UNIVERSITY OF CAMBRIDGE

Clinical Presence: Impact on Predictive Modelling and Algorithmic Fairness

Vincent Jeanselme

Under the supervision of
Dr Jessica Barrett and Dr Brian Tom

MRC Biostatistics Unit
School of Clinical Medicine

Hughes Hall

This thesis is submitted on April, 29th 2024
for the degree of *Doctor of Philosophy*

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Vincent Jeanselme
April 2024

Abstract

Hospitals routinely collect large amounts of data that may provide insights beyond what experimental studies can offer due to their practical, ethical, or budgetary limitations. However, these *observational* data present critical challenges: (i) the quantity and diversity of modalities make traditional statistical tools less amenable to model these data, and (ii) multiple factors influence what and when data are collected.

If not carefully considered, the observational process, referred to as *clinical presence* in this thesis, lessens the potential of observational data for predictive modelling. Particularly, clinical presence not only reflects medical expertise and patient deterioration but also socio-medical disparities deeply ingrained in healthcare practices. Failure to disentangle signal from bias in clinical presence risks perpetuating and amplifying socio-medical disparities.

In this thesis, we explore the understudied impact of clinical presence on predictive modelling and associated algorithmic fairness properties. We propose to use machine learning to tackle scalability and modelling flexibility while accounting for statistical biases and potential socio-medical disparities associated with clinical presence. By connecting methodologies from machine learning, biostatistics, and algorithmic fairness, we aim to provide insights into clinical presence and its impact on predictive models and algorithmic fairness.

In each chapter, we explore a different aspect of clinical presence and its impact on predictive modelling and fairness. First, we examine the challenges of missingness and how practitioners' use of imputation influences algorithmic fairness. Then, we aim to discover subgroups of patients under-served by current medical practices under non-random treatment assignment. Beyond observed covariates and treatment assignment, we show that the process associated with observed outcomes, particularly if preventing the observation of the outcome of interest, may impact groups with distinct risk profiles differently. Finally, we investigate the impact of the irregularities in longitudinal medical data and their impact on predictive models and their transportability.

Through this research, we aim to develop more equitable and accurate predictive models by addressing the complexities of clinical presence. By identifying and accounting for medical disparities deeply ingrained in medical history and, consequently, practices and data, we can ensure that the benefits of novel medical tools are equitably distributed to all.

Acknowledgements

There are simply too many people for whom I wish to express my appreciation. So, I want to extend my heartfelt thanks to each of you. I am incredibly grateful to have met so many amazing, bright and affectionate people throughout my journey from France to England via the United States.

To my supervisors, Jessica Barrett and Brian Tom, thank you for the opportunity to work on this exciting and challenging problem. Your guidance, insights and feedback have been invaluable, and I am grateful for your patience while I was discovering the biostatistics field.

To Artur Dubrawski and James Crowley, without whom this research journey would never have been possible. Thank you for believing in my early work, guiding it, and transmitting your love for teaching and research.

To all my collaborators, thank you for your invaluable input, feedback, and encouragement, which helped me improve the quality of my work and grow as a researcher. I appreciate your devoted time and effort in advising and commenting on my work.

To Maria De-Arteaga, thank you for your constant support, kind words of encouragement, and advice from Carnegie Mellon to Cambridge. I can not express how grateful I am for everything.

To the wonderful friends I have met on this journey and made it unforgettable, I am incredibly grateful and feel lucky to have met you. Thank you for every moment we have shared.

À ma famille, sans vous, rien n'aurait été possible. Merci pour vos appels, mots d'encouragement et l'exemple que vous m'avez donné chaque jour à travers votre travail, votre persévérance et vos cheminements qui ont été, et sont, une constante source d'inspiration.

我的盛盛, 尽管距离遥远，但你一直是我坚定的支持.

Contents

Acronyms

AUC	Area Under the Curve.
CI	Confidence Interval.
CIF	Cumulative Incidence Function.
CVD	Cardiovascular Disease.
DAG	Directed Acyclic Graph.
EM	Expectation Maximisation.
FNR	False Negative Rate.
GRU	Gated Recurrent Unit.
ICU	Intensive Care Unit.
ITE	Individualised Treatment Effect.
LSTM	Long Short Term Memory.
MAR	Missing At Random.
MCAR	Missing Completely At Random.
ML	Machine Learning.
MNAR	Missing Not At Random.
ODE	Ordinary Differential Equation.
RCT	Randomised Controlled Trial.
RNN	Recurrent Neural Network.

Chapter 1

Introduction

When a patient falls ill, they decide whether to visit their general practitioner or wait. The factors influencing this choice extend beyond the patient's condition alone. Elements such as the patient's medical coverage, proximity to the nearest practitioner, and work flexibility also play a role in this decision. Then, when interfacing with a medical professional, the medical expertise and patient's medical history impact which laboratory tests are requested, which diagnoses are made, and which treatments are recommended by the practitioner. These are only a few examples of the complex interaction between patients and the healthcare system, which inform data collection.

While ignorable in controlled settings where potential confounders are independent of the observational process, these factors are a crucial challenge in modelling data routinely collected in medical practice. These *observational data* do not reflect patients' health alone but also reflect interactions between the patient and the healthcare system. These observational patterns introduce statistical challenges, often overlooked in developing predictive models. Ignoring the nuances associated with the observational process results in a critical gap between the development of predictive models under simplifying assumptions and their application in medical practice [286]. This thesis contributes to bridging this gap by delving into this observational process, which we term *clinical presence*, and examines how it impacts predictive modelling and potentially biases medical models for groups who differ in their interaction with the healthcare system.

1.1 Observational data's promises

Models predictive of medical deterioration [366], the risk for a condition [197] or recommending treatments [365] have been proposed to assist medical practitioners in their decisions. These models are developed based on either of two types of medical data: experimental and observational. As discussed in the following section, experimental settings provide precise answers to a restricted set of questions. In contrast, observational settings offer potentially biased answers

to an unconstrained set of questions. As cost remains a barrier to address the limitations of experimentation, our work explores how to address the limitations of the large amounts of routinely collected data for improved predictive modelling.

Experimental. An experiment is executed in a controlled environment in which researchers investigate the effects of a particular intervention, treatment, or condition. Researchers carefully design this environment to mitigate biases from potential confounders [121] that could affect the study's outcome, thereby allowing researchers to study the causal link between intervention and outcome. Design considerations encompass randomisation, blinding and standardisation to avoid potential confounders. Randomisation ensures that a patient has an equal chance to receive the intervention, uninfluenced by medical expertise or the patient's decisions. When possible, one further aims to blind researchers, practitioners and patients with respect to who actually receives the intervention, to mitigate the impact of expectation on outcome. Finally, standardisation of intervention and measurement scheduling avoids differences in procedure and collected data. When meeting these conditions, researchers can rigorously evaluate the causal impact of an intervention on an outcome.

Experimentation is expensive however, often reducing the studied cohort and its representativeness of the population and setting in which the intervention will later be administrated [126]. This misrepresentation results in a skewed understanding of the causal impact of intervention on outcome, which may not generalise beyond a specific population and setting. These constraints render experiments less relevant to the development of predictive models for real-world settings. Further, experimental design is not always possible or ethical. For instance, experiments have long excluded pregnant women due to the risk to the fetus [209], and interventions with known risk are unethical to test, as the exploration of the causal relation between smoking and lung cancer would require.

Observational. The analysis of observational data, marked by the absence of controlled interventions, is valuable in medical research for studying associations between covariates, identifying disease risk factors, and generating hypotheses for further experiments. One can distinguish two broad types of observational data: cohort studies and routinely collected data. Cohort studies consist of following a given population over time. As these studies follow a precise design protocol, they suffer from the same limitations as experimental data: cost and population size. Routinely collected data differs as they result from standard clinical practice. While routinely collected data do not meet the design considerations of the previous studies, they provide essential insights into real-world practice on a diverse population. Further, these data have a limited additional cost, rendering large and long-term cohort follow-up possible. These two properties have allowed researchers to study long-term effects [29], rare conditions [95], and evaluate the real-world effectiveness of existing policies and guidelines [21].

However, because routinely collected data are not collected primarily for research purposes, these data lack the controlled design of experimental settings, leading to potential confounders that can affect the validity of findings.

1.2 Clinical presence's challenges

None of the described design considerations of the experimental setting hold in observational studies: (i) interventions are not randomised, (ii) patients and practitioners are aware of interventions and testing, and (iii) not all procedures and data collection are standardised across experts and hospitals. These different factors impact when and what data are collected. This thesis describes the data collection patterns as *clinical presence*. In other words, clinical presence encompasses the factors influencing why and when one observes data.

Definition 1.1 (Clinical presence)**.** Clinical presence is the decision process that results in medical data collection.

In this thesis, we argue that a deeper understanding of clinical presence is necessary for developing predictive modelling, as learning from data means analysing patterns of observations. Critically, we distinguish three different types of patterns in clinical presence:

1. *Signal:* Patterns may reflect medical expertise and the patient's condition, informing the outcome of interest. For instance, Agniel, Kohane, and Weber [4] show how using *missingness* in measured covariates improves 3-year-survival predictive performance, as the choice of measurements is associated with the patient's condition.

2. *Bias:* Irrelevant factors result in shared and flawed patterns of observations among practitioners. For instance, Bishara et al. [36] show that the day of discharge influences antibiotic recommendation even when controlling for the condition.

3. *Noise:* Kahneman, Sibony, and Sunstein [156] define noise as "unwanted variability", differing from bias in its randomness. For instance, medical practitioners often disagree among but also within themselves, as studied in [80]. In this study, the authors requested multiple readings of the same coronary angiogram from experts without their knowledge. Medical practitioners were inconsistent with their initial assessment across the different readings.

Identifying these patterns and their origins is essential to distinguish signal from noise and address existing biases. Despite the critical importance of these patterns, (i) the medical literature often disregards this information [108], (ii) the Machine Learning (ML) literature assumes informativeness of clinical presence, with multiple works using missingness as a predictor [116, 190], and (iii) the statistical field corrects for the process to estimate the

unbiased association between exposure and outcome [65, 245, 251, 308]. Integrating signals while mitigating biases is the critical challenge in handling clinical presence and developing tools to reduce the noise in these decisions. One of the sources of bias we focus on in this thesis is socio-medical disparities.

1.3 Impact on algorithmic fairness

Historically, medical research and practice have been biased against marginalised groups [161, 327], resulting in a lower quality of care and trust in the medical system for these groups. For instance, risk scores used in clinical practice were often developed on populations with low ethnic diversity, detrimentally impacting risk management for previously ignored groups. Kartoun et al. [161] identify miscalibration for Black patients in cardiovascular risk estimates and suggest their under-representation in risk score development as a potential explanation.

Clinical presence may reflect these socio-medical disparities. Using advanced modelling on observational data comes with the risk of perpetuating and amplifying these disparities [61] with life-threatening implications. For instance, the medical knowledge of heart disease has long focused on men [330] with guidelines aimed at this population [309]. Modelling of previous medical recommendations would result in a biased association between sex and testing and risk automating the already limited testing for women.

Algorithmic fairness, which aims to identify, measure and mitigate the risk of inequitable real-world deployment, has become an important concern in ML for healthcare [53]. By carefully designing algorithms, we can reduce the socio-medical disparities present in observed data, avoid automating these biases, and better assist practitioners in their decisions.

Definition 1.2 (Algorithmic Fairness)**.** Algorithmic fairness involves identifying and mitigating biases to ensure equitable outcomes for all individuals, irrespective of socio-economic characteristics.

This thesis focuses on the observational process and its critical impact on algorithmic fairness in healthcare. We explore how existing disparities may lead to systematic differences in interaction between patients and the healthcare system and show how ignoring these differences in model development can reinforce socio-medical biases.

1.4 Problem statement

Healthcare increasingly uses ML to inform patient care by analysing patterns in observational data. Risk prediction, prioritisation, and treatment response estimation have transformative potential to assist medical practitioners, driving down costs and elevating the quality of care. However, there remains a gap between developing models on observational data and informing

clinical practice. This thesis investigates critical dimensions of clinical presence and its impact on predictive modelling and algorithmic fairness by asking:

How should we handle clinical presence for fair predictive modelling?

Our work at the intersection between medical applications and ML aims to highlight the importance of the observational process in predictive modelling and offer novel tools to tackle its associated challenges. By studying the ever-evolving observational process, we seek to improve the integration of predictive models into clinical practice to improve healthcare delivery for all.

1.5 Plan

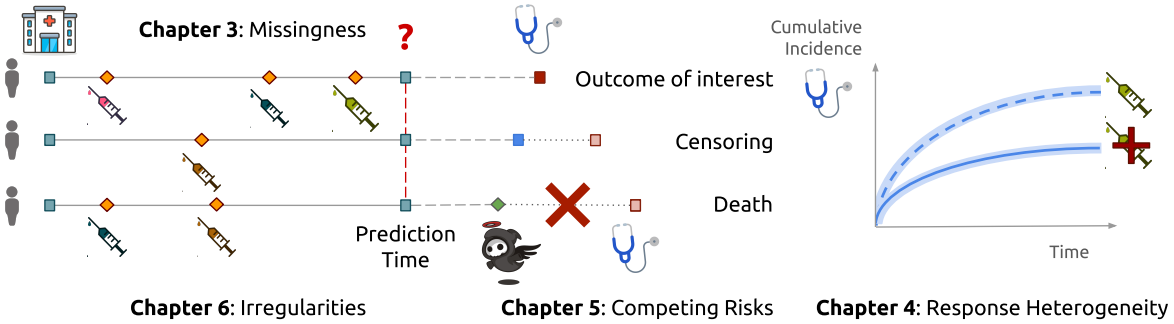


Figure 1.1: Visual summary of the challenges associated with clinical presence studied in this thesis.

This thesis is structured around four challenges associated with clinical presence: (i) missingness, (ii) treatment assignment bias, (iii) competing risk process, and (iv) irregular sampling. In Chapter 3, we analyse the impact of imputation on algorithmic fairness under group-specific missingness patterns. In Chapter 4, we study the heterogeneity in survival outcomes and treatment responses under observational treatment regimes to identify subgroups under-served by current treatment practice. In Chapter 5, we explore the impact of the outcome process on algorithmic fairness and demonstrate that the common practice of ignoring outcomes that preclude the one of interest, known as competing risks, biases modelling and increases the difference in performance between patients with distinct risk profiles. Finally, in Chapter 6, we propose a new methodology to model irregularities in the longitudinal observational process for improved transportability.

Each chapter is an independent contribution discussing a challenge associated with clinical presence and proposing a potential remedy for improved predictive modelling. As neural networks, survival modelling and algorithmic fairness are central concepts to this thesis, Chapter 2 first reviews the methodological foundations for our proposed work.

Chapter 2

Background

This chapter reviews the core methodological concepts associated with neural networks, their application to survival analysis, their associated evaluation procedure and the problem of algorithmic fairness.

Why neural networks? Our motivation for using neural networks is their capacity to uncover associations between covariates and outcomes from data when these relationships are unknown or difficult to specify in a parametric form. By carefully designing and training neural networks, researchers can extract insights from large amounts of data, addressing novel medical questions that traditional statistical models may not be able to answer. This property is particularly relevant for unstructured modalities, such as images and text, often unamenable to conventional approaches. Further, these models offer the flexibility and scalability to handle the ever-growing medical data and associated modalities, thereby presenting an opportunity for improved medical predictions.

2.1 Neural networks

This section introduces neural networks and some of their variations used throughout this thesis. For an exhaustive description of neural networks and their evolution, please refer to [37, 110].

