syndrome (F10); alcoholic polyneuropathy (G62.1); alcoholic cardiomyopathy (I42.6); alcoholic-induced acute (K85.2) and chronic (K86.0) pancreatitis; alcoholic liver disease (K70); alcoholic gastritis (K29.2); Anatomical Theraputic Chemical (ATC) Classification System codes: drugs used in alcohol dependence (N07BB)
Parental drug abuse	A parent diagnosed with an illness related to drug abuse or receiving a prescription of a drug used in treatment of drug addiction. ICD-8 codes: drug dependence (304); ICD-10 codes: opioids (F11); cannabinoids (F12); sedatives/hypnotics (F13); cocaine (F14); other stimulants (F15); hallucinogens (F16); other and multiple drugs (F18–F19); ATC codes: drugs used in opioid dependence (N07BC)
Parental separation	The mother no longer sharing address with a partner

Hospital Admission Data

Data on hospital admissions was derived from the National Patient Register by Rod et al. [Rod et al., 2021]. The hospitalisation variables used here differ slightly in that the code for double counting and the groupings of diagnostics are different. When an individual in Denmark visits the emergency department, inpatient, or outpatient services the primary discharge diagnostic code is recorded, any additional diagnoses and the admission and discharge dates. The data comprises contact with the emergency department, inpatient, and outpatient care. Outpatient care refers to specialist doctor appointments as well as surgeries and treatments that do not require overnight care. For outpatient care, it is common for doctors to open a case which may comprise of multiple outpatient appointments whilst a condition is treated, or the cause of illness is investigated which is recorded as one long admission. Moreover, in some regions of Denmark, it is common for individuals to visit emergency or outpatient care instead of a GP due to distance to care. For this reason, the input hospital admission data and the outcome examined in this analysis are the number of new admissions to emergency, inpatient or outpatient care each year. When an individual visits the emergency department and subsequently is admitted to inpatient care and transferred between different specialist units, each transfer is recorded in the same way as a new admission. Therefore, overlapping admissions or those within seven days of each other are recorded as a single admission. All records with a missing admission date were excluded from the analysis and records without a discharge date were given one seven days later than the admission date. The diagnostic codes for the primary diagnoses are recorded using the ICD system; between 1980 and

1994 the ICD-8 system is used and after 1994 it switches to the ICD-10 system.

Multiple hospital admission variables according to different disease groups were defined according to the World Health Organisation (WHO) Global Health Estimate (GHE) groups [Department of Data and Analytics (DNA), 2020]. These groups were initially designed to capture the global burden of disease in terms of morbidity and mortality. The ICD-10 code Chapters are also arranged by disease group; however, the Chapters do not separate communicable and non-communicable diseases and so the GHE categories were used to group diagnostic codes instead of ICD-10 Chapters. Using the ICD-10 disease diagnostic codes of the primary diagnosis, the hospital records were categorised into 23 different variables which are shown in Table 4.3. The resulting data set is composed of 23 variables indicating the total number of new hospital admissions with a primary diagnosis in a given disease category each year for every individual. The total number of admissions in the population were calculated and category Unintentional injuries had less than five and was excluded from the analysis. Within each GHE category there are several subcategories (Appendix Table 8.1). The higher-level categories were used because using individual ICD-10 codes would have created a very high dimensional problem. In addition, a decision was made to also include the 19th ICD-10 code Chapter that covers *Injury, poisoning and certain other consequences of external causes* as an input variable because the corresponding collection of diagnostic codes (S00–T98) do not fall within any of the GHE categories yet it is one of the most common reasons for hospital admission and these accidents may relate to life course patterns.

Table 4.3: **The WHO GHE categories.** The ICD-10 codes that correspond to each GHE variable and its definition. Based upon the GHE categorisations by the World Health Organization [Department of Data and Analytics (DNA), 2020].

GHE cause name	ICD-10 codes
I. Communicable, maternal, perinatal and nutritional conditions	
A. Infectious and parasitic diseases	A00-B99, G00-G04, G14, N70-N73, P37.3, P37.4
B. Respiratory infections	H65-H66, J00-J22, P23, U04
C. Maternal conditions	O00-O99
D. Neonatal conditions	P00-P96 (minus P23, P37.3, P37.4)
E. Nutritional deficiencies	D50-D53, D64.9, E00-E02, E40-E46, E50-E64
II. Noncommunicable diseases	
A. Malignant neoplasms	C00-C97
B. Other neoplasms	D00-D48
C. Diabetes mellitus	E10-E14 (minus E10.2, E11.2, E12.2, E13.2, E14.2)
D. Endocrine, blood, immune disorders	D55-D64 (minus D64.9), D65-D89, E03-E07, E15-E34, E65-E88
E. Mental and substance use disorders	F04-F99, G72.1, Q86.0, X41-X42, X44, X45
F. Neurological conditions	F01-F03, G06-G98 (minus G14, G72.1)
G. Sense organ diseases	H00-H61, H68-H93
H. Cardiovascular diseases	I00-I99
I. Respiratory diseases	J30-J98
J. Digestive diseases	K20-K92
K. Genitourinary diseases	E10.2-E10.29, E11.2-E11.29, E12.2, E13.2-E13.29, E14.2, N00-N64, N75-N76, N80-N98
L. Skin diseases	L00-L98

M. Musculoskeletal diseases	M00-M99
N. Congenital anomalies	Q00-Q99 (minus Q86.0)
O. Oral conditions	K00-K14
P. Sudden infant death syndrome	R95

III. Injuries

A. Unintentional injuries	V01-X40, X43, X46-59, Y40-Y86, Y88, Y89
B. Intentional injuries	X60-Y09, Y35-Y36, Y870, Y871

Outcome Definition

The output variable was defined as a binary indicator indicating whether an individual had a hospital admission with a diagnostic code corresponding to a respiratory infection between the ages of 16 and 21. The ICD-10 codes corresponding to this classification are those classed as respiratory infections using the GHE definition (Table 4.3). The upper age of 21 was chosen to encompass individuals entering adulthood and capture greater number of RTI cases without decreasing the cohort size. In this analysis, the outcome has been designed as a good test of the suitability of this method for this data to give insights into how these DL approaches can be used to answer public health questions within a life course framework. The DANLIFE cohort is updated periodically so the number of individuals reaching age 16 will increase, meaning that the analysis cohort size and the statistical power will increase, along with the ability to explore severe RTI illness which becomes more prevalent in later life.

Other outcomes were explored in order to assess the broader applicability of these methods to other disease types. The two additional outcomes explored were diabetes mellitus and neurological conditions. These variables are created in the same way a RTIs but each for a different GHE group of diagnostic codes, also shown in table 4.3. These two additional outcomes were designed in the same way as the RTI outcome and were equal to one if there was one or more hospital admissions with a related ICD-10 diagnostic code between the ages of 16 and 20 and zero otherwise.

Computational Limitations

The primary constraint when working with the DANLIFE cohort is that, whilst the data has been anonymised, it is still identifiable and therefore necessitates a high level of data security. The data is stored on a remote server on which all data analysis and modelling must be conducted. The memory and CPU power on the server is severely limited and prohibits the running of larger and more complex models. These restrictions mean that the ResNet model was chosen to be used as the DL model within subsequent analysis. It has the fastest running time of all the models tested within Chapter 2 and is the least memory intensive whilst also capable of obtaining good performance. Similarly, the computational limitations, impact the formulation of the Signature model, particularly the choice of the signature augmentation methods and maximum truncation level. This will be explained further in the methods section.

4.2.2 ResNet Model Training Procedure

Feature Scaling

The data comprises of fields on many different scales, such as year of birth ranging from 1994 to 2002 whilst the number of hospital admissions each year is in the form of counts ranging from 0 to outliers greater than 100. Without scaling, the larger values, such as year of birth, will dominate the input data meaning that data with smaller maximum values will be ignored and the small variance in the year of birth will be lost in the analysis. Therefore, it is a common approach to scale this type of data before inputting it into the DL model. This enables faster training and better use of all data. There are two common scaling methods: normalisation and standardisation.

- **Normalisation:** otherwise known as min-max scaling, the formula is

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4.1}$$

- **Standardisation:** this also known as z-score scaling with formula

$$x' = \frac{x - mean(x)}{\sqrt{var(x)}} \tag{4.2}$$

Normalisation transforms the variables so that they fall between two specified values, usually zero and one. Standardisation which instead rescales the features to ensure they have a mean of zero and standard deviation of one. Standardisation assumes that the data is normally distributed. Therefore, a decision was made to normalise the features as their means and variance are such that after standardisation the scales are still quite disparate. Each feature was normalised separately over all the time points with the scaler fitted only to the training set to prevent information leakage from the test set. Then the scaler was applied to both the training and test sets using the `MinMaxScaler` from the `sklearn` package [Pedregosa et al., 2011]. This is done as the first step in preparing the data for the DL and Signature models. All data was normalised to be between zero and one which helps ensure that all the features lie on the same scale and have similar ranges.

Hyperparameter Optimisation

Due to the considerable computational limitations, several aspects of these experiments were adjusted in order to obtain meaningful results. Hyperparameter optimisation involves fitting the model repeatedly for different combinations of the possible hyperparameter values to find the parameter values that optimise model performance. Consequently, any differences in training time between different DL models will be magnified when the model is trained multiple times during hyperparameter optimisation. The choice of ResNet means that the entire optimisation procedure takes half the time of MLSTM-FCN. ResNet was not consistently the best performing model within the simulation study, however it performed comparably to the others in Chapter 2, suggesting that this choice will have minimal impact on results. Furthermore, the hyperparameter optimisation procedure had to be simplified greatly in comparison with Chapter 2. It was not possible to conduct the hyperparameter optimisation in a single run as the machine was required by other researchers during the day, therefore, a grid search approach was chosen. This meant that a select number of parameters could be tested in batches each night in the out-of-hours computing time. The length of time taken to conduct a single hyperparameter trial as well as other factors such as other individuals running their programs in the out-of-hours time and the bi-weekly machine shutdowns that occurred overnight, meant that often these runs failed or were cut short. Therefore, the number of parameters that were varied and the number of different

values they could take was greatly reduced which is explained further below. This also meant that it was not possible to conduct feature selection due to the time constraints.

Each individual i has associated input data X_i which takes the form $X_i = \{x_{i,1}, \ldots, x_{i,n}\}$ where each $x_{i,j} \in \mathbb{R}^d$. In this instance, the number of features is $d = 44$, number of time points is $n = 16$. The data is divided into train and test sets with an 80:20 split. In order to ensure that the different features each had a similar scale and standard deviation, the data was normalised and to prevent information leak from the training data to the testing data the transformation is fitted to only the training data and then applied to both sets. The outcome is positive if an individual had at least one hospital admission between the ages of 16 and 21 with a primary diagnosis with a diagnostic code corresponding to respiratory infections (Chapter 4 Section 4.2.1).

Fitting the ResNet model is conducted using the `tsai` python package version 0.2.22 [Oguiza, 2022]. The basic formulation of the ResNet architecture is explained Chapter 2 Section 2.1.2. In this analysis, several modifications to the model and the hyperparameter optimisation scheme are justified. Notably, the computational limitations meant that the number of model parameters optimised via hyperparameter optimisation had to be reduced. The parameter values are presented below, with multiple options given for those that were optimised:

- **Number of filters in first block:** 32, 64 or 96
- **Dropout rate in fully connected layer:** between 0.1 and 0.5
- **Kernel size for each convolutional layers:** Three numbers taken from 1, 3, 5, 7 and 9
- **Batch size:** 32, 64 or 96
- **Alpha:** between 0 and 0.5
- **Gamma:** between 1.01 and 5
- **Weight decay:** between 0 and 0.1
- **Early stopping patience:** 2, 4 or 6
- **Epochs:** 10 or 20
- **Learning rate:** 1e-3 or 1e-4

The models are each compiled with Adam optimiser and the loss function was chosen to be the focal loss which is described in Chapter 2 Section 2.3.3. The metric that governed hyperparameter selection was the AUPRC, which has been changed from Chapter 2 due to the imbalanced nature of the data. The ResNet model had the activation function changed to LeakyReLU from ReLU to improve stability. Two callbacks were used which were early stopping with patience and learning rate scheduling. As a result, the learning rate does not need to be optimised as it is reduced by a factor of 2e-10 when learning plateaus and so it is set to an initial value of 1e-3. The batch size and the α and γ parameters in the focal loss were also tuned in the hyperparameter optimisation. Once the hyperparameters were chosen then the model is rerun with three random seeds and the metric values are averaged.

Three-fold cross validation is conducted to maximise model performance, it is coded using the `optuna` package [Akiba et al., 2019]. Given the computational limitations it was not possible to explore the high dimensional search space of some of the models with this number of trials and therefore, the grid search sampler was selected. Once hyperparameters had been selected, the model is trained once more using these values with 90 percent of the train data used to train the model and the remaining 10 percent used to validate the model. Then the model is applied to the test set to compute predictions and calculate the evaluation metrics. Cross-validation is stratified. Model performance of both models was tested with multiple evaluation metrics

to provide a holistic overview of both the model's discriminative performance as well as its calibration. The primary metrics examined in this evaluation are precision, recall, the AUROC score, the AUPRC score the Brier score as described in Chapter 2 Section 2.3.2.

For the two other outcomes that are explored in this analysis, the hyperparameter optimisation procedure is considerably scaled down to limit the computational requirements. Consequently, the performance across the three outcomes is not directly comparable and the performance of the two non-communicable outcomes just illustrates the baseline performance that has the potential to be far better if more optimisation were to be carried out. The only parameter that is optimised in the cross-validation is the alpha parameter in the focal loss function, gamma was set to 2 and weight decay was set to 0.1. For diabetes mellitus, alpha was varied from 0.1 to 0.4 in steps of 0.05. Whereas for neurological conditions, model performance is assessed at three values of 0.15, 0.25 and 0.35 only.

4.2.3 Signature Model Setup

The Signature method was setup in the same way as in the previous Chapter 3 Section 3.2.3. Again, two additional augmentation methods were used to generate four Signature model formulations: basic, basepoint, lead-lag and full. No additional interpolation was used as this was shown not to have an impact on model performance for the simulated data. As with the ResNet model, the data is initially split into the same training and test sets. The signature is calculated for each individual in the entire training set and then a logistic regression model is applied to the resulting signature coefficients. The logistic regression model employed class balancing in addition to l1 regularisation to prevent overfitting the majority class. The choice of truncation level was severely limited due to the CPU limitations on the remote server. When the model was run with K equal to or greater than three then the model would use up the entire CPU of the server calculating the signatures of the training data. As the server was shared with 150 other researchers and the model was not able to run within the out-of-hours time these models had to be discarded. Therefore, the model was run with the truncation level taking values of one and two only.

4.3 Experiment Design

First the performance of ResNet is examined for the three disease outcomes investigated using multiple evaluation metrics, including precision, recall, AUROC, AUPRC, Brier score and individual classifications. It is important to note that this is not a direct comparison as each outcome had a different sized hyperparameter grid. Then the optimisation surface for ResNet for the RTI outcome is shown and discussed. Subsequently, the Signature method is applied to the three disease outcomes and its performance is assessed using the same evaluation metrics as the ResNet model. The performance of the Signature model is investigated for four different augmentation formulations (basic, basepoint, lead-lag and full) and for two truncation levels ($K = 1, 2$). Finally, the computation time is compared for both the ResNet model and the eight Signature models for the RTI outcome. The computation time is divided into training time and preparation time. The training time for ResNet is the time taken to conduct the final training on the full train set and the preparation time is the time taken to conduct hyperparameter optimisation. For the signature models, the training time is the time taken to fit the logistic regression model to the signature coefficients and the preparation time is the time taken to compute the signature coefficients.

4.4 Results

4.4.1 Data Summary

This section of the results illustrates the structure and key characteristics of both the input data and the outcomes. The input data is composed of individual characteristics, ACEs, and hospitalisation data. Table 4.1 provides the percentages of various individual characteristics in the cohort, showing the demographic distribution and the prevalence of specific attributes among the study population. Table 4.4 presents the counts of ACEs over a lifetime. It is important to note that this table does not show which individuals are experiencing multiple different ACEs. Certain ACEs have a maximum limit per year—for instance, some can only occur once per year, others twice, and some have no upper limit depending on the number of siblings. For the death of a parent, the maximum number of times this can occur in the life course is twice. The majority of individuals in the cohort never experience any ACEs. For each different ACE, over 60 percent of the study cohort will never experience it. The most commonly experienced ACE is parental separation, affecting over a third of the population at least once. The second most common is parental long-term unemployment. This variable ranges between 0 and 2 each year to indicate the number of parents that are unemployed. Parental long-term unemployment is also one of the ACEs that are experienced by some individuals most frequently. Another ACE that is commonly experienced is family poverty, this pattern is even more dramatic as this can only be experienced once each year. Of all the ACEs, family poverty has the highest proportion of the cohort experiencing seven or more exposures. Conversely, very few individuals experience even one instance of the death of a sibling or sibling psychiatric illness.

There are two further model variables that reflect ACEs: parental average personal income and family disposable income. Table 4.5 details the continuous income data, averaged over the first 16 years of life, alongside the percentage of the population in each income bracket. This table does not show how the two income variables are correlated or differ at the individual-level. The data indicates that parental average income generally exceeds family disposable income. In some cases, the average income is negative due to parents losing money. There is a significant range in incomes, with some individuals earning substantially more than others. This demonstrates the importance of using continuous variables, as individuals just above the poverty threshold would otherwise be treated the same as those with an average annual parental income exceeding 5,000,000 DKK.

The remaining model variables are those derived from hospitalisation records. Table 4.6 shows the incidence of hospitalisations by age according to the GHE category that contains the primary diagnosis. For some diagnoses, there is a singular value that represents the total incidence over the entire range due to data protection policies, and missingness occurs for those conditions that were not included in the model due to insufficient numbers. Overall, Chapter 19 (injuries and poisonings) has the highest incidence across the population by a considerable margin. Respiratory infections and infectious diseases have far greater incidence in infants than older children. This pattern is also followed to a lesser extent by neonatal conditions, respiratory diseases, and congenital anomalies. In contrast, hospitalisation incidence for diabetes mellitus, endocrine disorders, and musculoskeletal diseases increases with age.

Table 4.7 highlights the prevalence of the three disease outcomes to which the model was fitted. Respiratory conditions are the most prevalent, followed by neurological conditions, which have about half the prevalence of respiratory conditions. Diabetes has a prevalence approximately seven times lower than that of respiratory infections. Notice that these outcomes each affect less than five percent of the cohort, demonstrating how common it is for class imbalance to occur when modelling health conditions.

Table 4.4: **Number of ACE exposures experienced.** The percentage of the population that experienced a given number of ACE exposure events during the first 16 years of life. The ACEs here are defined by Rod et al. and these definitions are given in Table 4.2 [Rod et al., 2021]. *The maximum number of parental deaths that can be experienced is two.

Adverse childhood experience (ACE)	**Number of exposures experienced**						
	0	**1**	**2**	**3**	**4-6**	**7-9**	**$\geq$10**
Family poverty	76.62	9.71	4.42	2.67	3.76	1.55	1.26
Parental long-term unemployment	71.36	10.68	7.42	4.06	4.84	1.22	0.41
Death of a parent*	97.26	2.73	0.01	–	–	–	–
Death of a sibling	99.48	0.51	0.01	0.00	0.00	0.00	0.00
Parental somatic illness	84.24	9.13	3.47	1.54	1.34	0.23	0.05
Sibling somatic illness	96.98	1.54	0.62	0.32	0.37	0.11	0.06
Foster care	97.20	0.55	0.46	0.33	0.53	0.32	0.61
Parental psychiatric illness	94.70	3.43	0.95	0.37	0.39	0.10	0.05
Sibling psychiatric illness	99.28	0.58	0.11	0.02	0.01	0.00	0.00
Parental alcohol abuse	91.56	3.78	1.59	0.95	1.36	0.48	0.28
Parental drug abuse	97.61	1.23	0.39	0.20	0.30	0.14	0.12
Parental separation	60.08	27.22	9.11	2.66	0.92	0.01	0.00

Table 4.5: **Distribution of income in population.** The percentage of the population in each average income bracket for two measures of income; family disposable income and parental average income. These form the remaining two ACE measures in the model. The income is averaged over the first 16 years of life and measured in thousands of DKK.

Average income (1000 DKK)	**Family disposable income**	**Parental average personal income**
≤ 0	0.20	0.18
0 - 200	5.87	0.98
200 - 400	62.10	12.39
400 - 600	26.64	39.24
600 - 800	3.70	29.83
800 - 1,000	0.84	9.94
1,000 - 2,000	0.56	6.86
2,000 - 3,000	0.06	0.40
3,000 - 4,000	0.02	0.10
4,000 - 5,000	0.01	0.04
$\geq$ 5,000	0.01	0.06

Table 4.6: **Incidence rate of hospitalisation exposure events.** The incidence rates of the hospitalisation variables in different age ranges given as percentages of the total cohort. Most of the hospitalisation variables are based upon the GHE categorisations by the World Health Organization [Department of Data and Analytics (DNA), 2020]. *To adhere to data security regulations these incidence rates are given across the entire age range. **These categories had too few instances so were not used in the model.

GHE cause name	**Age (years)**			
	0 to 3	**4 to 7**	**8 to 11**	**12 to 15**
Communicable, maternal, perinatal and nutritional conditions				
Infectious and parasitic diseases	8.36	2.12	1.25	1.40
Respiratory infections	20.91	4.33	2.08	1.95
Maternal conditions	*0.18			
Neonatal conditions	15.29	1.19	0.73	0.56
Nutritional deficiencies	0.16	0.03	0.04	0.12
Noncommunicable diseases				
Malignant neoplasms	0.65	0.53	0.25	0.23
Other neoplasms	0.94	0.67	0.77	1.29
Diabetes mellitus	0.13	0.42	0.86	1.40
Endocrine, blood, immune disorders	1.66	1.97	3.24	4.02
Mental and substance use disorders	0.52	1.31	1.95	2.18
Neurological conditions	1.30	1.78	2.64	3.42
Sense organ diseases	2.51	4.19	4.78	4.21
Cardiovascular diseases	0.35	0.36	0.41	0.81
Respiratory diseases	10.12	8.41	6.52	6.42
Digestive diseases	5.56	3.60	2.92	3.29
Genitourinary diseases	2.88	3.37	3.01	3.14
Skin diseases	3.13	2.02	1.56	2.72
Musculoskeletal diseases	2.84	3.09	4.58	11.68
Congenital anomalies	10.91	6.52	5.52	5.76
Oral conditions	0.48	0.72	0.80	2.46
Sudden infant death syndrome	**NA			
Injuries				
Unintentional injuries	**NA			
Intentional injuries	*0.05			
ICD-10 Chapters				
Chapter 19: Injury, poisoning and certain other consequences of external causes	53.21	49.31	61.99	72.26

Table 4.7: **Incidence rates of each disease outcome.** The proportion of individuals who experienced the outcome in the follow up period between age 16 and 22. The outcome is positive if an individual was hospitalised at least once with a diagnosis within the corresponding GHE category as defined by the World Health Organization [Department of Data and Analytics (DNA), 2020]. The incidence rates are provided per 100,000 individuals.

Outcome	**Incidence rate (per 100,000 individuals)**
Respiratory infections	3411
Diabetes mellitus	480
Neurological conditions	1785

4.4.2 ResNet Performance

Table 4.8 shows the model performance of ResNet across multiple metrics for the three disease outcomes. The results for RTIs exhibit poor performance, the AUPRC score is 0.049 which is very close to zero indicating that the model is not able to correctly predict positive and negative outcomes. This value is related to the precision and recall which are 0.272 and 0.062, respectively. The precision value shows that of all the positives that the model predicted, only 27.2% were true positives. The recall shows that only 6.2% of the true positives present in the data were correctly classified by the model. The AUROC score is 0.528 which is only slightly above 0.5 that is the AUROC score that corresponds with entirely random predictions. There is a very large proportion of false negatives and a small proportion of false positives, indicating that the model has not been able to learn any patterns that are associated with positive outcomes. Overall, the ResNet model has poor performance across all metrics.

Motivated by the poor performance of the DL model on the RTI outcome, model performance on two further outcomes was assessed in order to examine the utility of DL approaches to predicting future health outcomes (Table 4.8). Two non-communicable diseases were chosen based upon fitting the ResNet model without hyperparameter optimisation. Then limited hyperparameter optimisation was carried out only looking at the focal loss parameter alpha which helps the model architecture adapt to the differing levels of imbalance in the outcomes. The greatest level of imbalance is in the diabetes outcome which has only 1.35% of outcomes that are positive. The proportion of positive outcomes for RTIs is almost ten times greater with 9.58% and neurological conditions has 5.01% of positive outcomes which falls between the two. The number of hyperparameter trials carried out and the number of hyperparameters examined is far greater for the RTI outcome than the other two outcomes. Therefore, these results should be regarded as baseline performance and are not directly comparable. However, despite the more limited number of hyperparameters trialled, both of the two alternative outcomes have greater performance than RTIs.

The AUPRC metric is the most impacted and therefore reflective of class imbalance in the data which aligned with the results seen in Section 2.4.2. However, in this instance, despite the greater level of class imbalance, the diabetes outcome has the highest AUPRC of the three outcomes with a score of 0.686. This corresponds with the small proportion of false positives and a greater proportion of true positives. The precision and recall values for diabetes are 0.922 and 0.743, respectively, indicating that almost all positives predicted by the model are true positives and the majority of positives outcomes being correctly identified. These results are impressive in the face of substantial class imbalance. The AUROC indicated excellent discriminative performance with a value of 0.871 and the brier score suggests that the model is well-calibrated as it is close to zero. The model performance for neurological conditions is middling. There is an acceptable precision score of 0.614 but a low recall leading to a low AUPRC score of 0.204. The AUROC is 0.655 which is better than random performance but suggests that there are still many false negatives. Overall, ResNet has performed excellently for diabetes, passably for neurological conditions and poorly for RTIs.

The model was optimised across a grid, specifically examining the influence of the focal loss parameters

Table 4.8: **ResNet performance on the DANLIFE cohort for multiple outcomes.** Evaluation metrics for the ResNet model for three different outcomes. TN = true negative, TP = true positive, FN = false negative, FP = false positive.

Outcome	Precision	Recall	AUROC	AUPRC	Brier	TN	FP	FN	TP
Respiratory infections	0.272	0.062	0.528	0.049	0.136	89486	528	2982	197
Diabetes mellitus	0.922	0.743	0.871	0.686	0.017	92718	28	115	332
Neurological conditions	0.614	0.313	0.655	0.204	0.083	91203	327	1143	520

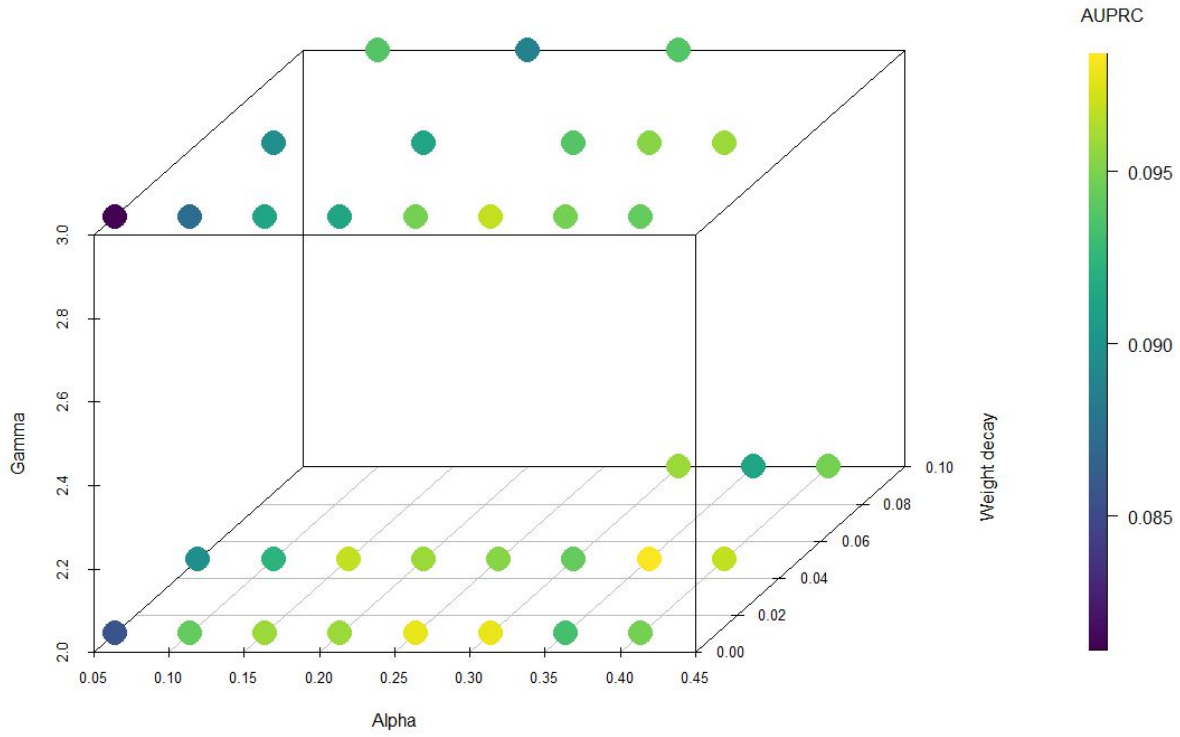


Figure 4.2: **Flat optimisation surface.** The average cross-validation AUPRC over the three folds at multiple hyperparameter points. The parameters alpha and gamma are the corresponding parameters from the focal loss function and weight decay is an additional model parameter.

and weight decay. Figure 4.2 shows the average cross-validation AUPRC values across the folds for RTIs at each point on the hyperparameter grid. The AUPRC values range between 0.08 and 0.1, demonstrating a very small range and little variation in model performance on the training data when these hyperparameters are altered. The gamma parameter is normally selected to be equal to two and often is not included in hyperparameter optimisation.

4.4.3 Signature Performance

The Signature model was fitted to the three different outcomes with four augmentation formulations and two values of the truncation parameter. Table 4.9 shows the model performance across a range of different evaluation metrics. The first point of note is that the AUPRC values for RTIs are very low, and each model formulation has a lower or equal AUPRC score to ResNet in the previous section. In contrast with ResNet, this is largely a result of a greater proportion of false positives which lowers the AUPRC score despite the generally lower numbers of false negatives. Another key result is that the full Signature model with truncation level two ($K = 2$) has an AUROC value of 0.628 which is greater than ResNet which has 0.528. In other experiments without hyperparameter optimisation, ResNet has been found to slightly outperform this model but as can be seen here, this comes at the cost of a large number of false positives. There is a slight trend towards the full augmentation models and particularly the basepoint augmentation leading to higher performance alongside a greater truncation level.

The diabetes outcome has dramatically different patterns in performance. The model with the highest AUROC and AUPRC scores is the full Signature model with truncation level $K = 1$. This is entirely contrary to what one would expect as the greater the truncation level, the more information about the time series is contained in its signature. In comparison with the ResNet model performance for diabetes, the Signature model has a marginally lower AUROC score and a reasonable drop in the AUPRC value. The best performing augmentation formulation is the full augmentation followed by the basepoint augmentation for truncation level $K = 1$. In the case of $K = 1$, the basic Signature model outperforms the lead-lag Signature model. The Brier score is low for all the models, signifying excellent calibration of the models.

In the case of neurological conditions, the highest AUROC score is not obtained by the same formulation that achieves the highest AUPRC score. The highest AUROC score is obtained by the full Signature model with truncation level $K = 2$, closely followed by the basepoint Signature model with truncation level $K = 2$. However, the AUPRC score is highest for the lead-lag Signature model with truncation level $K = 1$, followed by the basic Signature model with truncation level $K = 1$. The brier scores are higher for the neurological conditions outcome, indicating that the models are less well-calibrated. Within this group of models, the highest AUROC value is slightly greater than for the ResNet model, whilst the maximum AUPRC score obtained is considerably less than the ResNet model.

4.4.4 Computational Efficiency

It is important to consider the computational cost of different models and their formulations when deciding upon which is most suitable for conducting the analysis. Table 4.10, shows the times in seconds taken to prepare the data, model parameters and fit the model for the RTI outcome. The train time is the time taken to fit the logistic regression model for the signature and the time to train the final model in the case of ResNet. There is a substantial increase in the time taken to fit the model for the Signature model when the truncation level is doubled from one to two. The Signature model with the basic or basepoint formulation have the lowest training time since they have the fewest covariates in the model as the signature is shorter. The full and lead-lag model formulations taken almost double the time as their signature is about twice as long in length. Once the truncation level is increased, the time taken to train the model increases considerably due to the

Table 4.9: **Signature performance on the DANLIFE cohort for multiple outcomes.** Evaluation metrics for the Signature model for three different outcomes at two truncation levels (K) and four augmentation formulations. TN = true negative, TP = true positive, FN = false negative, FP = false positive.