2.1.1 Neuron

Neural networks aim to mimic biological brains composed of simple processing units — neurons — that, together, form complex systems. An artificial neuron is to a neural network as a biological neuron is to the brain. A neuron operates by receiving inputs that may trigger an output response. Formally, given a multidimensional *input* $x \in \mathbb{R}^m$ with $m \in \mathbb{N}$ its dimension, a neuron consists of a linear combination of its inputs — defined by a *weight matrix*

$W \in \mathbb{R}^m$ and a bias[1] $b \in \mathbb{R}$ that weigh the neuron's input connections — then passed through a non-linear *activation function* ϕ responsible for quantifying the neuron's response into an *output* response. The neural output o can be expressed as:

$$o(x) = \phi(x^T \cdot W + b)$$

The activation function determines whether an input triggers the associated neuron. Mathematically, this component ensures neural networks' non-linearity, and, consequently, their modelling flexibility. In its simplest form, an activation function can be a threshold function, such as the Heaviside [3] function:

$$\text{Heaviside}(x) = \mathbb{1}_{x>0}$$

with $\mathbb{1}$ representing the indicator function. In this context, the neuron responds to an input stimulus if the stimulus intensity is above zero. While simple, this function is unsuitable for common training procedures due to its null derivative at all points (except at the threshold where it is undefined).

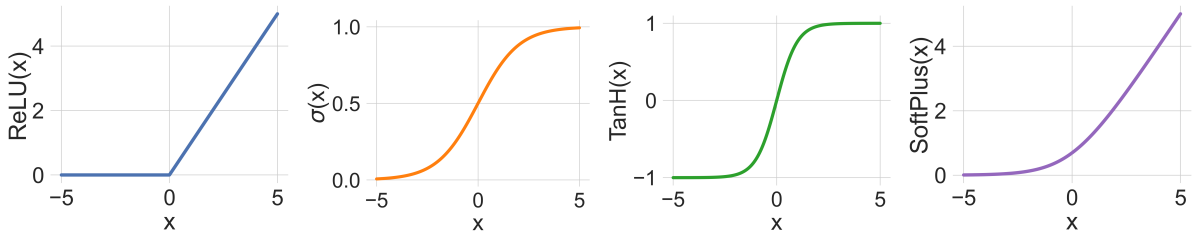
In our work, we consider four activation functions commonly used in the ML literature [239] that smoothly quantify how likely a neuron is to respond to a stimulus, namely the Rectified Linear Unit (ReLU) [228], the sigmoid (σ), the hyperbolic tangent (TanH) and the SoftPlus [85] functions defined as follows:

$$\text{ReLU}(x) = x \cdot \mathbb{1}_{x>0}, \quad \sigma(x) = \frac{1}{1+e^{-x}}, \quad \text{TanH}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \text{SoftPlus}(x) = \ln(1+e^x)$$

Figure 2.1 illustrates these different functions. The development of activation functions is often guided by empirical performance or to enforce desired properties upon the model. Note how the sigmoid and the hyperbolic tangent functions aim to smoothly approximate the Heaviside function. The motivation behind SoftPlus is to unconstrain the output values, and to increase the range of non-zero gradient [313] — critical for avoiding the problem of vanishing gradient in training. Finally, ReLU reduces the numerical cost and avoids the potential overflow of exponential computation, and by doing so, shows improved empirical convergence.

There exist activation functions that consider the outputs of multiple neurons simultaneously. Critically, these functions quantify a neuron's response relative to other neurons' activation. For instance, the SoftMax function [110] ensures the positivity and the summation to 1 of its outputs, through the transformation:

[1]While this denomination is common in the ML literature, this term corresponds to the statistical concept of *intercept*. To avoid confusion, we refer to both weights and bias as a neuron's *parameters* throughout this thesis.



(a) ReLU (b) Sigmoid (c) Hyperbolic tangent (d) SoftPlus

Figure 2.1: Four activation functions used in this thesis.

$$\text{SoftMax}(o_i) = \frac{e^{o_i}}{\sum_{j=1}^{n} e^{o_j}}$$

with o_i the output of neuron i out of the n considered. This function outputs the activation probability distribution over the set of input neurons.

2.1.2 Neurons' arrangement

Individually, neurons resemble generalised linear regressions [229] with the activation corresponding to the inverse of the link function. Interconnected, neurons form a network with complex behaviours. This section describes two types of neural architectures tackling static and temporal data: Multi-Layer Perceptron and Recurrent Neural Networks (RNN).

2.1.2.1 Multi-layer perceptron

A simple arrangement of independent neurons with the same inputs is known as a *layer*. Accumulating multiple layers results in a multi-layer perceptron. Inner layers l are referred to as *hidden* if their outputs serve as inputs to another layer of neurons.

For prediction, the input data is processed by each neuron of the first layer, forming outputs that are then used by the following layer as inputs. This process is iterated until reaching the final *output layer* that returns the predicted outcomes. This procedure from the inputs to the output layer is known as the *forward propagation*. For instance, consider a two-layer network as presented in Figure 2.2: a hidden layer takes the data as input, transforms it, and passes it to the output layer. This last layer aims to generate the target labels.

A central property of this simple arrangement is its ability to approximate *any* continuous functions as described by the following universal approximation theorem originally demonstrated in [134].

Theorem 2.1 (Neural networks are universal approximators [134])**.** *Consider* Γ*, the space of continuous functions from the input space* $\mathcal{X} \subseteq \mathbb{R}^m$ *to the output space* $\mathbb{R}$*,* $\forall g \in \Gamma, \forall \epsilon \in \mathbb{R}^+$*,*

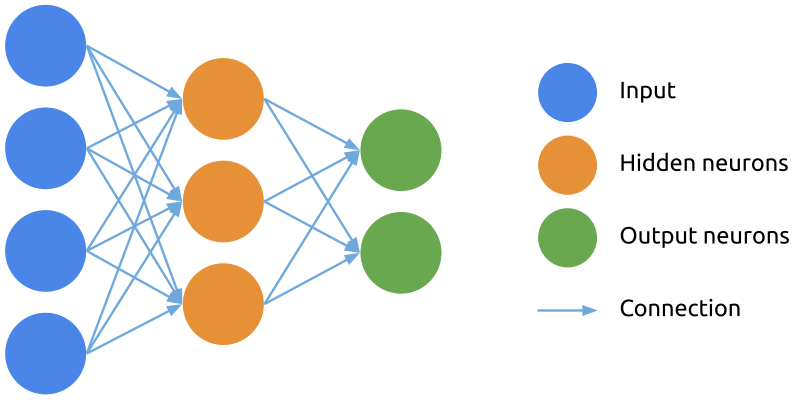


Figure 2.2: Fully connected neural network with one hidden layer.

there exists a one hidden layer network N *that can approximate* g *with error* ϵ*, i.e.*

$$\forall x, |g(x) - N(x)| < \epsilon$$

This theorem is the key motivation for using neural networks as it guarantees that there exists a neural network that represents any arbitrary continuous function. Nonetheless, the number of neurons necessary to satisfy this theorem with a one-hidden-layer network could be intractable and slow to converge, hence the use of multiple layers and alternative architectures.

2.1.2.2 Recurrent neural network

When considering repeated measurements, RNN take advantage of the sequential nature of the data. At each new time point, the model is applied to the new observation, while also considering past information through a *memory* state. Formally, a RNN N is applied on the observation x_t observed at time t using the memory state $h_{t-1} \in \mathbb{R}^d$ obtained at time $t-1$ with d the embedding dimension. This defines the recurrence:

$$h_t = N(x_t, h_{t-1})$$

An output layer can then be employed to predict the outcome $\hat{y}$ from any of the memory states, as shown in Figure 2.3.

While simple, this iterative model presents a limited capacity to capture long-term trends [130]. This phenomenon stems from the training procedure that updates the model's parameters given the gradient of the loss. The gradient associated with earlier events tends to fade at each new time point. This vanishing gradient favours recent events in the parameters' updates. Alternative architectures tackle this particular issue by adding connections to maintain the gradient's strength over time. Two such architectures used in this thesis are the Long Short Term Memory (LSTM [130]) and the Gated Recursive Unit (GRU [57]) architectures.

These recursive models, presented in Figure 2.3, differ from the simple RNN in their updating mechanisms of the memory state via multiple combinations of neurons and activation functions that balance the influence of the previous memory state and the new observation at each new time point[2]. These learnt mechanisms choose which information should be kept from a new observation and from the long-term memory. Specifically, the LSTM architecture introduces a new memory state C, which represents this long-term memory. GRU presents a similar idea in a simplified form by using a single memory state balancing both long-term and short-term information.

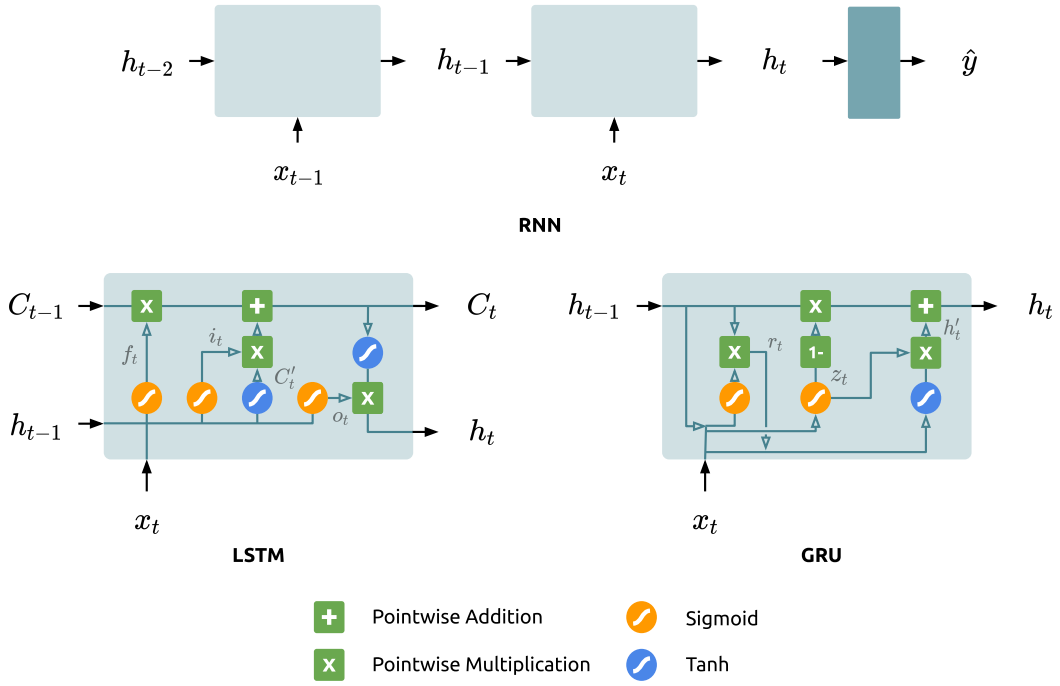


Figure 2.3: Architectures of Recurrent Neural Networks used in this thesis.

Mathematically, the LSTM can be described as the following set of equations:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1}) && \text{(Forget Gate)} \\
i_t &= \sigma(W_i x_t + U_i h_{t-1}) && \text{(Input Gate)} \\
o_t &= \sigma(W_o x_t + U_o h_{t-1}) && \text{(Output Gate)} \\
C'_t &= \text{TanH}(W_c x_t + U_c h_{t-1}) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot C'_t && \text{(Cell State)} \\
h_t &= o_t \cdot \text{TanH}(C_t)
\end{aligned}
$$

where W_g and U_g denote the set of parameters associated to the gate g. Similarly, the GRU is

[2]We invite the reader to refer to the original works [57, 130] for detailed descriptions of these mechanisms.

defined as:

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1}) && \text{(Update Gate)} \\
r_t &= \sigma(W_r x_t + U_r h_{t-1}) && \text{(Reset Gate)} \\
h'_t &= \mathsf{TanH}(W x_t + r_t \cdot U h_{t-1}) && \text{(Update)} \\
h_t &= (1 - z_t) \cdot h_{t-1} + z_t \cdot h'_t
\end{aligned}
$$

In this section, we introduced two standard recurrent neural networks dealing with repeated measurements. Note that other architectures exist to deal with different modalities and their specific challenges and could be used to adapt the methodologies introduced in this thesis.

2.1.2.3 Monotonic neural network

Not only can neural networks be organised in different ways to tackle different data modalities, but one can also enforce constraints to ensure a desired property. Such a subtype of *constrained* neural networks used throughout this thesis is the monotonic neural network. A monotonic neural network is a neural network that ensures the monotonicity of its output, given its inputs. Formally, a monotonic network M presents an increased output for any increase in the input:

$$\forall (t, t') \in \mathcal{X}^2, t' \geq_{\mathcal{X}} t \implies M(t') \geq_{\mathcal{Y}} M(t)$$

given the ordering relations $\geq_{\mathcal{X}}$ and $\geq_{\mathcal{Y}}$ for the input and output spaces. Constrained optimisations [358] or penalised loss [195] can achieve this property. However, these approaches increase the optimisation cost. An efficient alternative is to constrain all neurons' weights to be positive by applying a transformation over them, such as the absolute value [241] or square function [56, 266]. This last transformation redefines a neuron's output as $N(x) = \phi(x^T \cdot W^{\mathbf{2}} + b)$. The square function's differentiability improves the training convergence compared to the non-differentiable absolute value.

Crucially, Daniels and Velikova [77] and Lang [173] demonstrate that positively weighted networks are universal monotone approximators as summarised by the following Theorem:

Theorem 2.2 (Positively weighted neural network are universal monotonic approximations (adapted from [77]))**.** *Consider $k \in \mathbb{N}$ and Γ, the space of continuous monotone functions from a compact subset of $\mathbb{R}^k$ to the output space $\mathbb{R}$, $\forall g \in \Gamma, \forall \epsilon \in \mathbb{R}^+$, there exists a k-hidden layer network M_k with positive weights that can approximate g with error ϵ, i.e.*

$$\forall x, |g(x) - M_k(x)| < \epsilon$$

This property ensures the model's flexibility to model any monotonic function and motivates our use of monotonic neural networks for modelling the cumulative incidence function in survival

analysis.

2.1.3 Training

Training a neural network consists of estimating the optimal values of each parameter — weights and bias associated with each neuron — to minimise the difference between the predicted outputs $\hat{y}$ and target values y. Formally, one aims to optimise the following system:

$$\min_{\Omega} \mathcal{L}(y, N_{\Omega}(x))$$

in which Ω is the set of parameters characterising the neural network N, $\mathcal{L}$ is a loss function quantifying the error between the predicted value based on the input $x \in \mathcal{X}$ and the target label $y \in \mathcal{Y}$. For instance, common losses [337] are (i) the mean squared error (MSE) for regression that consists in the squared distance between the predicted and ground truth values, and (ii) the cross-entropy for binary classification, corresponding to the negative log-likelihood defined as the following sum over the datapoints i with an associated binary outcome y_i:

$$\mathcal{L}_{CE} = -\sum_{i} y_i \cdot \log(N_{\Omega}(x_i)) + (1 - y_i) \cdot \log(1 - N_{\Omega}(x_i)). \tag{2.1}$$

The dependencies between neurons make this optimisation particularly complex as one can not optimise for each parameter independently. An optimal parameter at the level of one neuron, assuming the rest of the network fixed, does not lead to the global optimum. Training a model requires finding the set of parameters that globally, not locally, solves the optimisation problem. An additional challenge comes from the model's non-linearity, which leads to the non-convexity of the loss given the parameters. This section introduces the multiple techniques available to efficiently train a neural network.

Gradient descent. The optimisation literature presents multiple strategies to approximate global optimal parameters in this context [249]. At its core, gradient descent optimisation remains a common training procedure to discover the optimal values of the neural network's parameters. Gradient descent minimises the loss function by iteratively updating the parameters' values in *the opposite direction of the gradient of the loss*. The gradient represents the change in the loss in the vicinity of the current parameters. Following this direction results in updating parameters towards a reduced loss. Formally, this optimisation procedure consists of the following steps:

1. Randomly initialise the parameters at iteration 0, Ω_0.

2. Compute the loss $\mathcal{L}$ over the training data by forward propagating the data and measuring the difference between the neural outputs and target labels.

3. Compute the gradient ∇_{Ω_t} of $\mathcal{L}$ given all parameters at iteration t.

4. Update the parameters following the opposite gradient direction:

$$\Omega_{t+1} = \Omega_t - \alpha \nabla_{\Omega_t} \mathcal{L}(y, N_{\Omega_t}(x))$$

   with $\alpha \in \mathbb{R}^+$ a hyperparameter known as the learning rate.

5. Repeat from (2) until convergence. Each of these iterations is known as an *epoch*.

The choice of an adequate learning rate α plays a central role in the training convergence: too small, the algorithm requires many epochs to converge; too large, the loss oscillates without reaching the optimum. Adaptive strategies to choose α, such as AdaGrad [84] and Adam [166], consist in updating this value given the change in the loss. Adam, adopted in our work, defines the learning rate as a function of the approximated loss second moment. This choice results in an increased learning rate when the speed of loss change is fast, and a decreased one when this change becomes slower. In other words, the algorithm uses large updates at the initialisation and smaller ones near the optimum, leading to faster convergence.

Backpropagation. The gradient descent algorithm relies on fast estimations of the gradient given *all* parameters. Leveraging the chain rule makes this computation efficient in a neural network. The loss computed at the level of the output layer is iteratively passed to the earlier layers. This process is therefore known as *backpropagation* as it follows the opposite direction of the forward propagation.

Specifically, for computing the gradient of the loss given the i^{th} weight of a given neuron j, denoted by $W_{i,j}$, one aims to compute:

$$\frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial W_{i,j}} = \frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_j} \frac{\partial o_j}{\partial W_{i,j}}$$

with o_j the output of the j^{th} neuron defined as $\phi_j(\sum_k z_k \cdot W_{k,j} + b_j)$, and z, the neuron's inputs. The quantity $\frac{\partial o_j}{\partial W_{i,j}}$ can be calculated at the neuron level given a differentiable activation function[3]. The remaining partial derivative $\frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_j}$ can recursively be computed:

- If j is a neuron of the output layer, its output o is the predicted value used to compute the loss, i.e. $N_\Omega(x)$, allowing the direct computation of $\frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_j}$ given a *differentiable* loss function.

- If the neuron is in an inner layer, one can compute this derivative by leveraging the gradient computed in the neurons receiving its output o_j as input. We denote this set of neurons as $\mathcal{R}_j = \{k \text{ such that } \exists i, z_{k,i} = o_j\}$; this means that $\forall k \in \mathcal{R}_j$, the neuron k

[3]Note that when using ReLU function, one assumes a fixed value at 0, the point of non-differentiability.

receive j's outputs as inputs. Using the total derivative given o_j, one can estimate the gradient for neuron j as:

$$\frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_j} = \sum_{k \in \mathcal{R}_j} \frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_k} \frac{\partial o_k}{\partial o_j}$$

In this expression, $\frac{\partial o_k}{\partial o_j}$ is the derivative of the activation function of neuron k and $\frac{\partial \mathcal{L}(y, N_\Omega(x))}{\partial o_k}$ is known by recurrence as the neurons in $\mathcal{R}_j$ are closer to the output layer than the neuron j (in a non-cyclic neural network).

The backpropagation algorithm is therefore a recurrence from the output layer to the input layer consisting of the following steps:

1. Compute gradient for each neuron in the output layer.
2. Compute gradient for neurons with an output connected to the previously computed layer, i.e. neurons belonging to the set $\mathcal{R}_j$, following the previously described computation.
3. Repeat 2 until all parameters' gradient is computed.

An additional complexity in recurrent neural networks is the repeated use of each neuron's weights at each longitudinal observation. This recurrence implies that, during backpropagation, the update must account for the change derived from each time point. In the literature, it is common to use the average of this change [347]. The resulting algorithm is known as backpropagation through time.

Automatic differentiation. Neural network training relies on the accurate and efficient estimation of all partial derivatives used in the backpropagation algorithm. Automatic differentiation aims to fulfil these two objectives. As any computer function is a succession of elementary operators with known derivatives, automatic differentiation combines each of these elementary derivatives [23] to estimate the quantity of interest[4]. Critically, this computation results in the exact computation — as precise as the computer function itself — without numerical approximation or symbolic differentiation of the derivative of interest.

Improving convergence. While gradient descent is guaranteed to recover the global minimum with a convex loss, it may converge towards local minima when this property does not hold. To avoid converging towards local minima, stochastic approaches encourage exploration beyond the current parameters' selection. The strategy adopted in our work is the Stochastic Gradient Descent. This algorithm estimates the gradient on a subset of the data, randomly drawn at

[4]For further implementation details for the Pytorch library used in this work, we invite the reader to refer to [243].

each iteration of the algorithm, instead of using all training data[5]. This estimation reduces the training computational cost, improving convergence speed. Additionally, the algorithm results in updates following gradient directions otherwise unexplored. Empirically, [83] showed the superiority of this approach in discovering global optima for a set of neural architectures.

Avoiding overfitting. Simple neural networks may not capture the complexity of the problem at hand. This phenomenon is known as under-fitting. Conversely, increased flexibility comes with the risk of overfitting to the data, i.e. learning a representation that only captures the complexity of the training data but does not generalise beyond. Figure 2.4 illustrates the problem of under- and over-fitting.

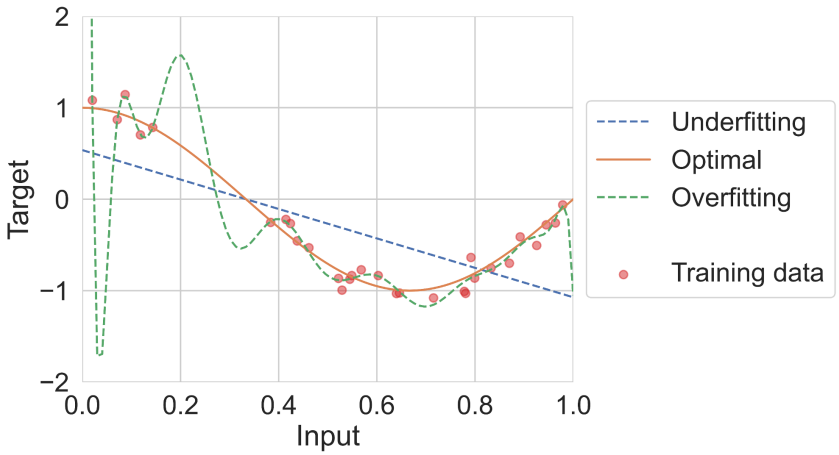


Figure 2.4: Complex models may overfit, while simple ones may underfit the training data. Finding an adequate model is to strike this balance.

When training a neural network, one aims to balance these risks. Our work relies on the following set of tools to achieve this balance:

- Hyperparameter selection: the choice of hyperparameters — such as the number of layers, neurons, and activation functions — plays a critical role in the model's flexibility, training and convergence. Measuring the model's performance on a left-aside subset of the training data, known as *validation* set, often guides hyperparameters' choice. In practice, one randomly selects a set of values for the hyperparameters — as [30] empirically showed the superiority of random search over exhaustive exploration — creates a model following the selected hyperparameters' values, trains it, and finally evaluates it on the left-aside data. After repeating multiple draws, one selects the neural network with the best validation performance.

- Early stopping: stopping the model's training when performance computed on a left-aside sample of the training set ceases to improve.

[5]The sample size used to approximate the gradient is selected as a hyperparameter known as batch size.

- Regularisation: penalising the loss by the model complexity, as proposed in ridge [131] and lasso regressions [320], avoids overfitting.
- Dropout: this neural network-specific form of regularisation, proposed in [305], consists of discarding a random subset of neurons at each training iteration to avoid over-relying on any particular neuron. At each epoch, one draws a subset of neurons and reduces their parameters' values to 0, then computes and backpropagates the resulting loss.

This section is a non-exhaustive introduction to the multiple tools used for neural networks' efficient optimisation and generalisation. Note that each of these concepts is an active research field beyond the scope of this thesis. Nonetheless, the concepts and tools introduced in this section are sufficient to understand the content of this thesis.

2.2 Survival analysis

Survival analysis involves modelling the time to an event of interest [68], which plays a critical role in medicine to understand disease manifestation, treatment outcomes, and the influence of different risk factors on patient health [285]. This analysis differs from standard regression settings as patients may not experience the outcome of interest over the study period. For instance, consider a 5-year-long trial to measure the impact of cholesterol-lowering drugs on the risk of cardiovascular disease as shown in Figure 2.5. Some of the patients may present with the condition over the study period, but some do not. Despite not knowing if and when these *censored* patients would experience the event of interest, they inform the survival regression as they participate event-free until exiting the study. Ignoring these patients would result in a biased survival estimate as, in this example, patients with longer survival are more likely to be ignored. To tackle this bias, multiple approaches propose to maximise the likelihood of all observed data.

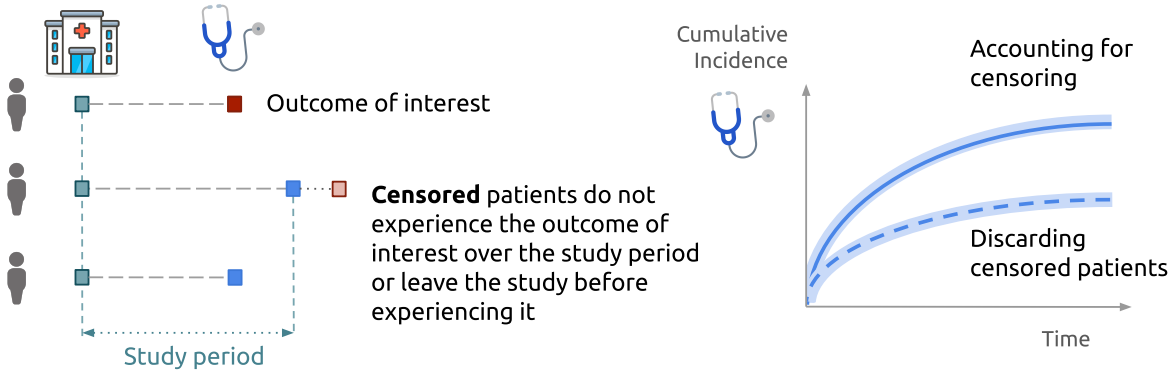


Figure 2.5: Censored patients inform the modelling of the outcome of interest. Note that the direction of the bias is only illustrative as it depends upon the censoring mechanism.

Formally, consider a set of observations defined as $\{x_i, t_i, d_i\}_i$ where x_i are the observed covariates for patient i, $t_i \in \mathbb{R}^+$ is the last time the patient was present in the study measured from the appropriate time origin, and d_i represents the observed event. In this review, we introduce the single event setting: the patient is either (right)-censored $d_i = 0$ or the event of interest is observed $d_i = 1$. Additionally, we assume non-informative censoring, i.e., if $d_i = 0$, the patient is right-censored for a cause *uncorrelated* with the outcomes of interest. In this context, the central quantity of interest is the survival function, which is the probability of surviving longer than t, defined as:

$$S(t) = \mathbb{P}(T \geq t)$$

with T being the random variable associated with the survival time. Equivalently, one can estimate the cumulative density function $F(t)$ of observing an event before time t, linked to the survival function through the relation:

$$F(t) = 1 - S(t)$$

The instantaneous hazard function $\lambda(t)$, which quantifies the risk of experiencing the event in the next instant given survival until time t, is often the modelled quantity as it offers an intuition of the patient's risk:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t \mid t < T)}{\delta t}$$

Using Bayes' theorem, this quantity can be expressed as a function of the survival function as follows:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t)}{S(t)\delta t} = \frac{dF(t)}{S(t) \cdot dt} = -\frac{d \ln S(t)}{dt}$$

Equivalently, one can express the survival as a function of the hazard function:

$$S(t) = \exp\left[-\int_0^t \lambda(u)du\right] = \exp\left[-\Lambda(t)\right]$$

with $\Lambda(t)$, known as the cumulative hazard function, is the integral of the instantaneous hazard corresponding to the cumulative risk of experiencing an event by time t.

These quantities are often the target of survival models. Estimating them may rely on maximising the likelihood of the observed data. Assumption concerning the censoring process is necessary to express this likelihood. In our work, we follow the common assumption of non-informative censoring, i.e., event and censoring times are independent given the covariates. Specifically, each patient i with an observed event contributes to the likelihood, the probability of experiencing the event at t_i without previous events, i.e., $\lambda(t_i)S(t_i)$. Note that this quantity is the negative derivative of S evaluated in t_i. The likelihood associated with each censored

patient is the probability of not experiencing the event until t_i, i.e., $S(t_i)$. Using the previous relation, this results in the following log-likelihood accounting for all the observed data:

$$l = \sum_{i, d_i \neq 0} \log \lambda(t_i) - \sum_i \Lambda(t_i) \tag{2.2}$$

In the following sections, we review both the traditional statistical and neural network approaches to model these quantities.

2.2.1 Statistical modelling

Survival analysis is an active field of research in the statistical community [162], we divide common approaches for modelling survival data into four categories: (i) discretisation, (ii) parametric, (iii) non-parametric, and (iv) semi-parametric modelling. For each category, we review the foundational statistical models and their associated assumptions.

Discretisation. A first approach to leverage standard regression models is to discretise the time to event outcome. If an analysis consists of estimating the risk of observing the outcome by a given time horizon, one can binarise the outcome: did the patient observe or not the event by the evaluation horizon? For instance, in modelling the 5-year risk of cardiovascular disease, one may classify patients with an event in that time window as positive and the rest as negative.

Many applications necessitate a finer resolution of the survival function estimate that such coarse binarisation does not provide. Further discretisations of the time horizon into finite time intervals can offer a finer survival estimate as originally proposed in [293]. For each discrete interval, a model assesses the conditional probability of observing the event in this interval, given survival until then. Similarly, piece-wise models assume constant hazard for each time interval. [133, 171] propose an equivalent Poisson regression to model this latter formulation.

Parametric. Assuming a parametric survival distribution results in a closed-form likelihood, leading to efficient likelihood maximisation, even when accounting for censoring. For instance, a flexible and common distribution choice is the Weibull distribution, defined by the cumulative distribution function:

$$F_{\text{Weibull}}(t) := 1 - e^{-\frac{t^v}{u}}$$

with u and v the distribution's scale and shape parameters. More flexible distributional choices such as generalised gamma [73] or F-distribution [72] can better capture the complexity of the survival function. One can further parameterise the distribution's parameters as functions of the observed covariates, e.g. u and v in the Weibull distribution can be defined as a function of x. Alternatively, one can consider a mixture of parametric distributions, as proposed in [206],

with a logistic model assigning a patient to the mixture according to its observed covariates.

Non-Parametric. To avoid parametric assumptions, one may use non-parametric estimators of the survival quantities. The Kaplan-Meier estimator [160] aims to directly approximate the survival function by estimating:

$$\hat{S}_{\mathsf{KM}}(t) = \prod_{i,t_i<t} \left(1 - \frac{d_i}{\sum_j \mathbb{1}(t_i < t_j)}\right)$$

Alternatively, one can estimate the cumulative hazard $\Lambda(t) = \int_0^t \lambda(u)du$ using the Nelson-Aalen estimator:

$$\hat{\Lambda}_{\mathsf{NA}}(t) = \sum_{i,t_i<t} \frac{d_i}{\sum_j \mathbb{1}(t_i < t_j)}$$

While these approaches do not make parametric assumptions over the survival distribution, they do not directly model the relation between covariates and outcomes.

Semi-Parametric. A commonly used model in the clinical literature relies on the flexibility of non-parametric models for estimating the population-level hazard and models the covariates' relation to the outcome through a parametric model. The Cox proportional hazards model [74] is composed of a linear combination of covariates parameterised by a vector β: $\eta(x) = \beta^T x$, to model deviations from a population's non-parametric baseline hazard function $\lambda_0(t)$. Formally, this approach models the hazard as:

$$\lambda(t \mid X) = \lambda_0(t)e^{\eta(X)}$$

Critically, this form reflects an assumption of constant proportionality between the baseline hazard and the individuals' evolutions. This simplifying assumption allows the parameters associated with η to be estimated independently of the non-parametric baseline hazard function, by maximising the partial log-likelihood defined as:

$$l_{\mathsf{CoxPH}} = \sum_{i,d_i=1} \left[\eta(x_i) - \log \sum_{j,t_j \geq t_i} e^{\eta(x_j)}\right]$$

Following a similar idea as the Nelson-Aalen estimator, Breslow [40] proposes to maximise the likelihood while fixing $\eta(x)$ to obtain the cumulative hazard estimator:

$$\hat{\Lambda}_{\mathsf{Breslow}} = \sum_{i,t_i<t} \frac{d_i}{\sum_j \mathbb{1}(t_i < t_j)e^{\eta(x_j)}}$$

Predicting survival at a given time horizon t is then possible by leveraging both the regression

parameters and the Breslow estimator:

$$\hat{S}(t \mid x) = \exp(-\hat{\Lambda}_{\text{Breslow}}(t)e^{\eta(x)})$$

Nonetheless, the proportional hazards assumption rarely holds in medical applications [306] and alternatives, such as stratified group baseline hazard functions, have been proposed to relax this assumption.