Augmentation	K	Precision	Recall	AUROC	AUPRC	Brier	TN	FP	FN	TP
Respiratory infections										
Basic	1	0.034	1.000	0.500	0.034	0.934	0	90014	0	3179
Lead-Lag	1	0.154	0.001	0.500	0.034	0.045	90003	11	3177	2
Basic	2	0.047	0.459	0.565	0.040	0.242	60357	29657	1720	1459
Basepoint	1	0.063	0.228	0.554	0.041	0.106	79317	10697	2455	724
Full	1	0.168	0.065	0.527	0.043	0.053	88991	1023	2973	206
Lead-Lag	2	0.053	0.448	0.584	0.043	0.235	64812	25202	1755	1424
Basepoint	2	0.056	0.620	0.625	0.048	0.223	56687	33327	1208	1971
Full	2	0.059	0.583	0.628	0.049	0.217	60699	29315	1327	1852
Diabetes mellitus										
Basepoint	2	0.029	0.774	0.825	0.024	0.099	81177	11569	101	346
Basic	2	0.031	0.770	0.828	0.025	0.110	82168	10578	103	344
Full	2	0.040	0.774	0.843	0.032	0.084	84488	8258	101	346
Lead-Lag	1	0.048	0.763	0.845	0.038	0.081	86056	6690	106	341
Lead-Lag	2	0.050	0.767	0.849	0.040	0.071	86291	6455	104	343
Basic	1	0.183	0.736	0.860	0.136	0.061	91277	1469	118	329
Basepoint	1	0.509	0.732	0.864	0.373	0.006	92430	316	120	327
Full	1	0.818	0.732	0.865	0.599	0.003	92673	73	120	327
Neurological conditions										
Basepoint	1	0.031	0.705	0.652	0.027	0.272	54762	36768	491	1172
Full	1	0.032	0.705	0.659	0.028	0.247	56148	35382	490	1173
Basepoint	2	0.038	0.667	0.681	0.031	0.211	63512	28018	553	1110
Basic	2	0.049	0.439	0.642	0.031	0.195	77340	14190	933	730
Full	2	0.040	0.662	0.686	0.032	0.193	64928	26602	562	1101
Lead-Lag	2	0.057	0.440	0.653	0.035	0.193	79367	12163	932	731
Basic	1	0.088	0.351	0.642	0.042	0.128	85446	6084	1079	584
Lead-Lag	1	0.091	0.360	0.647	0.044	0.146	85526	6004	1064	599

Table 4.10: **Time performance on the DANLIFE cohort for the RTI outcome.** Train time is the time taken to fit the model; either logistic regression for the Signature method or the ResNet once hyperparameters have been selected. preparation time is time taken to do augmentations and signature or to do hyperparameter optimisation.

Model name	AUROC	AUPRC	Brier	Train time	Preparation time
Signature Basepoint ($K = 1$)	0.554	0.041	0.106	44s	288s (4m 44s)
Signature Basic ($K = 1$)	0.500	0.034	0.934	47s	287s (4m 43s)
Signature Lead-Lag ($K = 1$)	0.500	0.034	0.045	68s (1m 8s)	500s (8m 20s)
Signature Full ($K = 1$)	0.527	0.043	0.053	77s (1m 17s)	519s (8m 39s)
Signature Basic ($K = 2$)	0.565	0.040	0.242	716s (11m 56s)	353s (5m 53s)
Signature Basepoint ($K = 2$)	0.625	0.048	0.223	1059s (17m 39s)	356s (5m 56s)
Signature Lead-Lag ($K = 2$)	0.584	0.043	0.235	2125s (35m 25s)	799s (13m 19s)
Signature Full ($K = 2$)	0.628	0.049	0.217	2512s (41m 52s)	815s (13m 35s)
ResNet	0.530	0.051	0.135	4170s (1h 9m 30s)	468305s (130h 5m 5s)

exponential length of the signature. It is especially long for the full model which takes 41m 52s to fit, which is around four times the length of time taken to fit the basic model. The trend shows that $K = 3$ would be prohibitively large to compute. The ResNet model has a training time of 41m 52s which is less than twice the time of the full Signature model with truncation level $K = 2$.

The preparation time is the time taken to compute the signatures and augmentations for the Signature models and the time taken to conduct hyperparameter optimisation for the ResNet model. There is little difference in the preparation time for the different Signature models. The inclusion of different augmentations has a greater impact on the time taken than the truncation level. This measure is where there is the greatest difference between the ResNet model and the Signature models with the slowest Signature model taking 13m 35s compared to 130h 5m 5s for the ResNet model. This is a considerable amount of time, particularly considering that it is for a very coarse grid of hyperparameter values and few hyperparameters. Conducting a full hyperparameter optimisation would take weeks or months. Additionally, due to the nature of the computer setup and the fact the server was shared between over 150 researchers, these hours had to be conducted outside of work hours and the machine frequently shut down unexpectedly. Therefore, the true length of time taken to conduct hyperparameter optimisation was much longer.

4.5 Discussion

4.5.1 Understanding the Results and their Implications

The ResNet model was considerably better at predicting some disease outcomes than others. Despite greater class imbalance, the model achieves greater AUPRC and AUROC scores on diabetes mellitus and neurological conditions than RTIs. This could be because of the way in which these diseases occur within the population. Diabetes is known to have genetic factors as well as to be brought on by the accumulation of lifestyle factors over time. The nature of risk increasingly slowly over time with additional exposures is true of most non-communicable diseases, however, whilst it impacts susceptibility to infectious diseases there are many other factors at play. For instance, with RTIs there is the randomness inherent in whether an individual will become infected which is largely independent of their unique life experiences. It may also be the case that the features contained within the model are more relevant for diabetes and neurological conditions than for RTIs and that several key variables are not included or poorly represented in the data. These include things like smoking status, air pollution levels and living conditions. Overall, the fact that the model performed very well for diabetes demonstrates the utility of this approach and how it can be used to predict rare medical

events.

The optimisation surface for hyperparameter optimisation of ResNet for the RTI outcome is very flat. This can indicate one of two things, either the maximum in the trial space has not been found or the data does not contain enough key variables that can predict RTIs. The reasons why the maximal point may not have been reached include that the wrong parameters were investigated or that the grid is too coarse so that the maximum fell between two values. Chapter 2 demonstrated that if the patterns in the data exist and take a similar form to the life course patterns then the ResNet model can achieve excellent performance. Therefore, it is possible to conclude that either these patterns linking the input with the outcome do not exist or that the relationships are more complex than those investigated. In addition, in Chapter 2 the results showed that ResNet was extremely sensitive to hyperparameter choice and required extensive optimisation schemes to ensure best performance. Due to the computational limitations, the optimisation procedure was reduced and so one possibility is that the coarse grid and limited number of parameters optimised meant that the maxima was missed.

When evaluating model performance with multiple metrics it is important to take into consideration their strengths and weaknesses. As discussed previously, they are impacted differently by class imbalance with AUROC thought to give overestimates of model performance in these circumstances. Furthermore, they each have different priorities when it comes to assessing performance. AUROC assigns the same cost to false negatives and false positives. For example, if ten percent of positives and ten percent of negatives are misclassified then you may obtain a very high AUROC score. However, the AUPRC penalises false positives more greatly and numbers of misclassified individuals instead of proportions. In the example above, the AUPRC would be much lower than the AUROC as if there are 100,000 negatives and 1,000 positives then there are 10,000 false positives compared to 100 false negatives. The large number of false positives impacts the precision and therefore, the AUPRC score. In this analysis, the AUPRC score is not lower for outcomes with greater levels of class imbalance. In Chapter 2, the AUPRC values were greatly decreased in cases of significant class imbalance. This suggests that when using the real data, the lower model performance for RTIs was due to higher levels of noise present in the data or substantially more complex life course patterns. On the other hand, it may instead be the case that the patterns for diabetes are very simplistic and there is little noise present.

The Signature method performed comparatively to ResNet when measured by AUROC, but lower when measured by AUPRC. The full Signature model with truncation level equal to one was the best performing model for diabetes, suggesting that the patterns in the data governing the occurrence of diabetes are simple and linear because the performance decreased for higher truncation levels so higher order interactions were not driving the outcome. The higher covariates are akin to variance and inter-variable interactions over time and therefore, these are not informative. It may be the case that if it were possible to use a higher truncation level than two then performance would improve again as the interactions are higher order. This is further confirmed as the introduction of the lead-lag augmentation to the basic Signature model decreases performance with truncation level equal to one. Another influential factor in this pattern is that the Signature method works best when there are only a small number of features. Feature selection would be useful in this case to reduce computational burden and lower the number of model parameters and improve performance. However, time limitations mean that this would be a first step in future research.

The computation time of the full Signature model at truncation level $K = 2$ was around half the model training time of the ResNet model. The signature is much faster overall because the time taken to compute the signature is far less than the time required to conduct hyperparameter optimisation. Given that the computer could only be used out of hours, the use of grid search meant that the computational time of the ResNet model hyperparameter optimisation could be split and run in smaller chunks. In some instances, it is possible to fit

the model “out of the box" without hyperparameter optimisation and this can be used to look at the influence of different hyperparameters in a less systematic way to reduce the number to optimise. It is also possible to conduct feature selection with a model that is not optimised to reduce the number of input features, minimise collinearity and therefore speed up hyperparameter optimisation when it is carried out. Another potential improvement is transfer learning that trains the model on a simpler problem and then transfers the model weights to the new problem to speed up training. All these methods would allow for the preparation time for the ResNet model to be reduced and make it more comparable to the signature with better performance.

Using Data for Good

The use of this highly identifiable and individual data is controversial. There is a growing body of research that aim to establish new ethics frameworks for the use of big data within medical ML applications [Mooney and Pejaver, 2018]. The primary concerns surrounding this type of research are informed consent, privacy, and data ownership [Mittelstadt and Floridi, 2016]. Informed consent relates to the idea that an individual consenting to their data being collected may consent to certain uses but not others. For instance, they may accept their data being used to improve population health but not for certain procedures or use by companies for marketing. Another idea is that of data ownership, the data can be sold and is owned in perpetuity without the consent of all individuals within the data set, which is particularly difficult within the EU due to new data regulations [Safdar et al., 2020]. It is possible that the aims of the study change over time which is of concern to individuals. A new consideration with big data is which groups have access to the considerable computational power required to analyse the data, this tends to be monopolised by big companies whose goals include maximising their profit whilst research institutions have limited capacity and cannot operate models on the same scale [Mittelstadt and Floridi, 2016]. Finally, it is possible that the use of large data sets ignores the bias that impacts small subgroups of the population [Giovanola and Tiribelli, 2023]. The Danish administrative systems mean that all individuals living in Denmark automatically consent to data use and storage. Therefore, it is important that regulations are established and enforced that help to further research that contributes to population health whilst considering the potential ethical harms caused by data misuse. The Nordic countries have already created their own ethics systems set up to ensure that the identities of individuals are protected, and the data is used for good [Ludvigsson et al., 2015]. It is imperative that these legal frameworks are kept up to date to future proof these resources from advancements in AI and their benefits for public health research.

4.5.2 Limitations & Future Work

There are multiple limitations that have impacted the results of this analysis and would need to be addressed within future research. One key consideration is the formulation of the outcome which may have had significant repercussions on model performance. The problem was designed as a binary TSC task, and the choice of a binary outcome introduces boundary cases. This means that there will be individuals who incurred an RTI-related hospitalisation just outside of the window between their 16th and 22nd birthday that are treated very differently by the model than someone hospitalised a few days earlier or later within the window. These boundary cases can reduce model performance whilst not impacting an individual’s disease risk. There are alternative formulations of the outcome that enable better model performance and learning. One possibility is to change the binary classification to a constructed measure of disease severity. This outcome could incorporate the length of the hospital stay, the number of admissions within the given period or deconstruct the RTI GHE category into smaller subgroups with ICD-10 codes reflecting severity. Another alternative is using a time series of new admissions each year within the follow up as an outcome. This formulation gives the DL model more outputs to predict and allows the performance to better represent reality. The outcome was

designed as a first exploration of how these methods can work on real data and going forward an assessment of outcome design on model performance is advisable.

In terms of the research questions adressed in this Chapter, it is important to be aware that the cohort is still very young and individuals between 16 and 22 are likely to have low morbidity and mortality rates. Therefore, the disease incidence within the population is very low and the result of the accumulation of risk will generally only result in negative health outcomes later in adulthood. Many of the non-communicable diseases develop over a long period of time and are therefore, more prevalent in older age. The nature of the cohort means that over time the number of participants will increase and allow for an in-depth exploration of more long-term temporal dynamics and exposure interactions. There is much scope for conducting further analyses using these DL methods with more time steps to make the most out of their impressive abilities to recognise patterns.

Computational Limitations

The most prominent limitation that impacted this analysis is the severe computational restrictions. This had far-reaching effects on several areas of the analysis. Due to the high memory requirements for the DL models the code had to be run outside of working hours. This meant that coarse grid search was used to conduct hyperparameter optimisation so that it could be carried out in batches and fewer hyperparameters were tested to reduce computational time. In addition, there were twice weekly server shutdowns that meant several runs were not able to complete. Other users who had more intensive jobs to run also ran their jobs out of work hours which meant that it was not possible for these jobs to run at the same time. As a result, these computational restrictions impacted several critical elements of this analysis, including but not limited to the choice of the ResNet model, the choice of optimisation scheme, the number of hyperparameters trialled, the truncation number for the signature, the values of the hyperparameters trialled. Therefore, the results presented in this analysis represent baseline model performance under sub-optimal conditions. It is highly possible that if this analysis was not limited by computation time, then both ResNet and the Signature models would achieve better performance. One key takeaway from the analysis in this chapter is that the development of personalised medicine using register data would require substantial investment in computational support by means of more powerful machines with access to the secure data.

The ResNet model was chosen as it achieved competitive performance within the simulations and had the lowest computation time of the DL models when applied to the DANLIFE data. However, it is very possible that the DANLIFE data differs from the simulated data sets in such a way that the LSTMAttention or MLSTM-FCN models may outperform ResNet. One reason for this possibility is that ResNet is designed for images and has been adapted for time series by treating the time points and features as points on the x and y axes. This means that inter-variable relationships are most likely to be found between the different features that have been placed next to each other in the input and those further away may be missed. The time component is treated in the same way as the different features and therefore, the ability for the model to totally represent temporal connections may be weaker than a model that explicitly has temporal dependencies within it, such as the MLSTM-FCN model. With greater computational power, the alternative DL models could be tested on the data to ensure optimum performance and potentially demonstrate better capacity to detect temporal patterns present in the data.

The hyperparameter optimisation procedure had to be simplified extensively from the ideal optimisation setup. The number of different hyperparameters that could be trialled in the computational time available was severely limited. The resulting optimisation surface showed little variation in AUPRC across the different parameter sets trialled. One possible reason behind the flat optimisation surface is that the maxima fell between the small number of hyperparameter values tested or that other hyperparameters should have been

altered as they would have had a greater impact. The other explanation is that the variables in the input data were not sufficient to generate reliable predictions of future RTIs. In addition, the computational limitations meant that it was not possible to carry out feature selection to limit the number of uninformative or collinear features within the input data. Therefore, it is unlikely that optimum performance had been reached.

The signature truncation level was not able to be made greater than two due to CPU requirements. Whilst the Signature model is less memory intensive, the computation requirements of the Signature model for a greater number of features and higher truncation level is substantial. Therefore, the model was not able to be fitted for truncation levels greater than two as it could not run over the course of a weekend. The level of truncation in the Signature model is the most influential model parameter since the greater its value, the more higher order temporal and inter-feature patterns in the data are represented by its signature. The simulated data sets generally showed little improvement between truncation levels two and three, however the DANLIFE data will likely contain more complicated patterns linking the exposures and the outcome which may require a greater truncation level to be captured by the signature. Consequently, the assessment of signature performance is greatly restricted. Whilst this was not shown to be the case for diabetes, it may be that RTIs have more complex temporal relationships within the data that are not appropriately represented by a truncation level less than three.

Considerations Related to Real Data

The use of registry data introduces several key features of this analysis that are important to consider when interpreting results. The DANLIFE cohort is composed solely of registry data that is generated as a collection of all recorded instances of an individual encountering a governmental administrative system. As such these results differ from other life course studies which predominately utilise data collected via questionnaires as part of a small scale study. One benefit of using registry data is that the data is already routinely collected and any interventions or risk prediction models developed from this analysis would be able to be used across the population without any further data collection costs. However, the main drawback of using data collected from registers as opposed to questionnaires is that many potentially relevant variables are not recorded in the registers, such as levels of air pollution, smoking status, or other lifestyle factors such as diet and exercise. When using questionnaires, it is possible to design the survey to ensure the data perfectly matches the research question. This is particularly relevant for infectious diseases which inherently involve underlying random processes of transmission and social networks that are difficult to reconstruct from registry data alone. Additionally, there is a trade-off between registers limiting missingness in data and virtually eliminating loss to follow-up but the finest granularity of variables being yearly. It is possible that other variables in the cohort data will act as proxy variables for those that the registry data does not contain, but this is unlikely to be the case for all related variables. The current DANLIFE cohort is being updated to incorporate additional variables, including environmental factors such as pollution levels, home addresses and prescription records that allow for the inclusion for less severe medical conditions.

The definitions of the ACE variables may lead to some instances of misclassification of individuals [Rod et al., 2021]. For example, parental alcohol abuse is derived from medical records and consequently will not include individuals whose parents did not receive any kind of treatment for alcohol abuse. The construction of the ACE variables can also lead to individuals who are boundary cases and treated differently by the model whilst having very similar traits. For instance, an individual may fall just above the poverty threshold and are classified as not living in poverty which means that their health trajectory may be very similar to those classified as living in poverty and may lessen the perceived impact of poverty on future health. The categorisation of continuous variables prior to model fitting is thought to negatively impact analysis [Naggara et al., 2011]. For this reason, DL models are better suited to continuous data as patterns within the data are not masked by

artificial categorisation. Two additional continuous income variables were introduced to combat this effect. However, it was not possible to conduct feature selection, due to the computational limitations, and therefore it was not possible to remove nuisance variables that would improve model performance and reduce model training time. Moreover, some variables, and particularly the three variables relating to income and poverty, may be collinear. Multicollinearity can severely impact model performance and may explain the lack of signal in the data. Identifying instances of multicollinearity is a significant problem with MTS data as the dependencies between different variables can alter across the time domain [Chan et al., 2022]. In future work, several of the ACE variables could be replaced with their pre-categorisation continuous forms and feature selection methods for time series coulg be explored to improve the DL model performance.

4.6 Conclusions

The findings of this analysis have direct implications for the development of risk prediction models, particularly in the context of predicting various health conditions. The superior performance of ResNet in predicting diabetes and neurological conditions suggests that this model holds promise for constructing risk prediction models tailored to non-communicable health outcomes. Its ability to discern patterns in data related to these conditions signifies its potential utility in identifying individuals at risk and enabling targeted interventions. On the other hand, the challenges encountered by ResNet in predicting RTIs demonstrates the complexity of modelling infectious diseases. The observed flat optimisation surface indicates that the dynamics of RTIs may require a more sophisticated modelling approach or a fine-tuning of hyperparameters. This insight is crucial for refining risk prediction models focused on RTIs, ensuring they adequately capture the intricate relationships within the data.

The suboptimal performance of the Signature model across all diseases, particularly in terms its lower AUPRC scores, raises questions about its suitability for risk modelling in the current context. It was also found that the Signature model needed significant CPU resources to compute the signature coefficients when there were more input features, resulting in this model having comparable computational requirements to ResNet. However, with enhanced computational resources and the introduction of feature selection, the Signature model may still have potential within risk prediction applications. Overall, the findings within this Chapter underscore the sensitivity of these models to the specific health condition modelled, suggesting that a one-size-fits-all approach may not be suitable and both the models and the data need to be selected according to the particular disease outcome.

The experiments conducted in this chapter have demonstrated that DL models are capable of achieving excellent performance for some health outcomes, particularly noncommunicable diseases, by incorporating the dynamic temporal effects of exposure event occurrence. The next step in this analysis is to examine how several explainability methods could be used to obtain insights into the most influential exposures and time periods in generating disease risk. This will form the final step in demonstrating how DL models can be used within risk factor analysis to identify high-risk individuals as well as potential risk factors. Additionally, combining DL models with explainability methods is essential for these methods to be used as risk prediction models within healthcare settings. This work will act as the foundation for further studies that explore how the life course mechanisms shape an individual's health trajectory.

Chapter 5

Explainability Approaches for Deep Learning Methods

5.1 Introduction

Deep learning (DL) models are often described as black-box models, meaning that the precise way in which the input data results in a given model prediction is difficult to quantify. This is especially undesirable in healthcare settings where it is crucial to justify the reasoning behind medical decisions, for example, deciding which individuals receive a specific treatment or identifying high-risk individuals for intervention programs [Shaban-Nejad et al., 2021]. DL models can contain hidden biases which, if not identified, can result in misdiagnoses, unnecessary loss of life years, and multiple other negative outcomes. Knowledge of the reasons behind DL model decisions, helps to avoid medical mistakes and ensure best practice. Whilst DL models have the ability to obtain very high accuracy scores when applied within such medical contexts, a desire to understand their decision-making process has resulted in a rise in explainable artificial intelligence (XAI) methods to make these models less opaque [Theissler et al., 2022]. There are other benefits of understanding the inner model workings, in addition to providing a greater understanding and therefore, trust in their decisions. Firstly, use of explainability methods enables the user to test for potential model biases and ensure cohesion with ethics guidelines. The model can also be checked for robustness, meaning if the model can obtain similarly good performance when exposed to slightly different conditions to those with which it was trained. Overall, the notion of explainability allows for these highly accurate and personalised models to be supplemented so that they can be safely used within healthcare.

The term explainability is often used interchangeably with the term interpretability within the literature, however, they are two distinct concepts with some key differences [Vilone and Longo, 2021]. The definition of explainability that will be used here is the notion of being able to understand the way in which each input variable impacts the model outcome. In contrast, interpretability generally refers to how well the models' workings can be understood by a human. Some models, such as linear regression, generalised linear models, and decision trees, are interpretable by design. For example, each coefficient in a linear regression model gives the magnitude of influence of each input variable on the output. Moreover, the sum of the coefficients multiplied by the input variables is equal to the output, so it is possible to identify the reasons behind every model decision. These kinds of models are frequently referred to as being transparent indicating that it is easy to see the model workings. Whereas DL models map the input variables to an output via multiple complex nonlinear interactions, so these models are not inherently interpretable or called opaque. The desire for justifiable

model decisions has led many researchers to use interpretable models, however, explainability aims to allow for more complex and accurate models to be used in these settings.

In the context of this body of work, explainability methods enable DL models to be used as risk prediction models. Traditionally risk prediction models take the form of additive risk scores, where a patient is asked whether they have experienced multiple exposures, such as family health, smoking status and their alcohol consumption. Then each positive answer is equated to a number of points, which are summed once the questionnaire has been completed. There may be multiple thresholds that the final score for a patient is compared to, these threshold each relate to a risk level and may correspond to some advice, treatment, or intervention. Within the landscape of RTIs, many medical professionals use the pneumonia severity index (PSI) to estimate the mortality risk of adults with community acquired pneumonia [Flanders et al., 1999]. The PSI score is calculated as the weighted sum of multiple indicators including age, sex, liver disease history and blood pressure. The resulting score, which lies between 8 and 400, indicates the level of mortality risk. However, this risk prediction model is very simplistic and is known to be overly impacted by the patients age and consequentially, this model may miss severe pneumonia cases in young adults. The DL models have the potential to be far more accurate and explainability methods allow for underlying biases to be identified and addressed. This framework means that accurate and personalised risk prediction models can be developed that are able to capture dynamic inter-variable interactions with the potential to be used for making medical decisions.

In addition to explainability, it is possible to use the DL model to estimate the relative risk, otherwise known as the risk ratio, that is associated with each exposure. The relative risk is used to quantify the influence of each exposure on the outcome and can which determine which exposures represent significant risk factors for respiratory tract infections (RTIs). Risk ratios measure the increase in risk associated with having a certain characteristic or experiencing a specific exposure event. These ratios are calculated by diving the risk in the exposed group by the risk in the unexposed group, with the risk in a group traditionally estimated using a logistic regression or Cox proportional hazard model. However, as discussed in Chapter 1, these methods do not account for the complex, nonlinear interactions that exist between the exposures and the outcome. Another possible approach is to use the fitted DL model to estimate the risk within both groups which has the potential to produce more accurate risk ratio estimates that reflect the dynamic nature of the interactions between exposures. Then the greater the risk ratio, the more likely it is to act as a risk factor that can determine RTI-related hospitalisation risk.

In this chapter, several explainability methods will be examined with a focus on how they can complement the ResNet model. The XAI approaches will allow a thorough investigation into which features and time points are most influential in predicting the disease outcomes with the aim of identifying developmental periods or potential risk factors. Moreover, the impact of exposure timing will be examined and in particular how the timing of different ACE exposures relates to future disease risk. Finally, risk ratios will be estimated for each exposure using the DL model predictions and these will be compared to estimates within the literature.

5.1.1 Overview of Explainability Methods

The XAI approaches can be broadly divided into two distinct categories: model-specific and model-agnostic methods. Model-specific XAI methods are explicitly integrated within the ML architecture and generate estimates of the influence of each feature each time the model is trained. In comparison, the model-agnostic XAI schemes are applicable to any ML model architecture and are applied as an additional step after the model has been trained, referred to as a post-hoc method. Examples of model-specific methods for neural networks (NNs) include saliency maps and layer activation maps [Sharma et al., 2017; Simonyan et al., 2014]. The limitations of model-specific methods are that these approaches can only be used alongside certain specific

families of ML models. Model-specific approaches that are designed for simple models lack modelling accuracy and those designed for DL methods tend to have high associated computational costs. In contrast, model-agnostic methods do not depend upon the specific model assumptions or architectures and are more versatile and generalisable to a wide range of frameworks. In this research, the ResNet model was chosen because it had good performance and was more computationally feasible than the other DL models. Therefore, the XAI methods that will be explored in this analysis are model-agnostic.

Another categorisation of XAI methods is whether the method outputs local or global explanations [Saluja et al., 2021]. Local explanations are individual-level explanations which highlight the most influential features in predicting a single outcome. An example of the type of question that this kind of explanation can be used to answer is *"Why has this individual been classified as someone who will be hospitalised with an RTI?"*. On the other hand, global explanations examine the most important features at the population-level, showing the overall impact of each feature on the model outcomes for the entire population. The kind of question which may be answered using global explanations is *"Which exposures have the greatest impact on whether people are hospitalised with an RTI?"*. In general, it is useful to study model explainability through both global and local lenses to gain an overarching understanding of the patterns detected by the model.

There are multiple, model-agnostic explainability methods that are appropriate for use with the ResNet architecture [Naser, 2021]. Global explainability methods include partial dependence plots (PDPs), individual conditional expectations (ICEs) and permutation feature importance (PFI). PDPs inspects the relationships between each input feature and the predicted outcomes and this approach depends upon the assumption that features act independently to each other [Friedman, 2001]. The procedure used to create PDPs inputs every plausible value of each feature into the model one by one and plots the resulting outputs to allow for comparison. ICEs are an extension of PDPs that break down the PDPs into individual observations to create a curve [Goldstein et al., 2013]. PFI on the other hand, is a more straightforward and computationally efficient technique that is more applicable to multivariate time series (MTS) data [Altmann et al., 2010]. It assesses feature importance by evaluating how the model performance deteriorates when feature values are randomly permuted. The advantages of PFI include its simplicity and speed, which are particularly useful for quickly assessing the relevance of features in MTS data. However, this method is primarily focused on global interpretability and may not capture complex feature interactions effectively, particularly the temporal relationships present in MTS data. Additionally, this approach is sensitive to the choice of hyperparameters, such as the number of permutations investigated, and consequently may exhibit some randomness in its results. Motivated by its computational efficiency and ease of calculation for MTS, the PFI method will be used within the following explainability analysis.

Local methods include Local Interpretable Model-agnostic Explanations (LIME) and SHAP. The main goal of LIME is to approximate the decision boundary for each individual in the population separately [Ribeiro et al., 2016]. This method works by randomly perturbing each input features for a specific individual in order, then the perturbed features are then fed to the model to produce new model predictions. Then a locally interpretable surrogate model, such as a linear regression or decision tree, is fitted to the perturbed instances and their corresponding predictions. Weights are assigned to the perturbed instances based on their distance from the original data with closer instances receiving higher weights. Another explainability approach, known as SHAP, provides a theoretically grounded framework for interpreting model predictions [Molnar, 2022].The theoretically underpinnings of SHAP have made it the gold standard for XAI in the community and therefore this approach will also be examined in this analysis. SHAP works by assigning each feature in an single prediction a Shapley value which represents its contribution to the prediction. SHAP values are able to offer both global and local interpretability, allowing users to understand feature importance across the entire dataset as well as for individuals. The key advantage of SHAP is its consistency and its ability to handle feature interac-

tions effectively, making it well-suited for identifying intricate relationships in MTS data. However, the method has the potential to be computationally intensive, which may be a drawback in certain applications.

5.1.2 Permutation Feature Importance

PFI is an example of a model-agnostic global explainability method [Altmann et al., 2010]. This method compares the relative reduction in the value of the evaluation metric when each feature in the input data is randomly shuffled with the resulting perturbed data input into the model. The rationale behind this approach is that when the model is highly dependent on a given feature, permutating that feature will lead to the biggest drop in the value of the evaluation metric because the connections that the model is depending on are severed. In the case of MTS data, it is possible to either permute each feature to identify influential exposures or to permute each time point to find the ages that most impact model predictions. The PFI method is generally performed across the entire cohort, with the selected feature being permuted for all the individuals in the data set at once, producing the feature significance at the population-level. This method was selected to be assessed in this analysis because functions to perform PFI already exist within the `tsai` package so implementation of this method is fast and straightforward [Oguiza, 2022].

The PFI algorithm was first introduced by Fisher et al. and the explanation of the approach given below is based upon their description [Fisher et al., 2018]:

1. Let the trained model be denoted by f, the input matrix of data by denoted by $X \in \mathbb{R}^{N \times M \times T}$ where N is the number of individuals in the population, M is the number of model features and T is the number of time steps. Additionally, let the model output be denoted by $y \in \mathbb{R}^{N}$.
2. The baseline value of the evaluation metric of interest can be written as $E = M(y, f(X))$.
3. For each feature i within the total number of features M:
   (a) For all individuals in population N, permute feature i in the input matrix to obtain a new input matrix $\tilde{X}$.
   (b) Estimate the new evaluation metric value with the new input data as $\tilde{E} = M(y, f(\tilde{X}))$.
   (c) Then feature importance for feature i is defined as the difference between the two values of the evaluation metric: $\text{PFI}_i = \tilde{E} - E$.
4. Compare the resulting feature importance values for each feature.

The above description explains the procedure for conducting PFI to identify influential features, however, it is also possible to perform PFI across the time steps. The procedure used to perform PFI for time steps is exactly the same as the procedure described above, except that the permutations are conducted in the time dimension of the input data instead of the feature dimension.

A key limitation of this method is that if interactions exist between multiple features, or indeed multiple time points, then the perturbation of either feature will sever these connections and as a result the value of the evaluation metric will drop by the same amount for both features. For example, if there is a relationship between experiencing poverty and a hospitalisation in the same year that magnifies the likelihood of the outcome then they will both have the same feature importance. This phenomenon has two main consequences; firstly, it is not possible to easily identify these kinds of interactions using the PFI method, and, secondly, the changes in the evaluation metric for each feature do not sum to one unless all exposures act independently. Whilst this knowledge can imply that there are inter-variable interactions that are related to the disease outcome, it is not generally possibly to pinpoint the contributing features or nature of their relationship using PFI. When investigating time step importance using PFI, this effect is particularly important to consider since there is usually

significant autocorrelation between adjacent time points meaning that the PFI can be biased by unrealistic data instances which may not be observable in reality.

5.1.3 SHAP

The SHAP (SHapley Additive exPlanations) method is perhaps the most widely used measure of feature importance and explainability. It provides local model-agnostic explanations for each feature at each time step individually. The approach originated in cooperative game theory as a way to determine the proportion of the groups' points that is attributable to each player [Shapley, 1952]. For instance, if a team have played a cooperative game achieve a certain score, then the SHAP method seeks to allocate which points were contributed by each player so that any game rewards can be shared accordingly. In the case of an application to MTS, the SHAP values for a single individual represent the extent to which each feature at each time point contributed towards the overall model prediction, i.e. whether the is predicted to have an RTI. These individual level insights can be used to identify the features that drive the individual's prediction or averaged across the population to provide a global-level understanding of the most important features overall.

SHAP values are approximations of Shapely values that are less computational expensive to calculate. The Shapely value, which represents the proportion of the outcome that is attributable to a given feature, is defined as the average marginal contribution of that feature across all possible subgroups of features [Molnar, 2022]. The interaction effect is the additional combined feature effect after accounting for the individual feature effects. Assuming that a subset of p features is denoted by S, then the Shapely value for a feature $m \in \{1, \ldots, p\}$ is given by [Lundberg and Lee, 2017]:

$$\phi_m(v) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{m\}} \frac{|S|!(p - |S| - 1)!}{p!} (v(S \cup \{m\}) - v(S))$$

where $v(S)$ is a function that gives the value of the DL model prediction for features in set S marginalised over features not included in subset S:

$$v_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

where x is the vector of feature values of the instance to be explained, S is the set containing all the possible subsets of the different feature values ignoring feature m, $v(S)$ is the value achieved by that subgroup S, and $v(S \cup \{m\})$ is the realised value after m joins S.