2.2.2 Neural approaches

Paralleling these different approaches, ML alternatives rely on neural networks' flexibility to model the survival outcome without explicit specification of the relation between covariates and outcomes.

Discretisation. Following a similar discretisation idea to [293], Deephit [179] uses neural networks for estimating the probabilities of survival in a given time interval. As shown in Figure 2.6, a neural network uses the covariates to predict the probabilities associated with the n discrete time intervals. A final Softmax layer ensures a properly defined probability distribution.

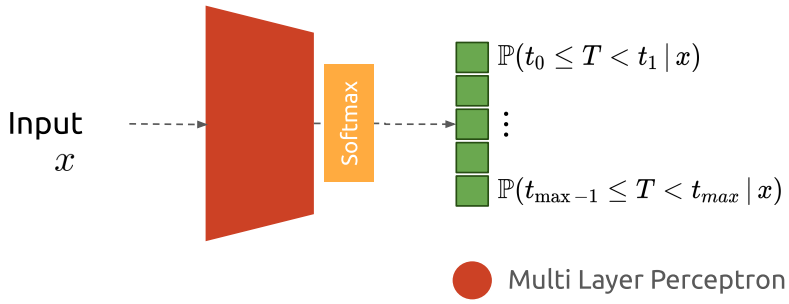


Figure 2.6: DeepHit Architecture: A multi-layer perceptron predicts the discretised conditional probabilities of event's occurrence.

A key difference with its traditional statistical alternative is the model's training procedure. The loss used to train DeepHit by backpropagation consists of the discrete log-likelihood penalised by the model capacity to rank patients' survival outcomes. Formally, the loss corresponds to $\mathcal{L}_{\text{DeepHit}} = -l_{\text{DeepHit}}(N) + \upsilon R(N)$ with $\upsilon \in \mathbb{R}^+$ a hyperparameter weighting the ranking penalisation R and l, the log-likelihood defined as:

$$l_{\text{DeepHit}} = \sum_{i,d_i=1} \log(N_{t_i}(x_i)) + \sum_{i,d_i=0} \log(1 - N_{\leq t_i}(x_i))$$

with $N_t(x)$, the neural network's output corresponding to the probability of having the event in the time interval containing t given the covariate x. Note the equivalence between the

log-likelihood evaluated in discrete times and the common cross-entropy loss used in binary classification tasks introduced in Equation (2.1). The penalty R encourages the model to have a larger predicted risk at time t_i for a patient i with an event at that time, than for any other patient with a later censoring or event, i.e.:

$$R(N) = \sum_{i,j,d_i=1,t_i<t_j} \gamma(N_{\mathbf{t_i}}(x_i), N_{\mathbf{t_i}}(x_j))$$

with γ a convex loss function. In the original work, the authors propose the use of an exponentiated difference for this loss function.

Parametric. Echoing the statistical parametric approaches, Nagpal, Li, and Dubrawski [221] introduced Deep Survival Machines (DSM) consisting of an assignment model and a mixture of parametric distributions parameterised through neural networks. Figure 2.7 displays how a first neural network assigns a patient to a given survival distribution given its covariates, and a second parameterises each of the Weibull or Log-Logistic distributions in the mixture.

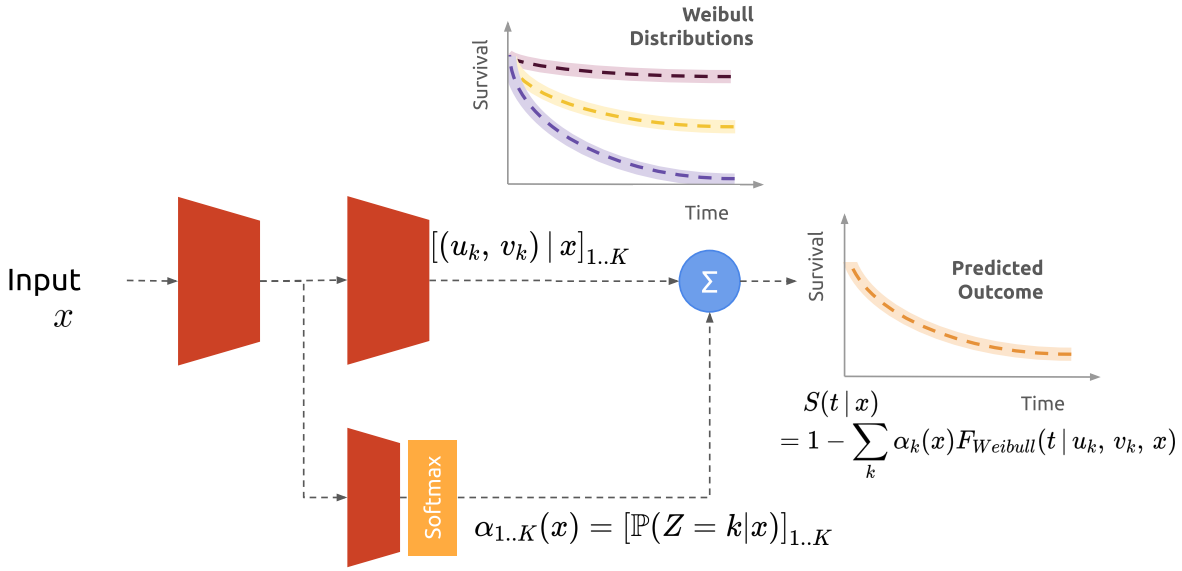


Figure 2.7: DSM Architecture with a mixture of Weibulls: A multi-layer perceptron parameterises each Weibull distribution in the mixture, while a second multi-layer perceptron assigns the covariates x to the different distributions.

The choice of parametric distributions results in a closed-form log-likelihood. Specifically, each patient contributes to the likelihood through the average of the different parametric distributions weighted by its assignment probability α. Formally, the training relies on the following log-likelihood:

$$l_{\text{DSM}} = \sum_i \sum_k \alpha_k(x_i) \left[\mathbb{1}(d_i = 1) \log \frac{\partial F(u \mid N(x_i))}{\partial u}\Big|_{u=t_i} + \mathbb{1}(d_i = 0) \log S(t_i \mid N(x_i)) \right]$$

with $N(x_i)$ the neural network's outputs parameterising the different distributions in the mixture. The authors propose the addition of a penalisation between the learnt distributions and expert-expected distributions. Similarly, [119] integrates a mixture of Gaussians to estimate the cumulative incidence function.

Semi-parametric. In the ML literature, DeepSurv [163] is an adaptation of the Cox model with non-linear covariate interactions, i.e. the previously described η is the output of a multi-layer perceptron $N(x) = \eta(x)$. This results in the same optimisation in which one aims to maximise the partial log-likelihood:

$$l_{\text{DeepSurv}} = -\sum_{i,d_i=0} N(x_i) - \log \sum_{j,t_j \geq t_i} e^{N(x_j)}$$

Predictions then follow the same procedure as the traditional Cox model relying on the Breslow cumulative hazard estimator.

Note that non neural network semi-parametric approaches — out of the scope of this thesis — have been introduced, such as survival trees [112] and their random survival forest extension [140]. In its simplest form, a survival tree iteratively divides the population based on their covariates into subgroups, then fits non-parametric estimators to each of the identified subgroups.

Other approaches. All the previously described strategies made simplifying approximations about the survival function to estimate the likelihood. Further away from the standard statistical approaches, the ML literature introduced alternative models to leverage the flexibility offered by neural networks for survival modelling[6] while avoiding these assumptions. Echoing statistical approaches to model the hazard [63] or cumulative hazard [315] as the solution of an ODE, DeSurv [78] proposes to solve an ODE whose differential equation is defined through a neural network to estimate the cumulative hazard. Closer to our work, Sumo-Net [266] uses monotonic neural networks to model the cumulative hazard function.

2.2.3 Evaluation

Following the training of these different models, one is interested in measuring the quality of the predicted survival probabilities to evaluate the model's clinical utility and allow model comparison. Consider a survival model outputting $\hat{S}(t \mid x)$, an estimate of $S(t \mid x)$. In this context, one aims to quantify the model's *discrimination* and *calibration*. Discrimination describes the model's capacity to maintain the time-of-event ordering in the predicted survival values. This means that a model should assign a higher predicted risk for a patient who observed

[6] A key advantage of using neural networks is their flexibility to handle different modalities: Deephit and DSM have been used for time series data [180, 220], and DeepSurv for images [183].

an event at time t_i than any other patients that survived longer than t_i, i.e. $\forall(i,j)$, if $d_i = 1, t_i < t_j \implies \hat{S}(t_i \mid x_i) < \hat{S}(t_i \mid x_j)$. While discrimination is essential for comparing patients, this metric does not capture the model's capability of quantifying the absolute risk. Calibration quantifies the discrepancy between predicted probabilities and observed probabilities of observing an event.

We first describe the Area Under the Receiver Operating Characteristic Curve, commonly used in ML. Then, we focus on the two time-dependent metrics to quantify discrimination and calibration at a given time horizon t, and then present the integrated metrics offering a concise overview of a model's performance.

Area Under the Receiver Operating Characteristic Curve (AUC). When evaluating a binary outcome, such as discretised survival at a specific time horizon, the Receiver Operating Characteristic curve (ROC) describes a model's discriminative performance. This curve displays the model's ability to distinguish between positive and negative cases at all possible thresholds applied to the model's predicted probabilities. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) across these threshold values. The AUC measures the area below this curve; higher values indicate better discrimination, with a perfect score of 1 and score under random performance of 0.5.

Truncated time-dependent C-Index. The truncated time-dependent C-index [101] is a generalisation of Area Under the Receiver Operating Characteristic Curve (AUC) to survival labels with right censoring evaluating performance for events up to time t. It captures the discriminative performance of a model by measuring the ordering of the survival predictions as follows:

$$CI(t) = \frac{\sum_{i,j;\, d_i=1,\, t_i<t,\, t_i<t_j} \omega(t_i) \mathbb{1}(\hat{S}(t \mid x_i) < \hat{S}(t \mid x_j))}{\sum_{i,j;\, d_i=1,\, t_i<t,\, t_i<t_j} \omega(t_i)}$$

where $\omega(t)$ is the inverse probability of censoring weights estimated through a Kaplan-Meier estimator of the time to censoring distribution. This weighting aims to statistically unbias the concordance estimate by attributing more weight to patients with observed outcomes. As these weights rely on a Kaplan-Meier estimate, the censoring mechanism is assumed to be independent of the outcome of interest.

Time-dependent Brier Score. Similarly, [114] introduces a calibration metric accounting for censoring as follows:

$$BS(t) = \sum_{i;\, t_i<t,\, d_i=1} \omega(t_i) \hat{S}(t \mid x_i)^2 + \sum_{i, t_i>t} \omega(t) \left[1 - \hat{S}(t \mid x_i)\right]^2$$

Overall metrics. Time-dependent metrics reflect the quality of the survival fit at a fixed horizon. These time-dependent metrics are often more relevant to understand the model's

behaviour during clinical deployment. These metrics, however, do not offer a succinct summary of the models' performance. The following metrics provide this overall summary by numerical integration of the Brier Score (IBS) [100] and evaluating the time-dependent C-Index (ICI) [7]:

$$ICI = \frac{\sum_{i,j;\, d_i=1,\, t_i<t_j} \mathbb{1}(\hat{S}(t_i \mid x_i) < \hat{S}(t_j \mid x_j))}{\sum_{i,j;\, d_i=1,\, t_i<t_j} 1}, \qquad IBS = \int_0^{t_{max}} BS(t)dw(t)$$

with t_{max} the last event time observed in the dataset.

Equipped with these models and evaluation tools, the remainder of this thesis investigates the impact of clinical presence on survival predictions.

2.3 Algorithmic fairness in medicine

For a given model, our work aims to quantify its algorithmic fairness to avoid reinforcing inequities. To quantify algorithmic fairness, three families of definitions emerge from the literature [210, 329].

- **Individual fairness** [86] deems an algorithm fair if similar individuals (according to a relevant metric) are treated similarly.
- **Causal fairness** deems an algorithm fair if the prediction would remain unchanged if an individual's group membership changed [169], or if group membership does not affect the prediction through inadmissible pathways [219].
- **Group fairness** defines fairness in terms of equal performance across groups, where the performance metric of interest may vary [60, 120].

Individual fairness requires access to a relevant, task-specific, distance metric to assess who is "similar", and notions of causal fairness require knowledge of the causal graph between all covariates and target labels. In practice, it is rare to have access to such distance metrics or causal graphs. As a result, group fairness definitions are the most widely used in practice and are the ones used throughout this thesis.

In healthcare, Rajkomar et al. [263] propose to quantify group fairness as the difference in (i) observed outcomes, (ii) model performance or (iii) care allocation. In this thesis, we focus on model performance, and in particular the "*equal performance*" definition of algorithmic fairness [263]. This definition evaluates if the model performs comparably across groups [62, 90, 237] by comparing group-level metrics, for relevant metrics of interest. This definition has been used to quantify if marginalised groups would be impacted differently by medical models' deployment [51, 52, 248, 287, 370]. For instance, Seyyed-Kalantari et al. [287] demonstrate

X-ray classifiers' performance gap between groups, and highlights the detrimental misdiagnosis for marginalised groups if the models were deployed.

Focusing on this measure of fairness, we propose to formalise the concept of equal performance: given patients member of a group g, we aim towards reducing the difference in a given performance metric between members of group g and non-members.

Definition 2.1 (Equal Performance)**.** A ML model $\mathcal{M}$ is fairer than another $\mathcal{N}$ with regard to group g if its absolute performance gap is the smaller, i.e. $|\Delta_g(\mathcal{M})| < |\Delta_g(\mathcal{N})|$, where $\Delta_g(\mathcal{M}) := d(\mathcal{M}(\{X_i\}_{G_i=g})) - d(\mathcal{M}(\{X_i\}_{G_i \neq g}))$ for some performance metric d, and (X_i, G_i) the associated covariates and group for patient i.

Note that practitioners should not only aim to satisfy this property but also aim to maximise performance, as random predictions would lead to equal performance but be detrimental to both groups.

Chapter 3

Missingness

Associated Publications. The work presented in this chapter is based on our publication: Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness [148] presented at ML4H 2022, IBC 2022, ISCB 2022, SCECR 2023 and CHIL 2023.

Problem statement. *How does the handling of group-specific missingness impact algorithm fairness?*

3.1 Motivation

Each observation, from orders of laboratory tests to treatment decisions, depends on access to medical care, patients' medical states, and practitioners' expert decisions. As a result, the collected medical records suffer from missing observations [289]. We refer to the missingness that stems from these clinical interactions as *clinical missingness*.

Clinical missingness is widespread in medical observational data [191, 343, 344]. Medical records reflect and inform treatment, and are not primarily gathered for scientific discovery and analysis. The prevalence of missing data is an issue because statistical analysis and ML often require complete data. Practitioners, therefore, rely on preprocessing strategies, such as imputation, to address missingness in their medical datasets. However, the importance of this step is often overlooked. In their literature review, Nijman et al. [236] note that 65% of ML papers on clinical applications mention the problem of missingness, but less than 10% report their assumptions on missing data, and 24% analyse how their choice of handling missing data impacts their conclusions.

Overlooking clinical missingness may have consequential repercussions on algorithmic fairness — a novel point that we raise and investigate in this chapter. This algorithmic fairness connection arises because clinical missingness patterns are often group-specific. In other words, the patterns and causes of missing data can vary between different population subgroups. Group-specific patterns of missingness are particularly notable in medical datasets.

They can occur due to historical healthcare biases or disparities, which subsequently influence healthcare access, treatment, and outcomes [54, 92, 146, 165, 238]. For instance, limited access to healthcare resources can translate into group disparities in available testing procedures. Additionally, medical guidelines and practice can also reinforce existing group inequalities by focusing primarily on populations considered high-risk. Consequently, these differences in medical interactions translate into group-specific missingness in testing and examination data. For instance, this is evidenced by [185], who show increased missingness in Black patients' family history records.

In making the connection between clinical missingness and algorithmic fairness, we raise a novel concern: in the presence of these group-specific missingness patterns, data imputation strategies can have different effects across groups, potentially resulting in significant algorithmic fairness impacts in ML pipelines. For example, an imputation technique may harm the downstream prediction performance for one group more than another.

The impact of the imputation step on algorithmic fairness has received limited attention in the literature. Our work is the first to investigate the link between clinical missingness and algorithmic fairness through three key contributions:

1. We mathematically characterise three different mechanisms through which disparities in the healthcare system may result in different types of clinical missingness. We justify their relevance through medical evidence. Using their causal representations, we show how these mechanisms can translate into group-specific missingness patterns.

2. We show that group-specific imputation — a common imputation practice — can be counter-intuitively detrimental to group fairness and harmful to a marginalised group (relative to non-group-specific imputation). This holds when considering either imputation reconstruction error or downstream predictive performance. We empirically show that this can occur under a variety of missingness patterns via simulation studies, and provide real-world evidence of this phenomenon in the widely used medical dataset MIMIC III.

3. We prove mathematically that, given an observed dataset, the appropriate choice of imputation between two frequently used approaches is under-determined. The superiority of an imputation strategy in data quality and group fairness gap depends upon the *unobserved* missingness process. *A priori* identification of the imputation strategy that would reduce the gap in reconstruction errors depends on the knowledge of this missingness process, i.e., no imputation strategy can be deemed superior regarding reducing downstream group fairness gaps.

While current imputation practices focus on improving the overall quality of the data available for modelling, this chapter shows that they can also affect algorithmic fairness. Current recommendations for imputation often focus on reducing reconstruction errors under

assumed simple missingness mechanisms; these simple missingness assumptions, however, are not adapted to the reality of clinical missingness. The absence of a theoretical understanding of fairness risks associated with imputation strategies has led practitioners to rely on intuition when handling missing data. As a result, current recommendations may have detrimental effects on algorithmic fairness. For example, studies often use a single imputation strategy with all likely confounders included to ensure the plausibility of the missingness assumptions [123, 184]. Because group missingness differences are a concern, imputation strategies frequently control for or stratify by group membership, as in [135], using group-specific mean imputation. The rationale is to improve subgroups' reconstruction errors. However, in our second major contribution noted above, our research demonstrates that such group-specific imputation can actually have harmful effects on downstream algorithmic fairness. Additionally, we show that reliance on a single imputation strategy without further consideration of alternative imputation strategies analysis is misleading. A sensitivity comparison between imputation strategies can reveal meaningfully different group fairness outcomes.

Our results call for careful consideration of missingness assumptions and imputation. Exploring the upstream missingness process should guide the selection of appropriate assumptions and, accordingly, imputation strategies. When missingness processes are unobserved (a frequent occurrence), practitioners should report and justify the assumptions made on the missingness process, and acknowledge the associated consequences on their analysis' conclusions. In particular, when dealing with missing data affecting covariates, reconstruction error cannot typically be measured because the missing data is unknown, but prediction performance can be evaluated. Evaluating different imputation strategies on downstream fairness prediction outcomes is paramount because different imputation strategies can result in distinct performance gaps between groups. In some cases, the choice of imputation can even reverse fairness outcomes: where one imputation strategy results in a favourable performance gap for a marginalised group relative to the majority, another imputation strategy that yields similar overall predictive performance results in a harmful performance gap for that same marginalised group.

To reach these conclusions, we first present in Section 3.2 the literature associated with missingness, fairness, and their intersection. Then, we introduce and formalise clinical missingness scenarios, their origins, and resulting group-specific patterns in Section 3.3. In Section 3.4, we theoretically prove how imputation strategies present different reconstruction fairness gaps under different patterns of clinical missingness. Through simulations, we empirically ascertain these results both at the level of reconstruction error and downstream performance in Section 3.5. Finally, in Section 3.6, we demonstrate how real-world study conclusions could be impacted under different imputation strategies in the widely used MIMIC III dataset.

3.2 Related work

This chapter explores the link between missingness and algorithmic fairness in ML for healthcare. In this section, we review related literature across domains.

3.2.1 Clinical missingness

Missingness naturally occurs in medical studies in which information is recorded for clinical decision-making [123]. Missing data may therefore present informative patterns. Current clinical understanding of missingness relies on the three well-studied patterns [193]: *Missing Completely At Random* (MCAR) — random subsets of patients and/or covariates are missing; *Missing At Random* (MAR) — missing data patterns are a function of observed variables; and *Missing Not At Random* (MNAR) — missing patterns depend on unobserved variables or on the missing values themselves. However, current missingness formalisations often ignore group-specific patterns of observation. Our work provides a finer formalisation of missingness in the medical context by taking into account membership to marginalised groups, and characterising how clinical presence may result in group-specific missingness patterns.

Missing data prohibits the use of traditional statistical models that require complete data. Ignoring patients with missing data, also known as complete case analysis, lowers statistical power [192] and statistically biases the coefficients of the studied association [348]. Thus, practitioners often replace missing data, selecting from a wide range of available imputation strategies. These include single imputation strategies, which replace missing data with a single value such as mean, median, or nearest neighbour value [22, 32], or multiple imputation strategies, which propose multiple possible values for each missing one [232, 272, 349] as a way to quantify the uncertainty associated with the missingness process. Typically, both types of imputation strategies assume MCAR and/or MAR patterns, and all associated theoretical guarantees depend on these assumptions.

Common imputation strategies may be ill-adapted to handle clinical missingness reflective of more complex patterns. Especially, MNAR and MAR are non-identifiable from observational data alone and require knowledge of the missingness process and domain expertise for adequate modelling [25]. The recommended strategy to tackle this non-identifiability issue is to control the imputation strategy on additional covariates to render these simplifying assumptions more plausible [123]. Our work shows potential shortcomings of this covariate-adjusted imputation strategy under clinical missingness patterns, providing theoretical and empirical evidence showing that this strategy may backfire when controlling or stratifying on group membership.

Previous literature has studied the consequences of making incorrect assumptions about the missingness process, with a focus on inference and how these may bias estimates of the parameters of interest, e.g. treatment effect or odds ratios [28]. Current recommendations to mitigate this risk call for exploring the estimates' robustness under different assumptions by

performing sensitivity analysis [317, 355]. Our work demonstrates that missingness patterns and their handling may not only statistically bias estimates. We show that these patterns may have a differential impact across groups' data quality and predictive performance, and may affect algorithmic fairness. Further, we show that sensitivity analysis should also be performed to assess the robustness of fairness in the conclusions regarding the choice of imputation.

3.2.2 Algorithmic fairness and missingness

A central thrust of research on algorithmic fairness has focused on developing methods to mitigate disparities, such as resampling [157], loss regularisation [158] or post-processing adjustment [120]. Such approaches, however, largely assume that data is complete. Furthermore, characterisations of sources of algorithmic bias [20, 61, 210] rarely focus on the potential impact of missing data.

An emerging body of research has begun to study the interplay between algorithmic fairness and missing data. On the statistical side, [93, 203] show that mean imputation presents better group fairness properties compared to complete case analysis. Getzen et al. [102] explore how perturbing medical events' frequency from observational data negatively impacts group-specific predictive performance. Wang and Singh [340] propose a group-specific weighting; adjusting each observation's weight as a function of the group's observation rate to improve the ratio of positive outcomes between groups. On the medical side, Ganju et al. [96] encourages the use of clinical decision support systems to improve data collection, as unfair medical decisions can emerge from missing standardised testing in marginalised groups.

Closer to our work, [47, 376] show that the choice of imputation may lead to distinct gaps in performance through simulations. However, these works do not discuss how the different missingness patterns may arise in medicine nor how a specific group may be impacted differently by different imputation strategies. In the portion of our work that focuses on simulations, we study different missingness patterns that may arise as a result of the data-generating process in healthcare. Furthermore, we study the impact of different imputation strategies through a theoretical lens, and empirically show the effects that imputation strategies may have using real-world data.

Recently, Jeong, Wang, and Calmon [151] studied the risks of relying on imputed data. The authors demonstrate (i) the non-conformality of imputation strategies, i.e., the optimal imputation strategy depends on the downstream task; and (ii) the potential statistical bias associated with imputation under MCAR patterns. Our work identifies non-random patterns of missingness and theoretically explores imputation biases beyond MCAR patterns. Importantly, we also show that the optimal strategy depends not only on the downstream task but on the data distribution and missingness patterns.

Finally, at the intersection of algorithmic fairness, missingness, and ML for health, [5, 104, 105, 211, 263] describe multiple challenges linked to medical data, among which they state

that historical biases may lead to missingness patterns that could impact fairness, but they do not empirically study this. While informative missingness has recently received revived attention [211], no work has studied its potential association with fairness and provided the theoretical framework to understand and handle group-specific patterns. Our work aims to address these gaps in the literature by characterising different types of group-specific missingness patterns in medicine, theoretically analysing the impact of imputation strategies on algorithmic fairness, and exploring the impact of different imputation strategies under different clinical presence scenarios.

3.3 Clinical missingness

The central motivation of this chapter is that the underlying — often unobserved — missingness process can reflect disparities and, therefore, have large, unanticipated impacts on group-specific performance that are not well understood. To better understand the implications of different forms of clinical missingness and their implications, we review the literature for historical examples of missingness processes. We then provide a formalisation of the identified types of missingness based on their causal structure, highlighting their group-specific nature.

In this review, we distinguish three forms of missingness processes (illustrated in Figure 3.1):

(S1) **Limited access to quality care**. If and when certain groups do not have access to the same health services and quality care as others, this may result in more missing covariates for disadvantaged groups.

(S2) **(Mis)-informed collection**. Often, medical research has focused on a subset of the population. The resulting guidelines may be ill-adapted to other groups, and relevant covariates may be missing due to standard recommendations of when to collect certain information.

(S3) **Confirmation bias**. The collection of certain types of data depends on practitioners' prognoses and informative proxies that are not recorded. These prognoses may be affected by group-specific expectations.

These scenarios have substantial medical evidence, which we summarise in Section 3.3.1. Furthermore, these three scenarios are mathematically distinct, as shown by the formalisation in Section 3.3.2.

3.3.1 Clinical evidence

Multiple studies in the medical literature provide evidence of different pathways through which the data-generating process may result in group-specific missingness patterns. We provide

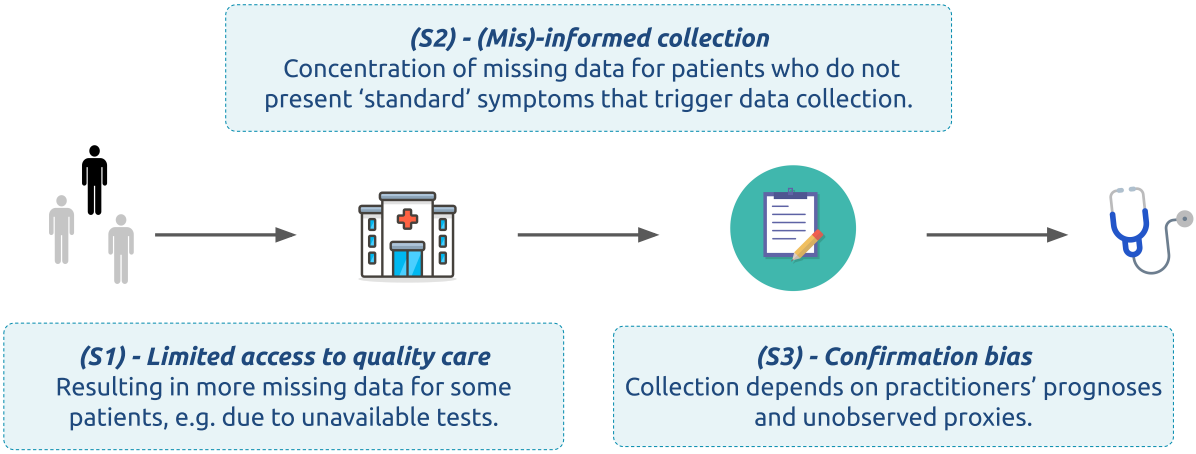


Figure 3.1: Examples of group-specific clinical presence mechanisms.

a succinct, structured review of the literature, taxonomised into the three distinct types of clinical missingness defined above.

Historical evidence of limited access to quality care. Socio-economic factors impact access to care and, consequently, missingness. For instance, education [19], urban residence [19], insurance [132], distance to hospitals [19] or mobility [314] have been shown to impact patients' interactions with the healthcare system. These differences in medical interactions may translate into inconsistent medical history [105], and limited access to advanced diagnostic tools [186], and may also impact behaviours such as additional waiting time before seeking care [345], and avoidance of preventive care [299]. Such reduced access to care for some subgroups of the population can result in missingness. For instance, the lack of medical history is in itself a problem of missing data. Avoidance or lack of access to care may translate into less frequent checkup data, and, therefore, a sparse record of patients' health evolution. Finally, limited access to more advanced diagnostic tools leads to absent tests in medical records.

Historical evidence of (mis)-informed collection. Historically, studies focused on perceived highest-risk groups and were constrained by the available and willing-to-participate patient population: breast cancer has been predominantly studied in women [8, 106], cardiovascular disease in men [330], skin cancers in whiter skins [107], and autism in men [113]. Resultant medical practices and guidelines target these groups. However, substantial evidence shows the prevalence of these conditions among a more diverse population: men experience 1% of breast cancers [362], 1 out of 3 women dies from cardiovascular disease [97], neoplasm can be cancerous in non-white populations, and autism has an estimated male-female ratio of 4.2:1 [369]. Stemming from social constructs and biological differences, distinct groups may present and express different symptoms for the same condition. Mauvais-Jarvis et al. [204] show how biological sex influences condition manifestation through genetics and how the associated

social construct of gender has epigenetic effects across a large set of cardiac conditions. The misalignment between condition manifestation in these groups and existing tests and guidelines can result in missing covariates necessary to identify the condition. Testing recommendations may only be prescribed conditioned on the observation of "standard" symptoms. If the symptoms considered do not include frequent symptoms for a marginalised subgroup, this will result in more missing tests for this group. For instance, women with heart failure may experience fatigue, while men are more likely to report chest pain [125]. As guidelines have focused on male patients [309], this difference in the expressed symptoms may not trigger further testing for women, as suggested by the difference in echocardiography between men and women [309].

Historical evidence of confirmation bias. Practitioners may perform or record a test only if they suspect that it will be informative. This phenomenon has been evidenced in the literature showing that the presence of tests in medical records is more closely related to the outcome than the actual test results [4, 296]. Wells et al. [346] also suggest that missing laboratory tests correspond to healthy results. Conversely, [275, 289, 342] show that sicker patients present more complete data. Under this data-generating mechanism, the way a condition affects different groups may result in group-specific testing patterns. For instance, general practitioners are more likely to record weights for underweight or overweight patients [233]. However, weights' distribution may differ depending on socio-economic characteristics [233] and, consequently, result in group-specific patterns of observation.