Shapely values can be positive or negative which signifies whether a particular feature increases or decreases the output value, i.e. the likelihood of hospitalisation with an RTI. The magnitude of the Shapely value reflects the strength of any association between the feature and the output. The resulting Shapley values represent the average contribution of each feature to the difference between the model's prediction and the average prediction. Shapely values, along with SHAP values, satisfy several key mathematical properties that are referred to as efficiency, symmetry, noncontribution and additivity. These characteristics are what give the Shapley Values the ability to provide explanations at both the global and local level [Lundberg and Lee, 2017]:

- Efficiency: the sum of the shapely values over all the different features is equal to the difference between the original prediction and the average prediction.
- Symmetry: if two features contribute equally to all subgroups, then their SHAP values will be the same.
- Non-contribution: if a feature does not contribute to any subgroups then it will have a corresponding SHAP value of zero.

- Additivity: The SHAP value of a feature at the final layer of an NN is equal to the sum of its SHAP values from the sub layers.

There are multiple ways to calculate SHAP values including KernelSHAP, TreeSHAP and GradientSHAP. KernelSHAP uses weighted linear regression to approximate Shapely values with sampling from the marginal distribution. This ignores the feature dependencies and means that the estimation also includes unlikely instances. This is a problem shared by many other explainability methods that depend upon permutation. TreeSHAP solves this problem by modelling conditional expected prediction. TreeSHAP is designed for tree-based ML models like RFs and XGBoost [Lundberg and Lee, 2017]. However, TreeSHAP can produce unintuitive feature attributions where features that have no influence on the prediction can get a TreeSHAP value different from zero. GradientSHAP is designed specifically for DL models and approximates the SHAP values using the gradients of the model predictions with respect to the input features. For a specific individual's MTS x, the gradients of the model output are $\nabla_x f(x)$. The approximation of the SHAP for each feature is calculated as:

$$\phi_i = \sum_{j=1}^{M} (x_j - E[x_j]) \cdot \nabla_{x_j} f(x)$$

where ϕ_i is the SHAP value for the ith feature, M is the number of features, x_j is the value of the jth feature for the input instance x, and $E[x_j]$ is the expected value of the jth feature.

5.1.4 Research Questions and Objectives

The main objective of this chapter is to introduce explainability methods to the DL model framework to allow for a deeper investigation into life course patterns that give rise to disease in the DANLIFE data set. Due to sub-optimal model performance for the RTI outcome, this analysis will focus on three outcomes with diabetes mellitus and neurological conditions also be investigated in addition to RTIs. Therefore, the results for the non-communicable methods will be able to demonstrate the utility of XAI methods for well-calibrated models that have good predictive performance. The overarching goal of this analysis is to demystify the complex predictions generated by the DL models, providing meaningful insights into the features and patterns contributing to disease risk. The specific objectives include identifying influential features, understanding the temporal dynamics captured by the models, and assessing the impact of different life course events on an individual's future health. Each XAI method, including SHAP and PFI, is applied to the DL models that have been trained on the DANLIFE data to interpret their predictions. Throughout the process, attention is given to the unique challenges posed by longitudinal health data and the implications of the findings for both model improvement and real-world applications. It is important to explore what insights explainability methods can give us about the inner workings of the model and any potential biases present. Using multiple explainability methods, the Chapter aims to uncover the black-box of DL models, making the decision-making process transparent and interpretable for both healthcare practitioners and researchers.

5.2 Methods

This section will describe the methods used to conduct PFI and calculate the SHAP values. Additionally, the way in which risk ratios can be estimated using DL models will be explained. Then the experimental design will be set out.

5.2.1 Permutation Feature Importance

There are inbuilt functions within the `tsai` package version 0.2.22 in python that can conduct PFI for both the input features and the time steps [Oguiza, 2022]. These analyses were conducted on the test set alone. Feature importance was measured using two evaluation metrics which are the AUROC and AUPRC scores. The influence of features or time steps will be measured in the proportion change in these metrics which is defined as

$$\text{Proportion change in metric} = \frac{\text{new metric value} - \text{baseline metric value}}{\text{baseline metric value}} \tag{5.1}$$

where the new metric value is with the feature permuted and the baseline metric value is without any permutation. Additionally, the raw AUROC scores are transformed so that this metric begins from zero instead of 0.5 to ensure that the proportion of change in AUROC values range between zero and one.

5.2.2 SHAP Values for Feature Attribution

SHAP values were computed in python using the GradientSHAP approach with package `shap` [Lundberg and Lee, 2017]. The explainer was fitted to all the training data and then SHAP values were calculated for 2000 individuals in the test set. The SHAP values were averaged across the population to ensure anonymity and comply with data protection agreements. The standard deviations of the SHAP values were also calculated to demonstrate the variability in the values.

5.2.3 Estimation of Risk Factor Scales

The method used to estimate risk ratios from the DL model outputs is based upon the work by [Dasgupta et al., 2014]. They find that ML models with binary outcomes can provide consistent probability estimates. This analysis is designed to reflect the potential epidemiological questions that can be answered using a well-calibrated DL model. Therefore, as the DL model for RTIs had poor predictive and calibration performance, the model with diabetes as the outcome is examined to highlight the utility of these methods. The individuals in the test set are split into two groups according to their MTS that records their individual characteristics and exposures. The risk of the disease outcome is calculated within each of the subgroups and divided. The risk ratio can be given as

$$\text{Risk ratio of exposure } X = \frac{\text{Risk in exposed group}}{\text{Risk in unexposed group}}$$

Often the risk ratios are calculated using the raw data or a fitted regression model, however, there are multiple benefits of using a DL model instead. Firstly, the logistic framework assumes linear relationships exist between the input features and the outcome whereas DL models can reflect more complex interactions. Consequently, the risk ratios allow a more in-depth exploration of the relationships that give rise to increased risk. It is important to consider that any bias present in the model will also be present in any estimates of the risk ratios. However, a well-calibrated DL model that has not overfit the training data will produce calibrated risk ratios. Several exposures were assessed based upon features that were found to be significant by PFI and SHAP. These are sex, parental origin, previous diabetes hospitalisation, poverty, parental alcohol abuse and maternal separation. It is also possible to split the population according to when the exposure occurred to explore the impact of timing.

5.2.4 Experimental Design

First, PFI will be conducted for the ResNet model fitted in turn to the three different outcomes, RTIs, diabetes mellitus and neurological conditions. This process will be conducted for two different dimensions to enable an examination of the most influential features and also time points for predicting the outcomes. The change in model performance is measured in terms of the proportion of change in both the AUROC and AURPC metrics. Then the SHAP values will be calculated for the simulated data described in Chapter 2. Two data sets will be explored which are Period2 and Timing1 which are described in Table 2.3. The ResNet model with no additional noise will be explained in each case. The nature of the simulated data means that individual level SHAP plots can be shown to demonstrate the key features of SHAP values and how they relate to exposure events. The corresponding average SHAP plots for the population of 500 individuals will be shown to illustrate the population-level insights that can be gained from the SHAP values. The SHAP values will be calculated for the ResNet model fitted to the three outcomes using the DANLIFE data. Due to the data protection rules, the individual level SHAP plots cannot be presented for the DANLIFE data and instead the average SHAP plots over 2000 individuals will be shown and their corresponding standard deviations. Finally, multiple risk ratios will be calculated demonstrate the power of this analysis in answering traditional epidemiological questions of interest. This will be carried out for the diabetes outcome since that model attained the highest performance and was also the best calibrated of the three outcomes.

5.3 Comprehensive Comparison of Interpretability Methods

5.3.1 Permutation Feature Importance

Features

Figure 5.1 shows the proportion change in the AUPRC and AUROC scores for the RTI outcome, which reflects how much the prediction of the output is attributable to each feature. In the case of the RTI outcome, the overall model performance was poor (Chapter 4) and therefore, a high proportion of change corresponds to a very small increase in the raw value of the evaluation metric. Consequently, these results are designed to compare the relative influence of the features and how the small amount of signal that does exist can be attributed to the different features. The baseline AUPRC and AUROC scores for the unpermuted input data are 0.105 and 0.676, respectively. There approximately 50% proportion change in the AUPRC score due to respiratory infections before age 16 which relates to a change of only 0.053 in the raw AUPRC score. Both evaluation metrics, pinpoint occurrences of respiratory infections prior to age 16 as being the most informative feature for predicting future RTI instances. The metrics also agree that ICD-10 Chapter 19 hospitalisations has is the second most influential feature in model predictions. The sex and year of birth variables also feature highly when measured with either metric, followed by age and infectious and parasitic diseases. The lower ranking features show more difference between the two metrics. The AUPRC metric suggests that the model has a greater dependence on parental unemployment, parental separation, and low birthweight. In contrast, the next most important features when measured by AUROC are parental origin, parental separation and parental education. Both metrics indicate that there are many features that do not contribute towards model performance and some that contribute negatively. Notice that the GHE category of respiratory diseases is different from respiratory infections, the diagnostic codes that make up each category are described in Table 8.1 which also breaks down the conditions that fall within the diabetes and neurological conditions as well.

The trends in the PFI plots for the diabetes mellitus outcome are noticeably different to the RTI outcome (Figure 5.2). The ResNet model achieved excellent performance with diabetes as the outcome, with base-

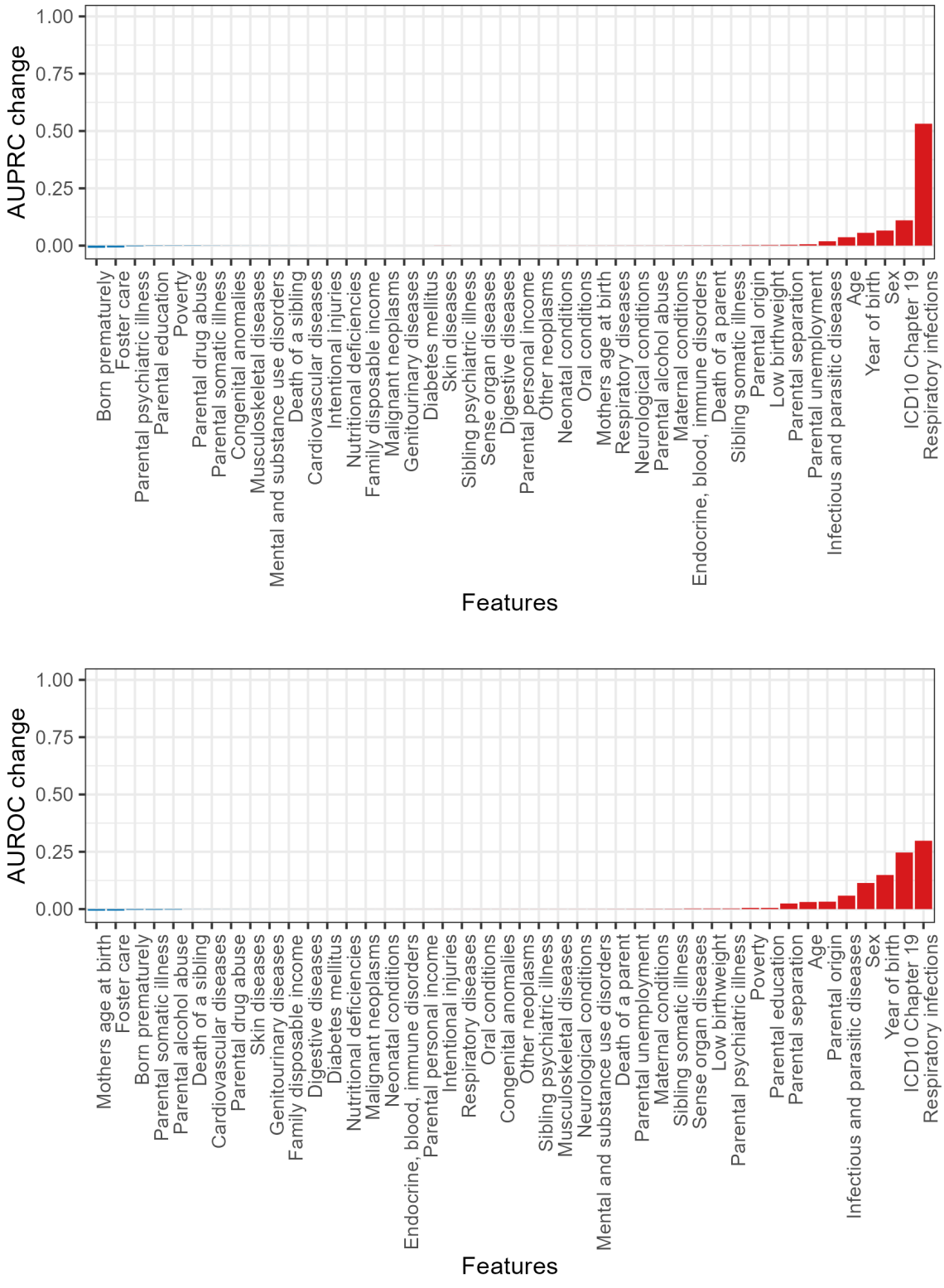


Figure 5.1: **Permutation feature importance with respiratory infections as the outcome.** The proportion change in the AUPRC (top) and AUROC (bottom) when each feature is permuted. Red indicates positive change in evaluation metric and blue indicates a negative change.

line AUPRC and AUROC scores of 0.741 and 0.897, respectively. Consequently, for the diabetes mellitus outcome, a large proportion of change implies a large change in the raw metric value. Following the same pattern as RTIs, the most influential feature identified by both metrics is previous hospitalisations with diabetes prior to age 16 and almost 100 percent of the value of both metrics is attributable to this feature. All the other features within the model are comparatively insignificant in their impact on model performance. The only other variables that noticeably improve model performance are the year of birth and age. There are a large number of remaining features that do not contribute to the signal and are uninformative. Several features decrease model performance when measured by AUROC, including poverty, mothers age at birth and low birthweight. Correspondingly, the AUPRC highlights parental separation, sex and parental origin as features that are potentially detrimental to performance. Therefore, when measured by AUPRC, inclusion of the sex variable is detrimental to model performance, whereas with the AUROC metric, sex is the fifth most influential variable in improving model performance.

The difference between the variables when measured with each of the two evaluation metrics is more pronounced for neurological conditions (Figure 5.3). The AUPRC metric picks out two features that are most influential with all other features having limited impact on model performance. Whereas the AUROC score appears to assign similar levels of influence to multiple features. The biggest predictor of neurological conditions in the follow up period is prior hospitalisations with neurological conditions as the main diagnosis. Accordingly, the proportion of change in the AUPRC value is close to one for this feature. The baseline metric values for neurological conditions are an AUPRC score of 0.299 and an AUROC score of 0.752. Both metrics highlight age as the next most important feature for model predictions, followed by year of birth. The remaining features have a minimal impact when measured using the AUPRC score and include sex, parental origin, ICD-10 Chapter 19 and premature birth. The AUROC assigns a higher importance to multiple features, particularly sex, followed by ICD-10 Chapter 19, parental education and parental origin.

Time steps

The results of PFI across the separate time steps for each of the three outcomes is displayed in Figure 5.4. The baseline values of model performance are the same for each outcome as given above. The AUPRC metric seems to indicate that, for all outcomes, the time step that directly precedes the outcome observation period is the most influential in model predictions, followed by the time step before it. For all outcomes, the early years of life have a limited impact on the model performance when measured by the AUPRC score. However, the proportions of change in the AUROC metric suggest that early years play an important role in generating an RTI outcome. Overall, the RTI outcomes is most influenced by the very recent years followed by the first three years of life. The AUROC scores for neurological conditions show a peak in the proportion of change around the ages of seven to nine as well as an increase at ages 15 and 16. Diabetes only depends on the year proceeding the outcome observation period when measured by AUROC. In fact, for the diabetes outcome, it appears that some time steps contribute slightly negatively towards model performance when measured by the AUROC metric.

5.3.2 SHAP

Simulated Data

Figure 5.5 shows individual level SHAP value plots for two individuals in the simulated Period2 data set; one true positive and one true negative. The Period2 data set is designed so that if four or more mild illness exposure events occur within the first ten years of an individual's life they get the disease outcome. The SHAP values for an individual indicate the proportion of the predicted outcome which is attributable to each feature at each time point. In the top figure, the individual has had six incidents of mild illness at ages 1, 4, 9,

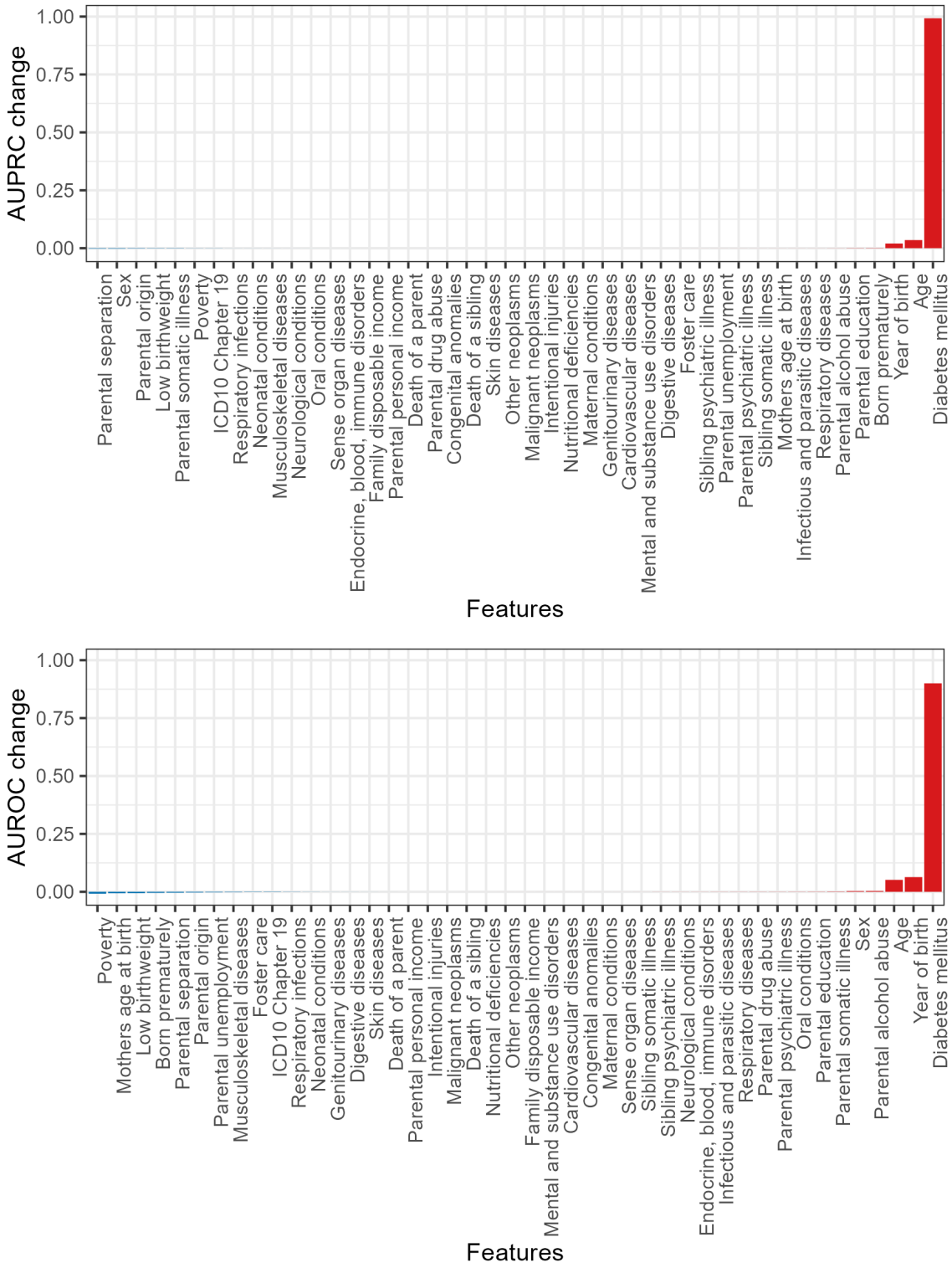


Figure 5.2: **Permutation feature importance with diabetes mellitus as the outcome.** The proportion change in the AUPRC (top) and AUROC (bottom) when each feature is permuted. Red indicates positive change in evaluation metric and blue indicates a negative change.

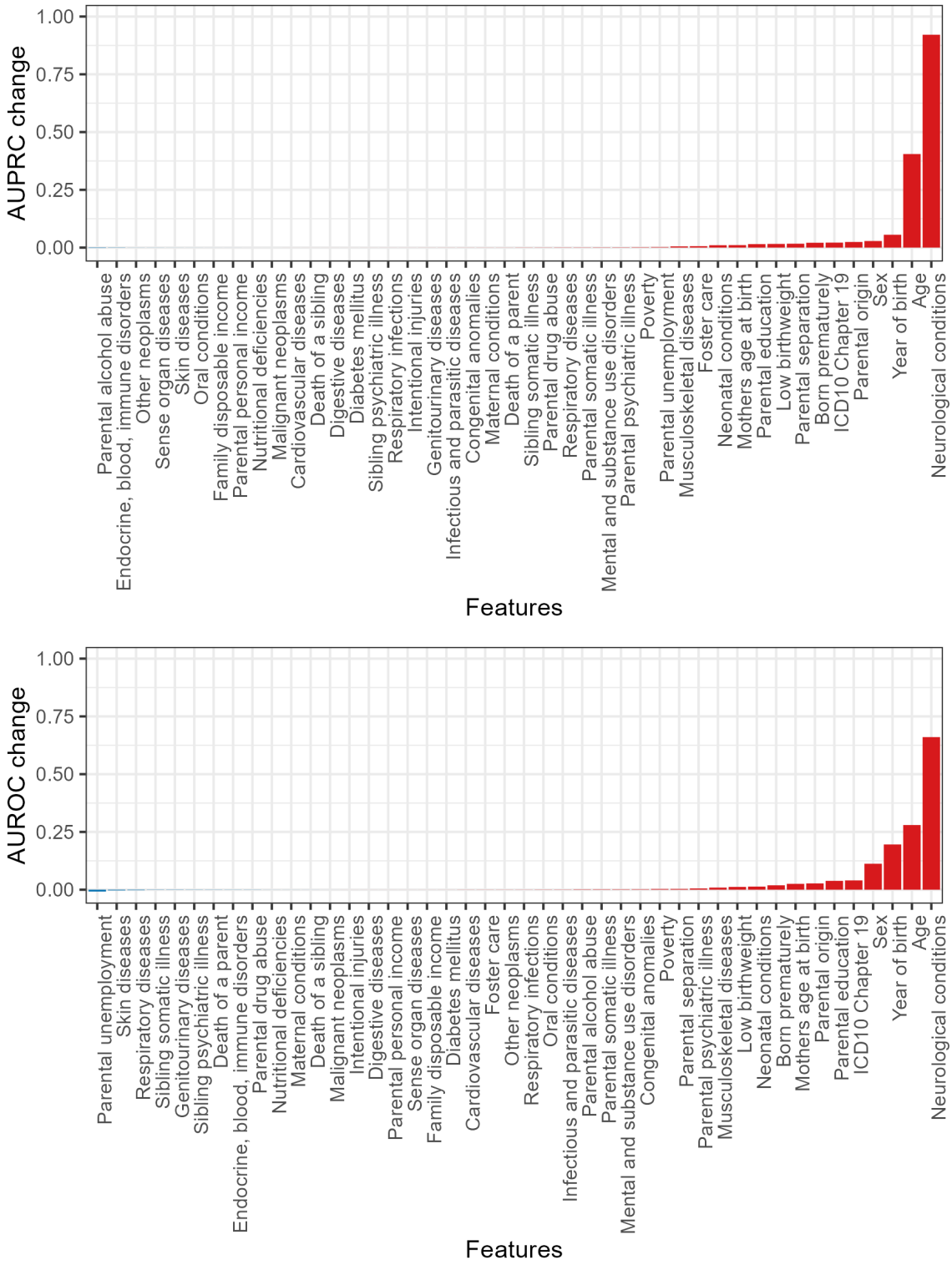


Figure 5.3: **Permutation feature importance with neurological conditions as the outcome.** The proportion change in the AUPRC (top) and AUROC (bottom) when each feature is permuted. Red indicates positive change in evaluation metric and blue indicates a negative change.

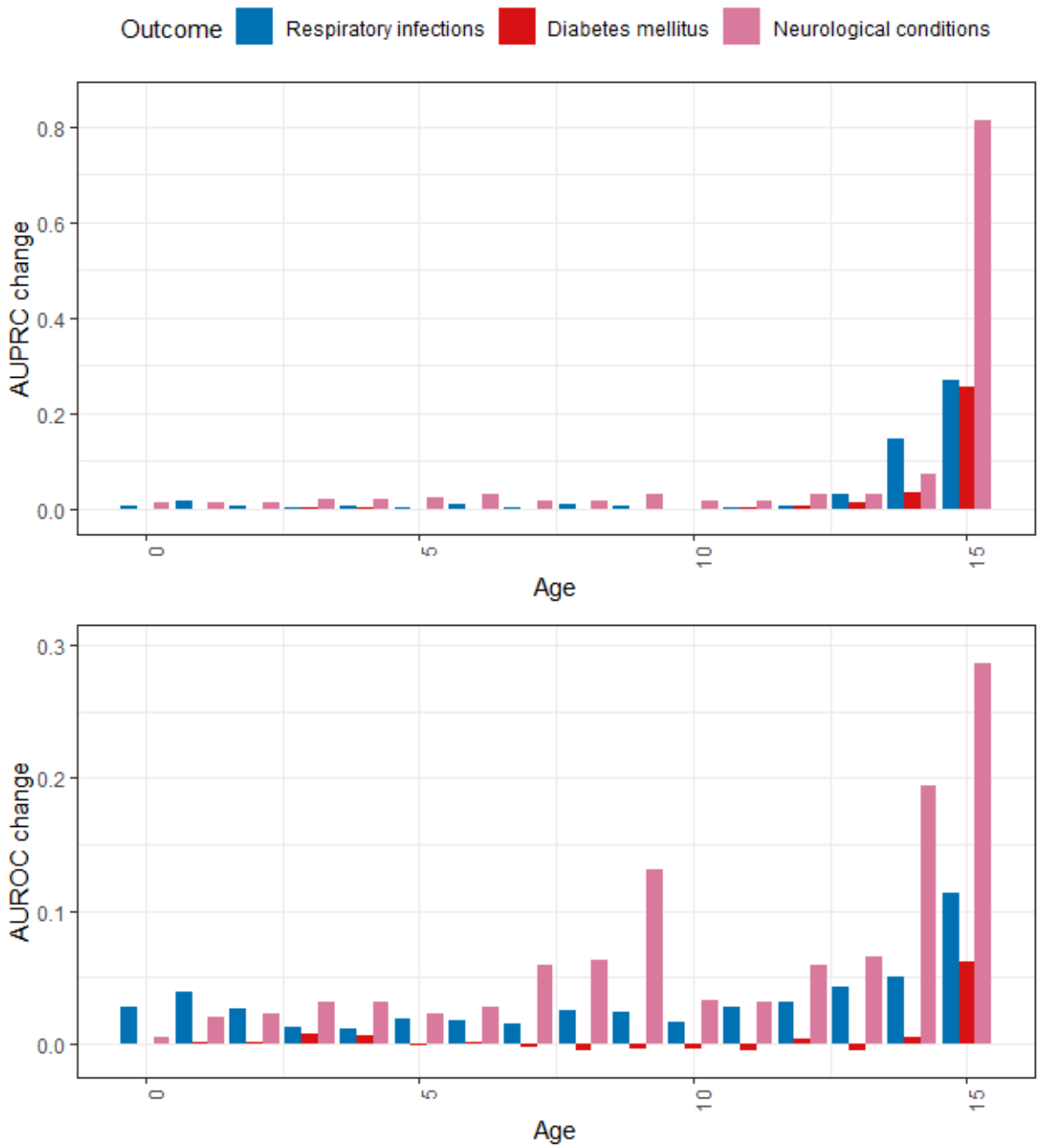


Figure 5.4: **Permutation feature importance for time steps.** The proportion change in AUPRC (top) and AUROC (bottom) metrics when conducting PFI.

10, 21 and 23. Four of these exposures are within the first ten years of life which led to the individual having a positive outcome. The SHAP values that correspond to each incident of mild illness prior to age ten are large positive values, indicating that these exposure occurrences had high influence on the model prediction of a positive outcome. No exposures occurred in the remaining years prior to age ten and correspondingly these time points have negative SHAP values that signify a negative association with the positive outcome. There are two further years where mild illness occurred and these are after age ten which have small negative associated SHAP values. Moreover, there are small positive SHAP values related to incidents of poverty and sibling illness before age ten and these variables increase the likelihood of mild illness in the data simulation. The bottom figure shows an individual who had three mild illness events in the first ten years of life which meant they had a negative outcome. As this individual was classified in the negative class their SHAP values have the opposite directions of associations. For example, the three mild illness exposures that occur prior to age ten have negative SHAP values and those that occur after the threshold have positive SHAP values. All the other features have small SHAP values as they have limited influence on the final classification.

The DANLIFE data contains the life courses of real individuals and therefore, the individual level SHAP plots cannot be shown. In this thesis it is only possible to present population level aggregations to ensure the data protection agreement is adhered to. Figure 5.6 shows the average SHAP values across the 500 individuals and the Period2 simulated data set, regardless of their actual or predicted classes. As a result, the average SHAP values indicate which features and time points most significantly and consistently contributed to model classification in general. As expected, the first ten years of the mild illness variable have high positive SHAP values, signifying that these variables have a big influence on model prediction. The average SHAP values for mild illnesses after this critical period are smaller and show associations in both directions which may be an artefact of the averaging procedure. All other variables have low average SHAP values indicating minimal contribution to classification.

The Period2 group is centred around a simple linear rule that generates the positive outcomes. A more complex pattern used to generate positive outcomes in the Timing1 simulated data set where positive outcome is generated if there are two occasions in the life course when a mild illness occurs in the same year as an exposure to parental alcohol abuse. Again, the life courses for two individuals, each correctly classified in one of the two classes are shown (Figure 5.7). The individual who is a true positive has a life course with two instances where the mild illness and parental alcohol abuse exposures occur at the same time. The SHAP values associated with these events have large positive SHAP values for both variables showing that these events have strong positive association with the positive class prediction. Also notice that the other parental alcohol abuse exposures that do not coincide with a mild illness exposure have small positive SHAP values. However, the corresponding non-occurrence of coinciding mild illness exposures are related to negative SHAP values of similar magnitudes that counteract the small positive SHAP values. The bottom figure shows the life course and corresponding SHAP values for a true negative individual. This individual only has one instance where the mild illness exposure and parental alcohol abuse exposure occur in the same year which is not enough to generate a positive outcome. This instance has large negative SHAP values for both features because this pattern is related to a positive outcome and is therefore, negatively associated with the negative classification predicted by the model. Similarly, the impact of the other unpaired mild illness or parental alcohol abuse exposure events also act in the opposite direction, with occurrences related to negative SHAP values and the corresponding non-occurrences assigned positive SHAP values. This indicates that the non-occurrence of the corresponding event is linked to the negative classification.

When the SHAP values are averaged across the population of 500 individuals, it can be seen that mild illness and parental alcohol abuse are correctly identified as being the most influential variables (Figure 5.8). Both features have SHAP values with large magnitudes and associations in both directions. The magnitudes of the

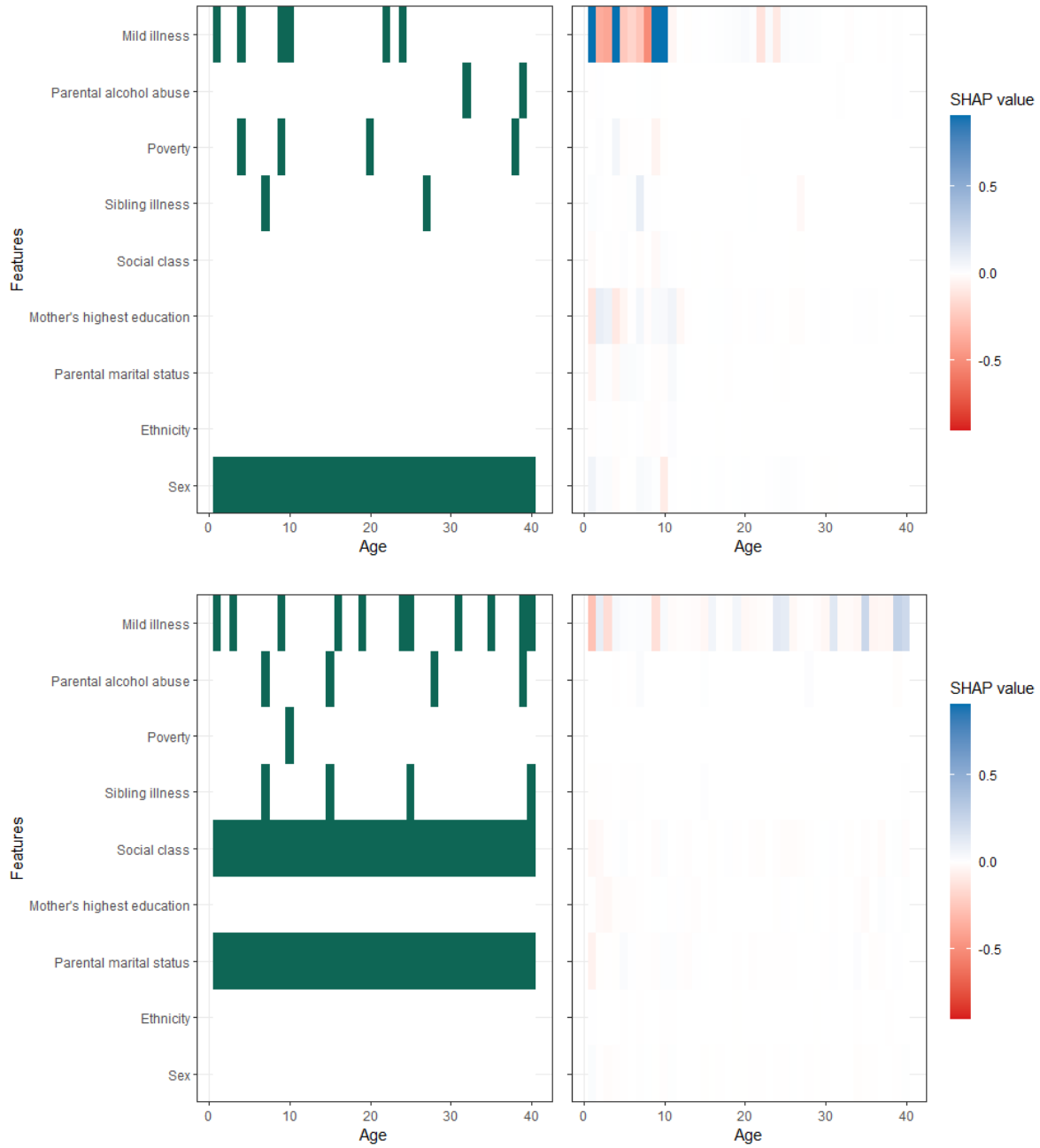


Figure 5.5: **Individual level SHAP values for simulated data set Period2.** The SHAP values for two example individuals in the Period2 simulated data set. The left figures show their exposure events which are binary with green values equal to one and zero otherwise. For each the x-axis shows the different time points in the life course and the y-axis gives the different features in the model. The right figures then show the corresponding SHAP values with blue values being positive and red being negative. The top individual is a correctly classified positive case, and the bottom individual is a correctly classified negative case.

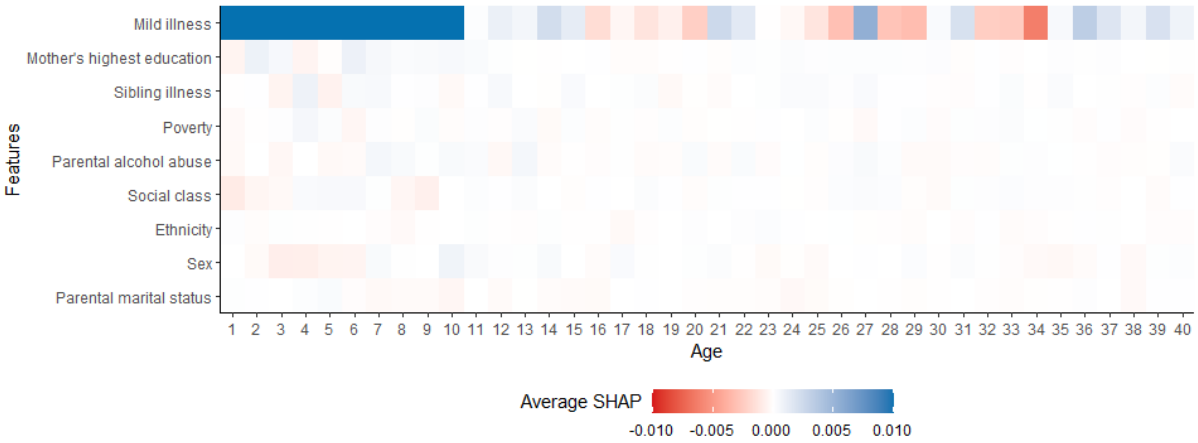


Figure 5.6: **Average SHAP values for simulated data set Period2.** The average SHAP values across 500 individuals in the Period2 simulated data set. Blue cells indicate positive average SHAP values and red shows the values are negative.

SHAP values of these features vary in part due to the small population size. It is reasonable to assume that these have been affected by values cancelling each other out as they act in opposite directions as shown in the individual level plots. There are some other time points and features that have small magnitude SHAP values, such as sex and sibling illness.

DANLIFE Data

The average SHAP plot for the RTI outcome is ordered according to the average magnitude and direction of each feature across all the time steps (Figure 5.9). The scale in this plot has been capped to improve interpretation and the uncapped plot can be seen in Appendix 8.4. Similarly to the PFI results, the variable with the greatest related average SHAP values is previous RTIs meaning it has the largest influence on model predictions across the population. The next most important variables in order are ICD-10 Chapter 19, year of birth and infectious and parasitic diseases. The SHAP values presented here are averaged over 2000 individuals and therefore, these variables have a definite population-level impact, but it is not necessarily the case that the variables with little population-level impact do not have substantial impact on individual level predictions. Individual level explanations can be produced but cannot be shown here due to data security. The RTI feature has different levels of importance at different time steps. The two most recent years have the greatest impact on model predictions as well as the early years. There are positive and negative average SHAP values associated with ICD-10 Chapter 19 hospitalisations and on average it is the second most influential variable. Additionally, several years of birth have been picked out as significant with average SHAP values greater than zero. Another influential feature is infectious and parasitic diseases which has varying directions of associations depending on the time period. One variable that has a negative association overall is poverty. The corresponding standard deviation plot shows that ICD-10 Chapter 19 has the greatest average value (Figure 5.10). This corresponds with the high level of variability shown on the average SHAP value plot. It picks up parental unemployment which had an average SHAP value close to zero but a greater level of variation in SHAP values. All other variables with high standard deviations correspond

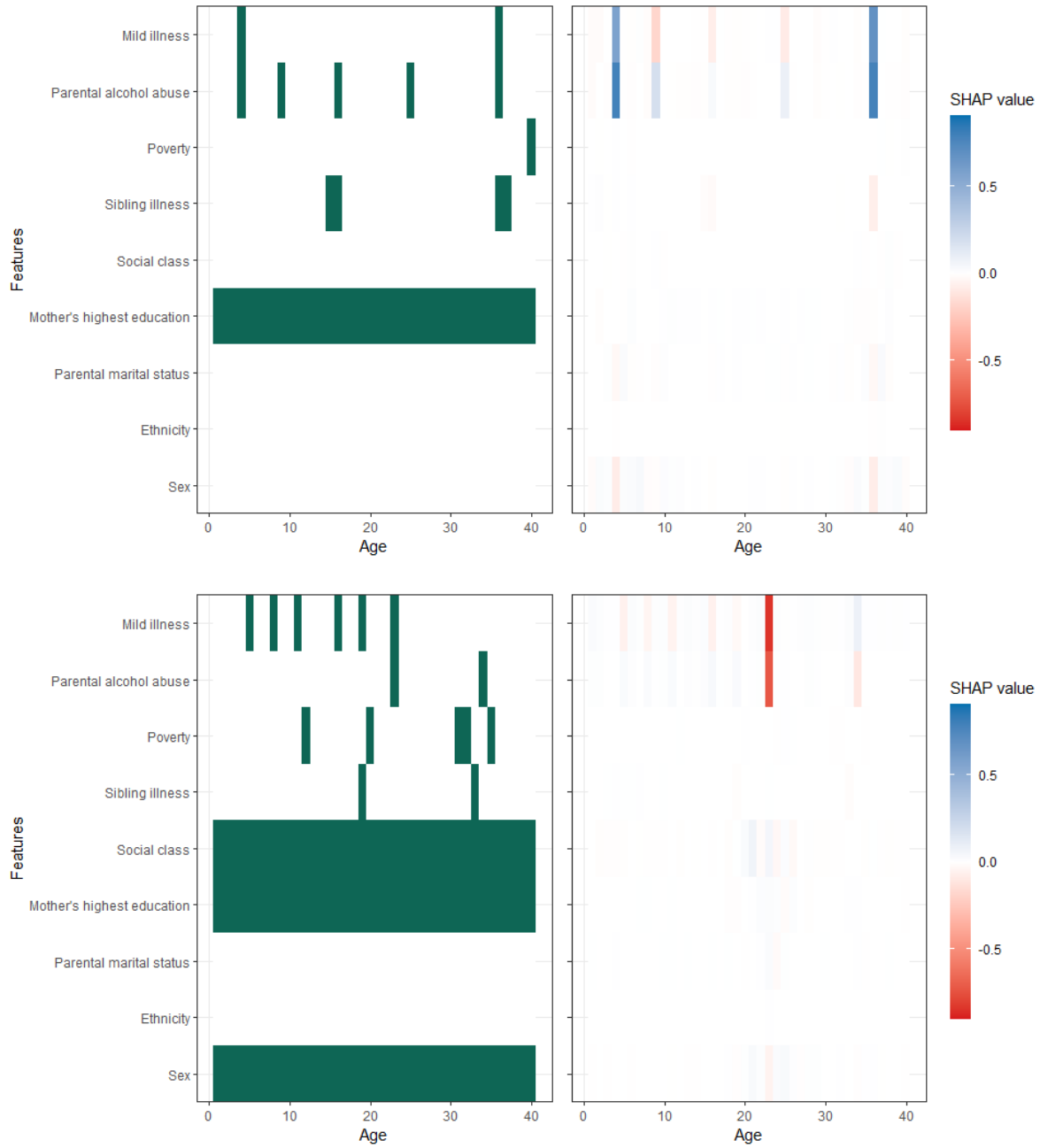


Figure 5.7: **Individual level SHAP values for simulated data set Timing1.** The SHAP values for two example individuals in the Timing1 simulated data set. The left figures show their exposure events which are binary with green values equal to one and zero otherwise. For each the x-axis shows the different time points in the life course and the y-axis gives the different features in the model. The right figures then show the corresponding SHAP values with blue values being positive, and red being negative. The top individual is a correctly classified positive case and the bottom individual is a correctly classified negative case.

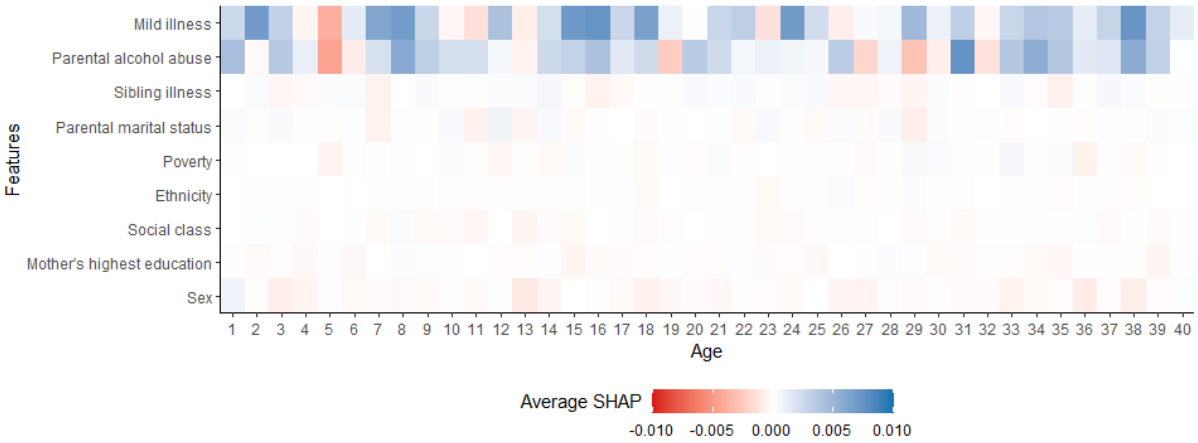


Figure 5.8: **Average SHAP values for simulated data set Timing1.** The average SHAP values across 500 individuals in the Timing1 simulated data set. Blue cells indicate positive average SHAP values and red shows the values are negative.

to the ones with the greatest or lowest average SHAP values over all the time steps, including RTIs, maternal separation, infectious and parasitic diseases, sex, and poverty.

The diabetes figure shows that previous diabetes, especially in the most recent years is most influential which aligns with the PFI patterns found previously (Figure 5.11). The plot is capped at 0.001, so the most positive values are in fact closer to 0.01 signifying a strong dependence on previous diabetes, especially relative to the other features. The second most influential feature overall is parental origin, with positive values indicating a positive association with the diabetes outcome. There also appears to be a temporal relationship present with later years of birth attributable to more positive predictions. Similarly, there is a window between the ages of three and six where exposure to parental alcohol abuse seems to have the greatest impact on future health. There are both positive and negative associations between maternal separation and diabetes in later life, this may be a result of the small cohort and a rare event or related to interactions with other exposures. The individual characteristics of parental education and mothers age at birth have the strongest negative associations with the outcome. These variables are both time invariant so the temporal patterns in significant arise entirely as a result of the averaging out of the true underlying effect. Finally, there is some protective effect if a child experiences parental unemployment between ages four and five. Many of the features have no overall association when averaged across the population of 2000 individuals. The plot of standard deviations highlights the features with most variability as those that had the greatest SHAP values (Figure 5.12). It does not pick out any additional features that may be influential on the individual level but masked by the averaging procedure, except for premature birth which has an average SHAP value close to zero but positive standard deviations.

Neurological conditions is the most influential variable (Figure 5.13). In fact, of all the outcomes this variable has the highest average SHAP value at 0.015 which signifies strong dependence of the outcome on this feature (uncapped figure available in Appendix 8.4). It is clear to see that the most recent years of neurological conditions are the most influential of these. Additionally, the year of birth feature has SHAP values with



Figure 5.9: **SHAP values for RTI outcome.** The SHAP values for the RTI outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association. The colour scale has been capped at 0.001, therefore, the darkest positive associations truly fall at around 0.004.

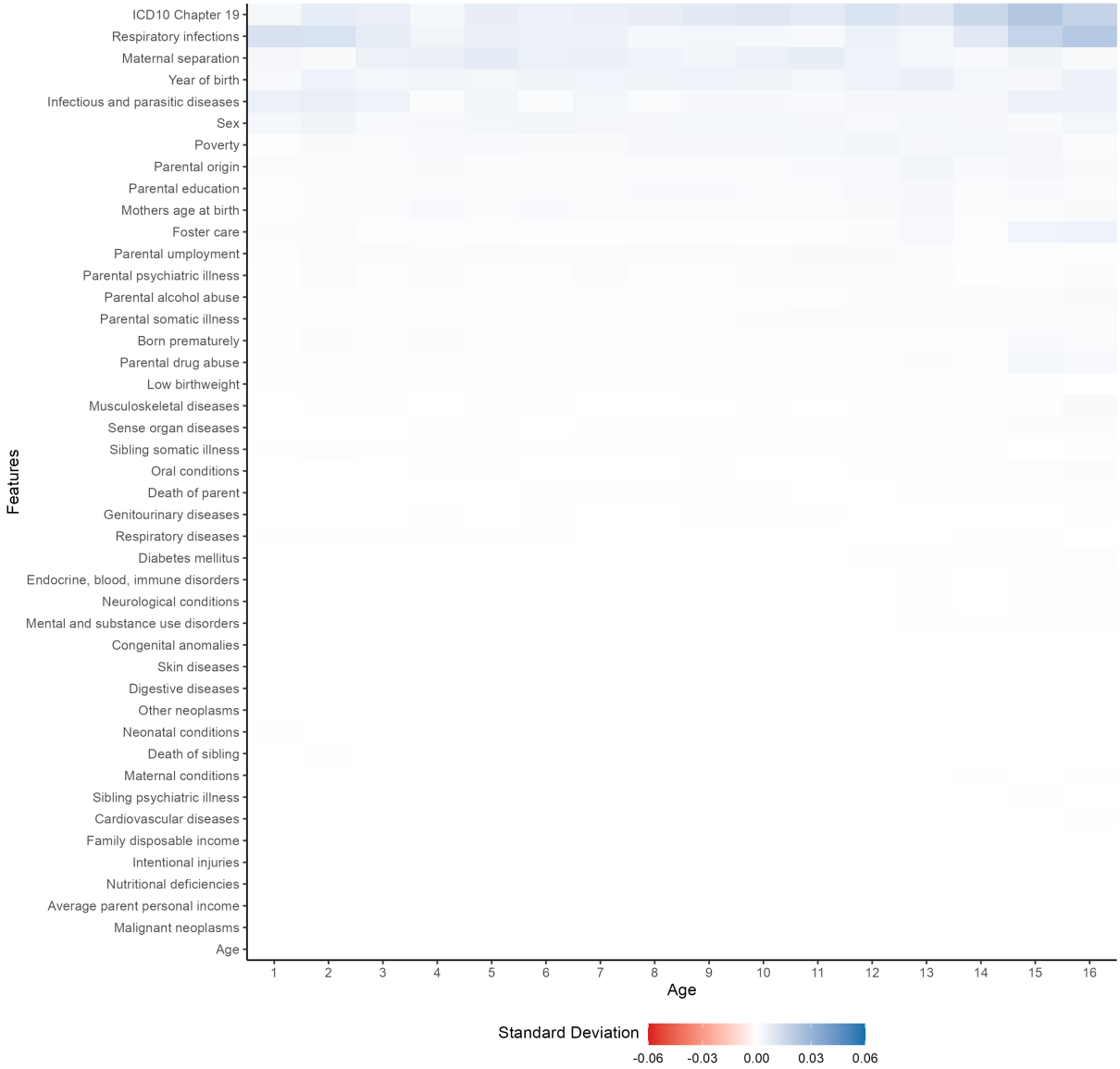


Figure 5.10: **SHAP standard deviation values for RTI outcome.** The SHAP standard deviation values for the RTI outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals.



Figure 5.11: **SHAP values for diabetes outcome.** The SHAP values for the diabetes outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association. The colour scale has been capped at 0.001, therefore, the darkest positive associations truly fall at around 0.01.

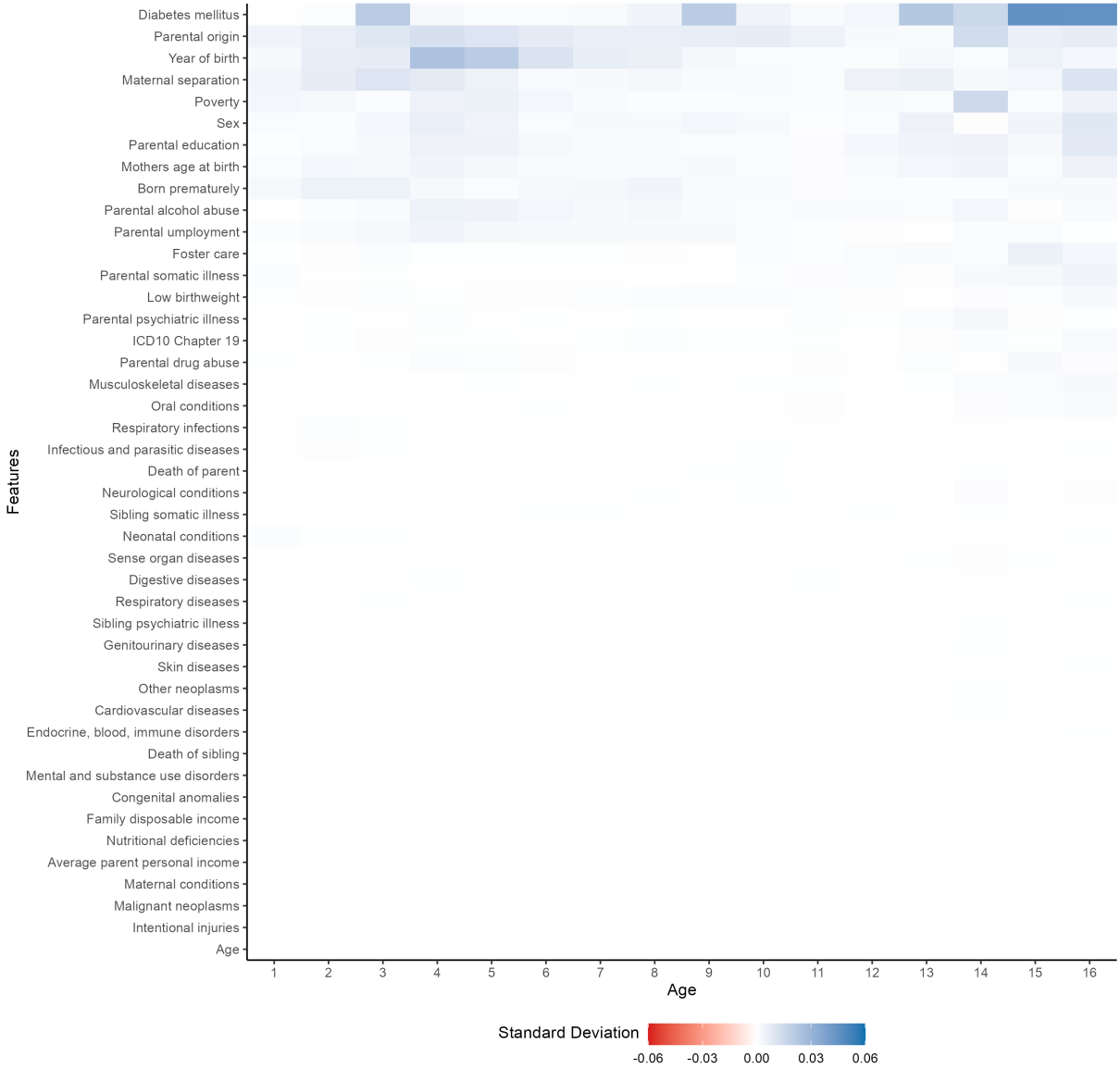


Figure 5.12: **SHAP standard deviation values for diabetes outcome.** The SHAP standard deviation values for the diabetes outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals.

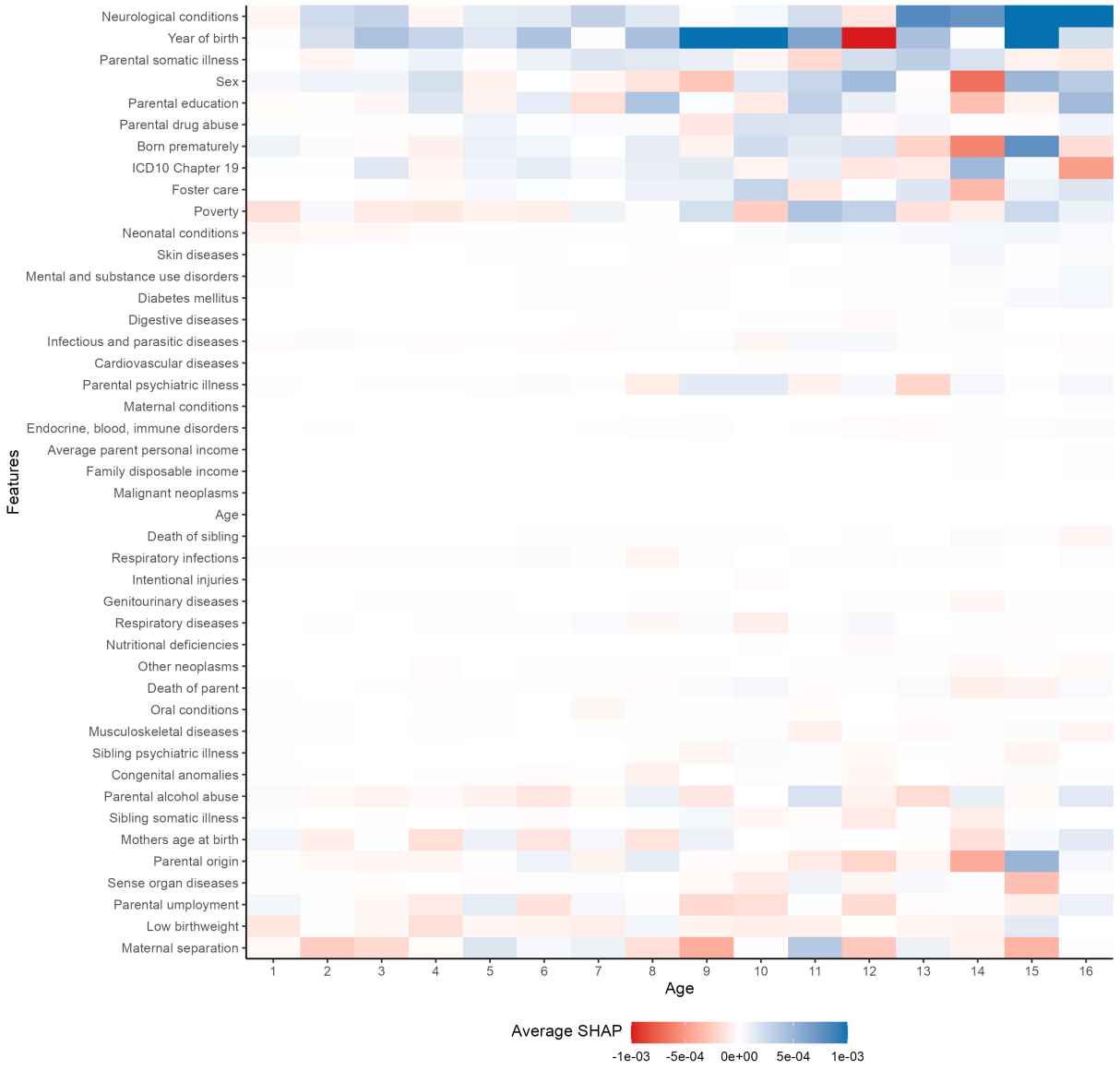


Figure 5.13: **SHAP values for neurological conditions outcome.** The SHAP values for the neurological conditions outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association. The colour scale has been capped at 0.001, therefore, the darkest positive associations truly fall at around 0.015.

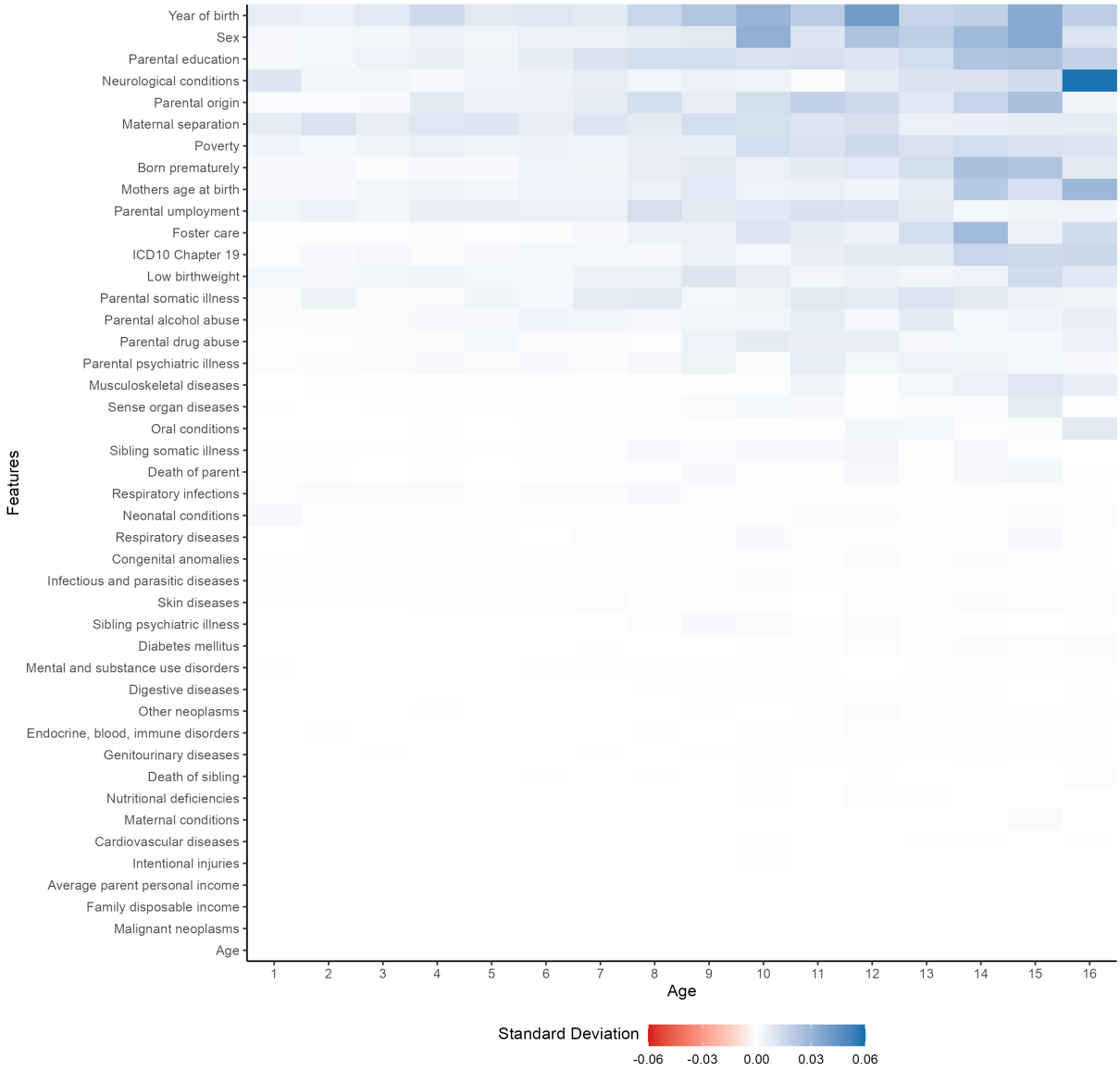


Figure 5.14: **SHAP standard deviation values for neurological conditions outcome.** The SHAP standard deviation values for the neurological conditions outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals.

large magnitudes in both directions. It is an individual level characteristic as so its value does not change over time. In comparison with the other outcomes, there are many more variables with greater positive and negative associations. Moreover, there is more variability across the different time steps of each variable in the direction of association. This pattern is also present in the standard deviations where many more features have values greater than zero (Figure 5.14). The third most impactful feature is exposure to parental somatic illness which is positively correlated with predictions of neurological conditions. The other variables with an average positive association are sex, parental education, premature birth, ICD-10 Chapter 19, foster care placement and poverty. The features with overall negative associations are maternal separation, low birthweight, parental unemployment, sense organ disorders, parental origin, mothers age at birth, sibling somatic illness and parental alcohol abuse. Several variables, but most notably parental psychiatric illness, have large average SHAP values at certain time points but when averaged over the time points these values are close to zero. This indicates that these variables have differing effects for different individuals.

5.3.3 Risk Factor Estimates

The risk ratio estimates for the diabetes outcomes for several exposures are presented in Table 5.1. The first result of note is how great the risk ratios are for previous diabetes hospitalisations. For diabetes under age 5, the risk ratio is 276.5 and for diabetes under 10 the risk ratio is 385.5. Therefore, these results show that the model is heavily dependent upon previous diagnoses when making predictions. Other exposures were investigated, including the ACEs of poverty, parental alcohol abuse and maternal separation. The condition to be in the exposed group for each of these was to ever have experienced the given exposure. The most influential exposure is parental alcohol abuse which has a predicted risk ratio of 2.062. This indicates that individuals who have experienced parental alcohol abuse are twice as likely to develop diabetes than those who have not. There is a small positive effect of maternal separation and a small negative effect of poverty which has a risk ratio of 0.975. Two individual level characteristics were also explored, these were sex and parental origin. Both characteristics had negative impact on risk with being male reducing the risk of RTIs with a risk ratio of 0.897 and having a parent of non-European origin having an risk ratio of 0.815. These both demonstrate a positive impact on disease risk for those with these traits.

Table 5.1: **Risk ratio estimates for diabetes mellitus outcome.** The risk ratios computed using model predictions on the test set with references as given.

Name	Definition	Categories	Risk Ratio
Sex	Sex given in the Danish Medical Birth Register	Female	1
		Male	0.8967
Parental origin	Defined as European if both parents are of European descent and non-European otherwise	European	1
		Non-European	0.8152
Diabetes under 5	At least one hospitalisation with main diagnoses of diabetes mellitus whilst under 5 years old	Not experienced	1
		Experienced	276.5108
Diabetes under 10	At least one hospitalisation with main diagnoses of diabetes mellitus whilst under 10 years old	Not experienced	1
		Experienced	385.4680
Poverty	At least one year of family income being below the poverty threshold during childhood	Not experienced	1
		Experienced	0.9754
Parental alcohol abuse	At least one incident of parental alcohol abuse during childhood	Not experienced	1
		Experienced	2.0623
Maternal separation	At least one event of maternal separation during childhood	Not experienced	1
		Experienced	1.0095

5.4 Discussion

In this chapter, multiple methods have been used to explore the impact of different exposures on the disease outcomes, helping to gain an understanding of the risk landscape. This analysis also demonstrates how risk ratios can be computed from the model predictions and provide further insight into the relationship linking

exposures experienced during childhood and disease risk in young adulthood. These methods form the first step towards understanding the pathways and patterns that connect multiple exposures, their timings, and the development of various health conditions.

This study has identified multiple important exposures that act as risk factors for RTIs as well as diabetes mellitus and neurological conditions. The methods used decompose the existing signal in the data meaning that although the model performance differed by disease outcome, the significant exposures and time periods identified using these methods are still informative. Therefore, despite poor predictive performance for the RTI outcome, the XAI approaches used in the analysis still pinpoint the key features that are most influential in improving model predictions. The results also emphasise that the DANLIFE data does not contain variables that are informative for this outcome because the proportions of change in the PFI analyses and the average SHAP values were close to zero for almost all variables. These findings further signify that the data requires additional variables to accurately model RTIs as an outcome.