3.3.2 Formalisation

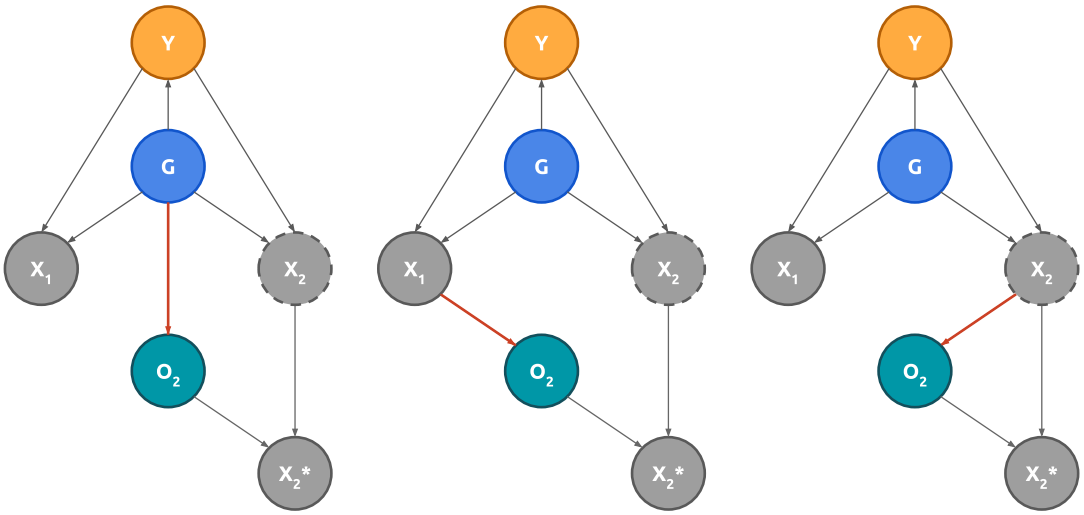
Each of the three scenarios above has different dependence structures between the observational processes and the resulting data. Formalisation and causal representation of these links permit further understanding of their dissimilarities, group-specific nature, and relation to standard patterns of missingness.

Consider a set of covariates X influenced by the underlying condition Y and the group membership G. Following the notations from [212], let O_i be the indicator of observation of X_i such that the observed value is defined as:

$$X_i^* = \begin{cases} \varnothing & \text{if } O_i = 0 \\ X_i & \text{otherwise} \end{cases}$$

We formalise the proposed scenarios in the bivariate case: X is the concatenation of two covariates (X_1, X_2). One covariate X_1 is always observed, while X_2 is potentially missing. Following these notations, Figure 3.2 displays the directed acyclic graphs (DAGs) associated with each scenario. Each DAG shows the dependencies between missingness, group membership, covariates, and outcome. Note that this formalisation generalises when X_1 and X_2 are sets of

covariates, but would not capture the potential dependencies between two partially missing covariates that could occur with more covariates.



(S1) (S2) (S3)

Figure 3.2: Directed Acyclic Graphs (DAGs) associated with the identified clinical missingness scenarios. Full circled covariates are observed, dashed ones are potentially unobserved. Y is the condition, G is the group membership, X_1 and X_2 are the two covariates — X_1 being always observed. O_2 is the observational process associated with X_2. Red arrows underline the dependency differences across scenarios.

The DAGs in Figure 3.2 highlight similarities and differences between the three scenarios. We assume the same dependence between group membership, condition, and covariates across the three scenarios. The condition *prevalence*, i.e., how often a patient may present the condition Y, may depend on G directly or through different mediators — merged with G for clarity. The condition *manifestation*, i.e., how the condition impacts covariates, may also depend on G, hence the connections both between Y and X, and G and X. As an example of these two distinct mechanisms, women and men present a similar risk of heart failure [172],— a similar condition prevalence. However, women express different symptoms than men with, for instance, more frequent symptoms of fatigue for women and chest pain for men [125] — different condition manifestations.

The DAGs also underline key differences in the missingness structure. In (S1), different groups have unequal access to care, thus the group membership G informs the missingness process O_2 because of differences in socio-economic factors. Meanwhile, the influence of medical covariates on the missingness patterns characterises both (S2) and (S3). In (S2), guidelines for measuring X_2 may depend on other observed covariates, whereas in (S3), measuring X_2 depends on the missing value itself — or unobservable covariates correlated with it. For instance, (S2) may consist of a guideline recommending to measure X_2 if X_1 is within a given range. (S3) differs as practitioners would measure X_2 only if this *same covariate* X_2 is expected to be in a given range.

Importantly, these representations highlight the connection between clinical missingness and group-specific patterns. Even under no direct dependence of the missingness process on group membership, there exist indirect pathways between G and O_2 through condition manifestation and prevalence. Missingness processes with dependencies on any of the covariates X would, therefore, present group-specific patterns of missingness.

This formalisation further embeds group-specific patterns in the traditional missingness framework: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Assuming observation of group membership and no unobserved confounders, (S1) and (S2) are sub-types of MAR patterns as missingness only depends on observed covariates. (S1) corresponds to what has been identified in the literature as *group-specific MCAR pattern* [151], in which each group-specific missingness process is independent of any other covariate, i.e., MCAR, given the group membership. (S3) corresponds to MNAR in which the dependence is on the missing covariate itself. Considering group membership highlights clinically relevant missingness subpatterns of the traditional framework.

3.4 Theoretical analysis of imputation and group fairness

In Section 3.3, we show that different types of clinical presence can lead to different missingness patterns. This section demonstrates why the properties of the missingness process matter. Specifically, we theoretically analyse the relationship between missingness patterns, imputation methodology, and group-fairness of the reconstruction error. We do so by focusing on two commonly used imputation strategies that are amenable to tractable analyses.

In the literature, the choice of imputation is guided by improving expected reconstruction quality, i.e., reducing the expected reconstruction error between the imputed value and ground truth. Yet, this expected reconstruction quality depends on the missingness process, and thus cannot be assessed in practice. The complexity in identifying the appropriate missingness assumptions and, therefore, optimal imputation for a given problem has led practitioners to often use simple imputation strategies such as mean imputation [236] — in which missing values are imputed with the mean of the observed data. However, while mean imputation is easy to implement, this strategy assumes (i) a MCAR process and (ii) a homogeneous population. Because these assumptions are unlikely to hold, intuition often leads researchers and practitioners to use mean imputation in subgroups of the population, as this is expected to better capture heterogeneity and improve reconstruction quality for subgroups. This *group-specific* mean imputation is a common imputation strategy [42, 64, 235], also used when fairness is a primary consideration [135]. Under this approach, each missing value is imputed with the mean values of the subgroup to which the missing datapoint belongs.

Motivated by these practices, we analyse group-specific mean imputation and compare it with population mean imputation. We first show how the reconstruction error under different

imputation strategies can be expressed in terms of the properties of the data and of the missingness process. A direct implication of this result is that, given an observed dataset, the problem of choosing the imputation strategy that reduces the fairness gap in reconstruction error is under-determined. We then show that under a substantial set of missingness patterns, group-specific mean imputation can actually hurt group fairness in reconstruction quality, compared with population mean imputation. In other words, stratifying by group membership can counter-intuitively deteriorate the reconstruction quality in the marginalised group that one aims to protect (Theorem 3.2) and increase the gap in reconstruction error between groups (Theorem 3.3). This is a novel finding that results from expressing the reconstruction error as a function of group-specific properties of the missingness process. We empirically study this in Section 3.5 and Section 3.6, and further show that this same phenomenon also occurs when the performance of imputation strategies is measured in terms of downstream prediction performance.

3.4.1 Problem setting

Population mean imputation replaces any missing value in a covariate with that covariate's mean, observed at the population level. The group-specific extension of this strategy replaces each missing value with the mean computed in the group to which the datapoint belongs. Formally, population mean imputation and group-specific mean imputation associate the imputed value $\tilde{x}_i$ to the point i as follows:

$$\tilde{x}_i^{pop} = \begin{cases} \mu^O & \text{if point } i \text{ is missing, i.e., } o_i = 0 \\ x_i & \text{if } o_i = 1 \end{cases}, \qquad \tilde{x}_i^{group} = \begin{cases} \mu_{g_i}^O & \text{if } o_i = 0 \\ x_i & \text{if } o_i = 1 \end{cases}$$

where μ^O is the observed population mean, $\mu^O = \frac{\sum_i o_i x_i}{\sum_i o_i}$, and μ_g^O is the observed group specific mean, $\mu_g^O = \frac{\sum_{i \in P_g} o_i x_i}{\sum_{i \in P_g} o_i}$. $P_g = \{i \mid g_i = g\}$ are the indices of g-members, with o_i indicating if point i is observed and g_i indicating group membership. Note that dimensions are omitted in this notation as these imputation strategies treat each covariate independently.

Given any group g, we consider two metrics of interest proposed in [151]. First is the reconstruction error for group g under imputation strategy $\mathcal{I}$, $L_g^{\mathcal{I}}$. Second is the reconstruction *error gap* between this group's error and the error in the rest of the population, $\Delta_g(\mathcal{I})$. Formally, these quantities are defined as follows:

Definition 3.1 (Reconstruction error)**.** The reconstruction error of an imputation strategy $\mathcal{I}$ in a group g is the average distance between the underlying true x_i and imputed values $\tilde{x}_i^{\mathcal{I}}$ over all missing data for that group:

$$L_g^{\mathcal{I}} = \mathbb{E}_i\Big[||\tilde{x}_i^{\mathcal{I}} - x_i||_2^2 \mid g_i = g, o_i = 0\Big] \tag{3.1}$$

An optimal imputation strategy assigns the true value to any missing value, resulting in reconstruction error $L_g = 0$.

To quantify algorithmic fairness, Jeong, Wang, and Calmon [151] adapt the equal performance definition introduced in Definition (2.1) to the imputation context:

Definition 3.2 (Equal Imputation Performance)**.** An imputation strategy $\mathcal{I}$ is fairer than another $\mathcal{J}$ with regards to group g if its absolute performance gap is the smaller, i.e. $|\Delta_g(\mathcal{I})| < |\Delta_g(\mathcal{J})|$, where $\Delta_g(\mathcal{I}) := d(\mathcal{I}(\{X_i\}_{G_i=g})) - d(\mathcal{I}(\{X_i\}_{G_i \neq g}))$ for some performance metric d, and (X_i, G_i) the associated covariates and group for patient i.

Connecting this measure of algorithmic fairness with reconstruction error leads to the measure of the reconstruction error gap, defined by the difference in a group's reconstruction error compared with the rest of the population:

$$\Delta_g^{\mathcal{I}} = L_g^{\mathcal{I}} - L_{\neg g}^{\mathcal{I}}$$

Under this definition, imputation strategy $\mathcal{I}$ is considered fairer than another $\mathcal{J}$ if its reconstruction gap is smaller, i.e., $|\Delta_g^{\mathcal{I}}| < |\Delta_g^{\mathcal{J}}|$. A null gap reflects equal error across groups. If $\Delta_g^{\mathcal{I}} > 0$, then this means group g has a larger reconstruction error than the rest of the population, a group fairness concern. Throughout our analysis, we compare the reconstruction errors of each group under different strategies, as well as the resulting performance gaps.

3.4.2 Relationship between missingness patterns and imputation strategies

Equipped with the two key metrics of group's reconstruction error $L_g^{\mathcal{I}}$ and group's reconstruction gap $\Delta_g^{\mathcal{I}}$, we now investigate the impacts of the imputation strategy $\mathcal{I}$ and the missingness process on these measures of algorithmic fairness. In Theorem 3.1, we express each reconstruction error, L_g^{group} and L_g^{pop}, as a function of (i) the underlying distribution of the covariate, and (ii) the missingness process. The missingness process is measured through α_g, the observation rate, and ρ_g, the correlation between observation indicators and the covariate values. For the same observed data, the studied imputation strategies can result in largely different reconstruction errors because of the missingness process and the influence of the underlying distribution.

Theorem 3.1 (Group and population mean imputations' reconstruction error)**.** *Assuming i.i.d. data points* $\{x_i\}$*, one can express the reconstruction error in group* g *resulting from group mean imputation as:*

$$L_g^{group} = \left(\overbrace{-\frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} \cdot \rho_g}^{B_g^{group}} \cdot \sigma_{X|G=g} \right)^2 + \sigma^2_{X|O=0,G=g} \tag{3.2}$$

where the missingness process is represented through (i) $\rho_g = \mathit{Corr}(O, X \mid G = g)$*, the unobserved correlation between the observation indicator and the ground truth covariate values, and (ii)* $\alpha_g = \mathbb{E}[O \mid G = g]$*, the observation rate in group* g*, which is observable. Other values impacting the reconstruction error are reflective of the underlying covariate distribution. This includes* $\sigma^2_{X|G=g} = \mathit{Var}(X \mid G = g)$*, the ground truth variance of the covariate in the group* g*; and* $\sigma^2_{X|O=0,G=g}$*, the variance of the unobserved values of this same group.*

Under the same assumptions, one can compute the reconstruction error in a group g *using population mean imputation as a function of the first term of Equation* (3.2)*, denoted as* B_g^{group}*:*

Group imputation bias

$$L_g^{pop} = \left(B_g^{group} + \mu_g^O - \mu^O \right)^2 + \sigma^2_{X|O=0,G=g} \tag{3.3}$$

Difference between group- and population- observed means

Detailed proof of the theorem is provided in Appendix A.1, Proof A.1.1.

Intuition. These reconstruction error expressions highlight that (i) reconstruction errors are functions of the missingness process and the data distribution, (ii) population mean reconstruction error is a function of group mean reconstruction error.

First, the expressions in Theorem 3.1 illustrate the dependence of reconstruction error both on the unobserved missingness process and the underlying covariate distribution. From the distribution, the covariate's standard deviation $\sigma_{X|G=g}$ directly affects the magnitude of the error because the more spread out the unobserved values are, the less accurate constant imputation strategies are. The distribution further influences the reconstruction error through the variance of the covariate's unobserved values $\sigma^2_{X|O=0,G=g}$, equally increasing the expected reconstruction errors for both imputation strategies.

Crucially, the expressions also show the influence of the missingness process on reconstruction error. The observed rate of missingness, α_g, affects the magnitude of the error as more extreme observation rates increase reconstruction error. The correlation ρ_g also plays a central role in imputation quality. For instance, if the missingness process follows (S1), the marginalised group presents a different rate of uniformly distributed missing data, which leads to $\rho_g = 0$ since data is MCAR in each group. Under (S2) and (S3) though, this correlation depends on the specific application, encompassing any possible value. Consider an example following (S2) in which general practitioners would only weigh patients based on a recorded family history of obesity. Similarly, consider an instance of (S3) in which practitioners only weigh patients that they perceive to be likely overweight. These two examples lead to a negative correlation between missingness and observed values as the patients with recorded weight are more likely to have an above-average weight.

Variance of unobserved values

Standard deviation

Missingness process

Thus, while these quantities express the quality of the studied imputation strategies, the reconstruction errors depend upon the characteristics of the observational process and underlying distribution. These quantities cannot be estimated from observed data alone.

Second, the population reconstruction error is a function of the group mean reconstruction error (L_g^{pop} depends on B_g^{group}, i.e., the first term from L_g^{group} expression). The only difference in the reconstruction error of the two imputation strategies is driven by the **difference between group- and population- observed means**. This expression underlines the balance between the difference in what each strategy imputes ($\mu_g^O - \mu^O$) and the group-imputation errors (B_g^{group}). Importantly, a larger absolute correlation ρ_g always increases the group imputation reconstruction error. However, this is not true in population reconstruction error in which the difference in means may counteract an increase in B_g^{group}.

Consider an example from (S3) in which practitioners only record the weights of underweight patients. In this case, ρ_g is negative and subsequently, B_g^{group} is positive. Further in this example, the underweight group presents a *smaller* mean weight than the overall population, i.e., $\mu_g^O - \mu^O < 0$. Then, the co-occurrence of these two phenomena results in $L_g^{group} > L_g^{pop}$ if $|B_g^{group}| > |\mu_g^O - \mu^O|$, and thus the population mean imputation outperforms the group mean imputation for group g.

This example shows that stratifying on group membership may hurt performance. We further investigate and formalise this hypothesis in Theorem 3.2 and 3.3 by exploring the reconstruction errors and resulting fairness gaps of these imputation strategies.

3.4.3 Fairness comparison between group-specific imputation and population imputation

The assumed superiority of group-specific imputation strategies with respect to reconstruction error relies on an oversimplification of the missingness process, assuming independence between missing values and the observational process, i.e., $\rho_g = 0$. When this assumption holds, the previous theorem shows that group-specific imputation will indeed have smaller reconstruction errors. However, as discussed earlier, this simplifying assumption is unlikely to hold under clinical missingness, where the correlation between the missing values and the missingness indicator can be positive or negative depending on the underlying covariate distribution and missingness process.