It is important to note that there is a difference between how SHAP values and PFI metric change are interpreted. The PFI results order the input features or time steps according to the magnitude of their impact on the evaluation metrics. Therefore, the inclusion of the variables with the greatest change in metric value are responsible for the largest improvement in outcome prediction. Furthermore, any interactions between multiple variables will be accounted for in both the proportion of metric changes related to each variable, meaning that the changes in evaluation metrics do not sum to the baseline evaluation metric value. In comparison to PFI, the interpretation of SHAP values differs greatly. The sign of a SHAP value gives the direction of any association between that variable and the outcome. Therefore, unlike PFI, large negative values do not indicate negative contribution to model performance, but instead improve performance. At the individual level, the SHAP values for an individual provide an explanation of why the model assigned the predicted class. When SHAP values are averaged at the population-level, the average values can indicate general trends. However, average SHAP values may mask the relationships between certain variables and the outcome when unique life histories cause the exposure to act in opposite directions.

A key finding from this analysis is that, for all outcomes investigated, the most significant variable in predicting the disease outcome is previous occurrences of the same disease. This is consistent with expectations, especially with the non-communicable outcomes as they are generally lifelong conditions with previous hospitalisations inferring the outcome almost with certainty. This is particularly true of diabetes, because the DANLIFE data is centred in childhood and therefore, any diabetes diagnoses are likely to be type-I diabetes which is genetic generally requires regular hospital visits. It has also been found in multiple studies that incidents of RTIs during childhood increase one's susceptibility to RTIs in later life due to decreased lung function [Dharmage et al., 2009; Dratva et al., 2016; Galobardes et al., 2008].

There are differing impacts of timing or reliance on different time periods for each disease. The PFI values do not indicate the importance of the timing of exposures, but more so the extent to which the inclusion of each time period improves model performance. All outcomes most heavily depend on the one or two years prior to the outcome observation period. For diabetes, these are the only important years with respect to model performance. This may be due to the more regular appointments related to this condition so that only the most recent year is required to achieve good model performance. Regular appointments for ongoing conditions may also be the reason that RTIs and neurological conditions also display dependence on the most recent and later years of childhood. The AUPRC and AUROC evaluation metrics show different patterns with the AUROC score displaying a greater dependence on earlier years in childhood. In addition to the most recent years, RTIs are also greatly influenced by the first three years of life which correlates with studies that have found that RTIs under the age of two leads to a lifelong effect on respiratory health [Stocks et al., 2013]. The predictions of neurological condition display dependence on the ages between seven and nine. This

aligns with the literature which shows there are many neurological conditions that are first diagnosed in young adolescents, for example idiopathic torsion dystonia which is often first diagnosed in children aged between seven and twelve [Macleod and Appleton, 2007].

Due to the data protection stipulations, it is not possible to show individual level SHAP values and instead the plots presented are averaged over 2000 individuals. As a result, it may be the case that exposures that act in different directions according to the other exposures experienced are lost in these plots. Any variability according to the time period or variables with large standard deviations but small average SHAP values indicate that they may have this kind of pattern. Features highlighted in this way are maternal separation for RTIs, poverty for diabetes and parental psychiatric illness for neurological conditions. There are multiple differences between the PFI and SHAP results. For example, parental origin is the feature with the second greatest SHAP values for diabetes, however, in the PFI plots its inclusion appears to have a slightly negative effect on model performance. Additionally, the population used to calculate the average SHAP values is far smaller than the total cohort and consequently, the rare medical exposures will have varying directions of associations according to the unique life courses of the small number of individuals that experienced them. The way in which the SHAP values have been averaged is designed to show which features and time points are most influential for inferring the predicted class. However, the use of both classes together and the nature of how the SHAP values directly relate to whether an exposure has occurred or not means that some values may cancel out. Therefore, as future research it would be worth exploring how one can best gain population level insights from the SHAP values by using different summary statistics or dividing the population by class.

Each of the outcomes is influenced by the age and year of birth variables. The individual characteristics whose effect alters over time impart the importance of time in the study and also showcase the effect of the dependencies being shared across time points. The age variable is the same for all individuals and monotonically increases to give their age from 0 to 15. The inclusion of this variable allows for more explicit temporal patterns to be found that aren't necessarily reflected in the other variables, such as changes in healthcare accessibility or social attitudes. The significance of this variable indicates that there are strong temporal patterns being picked up by the model. It also perhaps suggests that one of the DL models that explicitly has temporal linkage built in may achieve better performance which can be investigated in future work.

There are several notable differences between the most influential features identified for each of the three outcomes. Diabetes is almost entirely predicted by past instances of diabetes, with the PFI highlighting no further variables and the SHAP values suggesting that parental origin, parental alcohol abuse, maternal separation, mothers age at birth and parental education may contribute to model decision making. Three of these features seem to have a differing direction of association depending on the timing of the given exposure or the other exposures experienced during the life course. Notably, mothers age at birth and parental education have differing strengths of association, but they are both individual characteristics that do not change in time and therefore, this effect may arise due to the distributing of the association across the different time points. In contrast, non-European parental origin is consistently linked to increased diabetes risk which suggests that genetics play a key role in diabetes acquisition. No other medical exposures impact diabetes risk, unlike RTIs and neurological conditions. There is a higher likelihood of RTI diagnoses for individuals with a history of infectious and parasitic diseases and diagnoses with ICD-10 Chapter 19 codes. Infectious and parasitic diseases include tuberculosis as well as STDs, tropical diseases, and childhood diseases such as whooping cough, with many of these conditions having a lasting impact on the airway. ICD-10 Chapter 19 codes are defined as “injuries, poisoning and certain other consequences of external causes" which primarily cover multiple types of unintentional injury to the body, including bone fractures, burns, open wounds, nerve damage

and limb amputation. It may be that some of these wounds damage the respiratory system leaving individuals more susceptible to RTIs or that this variable acts as a proxy for other variables not included in the data set such as neglect, unsafe households, or neighbourhoods.

In contrast with the other two outcomes, neurological conditions seem to arise due to more intricate relationships between multiple exposures. The PFI analysis indicates a greater dependence on several features and the average SHAP plot shows many features with large magnitudes of association. The influential features include a mixture of individual characteristics, previous medical events, and ACEs. Notably, the risk of neurological conditions is increased by incidents of numerous ACEs such as parental somatic illness, parental drug abuse, being placed in foster care and poverty. There are also individual characteristics including sex, parental education and premature birth that increase risk. There is also evidence that ICD-10 Chapter 19 hospitalisations are linked in this case. The category neurological conditions encompasses Alzheimer's, Parkinson's, epilepsy, multiple sclerosis, and migraines. Therefore, it is understandable why there is such a great range of associated exposures. It is unlikely that any of the cases in this analysis are Parkinson's or Alzheimer's as the age of onset is generally 65 or later [Guerreiro and Bras, 2015]. Similarly, multiple sclerosis is most commonly diagnoses after age 20 [Liguori et al., 2000]. Therefore, the positive outcome cases in this analysis are most likely to be composed of cases of epilepsy and migraines, both of which are commonly first experienced in childhood. Known risk factors for epilepsy include genetic factors, instances of head trauma and previous infectious diseases, whilst migraines are more influenced by sex, genetics, and a range of lifestyle factors [Russo et al., 2016; Walsh et al., 2017]. It is highly likely that the ICD-10 Chapter 19 feature is reflecting the relationship between head trauma and epilepsy and the ACEs are correlated with the lifestyle factors that can increase the risk of migraines.

5.4.1 Risk Ratio Analysis

The risk ratio estimates for several exposures highlight multiple features that may act as risk factors for diabetes. Whilst risk ratios are very ubiquitous among risk studies, using the model predictions to calculate these allows for the representation of complex interactions between variables to inform their values. The highest risk ratio found was for previous diabetes occurrences. In fact, within the test set there are no individuals with positive outcomes who had not previously attended hospital for a diabetes-related condition. However, this analysis demonstrates how the time period may be split when calculating risk factors to help identify sensitive developmental periods.

The variable that has the second highest risk ratio is parental alcohol abuse with its occurrence at any point during childhood doubling one's risk of diabetes. This could be an indication of confounding or conversely arise due to the small number of individuals with diabetes in the test set. However, comparison with the literature also suggests a link between parental alcohol abuse and diabetes. It has been observed that those who drink during adolescence are more likely to become diabetic in adulthood [Liang and Chikritzhs, 2014]. Other studies have found that diabetes type-II is more common in men who experienced parental alcohol abuse as children [Upadhyaya et al., 2020]. The parental alcohol abuse variable was identified as being influential using the SHAP values, but it seems to have limited impact in the PFI results which demonstrates a limitation of solely depending upon PFI to conduct risk analysis. The same study examines the effects of divorce on diabetes prevalence and finds no association which matches the result found in this analysis with a risk ratio close to one for maternal separation [Upadhyaya et al., 2020]. Another notable finding is that those who experienced poverty during childhood were slightly less likely to develop diabetes. This is contrary to the existing literature as it is thought that poverty is a key social determinant of diabetes [Hwang and Shon, 2014]. The reason for this discrepancy could be that those living in poverty may have less access to healthcare and present less often at hospitals.

This analysis demonstrates the power of these methods and the insights that can be gained from using risk ratios. It completes the risk analysis pipeline, allowing for traditional epidemiological estimates to be produced that can be compared to the literature in addition to the extra understanding gained by using XAI approaches. Each of these methods can be used for any outcome for which there is a well-calibrated DL model, meaning that with additional model and data adjustments it may be possible to use this framework for RTI risk analysis. These results serve as examples to confer the utility of this framework with the condition that any further analysis utilising these methods must be carefully designed. It is essential that there is a comprehensive comparison of the risk ratios calculated across different time periods to ensure that no effects are masked by the formulation of the risk groups.

5.4.2 Limitations & Future Work

There are a number of limitations within this research that are essential to address. Firstly, neither the SHAP nor PFI methodology can reflect interactions between variables. Instead, any interactions get attributed to both features and this complicates the process of identifying the life course pathways and which features may be interacting with others. Some exposures may act to mitigate potential harmful effects of other features or to boost the magnitude of the impact on disease risk. Without tools to additionally consider these interactions, obtaining a concrete idea of associations by means of a DAG or similar is difficult. With MTS data it can be very difficult to test for feature interactions. There is a paucity of research into explainability methods for MTS data as the predominant focus has been on NLP and computer vision [Vielhaben et al., 2024]. Therefore, patterns such as the life course patterns simulated in Chapter 2 can be difficult to identify from SHAP plots alone without additional prior knowledge and human interpretation. There has been a recent interest in XAI for time series data, however methods for MTS data still lag behind [Di Martino and Delmastro, 2023; Mincu et al., 2021; Vielhaben et al., 2024]. There are methods that can be used to test for interactions between two time series features such as 2D PFI which perturbs two features at a time to aid in assessment of interactions. However, this method will again only look at the marginals across the time steps and does not improve the identification of significant time periods. Additionally, there are models that have built-in explainability such as explainable-by-design ensemble method (XEM); however, this method has not been extensively compared to other approaches for medical data [Fauvel et al., 2020]. Future research would involve an in-depth analysis of the existing XAI methods that may be applied to MTS data and potentially an investigation into possible improvements to these methods to ensure that temporal patterns and inter-variable interactions are able to be identified and understood without further knowledge. There is an increased interest in approaches for longitudinal medical data and as a result many new techniques are being introduced to help answer different questions. A framework for how these methods can be beneficial to understanding pathways and useful to healthcare practitioners is required.

Another point of note is that these methods only identify associations between different exposures, their timing, and the outcome. These associations are not evidence of a definite causal relationship. Whilst knowledge of associated exposures is important and can help to identify population subgroups that require additional support, this is not the same as being able to mitigate the exposure and have an impact on the number of individuals with the disease. Additionally, the results highlight that there are many variables which are uninformative. Further studies would need to be conducted that remove these features and assess the impact on model performance. It may be the case that once these features are removed the model is better able to identify links between the remaining variables and improve model performance and reduce model run time. Multiple experiments would need to be conducted to be certain that these features are truly uninformative as there are instances of features that decrease model performance in the PFI plots but have relatively large SHAP values. Therefore, the exclusion of these features would have to ensure that model performance improved. Several of the medical features encompass multiple diseases, such as infectious and parasitic

diseases which includes diseases as disparate as STDs, tuberculosis and meningitis, each of which have vastly differing levels of impact on someone's life. Therefore, in future work a full investigation into whether these features ought to be further broken down should be conducted.

SHAP is the most common method used for XAI and is thought to be the only explainability method that would be able to comply with legal regulations, such as the new requirement for "right to explanations" in the EU [Selbst and Powles, 2018]. This is because it is the only existing explainability method grounded in theory that is based upon mathematical principles that distribute the decision across the features equally. There have been several variations of the SHAP methodology proposed, notably causal SHAP values which are able to reflect causal relationships within the data [Heskes et al., 2020]. Despite this, there are concerns about the limitations of SHAP [Kumar et al., 2020]. The two main concerns are that their mathematical properties break down when they are used for feature selection and additionally that they may not relate to underlying causal structures. This motivates a redesigning of explainability approaches with medical healthcare data, such as the DANLIFE data, in mind. Future work would use the full range of explainability methods to aid with answering different questions that are most important to epidemiological researchers. The use of multiple explainability approaches would provide a more holistic overview of the associations and causative pathways present in the data.

5.5 Conclusions

The analysis within this chapter has demonstrated the advantages and effectiveness of DL models within risk analysis and life course studies. These models are capable of answering traditional epidemiological questions and generating more accurate risk ratios to explore the impact of certain risk factors at the population level. Additionally, using DL models alongside XAI methods enables individual insights into the key exposures in an individual's life that are associated with disease and are able to reflect the complex nonlinear relationships that generate diverse health trajectories. This model structure has the potential to be developed into accurate and powerful risk prediction models to both predict individual level risk scores and also open up the possibility of further exploration into the relationships linking exposures, their timings and disease risk. There is much scope for development of this research to create a robust framework to assess the pathways by which risk is accumulated throughout life. These models have the potential to change the landscape of epidemiological modelling and life course analysis, opening up avenues for more cost-effective public health spending, personalised healthcare approaches and a better understanding of disease mechanisms.

Chapter 6

Conclusions, Limitations and Future Research

The main aim of this work was to disentangle the way in which exposure events act as risk factors during childhood to raise the likelihood of hospitalisation related to respiratory tract infections (RTIs) in young adulthood. The first step in this analysis was conducting a simulation study that assessed the suitability of deep learning (DL) models for applications within life course analysis. Then these results were compared against an alternative modelling approach, the Signature method, which was designed for capturing patterns within time series data and has gained prominence within the wider research community. Following the simulation study, both machine learning (ML) modelling approaches were assessed on the DANLIFE cohort, which represents real longitudinal life course data. It was found that both models performed poorly for RTIs but achieved good results for some non-communicable disease outcomes. Finally, multiple explainable artificial intelligence (XAI) methods were used to obtain estimates of the influence of multiple different exposures on disease risk and identify potential risk factors within the DL modelling methodology. This study completed the framework for risk analysis that is able to reflect the dynamic relationships that link exposures to disease risk, forming a solid basis for future research. The key takeaways of the research contained within this thesis are explored below.

First, this analysis showed that DL models are effective and well-suited to life course epidemiology applications. In Chapter 2, the results showed that most DL models achieved excellent performance in scenarios simulated to reflect the key life course concepts. DL models were found to predict well in the simulated setting, have good discriminative and calibration performance, representing a considerable improvement on the logistic regression (LR) baseline model. This showed that the DL models are capable of learning the dynamic patterns between exposures when they exist in the simulated data. The main conclusion of the simulation study is that the DL models are well-suited to working with the simple patterns embedded in the simulated data, however, the relationships within the real world data may be far more complex. Additionally, the results showed that model performance degrades with noise and consequently, noise may severely impact model behaviour on real data where the noise level is unknown. In Chapter 4, ResNet was able to obtain an AUROC score of 0.871 and an AUPRC score of 0.686, when examining diabetes mellitus hospitalisations as an outcome. However, when examining RTIs as the outcome, ResNet performed poorly and got AUROC and AUPRC scores of 0.528 and 0.049, respectively. This shows that DL models are able to achieve high accuracy predictions of future disease depending on the characteristics of the disease in question. The key difference between the DL models for multivariate time series (MTS) data is that they incorporate notions of time which is not present in tabular data and the corresponding models. The inclusion of time is crucial in that it allows

for temporal dependencies to exist between variables, thereby being more capable of learning the life course patterns when present in the data. The passing of time and effects from past exposures are thought to be intrinsic in determining future health within the life course lens [Ben-Shlomo and Kuh, 2002]. Therefore, it is essential to understand the capabilities of models for MTS data.

In addition, this research demonstrated that the Signature method is not suitable for heavily imbalanced data which is often the case for medical outcomes such as RTIs. In Chapter 4, the results showed that the Signature method was able to obtain good AUROC performance and even outperform the ResNet model with hyperparameter optimisation with RTI outcomes. However, the AUPRC scores suggested that there were a large number of false positives that limited the utility of this approach for healthcare practitioners. This is true across all three outcomes investigated, even diabetes for which the ResNet model obtains a very high AUPRC score. Consequently, the Signature method is not suitable for applications when there is high class imbalance. It was found to achieve similar performance to the top performing DL models for some simulated scenarios that were investigated in Chapter 3. In addition, the model has significantly lower computational requirements, it is both faster to run and requires less memory. However, there is generally an improvement in performance for higher truncation levels, but these quickly need substantial CPU power and limit the complexity of the patterns that this approach can reflect. In contrast, DL methods have exceptional performance but require substantial computational resources and can be complex to fit.

Another key takeaway is that DL models are able to be used to identify influential exposures and obtain risk ratios to guide population health research and interventions. This has the potential to assist medical practitioners in guiding diagnoses, highlighting key exposures within the population and individual level disease risk. In Chapter 5, the use of explainability methods to gain insights into the processes that generate risk were explored. The various XAI approaches shed light on the different aspects of how exposures can impact an individual's disease risk. The permutation feature importance (PFI) analysis allowed the identification of the features and time points that resulted in the greatest improvement in model performance. The SHAP plots allowed for slightly different insights to be gained and showed how influential each feature at each time point was on the predicted outcome at the individual and population levels. Whilst within this thesis, the SHAP scores were only able to be shown at the individual level due to data protection rules and issues of identifiability, it is possible for healthcare practitioners to use the individual level result to justify medical decisions. This would enable doctors to use the DL model framework to predict RTI risk for individuals and flag the at-risk individuals to prioritise for treatments or within vaccination schemes. Additionally, population-level risk ratios were estimated using the model predictions which are comparable to the traditional epidemiology approaches used within risk analysis but better reflect the dynamic relationships present within the longitudinal data. These risk ratios can be used by decision makers to design interventions in the same way as existing risk ratios, however, as these risk ratios better encapsulate the effects of other exposures and dynamic factors, their use has the potential to improve the effectiveness of interventions.

6.1 Contribution to Life Course Epidemiology and RTI Research

This research constitutes a significant methodological advancement in life course analysis, completely redefining the ways in which health trajectories can be explored. Over the course of the thesis, each building block that forms part of this new DL-based framework for risk analysis through a life course lens has been introduced and assessed. The use of DL methods allows for a more in-depth study of the formulation of the patterns and pathways that accumulate across the life course to generate disease risk. The approaches explored include ResNet and the Signature method, both possessing a far greater capacity than traditional epidemiological models to capture complex interactions between different exposures and their timing. The simulation study allows for the embedding of patterns within the cohort that align with multiple life course

concepts. This is a novel approach that explores whether ML models can effectively learn and identify these patterns, thereby establishing a methodological foundation for exploring complex temporal patterns in health data. This research showcases the superiority of DL methods in comparison with traditional epidemiological modelling approaches. The knowledge gained in this research about the constraints and limitations associated with different ML models ensures a balanced understanding of their applicability and guides researchers in selecting suitable methodologies for specific life course inquiries. The methodological innovation introduced in this thesis lays the foundation for future studies aiming to unravel the complex dynamics of health trajectories over time.

This thesis culminates in a detailed investigation of the significant risk factors using a range of methods from risk ratios to multiple interpretability approaches. This approach allows for a deeper understanding of risk factors and sensitive developmental periods, presenting a pathway for further investigations into the interplay of variables influencing health outcomes. Most diseases are caused by multiple factors, yet risk factor estimates traditionally consider the effect of a single exposure on the outcome. This study allows for the influence of the risk factors to change depending on their timing and an individual's life events to date. The evaluation of DL models for predicting RTI-related hospitalisation risk contributes to the development of more accurate and nuanced risk prediction models. These newly developed methods promise improved individual level risk prediction models for diseases, paving the way for more personalised healthcare and efficient, cost-effective treatment. Precise individual level predictions can be extrapolated to estimate population-level risk, identifying significant exposures for specific diseases across the entire population. This knowledge is integral for identifying individuals at risk and designing targeted and effective public health interventions. This research innovatively relates interpretability with risk ratios, a relatively unexplored aspect in time series and healthcare that bridges the gap between DL approaches and traditional life course epidemiological theory. The inclusion and evaluation of explainability methods, such as PFI and SHAP, demonstrate a commitment to enhancing interpretability in epidemiological models, contributing to broader discussions on the transparency and interpretability of DL models in medical research. These advancements contribute to the creation of more robust risk prediction models that can account for complex relationships and provide actionable insights for healthcare practitioners.

This analysis offers valuable insights for those wishing to examine Nordic data, such as DANLIFE, using DL approaches with MTS data. This study represents the first of its kind to attempt to conduct life course analysis using the power of DL methodologies. The methodology developed in this thesis acts as a guide, outlining practical considerations and limitations for researchers interested in the application of advanced DL models on large-scale and high-security longitudinal data. The provided insights not only streamline the process but also enhance the efficiency of analysing complex life course patterns within large-scale datasets. By sharing the practical intricacies and potential pitfalls, this research facilitates the adoption of DL techniques in similar studies, fostering a collaborative and informed approach to exploring health trajectories in Nordic populations. This can also act as a guide for health policy and how administrative records are designed when they are being added in various countries to improve the data being collected, and its potential for research.

This thesis has made significant strides in the field of RTI research by investigating the differential predictive capabilities of DL models. Notably, the greater performance in predicting non-communicable diseases, such as diabetes, than infectious diseases provides critical insights into the nuanced nature of infectious disease modelling and limitations in this framework. The research has illuminated potential data limitations that may impact RTI predictions and prompts further considerations about the nature of data, the absence of relevant variables, and the potential impact of noise on the outcomes. By critically assessing the dataset's strengths and limitations, this research serves as a guide for future researchers aiming to leverage similar data sets for infectious disease studies. Knowledge of the key risk factors that impact RTI susceptibility and severity is

critical in instances such as the COVID-19 pandemic. Whilst RTIs related to SARS-COV-2 are not contained within DANLIFE as the data set does not contain individual records after 2018, it is an important example of an instance where knowledge of the key risk factors could have had a big impact on the outcomes. When interventions were designed to curb the disease progression, the knowledge of those most at risk was very limited. Knowledge of key risk factors or an individual level risk prediction model allow interventions to be implemented more precisely and limit the number of individuals impacted by big scale interventions, such as lockdowns. In addition, the ability to better identify high-risk individuals aids in the design of effective and efficient shielding schemes and vaccination programs. During the pandemic, the focus lay largely on the elderly and those with pre-existing conditions, however, other exposures, such as poverty, may have played an equally important role in generating risk. The versatility of these methods extends beyond health outcomes, encompassing various communicable and non-communicable diseases, as well as socioeconomic outcomes such as poverty rates.

6.2 Limitations and Scope for Future Research

One key consideration that would need to be addressed if these risk prediction models were to be put into practice is the availability of longstanding longitudinal data collection in other countries. Whilst there has been a recent shift towards the collection of more electronic health records (EHRs) and better linkage between different administrative records, there is still a scarcity of countries that have data sets capable of conducting this type of DL procedure on. The countries that do have cohorts are largely in the Nordic region and therefore, the result of any analysis using these cohorts alone would by nature be biased towards Nordic populations. The Nordic countries, and Denmark in particular, do not encompass a diverse range of individuals with for example the vast majority of Danes being white. There may be subtle differences in lifestyles and genetics by region that impact an individual's susceptibility to severe RTI outcomes that would render any estimates of risk factor sizes uninformative within other settings. Consequently, the generalisability of these results to new populations is limited and would require additional collection efforts and alterations to administrative record keeping in order to utilise similar DL frameworks within population risk analysis applications.

A very valuable role of explainability methods is the investigation of any biases in the data that are emulated by the DL model. Future work would include an in-depth analysis of potential biases propagated and perpetuated by the model. Using methods that have been utilised within this analysis to compare the prediction of risk within various sub-populations. For example, one could look at individual level SHAP plots to see if there are certain subgroups for whom risk accumulates differently. Model bias has been shown to be a serious concern in multiple DL applications when the model learns from biased data and then treats different population groups unfairly. For example, in Chicago a DL model was used to for predictive policing and direct patrols to certain areas of the city according to previous crime records [Jefferson, 2018]. A study conducted found that the police patrols were predominantly sent to majority Black neighbourhoods by the model. This was due to the bias present in the previous crime records which were used to train the model. Prior to the use of predictive policing, the police officers had demonstrated racial bias and had tended to patrol non-white neighbourhoods more often, leading to more recorded criminal incidents being encountered and recorded in those areas. Alongside the fact that police in the US have been found to be more likely to charge Black individuals more severely than white individuals who had committed the same crime, the data contained many more crimes reported in Black neighbourhoods and this led to bias within the DL model. Bias detection approaches that rely upon XAI methods were able to identify the differential treatment of populations leading to model improvements. This is also a concern within healthcare, there are many recorded incidents of AI algorithms assigning significantly lower risk scores to Black patients who exhibited the same level of illness as white patients [Obermeyer et al., 2019]. It may be the case that the DL models looking at RTI outcomes are

also propagating bias present in the data which ought to be investigated before employing the model within healthcare settings.

Another limitation of this research is that computational and time restrictions meant that it was not possible to conduct a full feature selection procedure. As a result, there may be input variables that are collinear, such as “Family disposable income” and “Average parental personal income”. Collinear variables can impact performance because they lead to model weights that have high variance and uncertainty making it difficult to determine the importance and statistical significance of different variables [Lafi and Kaneene, 1992]. Therefore, it is important to correct or remove them along with any nuisance variables that are uninformative [Chan et al., 2022]. This can help to improve model predictive performance in addition to reducing the length of time required to fit the model. Furthermore, as many variables, particularly medical exposures, are composed of multiple subcategories it is important to make sure that the patterns from one are not masked by the others. Feature selection for MTS data can be conducted using explainability methods, most notably PFI which may be used to iteratively fit the model, assess the least important feature which is then discarded before the model is refit to the reduced set of features. Additionally, it is possible to use measures of distance, such as in sequence analysis, to assess how collinear the different medical diagnoses features are to each other and remove collinear features prior to model fitting. The results of the thesis demonstrate that the hyperparameter optimisation surface is flat and therefore, the computational resources and time may be better used to instead focus on conducting feature selection which may result in a better performing model prior to optimisation

The most significant limitation of the work within this thesis were the extreme computational limitations. The use of a shared server which had low memory, limited CPU, limited working hours and frequent shutdowns meant that conducting a comprehensive DL study on the DANLIFE data was impossible. The hyperparameter optimisation had to be substantially scaled back, feature selection was not able to be conducted and the best performing DL architecture from the simulation study could not be assessed on the DANLIFE data. Moreover, the truncation factor for the Signature method was limited to low values, which potentially resulted in a dramatic reduction in model performance. The main takeaway message of this thesis for policymakers hoping to use these cutting-edge approaches is that the existence of these vast data sets alone is not enough without substantial computing resources that are capable of conducting DL analysis. With better infrastructure, the possibilities of these data sets and the potential insights are unparalleled, however without them these new health discoveries cannot be made.

A final point of note is that there are limitations of using the SHAP values to infer patterns and associations within the data and identify the key risk factors. The most prominent one for this analysis is that SHAP values do not convey temporal and inter-variable interactions. In terms of finding causal pathways that link exposures with disease, the SHAP values can only convey association and are not capable of reflecting causation. It would be very interesting to conduct further analysis to identify interactions between the variables and aid in the development of a DAG which would allow decision makers to identify the primary sources of disease risk as opposed to confounders or associated variables. There is currently a surge in research on improving model explainability methods. Suggested future work would incorporate a full exploration of XAI approaches and particularly focus on methodological developments that would enable the interactions occurring between the different exposures to be identified when using MTS data.

6.3 Final Remarks and Key Takeaways

In summary, this thesis represents a significant advancement in the field of life course epidemiology, particularly in the context of RTIs. Through the application of cutting-edge ML models, such as ResNet and

the Signature method, novel insights into the complex dynamics of health trajectories have been uncovered. The methodological innovations introduced, including the rigorous evaluations of ML models on simulated and real-world data, have fundamentally reshaped the landscape of life course analysis. By harnessing the capabilities of ML models to capture intricate temporal patterns and identify influential factors, a deeper understanding of the interplay between early-life exposures and health outcomes has been achieved.

Moving forward, these findings hold promise for healthcare practitioners and policymakers alike. While the DL models showed suboptimal performance in predicting RTIs, their success in identifying patterns associated with other health conditions underscores their potential for uncovering hidden relationships in longitudinal data. Moreover, the incorporation of explainability methods like SHAP and PFI has enhanced the transparency and interpretability of the models, enabling practitioners to identify key risk factors and pathways that lead to increased RTI risk. By providing interpretable explanations for DL models' decisions, this research enhances the understandability and trustworthiness of the results, which is crucial for model employment within medical decision-making. This comprehensive analysis of risk factor associations offers valuable insights for targeting interventions towards the most influential factors and high-risk groups, ultimately improving health outcomes and resource allocation in healthcare settings.

In conclusion, this research represents the first investigation of the utility of DL models within life course analysis and the study of influential risk factors. While limitations exist, particularly in XAI method development and feature selection, the framework laid out in this thesis provides a foundation for future research endeavours. By addressing these challenges, there is significant potential to transform the way life course research is conducted and improve our understanding of risk factors and their dynamic interactions leading to increased RTI risk. Looking ahead, continued exploration into the intricate dynamics of health trajectories promises to yield substantial insights, paving the way for more effective public health interventions and improved population health outcomes.

Chapter 7

References

CDC Seasonal Flu Vaccine Effectiveness Studies | CDC. URL https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm.

Statistics Denmark. URL http://www.dst.dk/en/Statistik/emner/priser-og-forbrug/forbrugerpriser.

A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo. Stock price prediction using the ARIMA model. *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, pages 106–112, 2014. doi: 10.1109/UKSIM.2014.67.

S. Aisenbrey and A. E. Fasang. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, 38(3):420–462, 2010. doi: 10.1177/0049124109357532. URL https://doi.org/10.1177/0049124109357532.

T. Akiba et al. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25rd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining*, 2019.

E. Akyildirim, M. Gambara, J. Teichmann, and S. Zhou. Applications of Signature Methods to Market Anomaly Detection, 2022.

A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 3 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq134. URL https://doi.org/10.1093/bioinformatics/btq134.

J. Alzubi, A. Nayyar, and A. Kumar. Machine Learning from Theory to Algorithms: An Overview Recent citations Analysis and classification of heart diseases using heartbeat features and machine learning algorithms Fajr Ibrahem Alarsan and Mamoon Younes-This content was downloaded from IP address 207 Machine Learning from Theory to Algorithms: An Overview. *J. Phys*, page 12012, 2018. doi: 10.1088/1742-6596/1142/1/012012.

K. G. Andersen et al. The proximal origin of SARS-CoV-2, 4 2020. ISSN 1546170X. URL https://doi.org/10.1038/s41591-020-0820-9.

A. J. Anderson, T. L. Snelling, H. C. Moore, and C. C. Blyth. Advances in Vaccines to Prevent Viral Respiratory Illnesses in Children. *Pediatric Drugs*, 19(6):523–531, 12 2017. ISSN 11792019. doi: 10.1007/s40272-017-0257-x. URL http://link.springer.com/10.1007/s40272-017-0257-x.

L. J. Anderson and S. Tong. Update on SARS research and other possibly zoonotic coronaviruses. *International Journal of Antimicrobial Agents*, 36(SUPPL. 1), 11 2010. ISSN 09248579. doi: 10.1016/j.ijantimicag.2010.06.016.

T. Aoki et al. Development and Validation of a Risk Scoring System for Severe Acute Lower Gastrointestinal Bleeding. *Clinical Gastroenterology and Hepatology*, 14(11):1562–1570, 11 2016. ISSN 1542-3565. doi: 10.1016/J.CGH.2016.05.042.

D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473, 2014. URL https://api.semanticscholar.org/CorpusID:11212020.

T. Bai and P. Tahmasebi. Attention-based LSTM-FCN for earthquake detection and location. *Geophysical Journal International*, 228(3):1568–1576, 12 2021. ISSN 0956-540X. doi: 10.1093/GJI/GGAB401. URL https://dx.doi.org/10.1093/gji/ggab401.

Z. Bai and S. C. Kremer. Sequence learning: Analysis and solutions for sparse data in high dimensional spaces. *2012 IEEE Symposium on Computational Intelligence and Computational Biology, CIBCB 2012*, pages 298–305, 2012. doi: 10.1109/CIBCB.2012.6217244.

N. Béatrice, G. Lise, Z. M. Victoria, and S. Louise. Longitudinal patterns of poverty and health in early childhood: Exploring the influence of concurrent, previous, and cumulative poverty on child health outcomes. *BMC Pediatrics*, 12(1):1–13, 9 2012. ISSN 14712431. doi: 10.1186/1471-2431-12-141/FIGURES/2. URL https://link.springer.com/articles/10.1186/1471-2431-12-141.