In this subsection, we illustrate that this is not an innocuous assumption. When violated, it affects group fairness, measured by the group reconstruction error and by the gap between the reconstruction errors in the marginalised group g and the rest of the population ($\Delta_g^{\mathcal{I}}$). Theorems 3.2 and 3.3 show that the advantage of group-specific imputation is not guaranteed under clinical missingness. Group mean imputation can actually lead to a larger reconstruction error than the population mean imputation for group g: $L_g^{group} > L_g^{pop}$ and a wider fairness

gap in reconstruction error: $\Delta_g^{group} > \Delta_g^{pop} > 0$. These findings question the assumed benefits of group-specific imputation on algorithmic fairness under complex missingness patterns. While practitioners aim to reduce the reconstruction error gap by controlling on group membership, Theorems 3.2 and 3.3 show there are settings where this practice may either increase the fairness gap or further harm the population they try to protect, where harm is defined as a worse reconstruction error for that group. In Section 3.5, we complement this result by empirically showing that such settings are not edge cases. The proof of Theorems 3.2 and 3.3 are presented in Appendix A.1, Proofs A.1.2 and A.1.3.

Theorem 3.2 (Comparison of group and population mean imputations' reconstruction error)**.** *The group reconstruction error resulting from group mean imputation is larger than the one resulting from population mean imputation, i.e.* $L_g^{group} > L_g^{pop}$*, iff one of the following conditions holds:*

Missingness process

$$\rho_g \cdot \frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} < \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} < 0 \text{ or } 0 < \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} < \rho_g \cdot \frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} \tag{3.4}$$

Distribution characteristics

Intuition. Theorem 3.1 introduced a connection between the reconstruction error, the covariate distribution, and the missingness process; Theorem 3.2 now explicitly analyses when group imputation counter-intuitively worsens the reconstruction error for a given group g in comparison to population imputation.

Consider the case when the observable group mean is larger than the population ($\mu_g^O > \mu^O$). In this setting, only the inequality on the right side of Theorem 3.2 can be satisfied. If the observational process, characterised by ρ_g and scaled by α_g, is sufficiently positive and pronounced relative to the normalised difference in means, then population mean imputation results in smaller reconstruction errors than group imputation. This is because the missing values are, on average, smaller than the observed group mean when the correlation is positive. Thus, the missing values are closer to μ^O than μ_g^O, resulting in the superiority of population mean imputation ($L_g^{group} > L_g^{pop}$). The more positively pronounced the observational process is, the more different the two imputation strategies can be where population imputation would still be superior. On the other hand, if the correlation ρ_g is negative, the missing values are on average larger than both the observed group and population means, and therefore consistently closer to the group mean. Thus, group imputation would then always be better in the case of negative correlation. (Similar reasonings follow when $\mu_g^O < \mu^O$).

Theorem 3.3 (Comparison of group and population mean imputations' fairness gaps)**.** *Under the simplifying assumptions* $\sigma^2_{X|\neg O,G} = \sigma^2_{X|\neg O,\neg G}$*, and* $\mu_g^O > \mu^O$*, both imputation strategies*

penalise the marginalised group and the reconstruction gap is larger for the group imputation than the population one (i.e., $\Delta_g^{group} > \Delta_g^{pop} > 0$) iff:

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) > \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot h(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot h(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

or

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) < \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot h(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot h(\alpha_{\neg g}, 1 - r_g, \alpha_g) < ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

with $r_g = \mathbb{P}[G = g]$, the ratio of the population belonging to group g, $f(\alpha_g, r_g, \alpha_{\neg g}) = \frac{2\alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1-r_g) - \alpha_g r_g)$, $e(\alpha_g) = \sqrt{\frac{\alpha_g}{1-\alpha_g}}$, and $h(\alpha_g, r_g, \alpha_{\neg g}) = \frac{\alpha_g r_g + \alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1 - r_g) - \alpha_g r_g)$.

Intuition. We are concerned about imputation harming group fairness. Consider a group g negatively impacted by imputation, where imputation increases its reconstruction error in comparison to the rest of the population ($\Delta_g > 0$). Under this consideration, Theorem 3.3 proves there exist settings where using population imputation reduces the absolute fairness gap $|\Delta_g|$ compared to group imputation. Appendix Figure A.1 illustrates an example in which these inequalities are satisfied.

Multiple combinations of missingness processes and underlying covariate distributions can lead to the same observed data. Some of these would satisfy the inequalities of Theorem 3.3, and others not. Knowledge of the observational process is, therefore, necessary to resolve the otherwise *under-determined* problem of identifying the most beneficial imputation strategy for the protected group. If this process is unknown, practitioners must acknowledge missingness assumptions and remember that reconstruction error properties demonstrated under MCAR settings do not generalise to more complex patterns.

While these theoretical results focus on two simple, mathematically-tractable imputation strategies, they underline how imputation choice is less intuitive than it seems. Importantly, controlling for group membership may be counter-intuitively harmful. Indeed, this theme is something that we investigate empirically in simulation and real-world data in the rest of this chapter, seeing evidence that this finding also holds when imputation performance is measured with respect to downstream prediction performance and fairness.

3.5 Empirical evidence of the impact of imputation on algorithmic fairness

The previous sections identify group-specific patterns of clinical missingness, and show that these may translate into disparities in reconstruction errors. In this section, we provide evidence of this phenomenon through simulations. First, we introduce the data generation in Section 3.5.1 and the handling of missingness in Section 3.5.2. By using simulations, we are able to control the underlying clinical missingness patterns and measure the impact of imputation on reconstruction errors for each group and the gap between them, as shown in Section 3.5.3. Imputation choices may also impact prediction disparities, thus we study the effect of different imputation strategies on algorithmic fairness of downstream predictions in Section 3.5.4. Figure 3.3 summarises how our experiments study algorithmic fairness concerns in both reconstruction error and downstream prediction performance. For reproducibility, code for all experiments is available on Github[1].

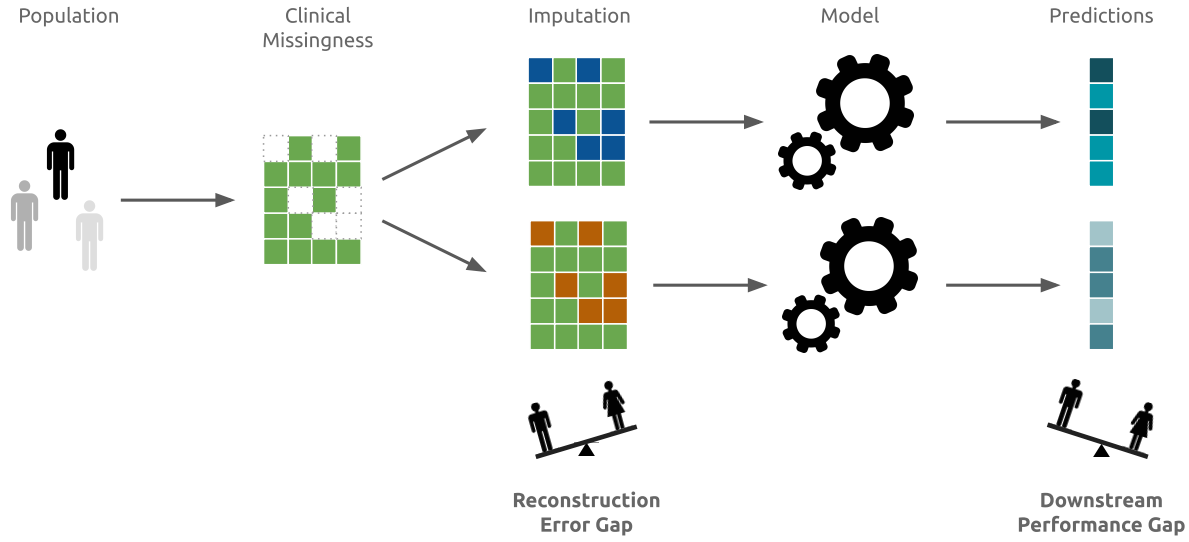


Figure 3.3: Impact of different imputation strategies on algorithmic fairness in the ML pipeline, given a population marked by group-specific missingness patterns. This chapter measures algorithmic fairness at two levels: (i) imputation, i.e., how different imputation strategies impact the quality of the reconstructed data for different groups, (ii) prediction, i.e., how different imputation strategies impact the downstream gap in performance.

3.5.1 Data generation

Our experiments rely on a population of N patients with associated covariates X, marginalised group membership G, and outcome of interest Y. Assume a simulated population consisting of a bivariate covariate set ($X \in \mathbb{R}^2$), with $N = 101,000$ individuals divided into two groups

[1] `https://github.com/Jeanselme/ClinicalPresenceFairness`

($G \in \{0, 1\}$), and consider the marginalised group ($G = 1$) is a minority in the population with $1,000$ patients for $100,000$ in the majority. We assume the two groups differ in condition manifestation, i.e., positive cases across groups differ in how they express the condition in the covariates X. Both groups present the same condition prevalence, with $2/3$ of the population presenting the condition. To enforce a difference in condition manifestation, the covariates associated with the negatives ($Y = 0$) from both groups are drawn from a shared bivariate normal distribution, while the covariates of patients affected by the condition are sampled from different bivariate normal distributions, depending on their group membership. This simulation therefore consists of three clusters illustrated in Figure 3.4, and the associated predictive task is to classify between positive and negative cases. (See Appendix A.2.1 for full data generation protocol and further simulations).

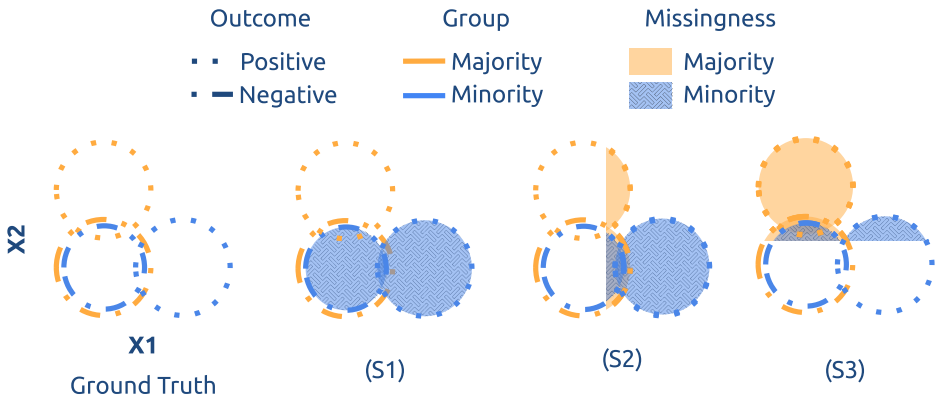


Figure 3.4: Graphical summary of clinical missingness in the simulation experiments. Missingness is enforced on X_2, affecting 50% of the shaded regions for the indicated group.

We then enforce the three clinical missingness patterns introduced in Section 3.3, by masking covariates on the second dimension X_2. The proposed missingness processes reflect the DAGs introduced in Figure 3.2: the missingness process has a direct dependence on G in (S1), while an indirect one in (S2) and (S3) through the difference in condition manifestation in these groups. Figure 3.4 provides a graphical summary of how clinical missingness is enforced on the synthetic data. Note how the different scenarios result in group-specific missingness patterns affecting group-specific clusters.

Formally, 50% of the data in the second dimension is removed in a given subgroup to enforce the three clinical presence scenarios as follows:

- Limited access to quality care (S1): $O_2 \mid [G = 1] \sim \text{Bernoulli}(0.5)$
- (Mis)-informed collection (S2): $O_2 \mid [X_1 > 0.5] \sim \text{Bernoulli}(0.5)$
- Confirmation bias (S3): $O_2 \mid [X_2 > 0.5] \sim \text{Bernoulli}(0.5)$

3.5.2 Handling missingness

Commonly, missing data are imputed before applying traditional statistical and ML models. While the theoretical results of Section 3.4 focus on two mean imputation strategies, our experiments also explore if these conclusions generalise over a larger set of common strategies to deal with missingness:

Population mean imputation (Mean). Missing data are replaced by the population mean of each covariate.

Conditional group-mean imputation (Group Mean). Missing data are replaced by the group mean of each covariate.

Multiple Imputation using Chained Equation (MICE). Missing data are iteratively drawn from a regression model built over all other available covariates after median initialisation. This approach is repeated I times with an associated predictive model for each imputed draw. At test time, the same imputation generates I imputed points for which models' predictions are averaged. MICE is recommended in the literature [143, 231, 349, 355, 381], based on the argument that it quantifies the uncertainty associated with the missingness process through Rubin's rules [273]. In the experiments, we used 10 iterations repeated 10 times, resulting in $I = 10$ datasets with associated predictive models.

Group MICE. The previous MICE strategy assumes a MAR mechanism. To make such an assumption more plausible, Haukoos and Newgard [123] recommend the addition of potentially informative covariates. In our experiment, we therefore adjust the regression for both group membership and all available covariates for imputing the missing data.

Group MICE Missing. Encoding missingness has been shown to improve performance when the patterns of missingness are informative [116, 190, 276, 304]. As clinical missingness can contain informative patterns [190], we concatenate missingness indicators to the imputed data from Group MICE (Appendix A.2.1 explores the concatenation of missingness indicators with the other strategies).

3.5.3 Impact on reconstruction error gaps

We analyse the impact of imputation on reconstruction quality over 100 repetitions of the proposed simulations. The three identified clinical presence scenarios are independently enforced. As the underlying distribution and missingness process are known, we can compute the reconstruction error gaps. Figure 3.5 (left panel) presents the gap in reconstruction error,

$\Delta_g^{\mathcal{I}}$, between the marginalised group and the rest of the population. A larger absolute value corresponds to a larger difference in data quality for each group. A negative sign expresses a better quality of data produced for the marginalised group relative to the majority, while a positive sign favours the majority. Figure 3.5 (right panel) presents the overall reconstruction errors as well as the reconstruction errors for each group. The results are presented for each imputation strategy across the three clinical missingness scenarios. In this figure, the dark-coloured bars represent group-specific imputation strategies, while the light-coloured bars are population-wide imputation strategies (MICE or Population Mean). Figure 3.5 empirically illustrates the theoretical results presented in Section 3.4 and provides new insights.

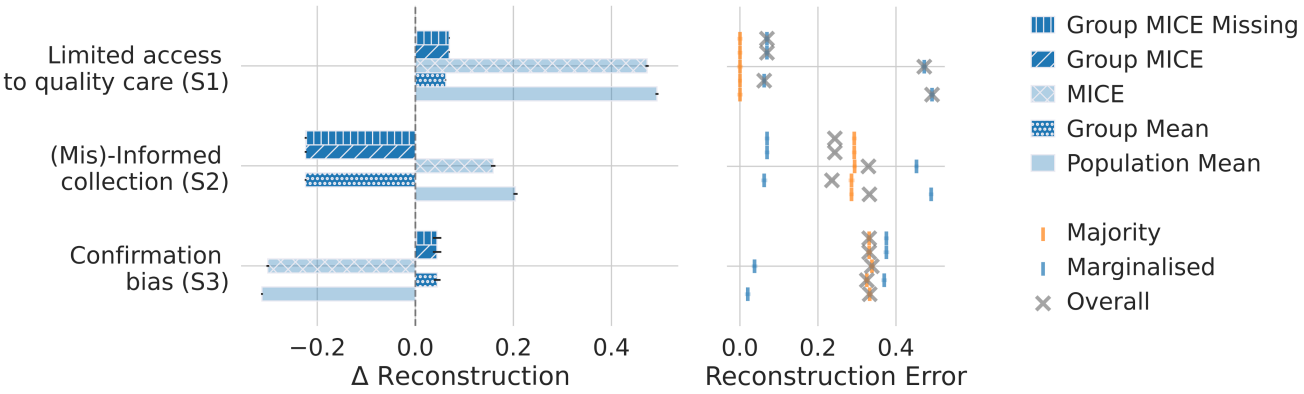


Figure 3.5: Reconstruction error gap (on the left) and group-specific reconstruction errors (on the right) across scenarios on 100 synthetic experiments. If $\Delta < 0$, the marginalised group has a smaller reconstruction error than the majority. Lower reconstruction error is better.