Y. Ben-Shlomo and D. Kuh. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31(2):285–293, 4 2002. ISSN 0300-5771. doi: 10.1093/IJE/31.2.285. URL https://academic.oup.com/ije/article/31/2/285/617688.

J. Bengtsson, N. Dich, A. Rieckmann, and N. Hulvej Rod. Cohort profile: the DANish LIFE course (DANLIFE) cohort, a prospective register-based cohort of all children born in Denmark since 1980. *BMJ open*, 9(9): e027217, 9 2019. ISSN 2044-6055. doi: 10.1136/bmjopen-2018-027217. URL http://www.ncbi.nlm.nih.gov/pubmed/31542736.

J. Bengtsson et al. Trajectories of Childhood Adversity and Type 1 Diabetes: A Nationwide Study of One Million Children. *Diabetes Care*, 44(3):740–747, 3 2021. ISSN 0149-5992. doi: 10.2337/DC20-1130. URL https://diabetesjournals.org/care/article/44/3/740/138624/Trajectories-of-Childhood-Adversity-and-Type-1.

A. Berbudi, N. Rahmadika, A. I. Tjahjadi, and R. Ruslami. Type 2 Diabetes and its Impact on the Immune System. *Current Diabetes Reviews*, 16(5):442, 10 2020. ISSN 15733998. doi: 10.2174/1573399815666191024085838. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7475801/.

F. C. Billari, J. Fürnkranz, and A. Prskawetz. Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population*, 22(1):37–65, 3 2006. ISSN 01686577. doi: 10.1007/S10680-005-5549-0/METRICS. URL https://link.springer.com/article/10.1007/s10680-005-5549-0.

P. Blanchard. Sequence Analysis for Political Science, 2011. URL https://papers.ssrn.com/abstract=1902086.

J. M. Bland and D. G. Altman. The odds ratio. *BMJ*, 320(7247):1468, 5 2000. ISSN 0959-8138. doi: 10.1136/BMJ.320.7247.1468. URL https://www.bmj.com/content/320/7247/1468.1.

K. Bønnelykke et al. Association between respiratory infections in early life and later asthma is independent of virus type. *Journal of Allergy and Clinical Immunology*, 136(1):81–86.e4, 2015. ISSN 10976825. doi: 10.1016/j.jaci.2015.02.024.

D. Borra, A. Andalo, S. Severi, and C. Corsi. On the Application of Convolutional Neural Networks for 12-lead ECG Multi-label Classification Using Datasets from Multiple Centers. *Computing in Cardiology*, 2020-Septe, 9 2020. ISSN 2325887X. doi: 10.22489/CINC.2020.349.

G. Box, G. Jenkins, G. Reinsel, and G. Ljung. *Time series analysis: forecasting and control*. Holden-Day, San Francisco, CA, 1970. URL https://books.google.com/books?hl=en&lr=&id=rNt5CgAAQBAJ&oi=fnd&pg=PR7&ots=DK28vQq-VF&sig=yb7JQmsLXJ-QLzkOHz6HeUH7z3I.

K. Boyd, K. H. Eng, and C. D. Page. Area under the precision-recall curve: Point estimates and confidence intervals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8190 LNAI(PART 3):451–466, 2013. ISSN 03029743. doi: 10.1007/978-3-642-40994-3{_}29/COVER. URL https://link.springer.com/chapter/10.1007/978-3-642-40994-3_29.

G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Department of Commerce*, 78(1), 1 1950.

J. Brown, A. Roy, R. Harris, and et al. Respiratory symptoms in people living with HIV and the effect of antiretroviral therapy: a systematic review and meta-analysis. *Thorax*, 72(4):355–366, 4 2017. ISSN 1468-3296. doi: 10.1136/thoraxjnl-2016-208657. URL http://www.ncbi.nlm.nih.gov/pubmed/27965402.

L. M. Bruce, C. H. Koger, and J. Li. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10):2331–2338, 10 2002. ISSN 01962892. doi: 10.1109/TGRS.2002.804721.

C. Brzinsky-Fay and U. Kohler. New developments in sequence analysis. *Sociological Methods and Research*, 38(3):359–364, 2009. ISSN 00491241. doi: 10.1177/0049124110363371. URL http://smr.sagepub.com.

M. R. Byrialsen and H. Raza. Assessing the Macroeconomic Effects of COVID-19 in an Empirical SFC Model for Denmark. *https://doi.org/10.1080/08911916.2021.2001194*, 50(4):318–350, 2021. ISSN 15580970. doi: 10.1080/08911916.2021.2001194. URL https://www.tandfonline.com/doi/abs/10.1080/08911916.2021.2001194.

A. Calderaro et al. Respiratory Tract Infections and Laboratory Diagnostic Methods: A Review with A Focus on Syndromic Panel-Based Assays. *Microorganisms*, 10, 9 2022. ISSN 2076-2607. doi: 10.3390/MICROORGANISMS10091856. URL https://www.mdpi.com/2076-2607/10/9/1856/htm.

T. Can. Introduction to bioinformatics. *Methods in Molecular Biology*, 1107:51–71, 2014. ISSN 10643745. doi: 10.1007/978-1-62703-748-8{_}4/FIGURES/10. URL https://link.springer.com/protocol/10.1007/978-1-62703-748-8_4.

J. Y.-L. Chan et al. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, 10(8), 2022. ISSN 2227-7390. doi: 10.3390/math10081283. URL https://www.mdpi.com/2227-7390/10/8/1283.

A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian. Neural Network Attributions: A Causal Perspective. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:1660–1676, 2 2019. URL http://arxiv.org/abs/1902.02302.

N. V. Chawla and D. A. Davis. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(SUPPL.3):660–665, 9 2013. ISSN 08848734. doi: 10.1007/S11606-013-2455-8/FIGURES/1. URL https://link.springer.com/article/10.1007/s11606-013-2455-8.

Q. Chen et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. *PLOS Medicine*, 17(11):e1003416, 11 2020. ISSN 1549-1676. doi: 10.1371/JOURNAL.PMED.1003416. URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003416.

E. R. Cheng, R. Steinhardt, and Z. B. Miled. Predicting Childhood Obesity Using Machine Learning: Practical Considerations. *BioMedInformatics 2022, Vol. 2, Pages 184-203*, 2(1):184–203, 3 2022. ISSN 2673-7426. doi: 10.3390/BIOMEDINFORMATICS2010012. URL https://www.mdpi.com/2673-7426/2/1/12/htm.

J. Cheng, L. Dong, and M. Lapata. Long Short-Term Memory-Networks for Machine Reading. *CoRR*, abs/1601.06733, 2016. URL http://arxiv.org/abs/1601.06733.

I. Chevyrev and A. Kormilitzin. A Primer on the Signature Method in Machine Learning, 2016.

E. Choi et al. RETAIN: Interpretable Predictive Model in Healthcare using Reverse }Time Attention Mechanism. *CoRR*, abs/1608.05745, 2016. URL http://arxiv.org/abs/1608.05745.

E. Choi et al. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 3 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocw112. URL https://academic.oup.com/jamia/article/24/2/361/2631499.

A. H. Cillessen and C. Borch. Developmental trajectories of adolescent popularity: A growth curve modelling analysis. *Journal of Adolescence*, 29(6):935–959, 12 2006. ISSN 0140-1971. doi: 10.1016/J.ADOLESCENCE.2006.05.005.

C. R. Clark et al. Predicting Self-Rated Health Across the Life Course: Health Equity Insights from Machine Learning Models. *Journal of General Internal Medicine*, 36(5):1181–1188, 5 2021. ISSN 15251497. doi: 10.1007/S11606-020-06438-1/TABLES/2. URL https://link.springer.com/article/10.1007/s11606-020-06438-1.

Colah's blog. Understanding LSTM Networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

N. A. Conners et al. Children of Mothers with Serious Substance Abuse Problems: An Accumulation of Risks. *The American Journal of Drug and Alcohol Abuse*, 29(4):743–758, 2003. ISSN 00952990. doi: 10.1081/ADA-120026258. URL https://www.tandfonline.com/doi/abs/10.1081/ADA-120026258.

J. Cortes-Ramirez et al. Environmental risk factors associated with respiratory diseases in children with socioeconomic disadvantage. *Heliyon*, 7(4):e06820, 4 2021. ISSN 24058440. doi: 10.1016/J.HELIYON.2021.E06820. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8093469/.

H. Coupland et al. Life course analysis using deep learning approaches: a simulation study. *PREPRINT (Version 1) available at Research Square*, 11 2023. doi: 10.21203/RS.3.RS-3601343/V1. URL https://www.researchsquare.com/article/rs-3601343/v1.

H. J. Crocker and A. W. Costall. An InceptionTime-Inspired Convolutional Neural Network to Detect Cardiac Abnormalities in Reduced-Lead ECG Data. *Computing in Cardiology*, 2021-Septe, 2021. ISSN 2325887X. doi: 10.23919/CINC53138.2021.9662678.

D. Dahiwade, G. Patle, and E. Meshram. Designing disease prediction model using machine learning approach. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, pages 1211–1215, 3 2019. doi: 10.1109/ICCMC.2019.8819782.

A. Dasgupta et al. Risk estimation using probability machines. *BioData Mining*, 7(1):1–17, 3 2014. ISSN 17560381. doi: 10.1186/1756-0381-7-2/FIGURES/7. URL https://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-7-2.

data.who.int. WHO Coronavirus (COVID-19) dashboard > Deaths [Dashboard]. .

J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series*, 148:233–240, 2006. doi: 10.1145/1143844.1143874.

A. P. De Filippis et al. Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *European Heart Journal*, 38(8):598–608, 2 2017. ISSN 0195-668X. doi: 10.1093/EURHEARTJ/EHW301. URL https://academic.oup.com/eurheartj/article/38/8/598/3056929.

J. Deb and T. K. Chakrabarty. Unsupervised Learning of the Sequences of Adulthood Transition Trajectories. In Johri Prashant, Verma Jitendra Kumar, and Paul Sudip, editors, *Applications of Machine Learning*, pages 293–319. Springer Singapore, Singapore, 2020. ISBN 978-981-15-3357-0. doi: 10.1007/978-981-15-3357-0{_}20. URL https://doi.org/10.1007/978-981-15-3357-0_20.

Department of Data and Analytics (DNA). WHO methods and data sources for country-level causes of death 2000-2019. Technical report, WHO, Geneva, 12 2020. URL http://www.who.int/gho/mortality_burden_disease/en/index.html.

S. C. Dharmage et al. Do childhood respiratory infections continue to influence adult respiratory morbidity? *European Respiratory Journal*, 33(2):237–244, 2 2009. ISSN 0903-1936. doi: 10.1183/09031936.00062907. URL https://erj.ersjournals.com/content/33/2/237.

W. D'Hoore, C. Sicotte, and C. Tilquin. Risk adjustment in outcome assessment: The Charlson comorbidity index. *Methods of Information in Medicine*, 32(5):382–387, 1993. ISSN 00261270. doi: 10.1055/S-0038-1634956/ID/JR1634956-14/BIB. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0038-1634956.

F. Di Martino and F. Delmastro. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315, 6 2023. ISSN 15737462. doi: 10.1007/S10462-022-10304-3/TABLES/11. URL https://link.springer.com/article/10.1007/s10462-022-10304-3.

J. L. Domingo and J. Rovira. Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environmental Research*, 187:109650, 8 2020. ISSN 10960953. doi: 10.1016/J.ENVRES.2020.109650. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7211639/.

J. Dratva et al. Early Life Origins of Lung Ageing: Early Life Exposures and Lung Function Decline in Adulthood in Two European Cohorts Aged 28-73 Years. *PLOS ONE*, 11(1):e0145127, 1 2016. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0145127. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0145127.

E. Dritsas and M. Trigka. Supervised Machine Learning Models for Liver Disease Risk Prediction. *Computers 2023, Vol. 12, Page 19*, 12(1):19, 1 2023. ISSN 2073-431X. doi: 10.3390/COMPUTERS12010019. URL https://www.mdpi.com/2073-431X/12/1/19.

M. E. Dupre, L. K. George, G. Liu, and E. D. Peterson. The Cumulative Effect of Unemployment on Risks for Acute Myocardial Infarction. *Archives of Internal Medicine*, 172(22):1731–1737, 5 2012. ISSN 0003-9926. doi: 10.1001/2013.jamainternmed.447. URL https://doi.org/10.1001/2013.jamainternmed.447.

S. R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6(3):361–365, 6 1996. ISSN 0959440X. doi: 10.1016/S0959-440X(96)80056-X.

E. P. Eggleston, J. H. Laub, and R. J. Sampson. Methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology*, 20(1):1–26, 3 2004. ISSN 07484518. doi: 10.1023/B:JOQC.0000016696.02763.CE/METRICS. URL https://link.springer.com/article/10.1023/B:JOQC.0000016696.02763.ce.

J. H. Ellenberg. Selection bias in observational and experimental studies. *Statistics in Medicine*, 13(5-7):557–567, 3 1994. ISSN 1097-0258. doi: 10.1002/SIM.4780130518. URL https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780130518.

C. Esteban et al. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. In *Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016*, pages 93–101. Institute of Electrical and Electronics Engineers Inc., 12 2016. ISBN 9781509061174. doi: 10.1109/ICHI.2016.16.

K. Fauvel et al. Local Cascade Ensemble for Multivariate Data Classification. *CoRR*, abs/2005.03645, 2020. URL https://arxiv.org/abs/2005.03645.

H. I. Fawaz et al. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 11 2020. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-020-00710-y. URL http://arxiv.org/abs/1909.04939.

A. M. Fendrick, A. S. Monto, B. Nightengale, and M. Sarnes. The Economic Burden of Non–Influenza-Related Viral Respiratory Tract Infection in the United States. *Archives of Internal Medicine*, 163(4):487–494, 2 2003. ISSN 0003-9926. doi: 10.1001/ARCHINTE.163.4.487. URL https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/215118.

A. Fisher, C. Rudin, and F. Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20, 1 2018. ISSN 15337928. URL https://arxiv.org/abs/1801.01489v5.

E. Flachs et al. Sygdomsbyrden i Danmark – sygdomme. Technical report, Statens Institut for Folke-sundhed, Syddansk Universitet, København : Sundhedsstyrelsen, 2015.

W. D. Flanders et al. Validation of the pneumonia severity index. *Journal of General Internal Medicine*, 14(6):333–340, 1999. ISSN 1525-1497. doi: 10.1046/j.1525-1497.1999.00351.x. URL https://doi.org/10.1046/j.1525-1497.1999.00351.x.

S. Flaxman et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: Technical description update, 2020.

J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29 (5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

J. Futoma et al. An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection, 11 2017. ISSN 2640-3498. URL https://proceedings.mlr.press/v68/futoma17a.html.

B. Galobardes, P. McCarron, M. Jeffreys, and G. Davey Smith. Association between early life history of respiratory disease and morbidity and mortality in adulthood. *Thorax*, 63(5):423–429, 5 2008. ISSN 0040-6376. doi: 10.1136/THX.2007.086744. URL https://thorax.bmj.com/content/63/5/423.

Y. d. Gao et al. Risk factors for severe and critically ill COVID-19 patients: A review. *Allergy*, 76(2):428–455, 2 2021. ISSN 1398-9995. doi: 10.1111/ALL.14657. URL https://onlinelibrary.wiley.com/doi/full/10.1111/all.14657.

B. Giovanola and S. Tiribelli. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI and Society*, 38(2):549–563, 4 2023. ISSN 14355655. doi: 10.1007/S00146-022-01455-6/METRICS. URL https://link.springer.com/article/10.1007/s00146-022-01455-6.

J. Golding. The Avon Longitudinal Study of Parents and Children (ALSPAC)–study design and collaborative opportunities. *European journal of endocrinology*, 151 Suppl 3(3), 11 2004. ISSN 0804-4643. doi: 10.1530/EJE.0.151U119. URL https://pubmed.ncbi.nlm.nih.gov/15554896/.

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 9 2013. ISSN 15372715. doi: 10.1080/10618600.2014.907095. URL https://arxiv.org/abs/1309.6392v2.

A. Gosselin, A. D. Du Loû, and E. Lelièvre. How to use sequence analysis for life course epidemiology? An example on HIV-positive Sub-Saharan migrants in France. *J Epidemiol Community Health*, 72(6):507–512, 6 2018. ISSN 0143-005X. doi: 10.1136/JECH-2017-209739. URL https://jech.bmj.com/content/72/6/507.

C. Goutte and E. Gaussier. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In *Lecture Notes in Computer Science*, volume 3408, pages 345–359. Springer Verlag, 2005. doi: 10.1007/978-3-540-31865-1{_}25. URL https://link.springer.com/chapter/10.1007/978-3-540-31865-1_25.

J. L. Gradus et al. Prediction of Sex-Specific Suicide Risk Using Machine Learning and Single-Payer Health Care Registry Data From Denmark. *JAMA psychiatry*, 77(1):25–34, 1 2020. ISSN 2168-6238. doi: 10.1001/JAMAPSYCHIATRY.2019.2905. URL https://pubmed.ncbi.nlm.nih.gov/31642880/.

T. Grant, E. P. Brigham, and M. C. McCormack. Childhood Origins of Adult Lung Disease as Opportunities for Prevention. *The journal of allergy and clinical immunology. In practice*, 8(3):849, 3 2020. ISSN 22132198. doi: 10.1016/J.JAIP.2020.01.015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7653544/.

A. Graves. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 37–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-24797-2. doi: 10.1007/978-3-642-24797-2{_}4. URL https://doi.org/10.1007/978-3-642-24797-2_4.

R. Guerreiro and J. Bras. The age factor in Alzheimer's disease. *Genome Medicine*, 7(1):1–3, 10 2015. ISSN 1756994X. doi: 10.1186/S13073-015-0232-5/METRICS. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0232-5.

A. K. Guitor and G. D. Wright. Antimicrobial Resistance and Respiratory Infections. *Chest*, 154(5):1202–1212, 11 2018. ISSN 1931-3543. doi: 10.1016/J.CHEST.2018.06.019. URL https://pubmed.ncbi.nlm.nih.gov/29959904/.

X. Guo et al. Bayesian inference for functional dynamics exploring in fMRI data. *Computational and Mathematical Methods in Medicine*, 2016, 2016. ISSN 17486718. doi: 10.1155/2016/3279050.

A. Gupta and B. Dhingra. Stock market prediction using Hidden Markov Models. *2012 Students Conference on Engineering and Systems, SCES 2012*, 2012. doi: 10.1109/SCES.2012.6199099.

E. G. H., M. K. Johnson, and C. Robert. The Emergence and Development of Life Course Theory. In J. T. Mortimer and M. J. Shanahan, editors, *Handbook of the Life Course*, pages 3–19. Springer US, Boston, MA, 2003. ISBN 978-0-306-48247-2. doi: 10.1007/978-0-306-48247-2{_}1. URL https://doi.org/10.1007/978-0-306-48247-2_1.

A. Haerskjold, K. Kristensen, M. Kamper-Jørgensen, and et al. Risk Factors for Hospitalization for Respiratory Syncytial Virus Infection: A Population-based Cohort Study of Danish Children. *The Pediatric Infectious Disease Journal*, page 1, 9 2015. ISSN 0891-3668. doi: 10.1097/INF.0000000000000924. URL http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006454-900000000-97594.

N. Halfon et al. Lifecourse health development: Past, present and future. *Maternal and Child Health Journal*, 18(2):344–365, 2 2014. ISSN 10927875. doi: 10.1007/S10995-013-1346-2/TABLES/1. URL https://link.springer.com/article/10.1007/s10995-013-1346-2.

B. Halpin. Sequence analysis of life-course data: a comparison of distance measures. *Limerick: University of Limerick, Department of Sociology (Working paper WP2012-02)*, 2012.

A. T. Harding and N. S. Heaton. Efforts to Improve the Seasonal Influenza Vaccine. *Vaccines 2018, Vol. 6, Page 19*, 6(2):19, 3 2018. ISSN 2076-393X. doi: 10.3390/VACCINES6020019. URL https://www.mdpi.com/2076-393X/6/2/19.

T. Hartonen et al. Nationwide health, socio-economic and genetic predictors of COVID-19 vaccination status in Finland. *Nature Human Behaviour 2023 7:7*, 7(7):1069–1083, 4 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01591-z. URL https://www.nature.com/articles/s41562-023-01591-z.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 12 2015. ISSN 10636919. doi: 10.48550/arxiv.1512.03385. URL https://arxiv.org/abs/1512.03385v1.

M. Hermans and B. Schrauwen. Training and Analysing Deep Recurrent Neural Networks. In C. J. Burges et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1ff8a7b5dc7a7d1f0ed65aaa29c04b1e-Paper.pdf.

T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual
}Predictions of Complex Models. *CoRR*, abs/2011.01625, 2020. URL https://arxiv.org/abs/2011.01625.

S. Hochreiter. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998. doi: 10.1142/S0218488598000094. URL https://doi.org/10.1142/S0218488598000094.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735. URL http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

R. Hoffmann. Handb. of health inequalities across the life course. *Handbook of Health Inequalities Across the Life Course*, pages 1–434, 4 2023. doi: 10.4337/9781800888166.

M. Hoogendoorn, R. T. Hoogenveen, M. P. Rutten-van Mölken, and et al. Case fatality of COPD exacerbations: A meta-analysis and statistical modelling approach. *European Respiratory Journal*, 37(3):508–515, 3 2011. ISSN 09031936. doi: 10.1183/09031936.00043710.

J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 3 2005. ISSN 10414347. doi: 10.1109/TKDE.2005.50.

T. Huang et al. A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems. *Artificial Intelligence Review*, 55(2):1289–1315, 2 2022. ISSN 15737462. doi: 10.1007/S10462-021-09993-Z/TABLES/8. URL https://link.springer.com/article/10.1007/s10462-021-09993-z.

K. Hughes et al. The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *The Lancet Public Health*, 2(8):e356–e366, 8 2017. ISSN 24682667. doi: 10.1016/S2468-2667(17)30118-4. URL http://www.thelancet.com/article/S2468266717301184/fulltext.

E. D. Hutchison. An Update on the Relevance of the Life Course Perspective for Social Work:. *https://doi.org/10.1177/1044389419873240*, 100(4):351–366, 9 2019. ISSN 19451350. doi: 10.1177/1044389419873240. URL https://journals.sagepub.com/doi/full/10.1177/1044389419873240.

J. Hwang and C. Shon. Relationship between socioeconomic status and type 2 diabetes: results from Korea National Health and Nutrition Examination Survey (KNHANES) 2010–2012. *BMJ Open*, 4(8):e005710, 8 2014. ISSN 2044-6055. doi: 10.1136/BMJOPEN-2014-005710. URL https://bmjopen.bmj.com/content/4/8/e005710.

Institute for Health Metrics and Evaluation (IHME). GBD Compare. Technical report, University of Washington, Seattle, WA: IHME, 2019.

E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax, 2017.

B. J. Jefferson. Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race. *Annals of the American Association of Geographers*, 108(1):1–16, 2018. doi: 10.1080/24694452.2017.1293500. URL https://doi.org/10.1080/24694452.2017.1293500.

W. Jia et al. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Computers and Electronics in Agriculture*, 172:105380, 5 2020. ISSN 0168-1699. doi: 10.1016/J.COMPAG.2020.105380.

D. Jin et al. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3):e1548, 5 2022. ISSN 2692-9368. doi: 10.1002/WSBM.1548. URL https://onlinelibrary.wiley.com/doi/full/10.1002/wsbm.1548.

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 12 2019. ISSN 21961115. doi: 10.1186/S40537-019-0192-5/TABLES/13. URL https://link.springer.com/articles/10.1186/s40537-019-0192-5.

L. Johnson, R. Cornish, A. Boyd, and J. MacLeod. Socio-demographic patterns in hospital admissions and accident and emergency attendances among young people using linkage to NHS Hospital Episode Statistics: Results from the Avon Longitudinal Study of Parents and Children. *BMC Health Services Research*, 19(1):1–13, 2 2019. ISSN 14726963. doi: 10.1186/S12913-019-3922-7/TABLES/2. URL https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-3922-7.

H. J. Jones and L. De Cossart. Risk scoring in surgical patients. *British Journal of Surgery*, 86(2):149–157, 2 1999. ISSN 1365-2168. doi: 10.1046/J.1365-2168.1999.01006.X. URL https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-2168.1999.01006.x.

N. L. Jones et al. Life Course Approaches to the Causes of Health Disparities. *American Journal of Public Health*, 109(Suppl 1):S48–S55, 1 2019. ISSN 0090-0036. doi: 10.2105/AJPH.2018.304738. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6356123/.

H. Jung et al. ResNet-Based Vehicle Classification and Localization in Traffic Surveillance Systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 934–940, 2017. doi: 10.1109/CVPRW.2017.129.

K. A. Kalmakis and G. E. Chandler. Adverse childhood experiences: towards a clear conceptual meaning. *Journal of Advanced Nursing*, 70(7):1489–1501, 7 2014. ISSN 1365-2648. doi: 10.1111/JAN.12329. URL https://onlinelibrary.wiley.com/doi/full/10.1111/jan.12329.

E. R. Kanasewich. *Time Sequence Analysis in Geophysics: Third Edition*. University of Alberta Press, 1981. ISBN 9780888640741. URL https://books.google.co.uk/books?id=k8SSLy-FYagC.

F. Karim, S. Majumdar, H. Darabi, and S. Chen. LSTM Fully Convolutional Networks for Time Series Classification. *CoRR*, abs/1709.05206, 2017. URL http://arxiv.org/abs/1709.05206.

A. Kastrati, M. B. Plomecka, R. Wattenhofer, and N. Langer. Using Deep Learning to Classify Saccade Direction from Brain Activity. *Eye Tracking Research and Applications Symposium (ETRA)*, PartF16925, 5 2021. doi: 10.1145/3448018.3458014. URL https://doi.org/10.1145/3448018.3458014.

N. Kaushal et al. Mutational Frequencies of SARS-CoV-2 Genome during the Beginning Months of the Outbreak in USA. *Pathogens 2020, Vol. 9, Page 565*, 9(7):565, 7 2020. ISSN 2076-0817. doi: 10.3390/PATHOGENS9070565. URL https://www.mdpi.com/2076-0817/9/7/565.

H. Kim, R. G. Webster, and R. J. Webby. Influenza Virus: Dealing with a Drifting and Shifting Pathogen. *Viral Immunology*, 31(2):174–183, 3 2018. ISSN 15578976. doi: 10.1089/VIM.2017.0141/ASSET/IMAGES/LARGE/FIGURE2.JPEG. URL https://www.liebertpub.com/doi/10.1089/vim.2017.0141.

L. C. Kleinman and E. C. Norton. What's the Risk? A Simple Approach for Estimating Adjusted Risk Measures from Nonlinear Models Including Logistic Regression. *Health Services Research*, 44(1):288–302, 2 2009. ISSN 1475-6773. doi: 10.1111/J.1475-6773.2008.00900.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1475-6773.2008.00900.x.

T. Kreimeier et al. Towards the Solution of Robust Gas Network Optimization Problems Using the Constrained Active Signature Method. In *Proceedings of the 10th International Network Optimization Conference (INOC)*, pages 7–10, Aachen, 6 2022. ISBN 9783893180905. doi: 10.48786/inoc.2022.09.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 2012. URL http://code.google.com/p/cuda-convnet/.

D. Kuh, Y. Ben-Shlomo, J. Lynch, and J. Hallqvist. Life course epidemiology. *J Epidemiol Community Health*, 57:778–783, 2003. URL www.jech.com.

I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR, 3 2020. URL https://proceedings.mlr.press/v119/kumar20e.html.

S. Q. Lafi and J. B. Kaneene. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Preventive Veterinary Medicine*, 13(4):261–275, 1992. ISSN 0167-5877. doi: https://doi.org/10.1016/0167-5877(92)90041-D. URL https://www.sciencedirect.com/science/article/pii/016758779290041D.

A. D. Lauritzen et al. Identifying recurrent breast cancer patients in national health registries using machine learning. *Acta Oncologica*, 62(4):350–357, 2023. ISSN 1651226X. doi: 10.1080/0284186X.2023.2201687. URL https://www.tandfonline.com/doi/abs/10.1080/0284186X.2023.2201687.

D. A. Lawlor et al. The second generation of The Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. *Wellcome Open Research*, 4:36, 12 2019. doi: 10.12688/WELLCOMEOPENRES.15087.2.

A. M. Lehrmann, P. V. Gehler, and S. Nowozin. Efficient Nonlinear Markov Models for Human Motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014.

N. H. Leung. Transmissibility and transmission of respiratory viruses. *Nature Reviews Microbiology 2021 19:8*, 19(8):528–545, 3 2021. ISSN 1740-1534. doi: 10.1038/s41579-021-00535-6. URL https://www.nature.com/articles/s41579-021-00535-6.

Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, 7 2017. ISSN 23318422. URL https://arxiv.org/abs/1707.01926v3.

W. Liang and T. Chikritzhs. Alcohol consumption during adolescence and risk of diabetes in young adulthood. *BioMed research international*, 2014, 2014. ISSN 2314-6141. doi: 10.1155/2014/795741. URL https://pubmed.ncbi.nlm.nih.gov/24757678/.

M. Liguori et al. Age at onset in multiple sclerosis. *Neurological Sciences 2000 21:2*, 21(2):S825–S829, 12 2000. ISSN 1590-3478. doi: 10.1007/S100720070020. URL https://link.springer.com/article/10.1007/s100720070020.

H. Lin et al. Epidemiological characteristics and forecasting incidence for patients with breast cancer in Shantou, Southern China: 2006–2017. *Cancer Medicine*, 10(8):2904, 4 2021. ISSN 20457634. doi: 10.1002/CAM4.3843. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8026945/.

T.-Y. Lin et al. Focal Loss for Dense Object Detection, 2017.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021, Vol. 23, Page 18*, 23(1):18, 12 2020. ISSN 1099-4300. doi: 10.3390/E23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with LSTM recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2016.

L. Liu, M. Han, Y. Zhou, and Y. Wang. LSTM recurrent neural networks for influenza trends prediction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10847 LNBI, pages 259–264. Springer Verlag, 6 2018. ISBN 9783319949673. doi: 10.1007/978-3-319-94968-0{_}25. URL http://link.springer.com/10.1007/978-3-319-94968-0_25.

Z. Lu, X. Jiang, and A. Kot. Deep Coupled ResNet for Low-Resolution Face Recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 4 2018. ISSN 10709908. doi: 10.1109/LSP.2018.2810121.

J. F. Ludvigsson et al. Ethical aspects of registry-based research in the Nordic countries. *Clinical Epidemiology*, 7:491–508, 11 2015. ISSN 11791349. doi: 10.2147/CLEP.S90589. URL https://doi.org/10.2147/CLEP.S90589.

S. M. Lundberg and S. I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-Decem:4766–4775, 5 2017. ISSN 10495258. doi: 10.48550/arxiv.1705.07874. URL https://arxiv.org/abs/1705.07874v2.

M.-T. Luong, H. Pham, and C. D. Manning. Effective Approaches to Attention-based Neural Machine Translation. *CoRR*, 2015. URL http://nlp.stanford.edu/projects/nmt.

T. Lyons. Rough paths, Signatures and the modelling of functions on streams, 2014.

T. Lyons and A. D. Mcleod. Signature Methods in Machine Learning, 2023.

F. Ma et al. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional }Recurrent Neural Networks. *CoRR*, abs/1706.05764, 2017. URL http://arxiv.org/abs/1706.05764.

I. MacDonald and W. Zucchini. *Hidden Markov and Other Models for Discrete- valued Time Series*. CRC Press, 1997. URL https://books.google.es/books?hl=en&lr=&id=ji0LKM5_kWQC&oi=fnd&pg=PR13&dq=hidden+markov+model+time+series&ots=8z0KQQ7tjh&sig=ytV2nUfW1AbNVsKDUpA02XqjWtk&redir_esc=y#v=onepage&q=hiddenmarkovmodeltimeseries&f=false.

S. Macleod and R. E. Appleton. Neurological disorders presenting mainly in adolescence. *Archives of Disease in Childhood*, 92(2):170, 2 2007. ISSN 00039888. doi: 10.1136/ADC.2005.088070. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2083313/.

E. Mahase. Covid-19: What do we know about "long covid"?, 7 2020. ISSN 17561833. URL http://dx.doi.org/10.1136/bmj.m2815.

M. A. Mansournia and D. G. Altman. Population attributable fraction. *BMJ*, 360, 2 2018. ISSN 0959-8138. doi: 10.1136/BMJ.K757. URL https://www.bmj.com/content/360/bmj.k757.

M. Marani, G. G. Katul, W. K. Pan, and A. J. Parolari. Intensity and frequency of extreme novel epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 118(35):e2105482118, 8 2021. ISSN 10916490. doi: 10.1073/PNAS.2105482118/SUPPL{_}FILE/PNAS.2105482118.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105482118.

A. Maruotti. Mixed {Hidden} {Markov} {Models} for {Longitudinal} {Data}: {An} {Overview}. *International Statistical Review*, 79(3):427–454, 12 2011. ISSN 03067734. doi: 10.1111/j.1751-5823.2011.00160.x. URL http://doi.wiley.com/10.1111/j.1751-5823.2011.00160.x.

H. T. May et al. The association of depression at any time to the risk of death following coronary artery disease diagnosis. *European Heart Journal - Quality of Care and Clinical Outcomes*, 3(4):296–302, 10 2017. ISSN 2058-5225. doi: 10.1093/ehjqcco/qcx017. URL https://doi.org/10.1093/ehjqcco/qcx017.

N. I. Mazur et al. Respiratory syncytial virus prevention within reach: the vaccine and monoclonal antibody landscape. *The Lancet Infectious Diseases*, 23(1):e2–e21, 1 2023. ISSN 14744457. doi: 10.1016/S1473-3099(22)00291-2/ATTACHMENT/0A07C7B7-C1F2-4A0D-93BD-9051B3290CEA/MMC1.PDF. URL http://www.thelancet.com/article/S1473309922002912/fulltext.

I. C. McDowell et al. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLOS Computational Biology*, 14(1):e1005896, 1 2018. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1005896. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005896.

L. A. McNutt, C. Wu, X. Xue, and J. P. Hafner. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *American Journal of Epidemiology*, 157(10):940–943, 5 2003. ISSN 0002-9262. doi: 10.1093/AJE/KWG074. URL https://dx.doi.org/10.1093/aje/kwg074.

G. C. Meier, J. Watkins, P. McEwan, and R. D. Pockett. Resource use and direct medical costs of acute respiratory illness in the UK based on linked primary and secondary care records from 2001 to 2009. *PLOS ONE*, 15(8):e0236472, 8 2020. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0236472. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236472.

T. Mikolov et al. Extensions of recurrent neural network language model. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5528–5531, 2011. ISBN 9781457705397. doi: 10.1109/ICASSP.2011.5947611.

D. Mincu et al. Concept-based model explanations for electronic health records. *ACM CHIL 2021 - Proceedings of the 2021 ACM Conference on Health, Inference, and Learning*, pages 36–46, 4 2021. doi: 10.1145/3450439.3451858. URL https://dl.acm.org/doi/10.1145/3450439.3451858.

G. D. Mishra, R. Cooper, and D. Kuh. A life course approach to reproductive health: Theory and methods. *Maturitas*, 65(2):92–97, 2 2010. ISSN 0378-5122. doi: 10.1016/J.MATURITAS.2009.12.009.

B. D. Mittelstadt and L. Floridi. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Law, Governance and Technology Series*, 29:445–480, 2016. ISSN 23521910. doi: 10.1007/978-3-319-33525-4{_}19/TABLES/2. URL https://link.springer.com/chapter/10.1007/978-3-319-33525-4_19.

H. T. Mocelin, G. B. Fischer, and A. Bush. Adverse early-life environmental exposures and their repercussions on adult respiratory health. *Jornal de Pediatria*, 98:86–95, 4 2022. ISSN 0021-7557. doi: 10.1016/J.JPED.2021.11.005. URL https://www.scielo.br/j/jped/a/Fjh7rYqczsXF5bJRw8dRtBC/.

C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

S. J. Mooney and V. Pejaver. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *https://doi.org/10.1146/annurev-publhealth-040617-014208*, 39:95–112, 4 2018. ISSN 15452093. doi: 10.1146/ANNUREV-PUBLHEALTH-040617-014208. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-publhealth-040617-014208.

J. Morrill, C. Salvi, P. Kidger, and J. Foster. Neural Rough Differential Equations for Long Time Series. *CoRR*, 2020a.

J. Morrill, A. Fermanian, P. Kidger, and T. Lyons. A Generalised Signature Method for Multivariate Time Series Feature Extraction. *CoRR*, 2021.

J. H. Morrill et al. Utilization of the Signature Method to Identify the Early Onset of Sepsis From Multivariate Physiological Time Series in Critical Care Monitoring. *Critical Care Medicine*, 48(10):E976–E981, 10 2020b. ISSN 15300293. doi: 10.1097/CCM.0000000000004510.

J. Mukhoti et al. Calibrating Deep Neural Networks using Focal Loss. *CoRR*, abs/2002.09437, 2020. URL https://arxiv.org/abs/2002.09437.

A. Muratova, E. Mitrofanova, and R. Islam. Comparison of Machine Learning Methods for Life Trajectory Analysis in Demography. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12672 LNAI:630–642, 2021. ISSN 16113349. doi: 10.1007/978-3-030-73280-6{_}50/TABLES/8. URL https://link.springer.com/chapter/10.1007/978-3-030-73280-6_50.

O. Naggara et al. Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. *AJNR: American Journal of Neuroradiology*, 32(3):437, 3 2011. ISSN 01956108. doi: 10.3174/AJNR.A2425. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8013096/.

D. S. Nagin and C. L. Odgers. Group-Based Trajectory Modeling in Clinical Research. *https://doi.org/10.1146/annurev.clinpsy.121208.131413*, 6:109–138, 3 2010. ISSN 15485943. doi: 10.1146/ANNUREV.CLINPSY.121208.131413. URL https://www.annualreviews.org/doi/abs/10.1146/annurev.clinpsy.121208.131413.

D. S. Nagin, B. L. Jones, V. L. Passos, and R. E. Tremblay. Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, 27(7):2015–2023, 7 2018. ISSN 14770334. doi: 10.1177/0962280216673085/ASSET/IMAGES/LARGE/10.1177{_}0962280216673085-FIG2.JPEG. URL https://journals.sagepub.com/doi/full/10.1177/0962280216673085?casa_token=BfNcKSgjp_8AAAAA%3AKEYOeVosNBmcIlql1_dftG1k_UANzGQYbMlbouJI4-SOLNdeVysA40yk0icv1ReCmDaDR-hY2GMAuQ.

M. Z. Naser. An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction*, 129:103821, 9 2021. ISSN 0926-5805. doi: 10.1016/J.AUTCON.2021.103821.

A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pages 615–622, 2004. doi: 10.1145/1015330.1015435.

Y. Ngamwong et al. Additive Synergism between Asbestos and Smoking in Lung Cancer Risk: A Systematic Review and Meta-Analysis. *PLOS ONE*, 10(8):e0135798, 8 2015. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0135798. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135798.

H. L. Nguena Nguefack et al. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. *Clinical Epidemiology*, 12:1205, 2020. ISSN 11791349. doi: 10.2147/CLEP.S265287. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7608582/.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342.

I. Oguiza. tsai - A state-of-the-art deep learning library for time series and sequential data, 2022.

B. Paes, S. Steele, M. Janes, and J. Pinelli. Risk-Scoring Tool for respiratory syncytial virus prophylaxis in premature infants born at 33–35 completed weeks' gestational age in Canada. *https://doi.org/10.1185/03007990902929112*, 25(7):1585–1591, 7 2009. ISSN 03007995. doi: 10.1185/03007990902929112. URL https://www.tandfonline.com/doi/abs/10.1185/03007990902929112.

J. Y. Park et al. Self-Attention LSTM-FCN model for arrhythmia classification and uncertainty assessment. *Artificial Intelligence in Medicine*, 142:102570, 8 2023. ISSN 0933-3657. doi: 10.1016/J.ARTMED.2023.102570.

P. K. Pattemore and L. C. Jennings. Epidemiology of Respiratory Infections. *Pediatric Respiratory Medicine*, page 435, 1 2008. doi: 10.1016/B978-032304048-8.50035-9. URL /pmc/articles/PMC7151775/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7151775/.

A. Pedersen, R. Zachariae, and D. H. Bovbjerg. Influence of Psychological Stress on Upper Respiratory Infection—A Meta-Analysis of Prospective Studies. *Psychosomatic Medicine*, 72(8):823–832, 10 2010. ISSN 0033-3174. doi: 10.1097/PSY.0b013e3181f1d003. URL https://insights.ovid.com/crossref?an=00006842-201010000-00014.

F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

I. Perez Arribas, G. M. Goodwin, J. R. Geddes, and et al. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational Psychiatry*, 8(1):1–7, 12 2018. ISSN 21583188. doi: 10.1038/s41398-018-0334-0. URL https://www.nature.com/articles/s41398-018-0334-0.

M. Pierce et al. Mental health responses to the COVID-19 pandemic: a latent class trajectory analysis using longitudinal UK data. *The Lancet Psychiatry*, 8(7):610–619, 7 2021. ISSN 22150374. doi: 10.1016/S2215-0366(21)00151-6. URL http://www.thelancet.com/article/S2215036621001516/fulltext.

R. Poulton et al. Association between children's experience of socioeconomic disadvantage and adult health: a life-course study. *The Lancet*, 360(9346):1640–1645, 11 2002. ISSN 0140-6736. doi: 10.1016/S0140-6736(02)11602-3.

L. R. Rabiner and B. H. Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. ISSN 07407467. doi: 10.1109/MASSP.1986.1165342.

F. Radenovic, G. Tolias, and O. Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 7 2019. ISSN 19393539. doi: 10.1109/TPAMI.2018.2846566.

L. Raffington et al. Associations of socioeconomic disparities with buccal DNA-methylation measures of biological aging. *Clinical Epigenetics*, 15(1):1–9, 12 2023. ISSN 18687083. doi: 10.1186/S13148-023-01489-7/TABLES/4. URL https://link.springer.com/articles/10.1186/s13148-023-01489-7.

V. Rajpaul et al. A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291, 9 2015. ISSN 0035-8711. doi: 10.1093/MNRAS/STV1428. URL https://dx.doi.org/10.1093/mnras/stv1428.

A. M. Rather, A. Agarwal, and V. N. Sastry. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, 42(6):3234–3241, 4 2015. ISSN 09574174. doi: 10.1016/j.eswa.2014.12.003.

A. H. Reid, T. G. Fanning, J. V. Hultin, and J. K. Taubenberger. Origin and evolution of the 1918 "Spanish" influenza virus
} hemagglutinin gene. *Proceedings of the National Academy of Sciences of the United States of America*, 96(4):1651, 2 1999. ISSN 00278424. doi: 10.1073/PNAS.96.4.1651. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC15547/.

A. H. Reid, J. K. Taubenberger, and T. G. Fanning. The 1918 Spanish influenza: Integrating history and biology, 1 2001. ISSN 12864579.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101, 2 2016. doi: 10.18653/v1/n16-3020. URL https://arxiv.org/abs/1602.04938v3.

A. Rieckmann et al. Causes of Outcome Learning: a causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome. *International Journal of Epidemiology*, 51(5):1622, 10 2022. ISSN 14643685. doi: 10.1093/IJE/DYAC078. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9799206/.

S. Roberts et al. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A*, 371, 2013. ISSN 1364503X. doi: 10.1098/rsta.2011.0550. URL https://royalsocietypublishing.org/doi/full/10.1098/rsta.2011.0550.

N. H. Rod et al. Hospitalisation patterns among children exposed to childhood adversity: a population-based cohort study of half a million children. *The Lancet Public Health*, 6(11):e826–e835, 11 2021. ISSN 24682667. doi: 10.1016/S2468-2667(21)00158-4. URL http://www.thelancet.com/article/S2468266721001584/fulltext.

M. K. Ross, W. Wei, and L. Ohno-Machado. "Big data" and the electronic health record, 8 2014. ISSN 23640502. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287068/.

M. Rottoli, P. Bernante, A. Belvedere, and et al. How important is obesity as a risk factor for respiratory failure, intensive care admission and death in hospitalised COVID-19 patients? Results from a single Italian centre. *European Journal of Endocrinology*, 183(4):389–397, 10 2020. ISSN 0804-4643. doi: 10.1530/EJE-20-0541. URL https://eje.bioscientifica.com/view/journals/eje/183/4/EJE-20-0541.xml.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0.

A. Russo et al. Lifestyle Factors and Migraine in Childhood. *Current Pain and Headache Reports*, 20(2):1–8, 1 2016. ISSN 15343081. doi: 10.1007/S11916-016-0539-Y/METRICS. URL https://link.springer.com/article/10.1007/s11916-016-0539-y.

N. M. Safdar, J. D. Banja, and C. C. Meltzer. Ethical considerations in artificial intelligence. *European Journal of Radiology*, 122:108768, 1 2020. ISSN 0720-048X. doi: 10.1016/J.EJRAD.2019.108768.

M. Sahadulla. Infections of the Respiratory System. In *Concise Handbook of Infectious Diseases*, pages 65–65. Jaypee Brothers Medical Publishers (P) Ltd., 2018. ISBN 0963117211. doi: 10.5005/jp/books/14120{_}12.

T. Saito and M. Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3), 3 2015. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0118432. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/.

R. Saluja et al. Towards a Rigorous Evaluation of Explainability for Multivariate Time Series. *arXiv*, 2021.

G. Savcisens et al. Using sequences of life-events to predict human lives. *Nature Computational Science 2023*, pages 1–14, 12 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00573-5. URL https://www.nature.com/articles/s43588-023-00573-5.

M. Schmidt et al. The Danish health care system and epidemiological research: From health care contacts to database records. *Clinical Epidemiology*, 11:563–591, 2019. ISSN 11791349. doi: 10.2147/CLEP.S179083. URL http://doi.org/10.2147/CLEP.S179083.

G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. John Wiley & Sons, New Jersey, 2003. URL https://books.google.dk/books?hl=en&lr=&id=X2Y6OkXl8ysC&oi=fnd&pg=PR5&dq=linear+regression&ots=semRE5rPhs&sig=i38wdFx8EPpNoA_pbVQJzgJQAdU&redir_esc=y#v=onepage&q=linearregression&f=false.

A. D. Selbst and J. Powles. Meaningful information and the right to explanation. *Philos. Technol.*, 31:543–556, 2018. doi: 10.1007/s13347-017-0263-5. URL http://ec.europa.eu/justice/data-protection/reform/.

R. R. Selvaraju et al. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks }via Gradient-based Localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

S. Selvin et al. Stock price prediction using LSTM, RNN and CNN-sliding window model. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, 2017-January: 1643–1647, 11 2017. doi: 10.1109/ICACCI.2017.8126078.

Y. Sha and M. D. Wang. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM-BCB '17, pages 233–240, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347228. doi: 10.1145/3107411.3107445. URL https://doi.org/10.1145/3107411.3107445.

A. Shaban-Nejad, M. Michalowski, J. S. Brownstein, and D. L. Buckeridge. Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2374–2375, 2021. doi: 10.1109/JBHI.2021.3088832.

M. S. Shanker, M. Y. Hu, and M. S. Hung. Effect of data standardization on neural network training. *Omega*, 24(4):385–397, 8 1996. ISSN 0305-0483. doi: 10.1016/0305-0483(96)00010-2.

L. S. Shapley. A Value for N-Person Games. *A Value for N-Person Games*, 4 1952. doi: 10.7249/P0295. URL https://www.rand.org/pubs/papers/P295.html.

S. Sharma, S. Sharma, and A. Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12): 310–316, 2017.

A. Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term }Memory (LSTM) Network. *CoRR*, abs/1808.03314, 2018. URL http://arxiv.org/abs/1808.03314.

R. H. Shumway and D. S. Stoffer. ARIMA Models. In *Time Series Analysis and Its Applications: With R Examples*, pages 75–163. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-52452-8{_}3. URL https://link.springer.com/chapter/10.1007/978-3-319-52452-8_3.

S. Siami-Namini, N. Tavakoli, and A. Siami Namin. A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401, 2018. doi: 10.1109/ICMLA.2018.00227.

E. A. Simoes. Environmental and demographic risk factors for respiratory syncytial virus lower respiratory tract disease. *Journal of Pediatrics*, 143(5 SUPPL.):118–126, 11 2003. ISSN 00223476. doi: 10.1067/s0022-3476(03)00511-0. URL https://www.sciencedirect.com/science/article/pii/S0022347603005110.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014.

S. Smith Jervelund and C. J. De Montgomery. Nordic registry data: value, validity and future. *https://doi.org/10.1177/1403494819898573*, 48(1):1–4, 1 2020. ISSN 16511905. doi: 10.1177/1403494819898573. URL https://journals.sagepub.com/doi/full/10.1177/1403494819898573.

J. Sola and J. Sevilla. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, 44(3 PART 3):1464–1468, 1997. ISSN 00189499. doi: 10.1109/23.589532.

C. Sommer, B. Resch, and E. Simões. Risk factors for severe respiratory syncytial virus lower respiratory tract infection. *The Open Microbiology Journal*, 5(1):144–154, 12 2011. ISSN 18742858. doi: 10.2174/1874285801105010144.

J. B. Soriano et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 6 2020. ISSN 22132619. doi: 10.1016/S2213-2600(20)30105-3. URL http://www.thelancet.com/article/S2213260020301053/fulltext.

S. L. Spruance, J. E. Reid, M. Grace, and M. Samore. Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792, 8 2004. ISSN 00664804. doi: 10.1128/AAC.48.8.2787-2792.2004/ASSET/CCF20F12-5BEF-4196-BB8A-B9B581311E42/ASSETS/GRAPHIC/ZAC0080442350003.JPEG. URL https://journals.asm.org/doi/10.1128/aac.48.8.2787-2792.2004.

G. W. Stewart. On the Early History of the Singular Value Decomposition. *https://doi.org/10.1137/1035134*, 35(4):551–566, 7 2006. ISSN 00361445. doi: 10.1137/1035134. URL https://epubs.siam.org/doi/10.1137/1035134.

J. Stocks, A. Hislop, and S. Sonnappa. Early lung development: lifelong effect on respiratory health and disease. *The Lancet. Respiratory medicine*, 1(9):728–742, 11 2013. ISSN 2213-2600. doi: 10.1016/S2213-2600(13)70118-8. URL https://pubmed.ncbi.nlm.nih.gov/24429276/.

C. Szegedy et al. Going Deeper with Convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1–9, 9 2014. ISSN 10636919. doi: 10.48550/arxiv.1409.4842. URL https://arxiv.org/abs/1409.4842v1.

C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2 2017. doi: 10.1609/aaai.v31i1.11231. URL https://ojs.aaai.org/index.php/AAAI/article/view/11231.

J. K. Taubenberger. The Origin and Virulence of the 1918 "Spanish" Influenza Virus. *Proceedings of the American Philosophical Society*, 150(1):86, 2006. ISSN 0003049X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2720273/.

A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti. Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. *IEEE Access*, 10:100700–100724, 2022. doi: 10.1109/ACCESS.2022.3207765.

O. M. Todd et al. New Horizons in the use of routine data for ageing research. *Age and Ageing*, 49(5):716, 8 2020. ISSN 14682834. doi: 10.1093/AGEING/AFAA018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7444666/.

K. Tokuda et al. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, 101(5): 1234–1252, 2013. ISSN 00189219. doi: 10.1109/JPROC.2013.2251852.

N. Tomes. "Destroyer and Teacher": Managing the Masses During the 1918–1919 Influenza Pandemic. *Public Health Reports*, 125(Suppl 3):48, 2010. ISSN 14682877. doi: 10.1177/00333549101250S308. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2862334/.

M. L. Trinhammer, A. C. Merrild, J. F. Lotz, and G. Makransky. Predicting crime during or after psychiatric care: Evaluating machine learning for risk assessment using the Danish patient registries. *Journal of Psychiatric Research*, 152:194–200, 8 2022. ISSN 0022-3956. doi: 10.1016/J.JPSYCHIRES.2022.06.009.

S. Upadhyaya et al. Parental Alcohol Problems, Parental Divorce, and Type 2 Diabetes in Adulthood: A Longitudinal Prospective Cohort Study in Middle-Aged Men. *Psychosomatic Medicine*, 82(9):817–822, 11 2020. ISSN 15347796. doi: 10.1097/PSY.0000000000000865. URL https://journals.lww.com/psychosomaticmedicine/fulltext/2020/11000/parental_alcohol_problems,_parental_divorce,_and.3.aspx.

A. Vabalas et al. Deep learning-based prediction of one-year mortality in the entire Finnish population is an accurate but unfair digital marker of aging. *medRxiv*, 2023. doi: 10.1101/2023.09.18.23295726. URL https://doi.org/10.1101/2023.09.18.23295726.

B. Van Calster et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1466-7. URL https://doi.org/10.1186/s12916-019-1466-7.

R. Van Den Goorbergh, M. Van Smeden, D. Timmerman, and Ben Van Calster. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association : JAMIA*, 29(9):1525–1534, 9 2022. ISSN 1527-974X. doi: 10.1093/JAMIA/OCAC093. URL https://pubmed.ncbi.nlm.nih.gov/35686364/.

P. Vartiainen et al. Risk factors for severe respiratory syncytial virus infection during the first year of life: development and validation of a clinical prediction model. *medRxiv*, 3(1):2023.02.23.23286237, 2 2023. doi: 10.1101/2023.02.23.23286237. URL https://www.medrxiv.org/content/10.1101/2023.02.23.23286237v1.

A. Vaswani et al. Attention is All you Need. In I. Guyon et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

L. M. Venmans et al. Prediction of complicated lower respiratory tract infections in older patients with diabetes. *British Journal of General Practice*, 58(553):564–568, 8 2008. ISSN 0960-1643. doi: 10.3399/BJGP08X319620. URL https://bjgp.org/content/58/553/564.

J. Vielhaben, S. Lapuschkin, G. Montavon, and W. Samek. Explainable AI for time series via Virtual Inspection Layers. *Pattern Recognition*, 150:110309, 6 2024. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2024.110309.

G. Vilone and L. Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 12 2021. ISSN 1566-2535. doi: 10.1016/J.INFFUS.2021.05.009.

M. Wahrendorf and T. Chandola. A Life Course Perspective on Work Stress and Health. In M. Siegrist Johannes
}and Wahrendorf, editor, *Work Stress and Health in a Globalized Economy: The Model of Effort-Reward Imbalance*, pages 43–66. Springer International Publishing, Cham, 2016. ISBN 978-3-319-32937-6. doi: 10.1007/978-3-319-32937-6{_}3. URL https://doi.org/10.1007/978-3-319-32937-6_3.

C. A. Walsh et al. Mapping the use of Group-Based Trajectory Modelling in medication adherence research: A scoping review protocol. *HRB Open Research*, 3, 2020. ISSN 25154826. doi: 10.12688/HRBOPENRES.13056.2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7265572/.

S. Walsh et al. A systematic review of the risks factors associated with the onset and natural progression of epilepsy. *NeuroToxicology*, 61:64–77, 7 2017. ISSN 0161-813X. doi: 10.1016/J.NEURO.2016.03.011.

L. Wang, W. Zhang, X. He, and H. Zha. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2447–2456, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219961. URL https://doi.org/10.1145/3219819.3219961.

Y. Wang, H. Yao, and S. Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 4 2016a. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2015.08.104.

Z. Wang, W. Yan, and T. Oates. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:1578–1585, 11 2016b. doi: 10.48550/arxiv.1611.06455. URL https://arxiv.org/abs/1611.06455v4.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 8 1987. ISSN 0169-7439. doi: 10.1016/0169-7439(87)80084-9.

World Health Organization and International Longevity Centre-UK. The Implications for Training of Embracing a Life Course Approach to Health. Technical report, World Health Organization, Geneva, 2000.

J. Xu et al. Reluplex made more practical: Leaky ReLU. *Proceedings - IEEE Symposium on Computers and Communications*, 2020-July, 7 2020. ISSN 15301346. doi: 10.1109/ISCC50000.2020.9219587.

W. Xu, Y. L. Fu, and D. Zhu. ResNet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine*, 240:107660, 10 2023. ISSN 0169-2607. doi: 10.1016/J.CMPB.2023.107660.

W. Yang, L. Jin, H. Ni, and T. Lyons. Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network. In *Proceedings - International Conference on Pattern Recognition*, volume 0, pages 4083–4088. Institute of Electrical and Electronics Engineers Inc., 1 2016. ISBN 9781509048472. doi: 10.1109/ICPR.2016.7900273.

G. Zerveas et al. A Transformer-based Framework for Multivariate Time Series Representation Learning. *CoRR*, abs/2010.02803, 2020. URL https://arxiv.org/abs/2010.02803.

K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning Deep CNN Denoiser Prior for Image Restoration, 2017.

Z. Zhang and M. R. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *Advances in Neural Information Processing Systems*, 31, 2018.

B. Zhao et al. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2 2017. ISSN 16711793. doi: 10.21629/JSEE.2017.01.18.

Z. Zhao Yu Zheng Zixing Zhang Haishuai Wang et al. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. *Interspeech*, 2018-September:272–276, 1 2018. ISSN 19909772. doi: 10.21437/Interspeech.2018-1477. URL https://digitalcommons.fairfield.edu/engineering-facultypubs/181.

J. Zhou, X. Song, and L. Sun. Continuous time hidden {Markov} model for longitudinal data. *Journal of Multivariate Analysis*, 179:104646, 9 2020. ISSN 0047-259X. doi: 10.1016/j.jmva.2020.104646. URL https://www.sciencedirect.com/science/article/pii/S0047259X2030227X.

A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, 1994. ISSN 1558254X. doi: 10.1109/42.363096.

Chapter 8

Appendices

8.1 Simulated Data Distributions

This section, displays plots relating to the distributions of the time dependent variables in the input data. Figure 8.1 displays the time series for ten individuals related to the x_6 variable which is designed to reflect sibling illness, it can be seen that most individuals experience sibling illness at some point in their life course and several experience it multiple times. The distribution of x_6 exposure events across the life course is shown in Figure 8.2, notice that the majority of events are concentrated in early life. The time series of poverty, x_7, for ten individuals is shown in Figure 8.3 alongside their education and class upon which this variable depends upon. Figure 8.4 shows the distribution of poverty exposures according to education and social class. It can be seen that individuals with lower maternal education levels and lower social class tend to experience more incidents of poverty. The time series of parental alcohol abuse exposures for ten individuals is shown in Figure 8.5 as well as their sex and ethnicity. Figure 8.6 shows how the incidents of parental alcohol abuse are distributed within the population according to their sex and ethnicity. Notice that non-white individuals and male individuals experience are more likely to experience incidents of parental alcohol abuse. This variable is purely representative and is not designed to represent any existing trends but to embody some representative pattern of an ACE. Figure 8.7 displays the time series of mild illness for three individuals as well as their associated time series for sibling illness and poverty which are related. The distribution of mild illness events across the life span is shown in Figure 8.8. This shows that this exposure event is significantly more likely to occur in the first years of life.

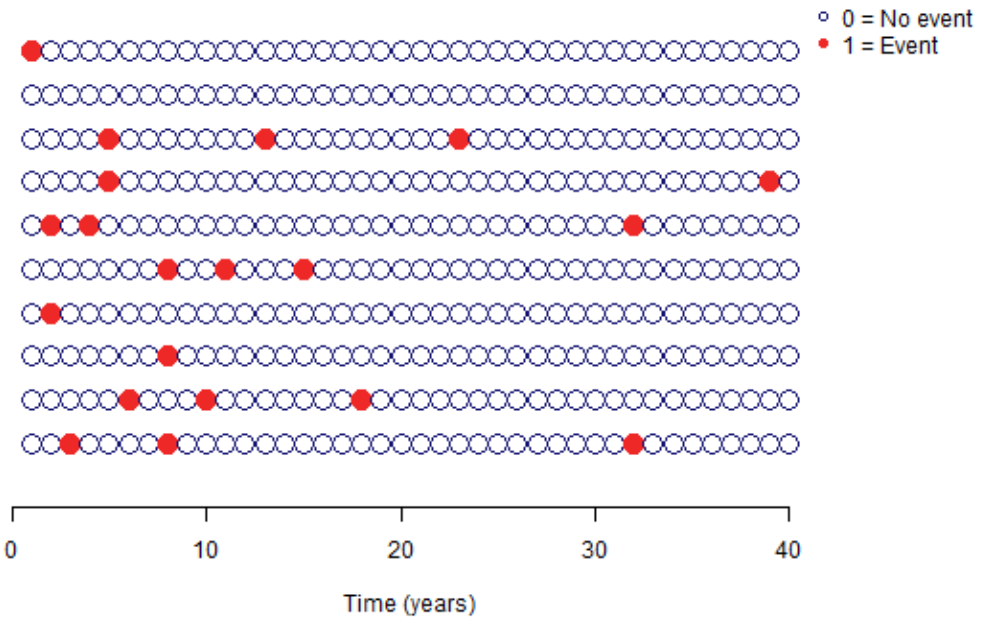


Figure 8.1: **Sibling illness exposure events.** Value of variable x_6 for first ten individuals across the 40 year time frame. Circles show a year and those that are red show the occurrence of the exposure.

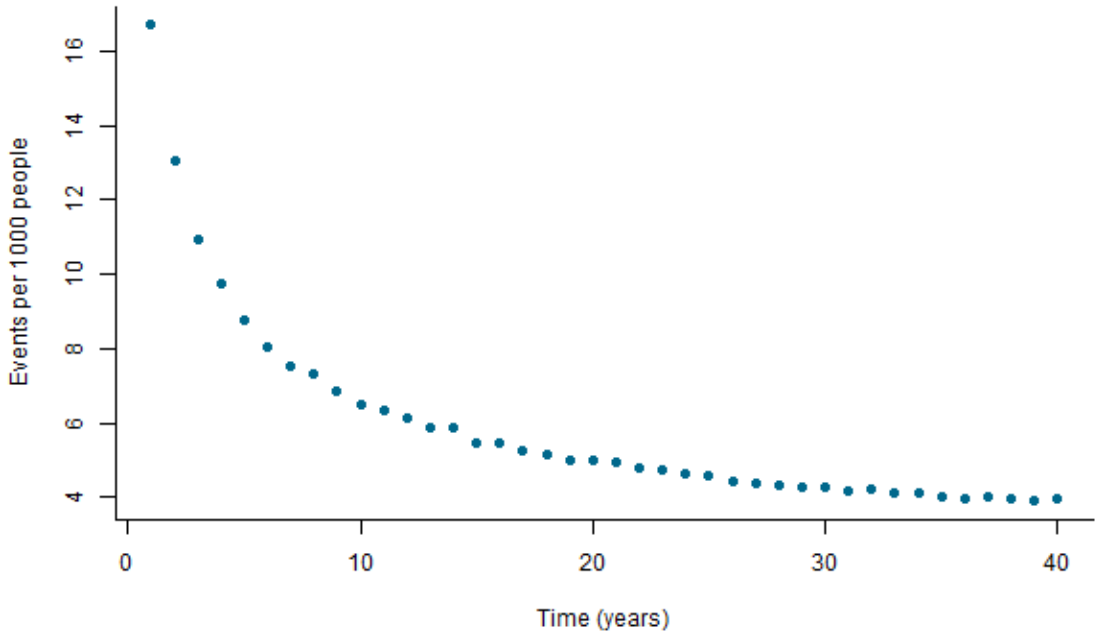


Figure 8.2: **Distribution of sibling illness exposure events across the life course.** Average number of x_6 variable exposure events per 1000 individuals over the course of the 40 year life span.

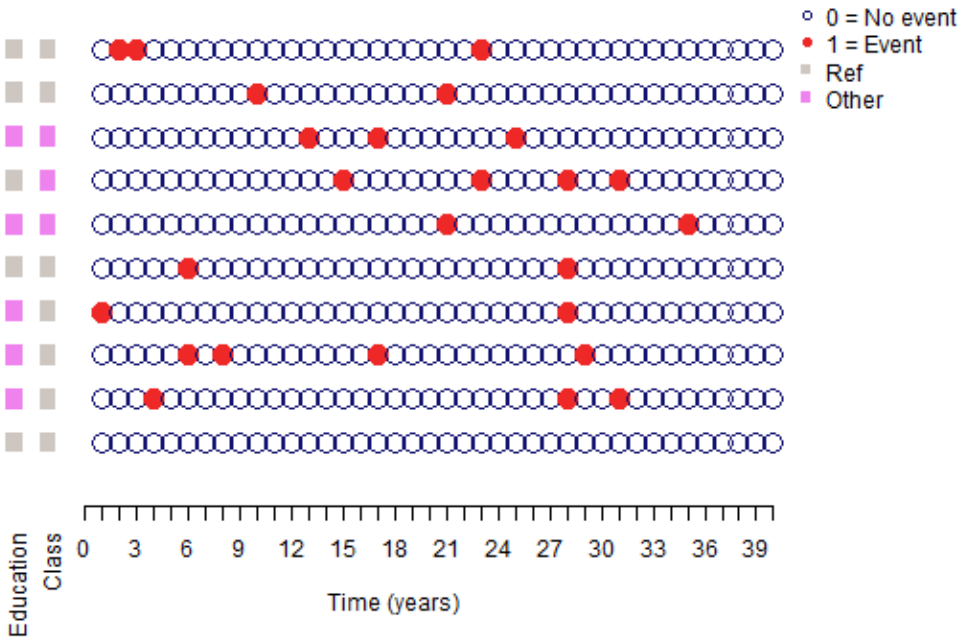


Figure 8.3: **Poverty exposure events within the population.** Value of variable x_7 for first ten individuals across the 40 year time frame. Circles show a year and those that are red show the occurrence of the exposure. The boxes on the left of the plot represent each individual's highest maternal education level and social class status with grey indicating the reference value and pink reflecting the alternative. The definitions of these variables alongside their reference categories are shown in Table 2.1.

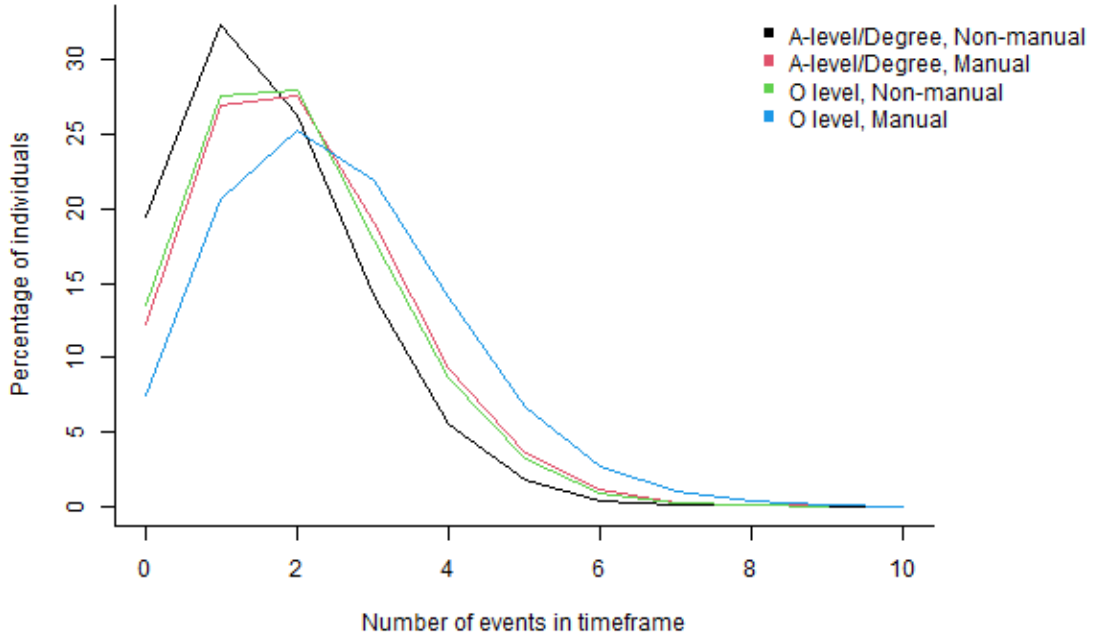


Figure 8.4: **Total number of instances of poverty according to education and social class.** The percentage of individuals who experience given numbers of incidents of poverty aggregated according to their maternal highest education attained and social class.

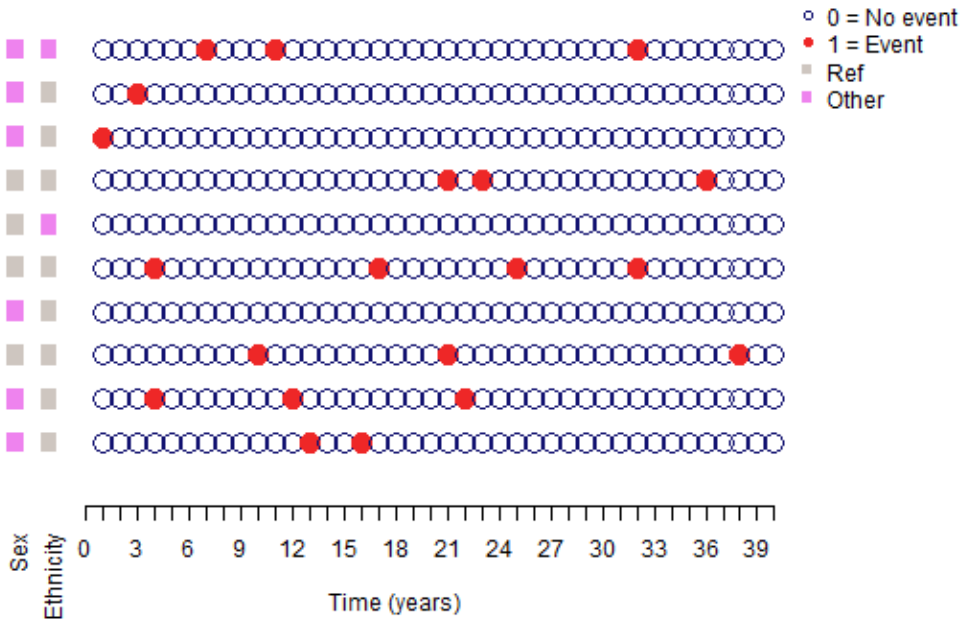


Figure 8.5: **Parental alcohol abuse exposure events within the population.** Value of variable x_8 for first ten individuals across the 40 year time frame. Circles show a year and those that are red show the occurrence of the exposure. The boxes on the left of the plot represent each individual's sex and ethnicity with grey indicating the reference value and pink reflecting the alternative. The definitions of these variables alongside their reference categories are shown in Table 2.1.

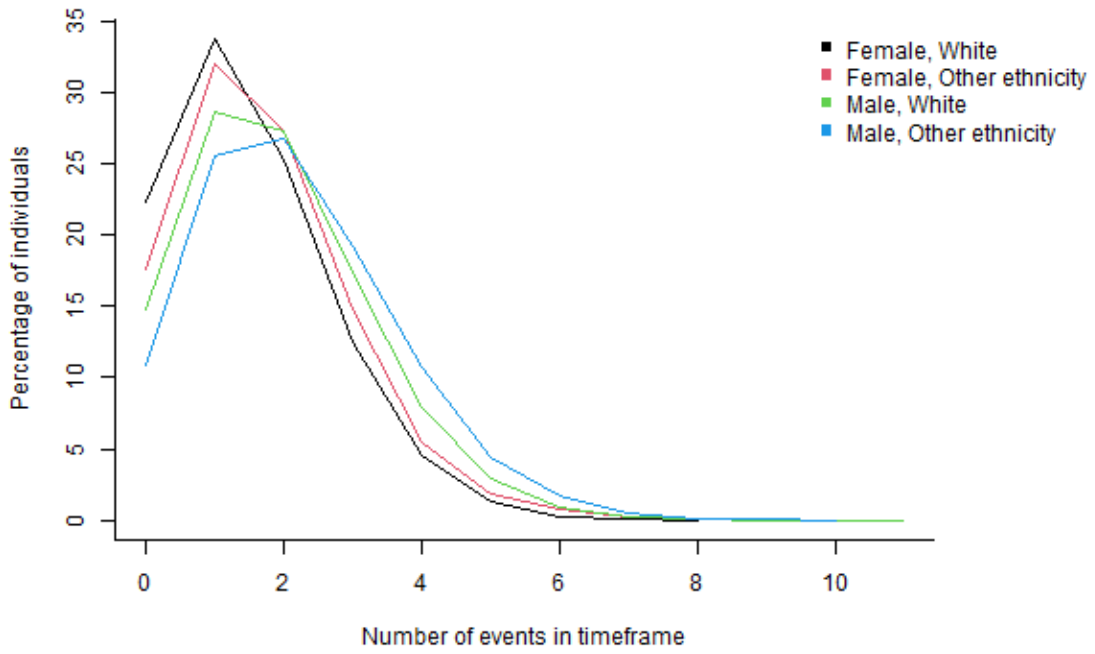


Figure 8.6: **Total number of instances of parental alcohol abuse according to their sex and ethnicity.** The percentage of individuals who experience given numbers of incidents of poverty aggregated according to their sex and ethnicity.

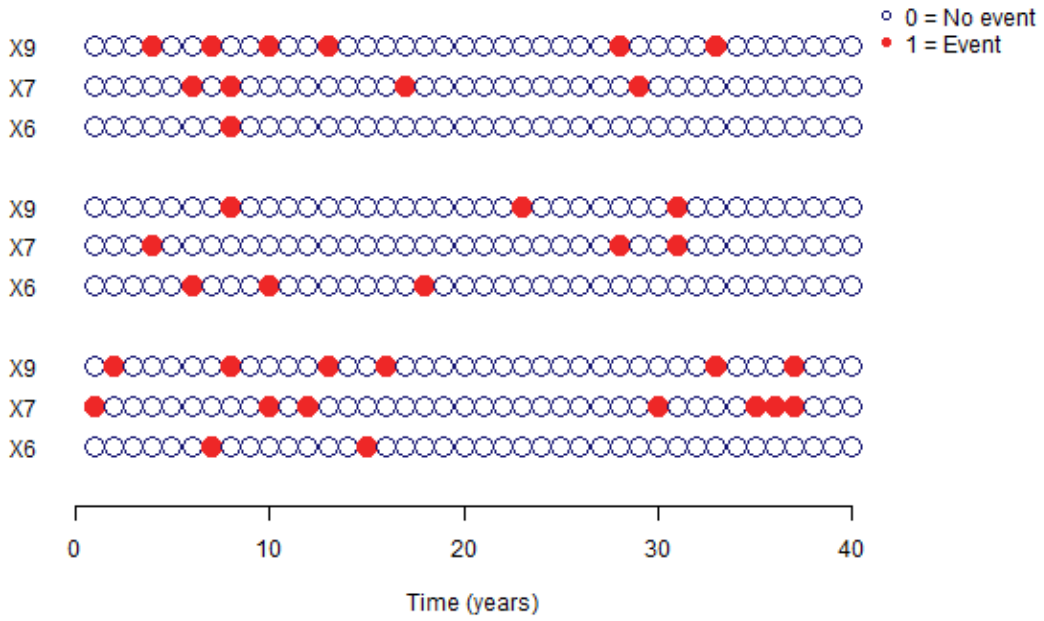


Figure 8.7: **Mild illness exposure events within the population.** Value of variable x_9 for first three individuals across the 40 year time frame. Circles show a year and those that are red show the occurrence of the exposure. The corresponding values of the variables x_6 and x_7 are shown.

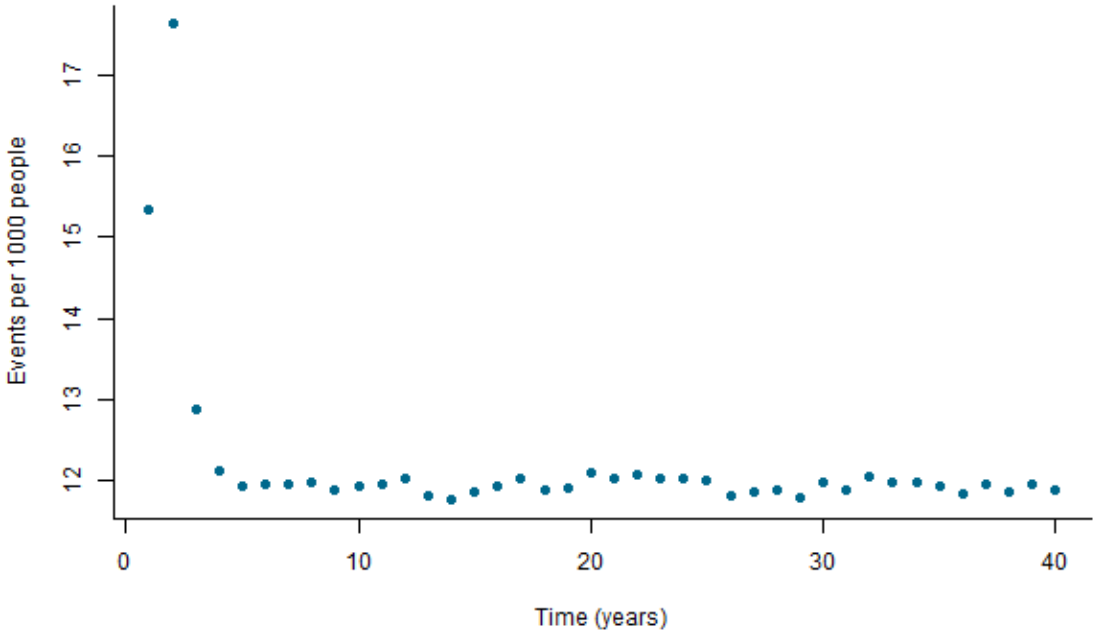


Figure 8.8: **Distribution of mild illness events across the life span.** The number of mild illness events experienced per 1000 people in the population over the course of the 40 year time frame.

8.2 GHE Causes

Table 8.1 shows the full list of GHE cause names alongside the corresponding ICD-10 codes that fall within each diagnosis. The table gives the broader groupings of infectious diseases, non-communicable diseases and injuries. There are also subcategories within each wider category demonstrating the range of diseases within each subheading.

Table 8.1: **ICD-10 codes associated with GHE categories.** The full table of GHE categories and corresponding ICD-10 codes. Taken directly from the World Health Organization Department of Data and Analytics (DNA) [2020]

GHE cause name	ICD-10 codes
I. Communicable, maternal, perinatal and nutritional conditions	A00-B99, D50-D53, D64.9, E00-E02, E40-E46, E50-E64, G00-G04, G14, H65-H66, J00-J22, N70-N73, O00-O99, P00-P96, U04
A. Infectious and parasitic diseases	A00-B99, G00-G04, G14, N70-N73, P37.3, P37.4
1. Tuberculosis	A15-A19, B90
2. STDs excluding HIV	A50-A64, N70-N73
a. Syphilis	A50-A53
b. Chlamydia	A55-A56
c. Gonorrhoea	A54
d. Trichomoniasis	A59
e. Genital herpes	A60
f. Other STDs	A57-A58, A63-A64, N70-N73
3. HIV/AIDS	B20-B24
a. HIV resulting in TB	B20.0
b. HIV resulting in other diseases	B20-B24 (minus B20.0)
4. Diarrhoeal diseases	A00, A01, A03, A04, A06-A09
5. Childhood-cluster diseases	A33-A37, B05
a. Whooping cough	A37
b. Diphtheria	A36
c. Measles	B05
d. Tetanus	A33-A35
6. Meningitis	A39, G00, G03
7. Encephalitis	A83-A86, B94.1, G04
8. Hepatitis	B15-B19 (minus B17.8)
a. Acute hepatitis A	B15
b. Acute hepatitis B	B16-B19 (minus B17.1, B17.2, B18.2, B18.8)
c. Acute hepatitis C	B17.1, B18.2
d. Acute hepatitis E	B17.2, B18.8
9. Parasitic and vector diseases	A71, A82, A90-A91, A95, B50-B57, B65, B67, B69, B73, B74.0-B74.2, P37.3-P37.4
a. Malaria	B50-B54, P37.3, P37.4
b. Trypanosomiasis	B56
c. Chagas disease	B57
d. Schistosomiasis	B65
e. Leishmaniasis	B55
f. Lymphatic filariasis	B74.0-B74.2

g. Onchocerciasis	B73
h. Cysticercosis	B69
i. Echinococcosis	B67
j. Dengue	A90-A91
k. Trachoma	A71
l. Yellow fever	A95
m. Rabies	A82
10. Intestinal nematode infections	B76-B81
a. Ascariasis	B77
b. Trichuriasis	B79
c. Hookworm disease	B76
d. Food-bourne trematodes	B78, B80, B81
11. Leprosy	A30
12. Other infectious diseases	A02, A05, A20-A28, A31, A32, A38, A40-A49, A65-A70, A74-A79, A80-A81, A87-A89, A92-A99, B00-B04, B06-B09, B17.8, B25-B49, B58-B60, B64, B66, B68, B70-B72, B74.3-B74.9, B75, B82-B89, B91-B99 (minus B94.1), G14
B. Respiratory infectious	H65-H66, J00-J22, P23, U04
1. Lower respiratory infections	J09-J22, P23, U04
2. Upper respiratory infections	J00-J06
3. Otitis media	H65-H66
C. Maternal conditions	O00-O99
D. Neonatal conditions	P00-P96 (minus P23, P37.3, P37.4)
1. Preterm birth complications	P05, P07, P22, P27-P28
2. Birth asphyxia and birth trauma	P03, P10-P15, P20-P21, P24-P26, P29
3. Neonatal sepsis and infections	P35-P39 (minus P37.3, P37.4)
4. Other neonatal conditions	P00-P02, P04, P08, P50-P96
E. Nutritional deficiencies	D50-D53, D64.9, E00-E02, E40-E46, E50-E64
1. Protein-energy malnutrition	E40-E46
2. Iodine deficiency	E00-E02
3. Vitamin A deficiency	E50
4. Iron-deficiency anaemia	D50, D64.9
5. Other nutritional deficiencies	D51-D53, E51-E64
II. Noncommunicable diseases	C00-C97, D00-D48, D55-D64 (minus D 64.9), D65-D89, E03-E07, E10-E34, E65-E88, F01-F99, G06-G98 (minus G14), H00-H61, H68-H93, I00-I99, J30-J98, K00-K92, L00-L98, M00-M99, N00-N64, N75-N98, Q00-Q99, X41-X42, X44, X45, R95
A. Malignant neoplasms	C00-C97
1. Mouth and oropharynx cancers	C00-C14
a. Lip and oral cavity	C00-C08
b. Nasopharynx	C11
c. Other pharynx	C09-C10, C12-C14
2. Oesophagus cancer	C15
3. Stomach cancer	C16

4. Colon and rectum cancers	C18-C21
5. Liver cancer	C22
6. Pancreas cancer	C25
7. Trachea, bronchus, lung cancers	C33-C34
8. Melanoma and other skin cancers	C43-C44
a. Malignant skin melanoma	C43
b. Non-melanoma skin cancer	C44
9. Breast cancer	C50
10. Cervix uteri cancer	C53
11. Corpus uteri cancer	C54
12. Ovary cancer	C56
13. Prostate cancer	C61
14. Testicular cancer	C62
15. Kidney, renal pelvis and ureter cancer	C64-C66
16. Bladder cancer	C67
17. Brain and nervous system cancers	C70-C72
18. Gallbladder and biliary tract cancer	C23-C24
19. Larynx cancer	C32
20. Thyroid cancer	C73
21. Mesothelioma	C45
22. Lymphomas, multiple myeloma	C81-C90, C96
a. Hodgkin lymphoma	C81
b. Non-Hodgkin lymphoma	C82-C86, C96
c. Multiple myeloma	C88, C90
23. Leukaemia	C91-C95
24. Other malignant neoplasms	C17, C26-C31, C37-C41, C46-C49, C51, C52, C57-C60, C63, C68, C69, C74-C75, C77-C79
B. Other neoplasms	D00-D48
C. Diabetes mellitus	E10-E14 (minus E10.2, E11.2, E12.2, E13.2, E14.2)
D. Endocrine, blood, immune disorders	D55-D64 (minus D64.9), D65-D89, E03-E07, E15-E34, E65-E88
1. Thalassaemias	D56
2. Sickle cell disorders and trait	D57
3. Other haemoglobinopathies and haemolytic anaemias	D55, D58-D59
4. Other endocrine, blood and immune disorders	D60-D64 (minus D64.9), D65-D89, E03-E07, E15-E34, E65-E88
E. Mental and substance use disorders	F04-F99, G72.1, Q86.0, X41-X42, X44, X45
1. Depressive disorders	F32-F33, F34.1
a. Major depressive disorder	F32-F33
b. Dysthymia	F34.1
2. Bipolar disorder	F30-F31
3. Schizophrenia	F20-F29
4. Alcohol use disorders	F10, G72.1, Q86.0, X45
5. Drug use disorders	F11-F16, F18-F19d , X41-X42, X44d
a. Opioid use disorders	F11, X42
b. Cocaine use disorders	F14

c. Amphetamine use disorders	F15
d. Cannabis use disorders	F12
e. Other drug use disorders	F13, F16, F18, X41
6. Anxiety disorders	F40-F44
7. Eating disorders	F50
8. Autism and Asperger syndrome	F84
9. Childhood behavioural disorders	F90-F92
a. Attention deficit/hyperactivity syndrome	F90
b. Conduct disorder	F91-F92
10. Idiopathic intellectual disability	F70-F79
11. Other mental and behavioural disorders	F04-F09, F17, F34-F39 (minus F34.1), F45-F48, F51-F69, F80-F83, F88-F89, F93-F99
F. Neurological conditions	F01-F03, G06-G98 (minus G14, G72.1)
1. Alzheimer disease and other dementias	F01-F03, G30-G31
2. Parkinson disease	G20-G21
3. Epilepsy	G40-G41
4. Multiple sclerosis	G35
5. Migraine	G43
6. Non-migraine headache	G44
7. Other neurological conditions	G06-G12, G23-G25, G36-G37, G45-G98 (minus G72.1)
G. Sense organ diseases	H00-H61, H68-H93
1. Glaucoma	H40
2. Cataracts	H25-H26
3. Uncorrected refractive errors	H49-H52
4. Macular degeneration	H35.3
5. Other vision loss	H30-H35 (minus H35.3), H53-H54
6. Other hearing loss	H90-H91
7. Other sense organ disorders	H00-H21, H27, H43-H47, H55-H61, H68-H83, H92-H93
H. Cardiovascular diseases	I00-I99
1. Rheumatic heart disease	I01-I09
2. Hypertensive heart disease	I11-I15
3. Ischaemic heart disease	I20-I25
4. Stroke	I60-I69
5. Cardiomyopathy, myocarditis, endocarditis	I30-I33, I38, I40, I42
6. Other circulatory diseases	I00, I26-I28, I34-I37, I44-I51, I70-I99
I. Respiratory diseases	J30-J98
1. Chronic obstructive pulmonary disease	J40-J44
2. Asthma	J45-J46
3. Other respiratory diseases	J30-J39, J47-J98
J. Digestive diseases	K20-K92
1. Peptic ulcer disease	K25-K27
2. Cirrhosis of the liver	K70, K74
3. Appendicitis	K35-K37
4. Gastritis and duodenitis	K29
5. Paralytic ileus and intestinal obstruction	K56

6. Inflammatory bowel disease	K50-K52, K58.0
7. Gallbladder and biliary diseases	K80-K83
8. Pancreatitis	K85-K86
9. Other digestive diseases	K20-K22, K28, K30-K31, K38, K40-K46, K55, K57, K58.9, K59-K66, K71-K73, K75-K76, K90-K92
K. Genitourinary diseases	E10.2-E10.29, E11.2-E11.29, E12.2, E13.2-E13.29, E14.2, N00-N64, N75-N76, N80-N98
1. Kidney diseases	N00-N19, E10.2,E11.2,E12.2,E13.2,E14.2
a. Acute glomerulonephritis	N00-N01
b. Chronic kidney disease due to diabetes	E10.2, E11.2, E12.2, E13.2, E14.2
c. Other chronic kidney disease	N02-N19
2. Benign prostatic hyperplasia	N40
3. Urolithiasis	N20-N23
4. Other urinary diseases	N25-N39, N41-N45, N47-N51
5. Infertility	N46, N97
6. Gynecological diseases	N60-N64, N75-N76, N80-N96, N98
L. Skin diseases	L00-L98
M. Musculoskeletal diseases	M00-M99
1. Rheumatoid arthritis	M05-M06
2. Osteoarthritis	M15-M19
3. Gout	M10
4. Back and neck pain	M45-M48, M50-M54
5. Other musculoskeletal disorders	M00, M02, M08, M11-M13, M20-M43, M60-M99
N. Congenital anomalies	Q00-Q99 (minus Q86.0)
1. Neural tube defects	Q00, Q05
2. Cleft lip and cleft palate	Q35-Q37
3. Down syndrome	Q90
4. Congenital heart anomalies	Q20-Q28
5. Other chromosomal anomalies	Q91-Q99
6. Other congenital anomalies	Q01-Q04, Q06-Q18, Q30-Q34, Q38-Q89 (excluding Q86.0)
O. Oral conditions	K00-K14
1. Dental caries	K02
2. Periodontal disease	K05
3. Edentulism	not available
4. Other oral disorders	K00, K01, K03, K04, K06-K14
P. Sudden infant death syndrome	R95
III. Injuries	V01-Y89 (minus X41-X42, X44, X45)
A. Unintentional injuries	V01-X40, X43, X46-59, Y40-Y86, Y88, Y89
1. Road injury	V01-V04, V06, V09-V80, V87, V89, V99
2. Poisonings	X40, X43, X46-X48, X49
3. Falls	W00-W19
4. Fire, heat and hot substances	X00-X19
5. Drowning	W65-W74
6. Exposure to mechanical forces	W20-W38, W40-W43, W45, W46, W49-W52, W75, W76

7. Natural disasters	X33-X39
8. Other unintentional injuries	Rest of V, W39, W44, W53-W64, W77-W99, X20-X32, X50-X59, Y40-Y86, Y88, Y89
B. Intentional injuries	X60-Y09, Y35-Y36, Y870, Y871
1. Self-harm	X60-X84, Y870
2. Interpersonal violence	X85-Y09, Y871
3. Collective violence and legal intervention	Y35-Y36

8.3 The Signature Method Window Transformations

For those interested, below are the formulas related to two types of window transformation. The formulations are taken directly from Morrill et al. Morrill et al. [2021].

1. **Sliding window** For length l and step s, with $Z_{i:j} = (Z_{t_i}, \ldots, Z_{t_j})$

$$W(Z) = (Z_{1:l}, Z_{s+1:s+l}, Z_{2s+1:2s+l}, \ldots) \tag{8.1}$$

2. **Expanding window** With initial length l and step s,

$$W(Z) = (Z_{1:l}, Z_{1:s+l}, Z_{1:2s+l}, \ldots) \tag{8.2}$$

8.4 SHAP Values Without Scale Limit

The average SHAP values across populations of 2000 individuals are shown in Figures 8.9, 8.10 and 8.11 for the RTIs, diabetes and neurological conditions outcomes, respectively. Unlike the plots shown in Chapter 5, there is no cap on the scale and therefore it is possible to see the relative scale of influence of previous diagnoses of the outcome compared to the other input features. It is possible to see that the ICD-10 Chapter 19 and year of birth variables are still comparatively influential for the RTI outcome. This contrasts heavily to diabetes where diabetes hospitalisations within the past year have the greatest impact on predictions. Finally, neurological conditions are greatly impacted by previous hospitalisations with neurological conditions and somewhat impacted by year of birth.



Figure 8.9: **SHAP values for RTI outcome without scale limit.** The SHAP values for the RTI outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association.

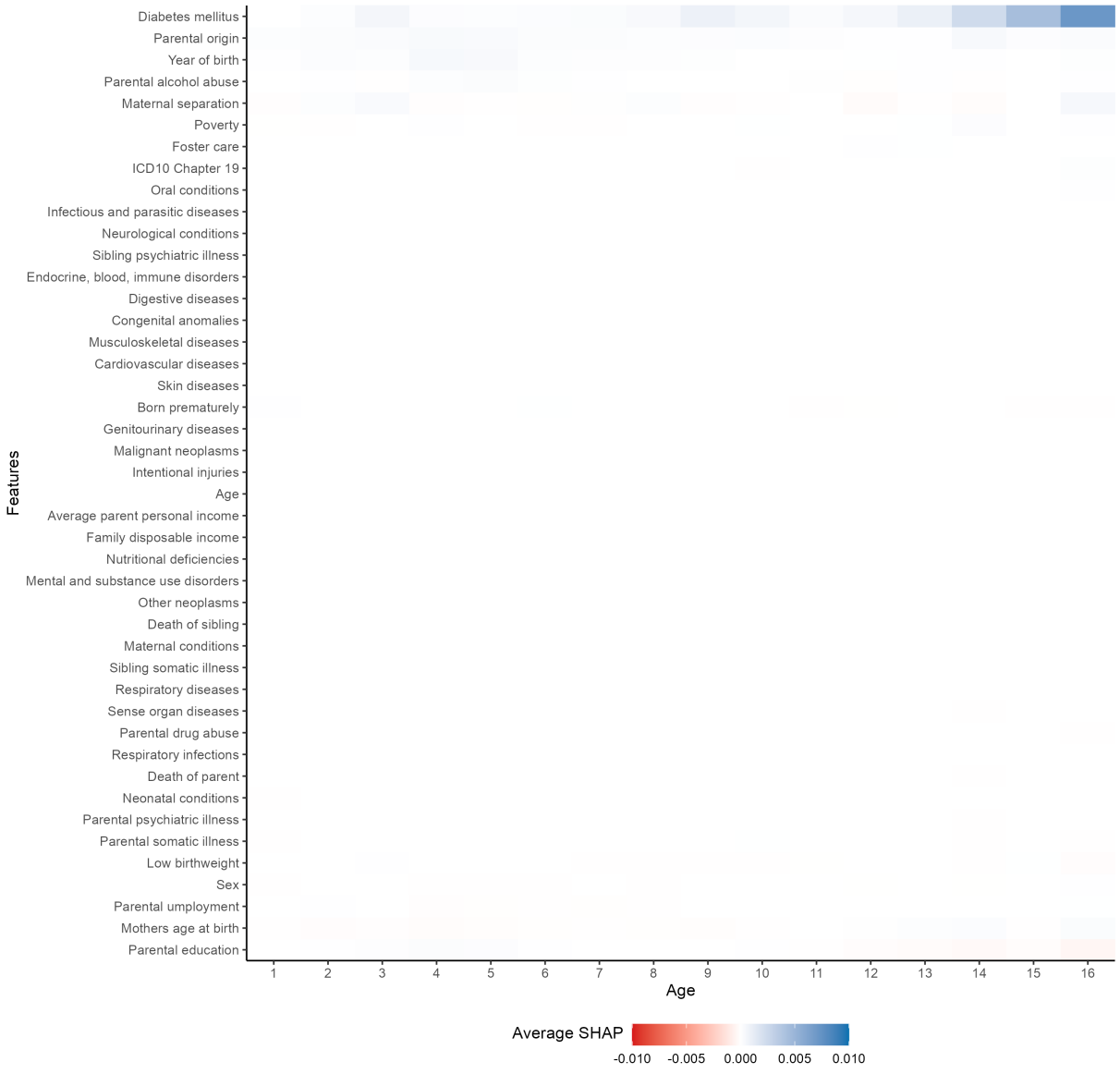


Figure 8.10: **SHAP values for diabetes outcome without scale limit.** The SHAP values for the diabetes outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association.

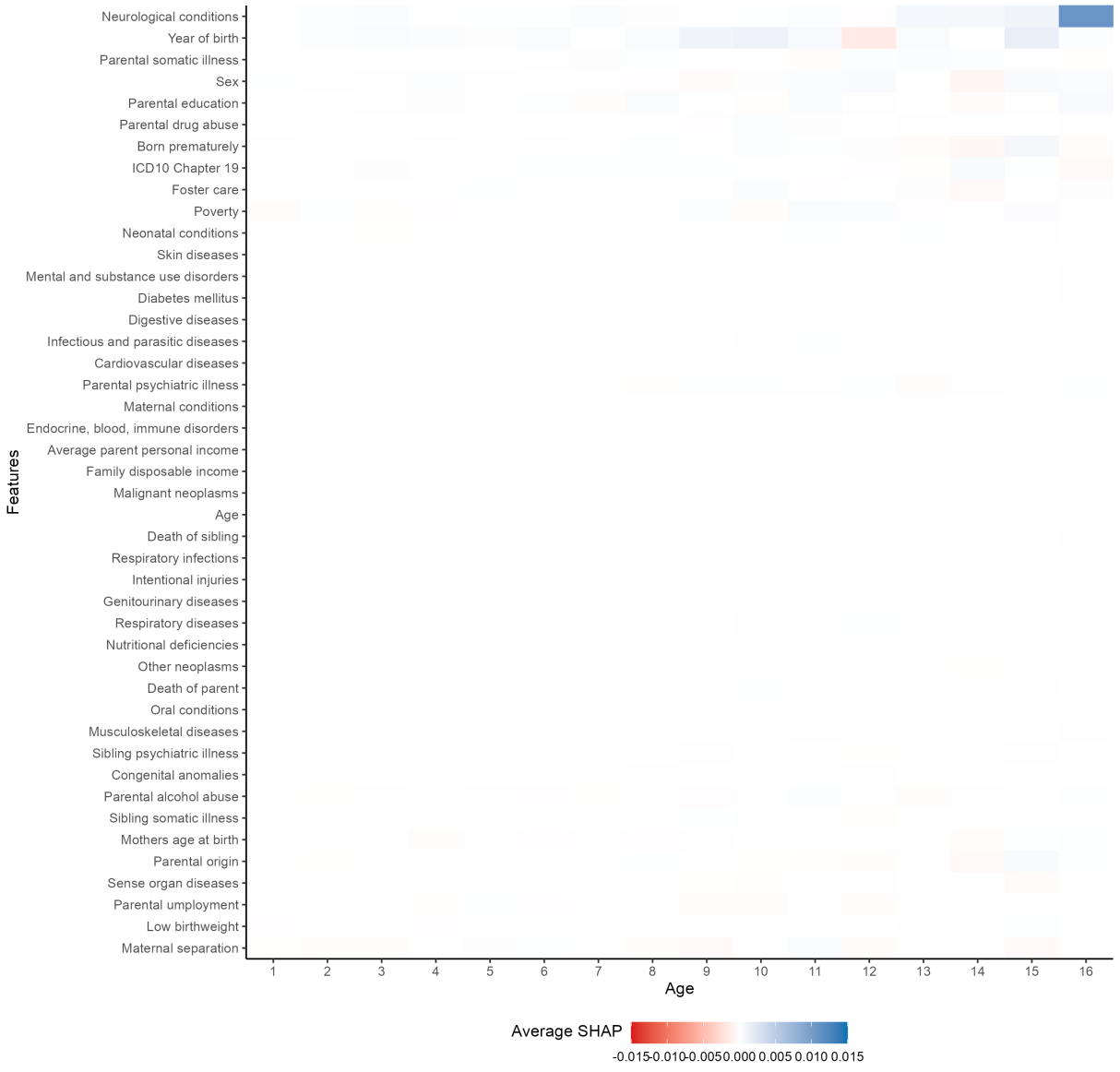


Figure 8.11: **SHAP values for neurological conditions outcome without scale limit.** The SHAP values for the neurological conditions outcome for each feature (y axis) at each time point (x axis) averaged over 2000 individuals. The colour scale displays the direction and magnitude of any association.