Insight 1.1: Different imputation strategies may have similar reconstruction errors at the population level while having different group reconstruction gaps. The right panel in Figure 3.5 shows that, for scenario (S3), multiple imputation strategies have the same overall reconstruction errors (marked by grey crosses) but different subgroups' errors (in colour). Consequently, different reconstruction gaps arise for different strategies under the same observed data, as shown by the left panel. This observation echoes the results from Theorem 3.1, showing how two imputation strategies may result in different group reconstruction errors.

Insight 1.2: No imputation strategy consistently outperforms the others across clinical presence scenarios. Figure 3.5 shows that under (S1), group imputation presents minimal reconstruction gaps (left panel), and minimal overall reconstruction error (right panel). In this context, group imputation strategies should be preferred. However, the choice is less clear in (S2), as group imputation strategies present overall lower reconstruction errors but larger absolute reconstruction gaps compared with their population imputation alternatives. Specifically, if practitioners prioritise algorithmic fairness, MICE imputation should be preferred under (S2), as it has the smallest absolute gap in reconstruction error. On the other hand,

if the overall error is the main concern, group mean imputation should be preferred. As in Theorems 3.2 and 3.3, this observation underlines that no imputation from those considered either consistently improves algorithmic fairness or consistently reduces overall reconstruction error across different missingness processes.

Insight 1.3: Current recommendations for group-specific imputation can increase the reconstruction gap and yield a worse reconstruction error for the marginalised group. Note how in the simulation (S2), the group mean imputation presents a lower marginalised group reconstruction error than its population alternative, as illustrated by the blue vertical bars on the right panel of Figure 3.5. However, an opposite observation can be made under (S3), in which the population mean imputation presents a smaller reconstruction error for the marginalised group. This means that group imputation benefits the marginalised group in (S2) but negatively impacts this group under (S3). This empirically illustrates the results of Theorem 3.2 demonstrating that group imputation may have adverse effects on the marginalised group that one aims to protect. Further, the reconstruction gaps in the left panel show that the population mean presents a lower reconstruction gap under (S2) but a larger one under (S3) in comparison to the group mean imputation. This observation echoes Theorem 3.3 showing that group imputation may present a larger fairness gap under some observational processes. Importantly, Group MICE and MICE present similar trends as group mean and population mean across (S2) and (S3), indicating that these results extend beyond mean imputation.

3.5.4 Impact on downstream algorithmic fairness

While reconstruction error gaps demonstrate how imputation impacts data quality, its estimation involves knowledge of the missingness process. Furthermore, imputation is often an intermediate step towards a different end goal, such as predictive modelling. Different imputation strategies impact the data available for modelling and therefore, may impact performance and algorithmic fairness in downstream predictions. Researchers and practitioners commonly explore algorithmic fairness in predictions to assess and mitigate the risks associated with a model. We propose to similarly measure how the choice of imputation affects downstream group-specific performance and potentially reinforces disparities in data marked by clinical missingness.

To this end, we complement the previously described imputation pipelines with a logistic regression. The prediction task is to differentiate between positive and negative cases. Note that we explore a single model since our goal is not to quantify how different prediction models may mitigate disparities in data quality; instead, we want to assess the downstream impact of imputation strategies on prediction.

Analogous to the quantification of algorithmic fairness in reconstruction error, we adopt the *equal predictive performance across groups* definition of algorithmic fairness ([263], see Definition 3.2). We use the AUC, defined in Section 2.2.3, as metric d in Definition 3.2 as the

outcome is binary in this chapter. The AUC measures each group's discriminative performance and is commonly used as a measure of algorithmic fairness in ML for healthcare [174, 267, 371]. A smaller absolute gap corresponds to less difference in discriminative performance between groups. A negative gap corresponds to a worse-performing model for the marginalised group. Figure 3.6 presents the gap between the majority and the marginalised group's performance (left panel) and the group-specific AUCs (right panel). These results are computed on a 20% test set and averaged over the 100 simulations.

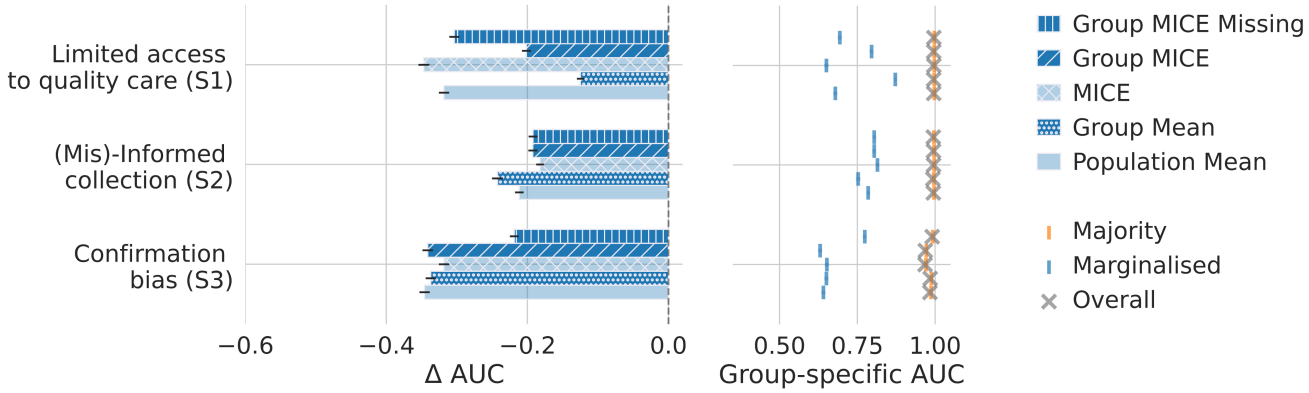


Figure 3.6: AUC performance gaps Δ (on the left) and group-specific AUCs (on the right) across scenarios on 100 synthetic experiments. If $\Delta < 0$, the marginalised group has a lower AUC than the majority. A higher AUC is better.

This analysis on downstream performance echoes the insights described in Section 3.5.3. Imputation strategies impact the data distribution available for modelling. Group-specific reconstruction qualities are then transposed into performance differences.

Insight 2.1: Different imputation strategies may result in similar downstream performance at the population level while having different group performance gaps. Figure 3.6's right panel indicates that all imputation strategies present similar AUC performances at the population level. They also present similar performances for the majority. However, performance differences for the marginalised group can be stark. Note how the AUC evaluated on the marginalised group under (S1) presents a gap of more than 0.1 between the population and group mean imputation strategies. Similarly to Insight 1.1 for reconstruction errors, equal overall performances do not translate into equal group-specific performances. This observation is also observed for the other strategies, as shown by the similar gap for MICE imputation strategies.

Insight 2.2: No strategy consistently outperforms the others across clinical presence scenarios. When considering the AUC gaps presented in Figure 3.6, note how the smallest gap is achieved by Group Mean in (S1), MICE in (S2) and Group MICE Missing in (S3). Despite similar overall performances, different imputation strategies should be preferred under

these three scenarios to reduce the performance gap. While the result of which imputation strategy is best for each scenario is specific to this simulation, this exemplifies how no strategy consistently reduces the performance gap across groups. Importantly, different imputation strategies should be preferred under the same underlying distributions but different missingness processes.

Insight 2.3: Current recommendations for group-specific imputation and use of missingness indicators can increase the performance gap and yield a worse performance for the marginalised group. In Figure 3.6's left panel, Group MICE presents a *larger* performance gap than MICE in (S3). Controlling for group membership to make the MAR assumption more plausible is not always suitable as it may add *noise*, or impact data distributions. The resulting distributions may be less separable by the model and, therefore, lead to poorer performance. Likewise, see how the model considering missingness provides an edge in (S3) compared to Group MICE, but hurts algorithmic fairness and marginalised group performance in (S1). This observation reinforces the necessity of conducting a sensitivity analysis on the choice of imputation. Additionally, it underlines how the benefit of controlling on relevant covariates depends on the missingness process.

Through simulations, this section provides empirical evidence for rejecting a series of hypotheses that underlie assumptions and recommendations made by common practice for handling missingness. A summary of these hypotheses is presented in Table 3.1. First, overall equal performance does not correspond to equal subgroups' performance; two imputation strategies may yield the same overall performance but different algorithmic fairness properties. Second, which imputation has superior performance depends on the missingness process, and therefore, there is no consistent "best" strategy across settings. Third, group imputation strategies can counter-intuitively increase reconstruction errors and reduce downstream predictive performance. Lastly, group imputation strategies do not necessarily reduce the fairness gaps. These observations converge towards the necessity of measuring the impact of different imputation strategies on fairness properties' robustness.

3.6 Case study: Handling missingness in MIMIC III

Section 3.5 uses simulations to illustrate the algorithmic fairness consequences of imputation on both reconstruction error and downstream performance. To further demonstrate that these insights have important practical implications, we show evidence of these concerns in a real-world setting, on a widely used observational dataset.

Hypotheses	Imputation quality	Predictive performance
Equally performing approaches at the population level have similar algorithmic fairness properties	✗	✗
Imputation properties are consistent across missingness processes	✗	✗
Controlling/stratifying on group results in improved group performance	✗	✗
Controlling/stratifying on group reduces group disparities	✗	✗

Table 3.1: Summary of the refuted hypotheses by the proposed simulations.

3.6.1 Dataset and empirical setup

We model short-term survival using the laboratory tests from the widely studied Medical Information Mart for Intensive Care (MIMIC III) dataset [155]. Following data harmonisation (as in [338]), we select adults who survived 24 hours or more after admission to the intensive care unit, resulting in a set of 36,296 patients sharing 67 laboratory tests. The goal is to predict short-term survival (7 days after the observation period — Y) using the most recent value of each laboratory test observed in the first 24 hours of observation (X). We select short-term survival as it is a standard task in the ML literature [220, 322, 360] and the associated labels are less likely to suffer from group-specific misdiagnosis, and, therefore, disentangles our analysis from potential biases in labelling.

There is a large amount of missing data in MIMIC III data because all possible laboratory tests are not performed. Using the same imputation strategies presented in Section 3.5.2, we first impute missing data, resulting in $\tilde{X}$, to enable modelling. After this preprocessing, our analysis consists of a logistic regression model —a pillar in the medical literature [108, 234]— to discriminate between positive and negative cases ($\text{logit}(Y) \sim \tilde{X}$). To avoid overfitting, we use a logistic regression with the strength of the $L2$ penalty, a tuning parameter selected through cross-validation among $[0.1, 1, 10, 100]$. Patients are split into three sets: 80% for training, 10% for hyper-parameter tuning and 10% for testing.

This model could have important consequences on patients' lives. Deploying this model can inform care prioritisation of patients with predicted elevated risks. Thus, ensuring equitable care is central to this problem. Following the same evaluation procedure as in the simulations, we measure the gap between group AUC performance as a group fairness metric.

However, this metric does not directly quantify how deployment can hurt subgroups if a hospital were to use a fixed allocation of resources to treat patients (such as beds or staff limits). Thus, as a second algorithmic fairness metric, we assume the availability of priority care for 30% of the population, and evaluate the False Negative Rate (FNR) given this resource

threshold. In other words, higher FNR in this setting measures how often the model incorrectly deprioritises high-risk patients. A gap in FNR between groups illustrates differences in the rates of missed patients between groups. (Additional experimental design descriptions and results are provided in Appendix A.2. Sensitivity to the prioritisation threshold is analysed in Appendix A.2.2).

3.6.2 Downstream algorithmic fairness consequences

We find that imputation has algorithmic fairness consequences in real-world medical data. The presented experiments focus on groups defined by the following attributes: ethnicity (Black vs non-Black)[2], sex (female vs male), and insurance (publicly vs privately insured)[3]. We first investigate if MIMIC presents distinct group-specific missingness patterns. Then, we explore the impact of imputation strategies on downstream predictive performance, both at the population and group levels, echoing Insights 2.1 to 2.3. This analysis provides the following insights that parallel and enrich the synthetic experiments' results.

	Orders		Distinct tests			Orders		Distinct tests	
Alive$^+$	5.68 (4.64)	$*$	40.80 (6.73)	$*$	Female	5.54 (4.45)	$*$	40.75 (6.89)	$*$
Dead$^+$	7.57 (5.44)		37.22 (7.50)		Male	6.03 (4.91)		40.41 (6.80)	
Black	5.24 (4.08)	$*$	40.94 (6.94)	$*$	Public	5.67 (4.57)	$*$	40.46 (6.76)	$*$
Other	5.86 (4.77)		40.52 (6.84)		Private	6.11 (5.01)		40.75 (7.01)	

$^+$ By the 8^{th} day after admission. $*$ Significant t-test p-value (< 0.001).

Table 3.2: Mean (std) number of orders and observed tests performed during the first post-admission stratified by marginalised groups and outcomes.

Insight 3.0: Real-world data presents group-specific clinical missingness. MIMIC data collection is the product of structured guidelines, medical intuition and prioritisation rules used in intensive care units. This process likely consists of an *unknown* mixture of the clinical missingness scenarios described in Section 3.3 that may impact subgroups differently. While the causes of clinical missingness *cannot* be conclusively determined from observational data alone, one can examine missing data and identify evidence of group-specific patterns. Table 3.2 shows the number of orders and the number of distinct laboratory tests (out of the 67 possible tests) performed during the first day post-admission for different subgroups.

In Table 3.2, we observe that there are a larger number of orders for patients who die during their stay compared with the ones who survive, and on average, fewer distinct tests.

[2]MIMIC's reported ethnicity contains 40 different categories, several referring to Black ethnicities such as Black African American, Black Haitian or Black African, with Black African American representing 91.8% of all Black ethnicities. In our analysis, for computing the proposed fairness metrics considering two groups, we consider the binarisation: all Black ethnicities as one group vs the rest of the population.

[3]For our analysis, we considered private insurance versus all other types of reported insurance.

This pattern could be consistent with a possible *confirmation bias* scenario (S3). Doctors may monitor more closely sicker patients, or patients with conditions with a higher risk of mortality. Another example of non-random missingness is that there are fewer test orders for female, Black, and publicly insured patients, but little difference in the number of distinct tests prescribed. This may be explained by the underlying conditions or other medically relevant factors, which may be different across groups. Importantly, the combination of a similar diversity of tests but less frequent observations results in a less up-to-date patient's health status available for modelling. Thus, even though the cause of testing differences is unclear, these observations show the connection between group membership, testing patterns, and outcomes.

		Group-specific		
Strategy	Population	Ethnicity	Sex	Insurance
Mean	0.745 (0.011)	0.744 (0.011)	0.747 (0.011)	0.745 (0.011)
MICE	0.736 (0.012)	0.738 (0.012)	0.734 (0.012)	0.730 (0.011)
MICE Missing	0.785 (0.010)	0.787 (0.009)	0.787 (0.009)	0.782 (0.010)

Table 3.3: Population-level AUC under different imputation strategies. Mean (std) computed on the test set bootstrapped 100 times.

Insight 3.1: Different imputation strategies may have similar prediction performance at the population level while having opposite group performance gaps. All group-specific mean imputation and population mean perform similarly at the population level, measured in terms of AUC, as shown in Table 3.3. However, Figure 3.7 highlights how these strategies impact marginalised groups differently. Both the amplitudes and signs of the gaps vary across imputation strategies, and this occurs both when considering gaps in AUC and gaps in FNR. Consider the ethnicity partition: these strategies have opposite algorithmic fairness consequences, measured in terms of FNR. Group mean imputation would result in a larger FNR gap, with a smaller FNR for non-Black than Black patients; by contrast, population mean imputation would reduce the gap in FNR, and yield a smaller FNR for Black than non-Black patients. This observation echoes the results from Theorem 3.3 and Insight 2.1, and shows that the group mean imputation may widen the fairness gap in comparison to its population imputation alternative. Additionally, Table A.5 in Appendix echoes Theorem 3.2 with a smaller absolute FNR for Black patients when using the population imputation. Crucially, this highlights how two pipelines, solely differing in their handling of missingness, can harm or favour the marginalised group's performance relative to the rest of the population, drastically shifting a model's algorithmic fairness properties.

Insight 3.2: No imputation strategy consistently outperforms the others across groups. In Figure 3.7, Group MICE Missing imputation would have the smallest AUC gap between ethnicity subgroups. However, note that using this strategy would result in the *largest*

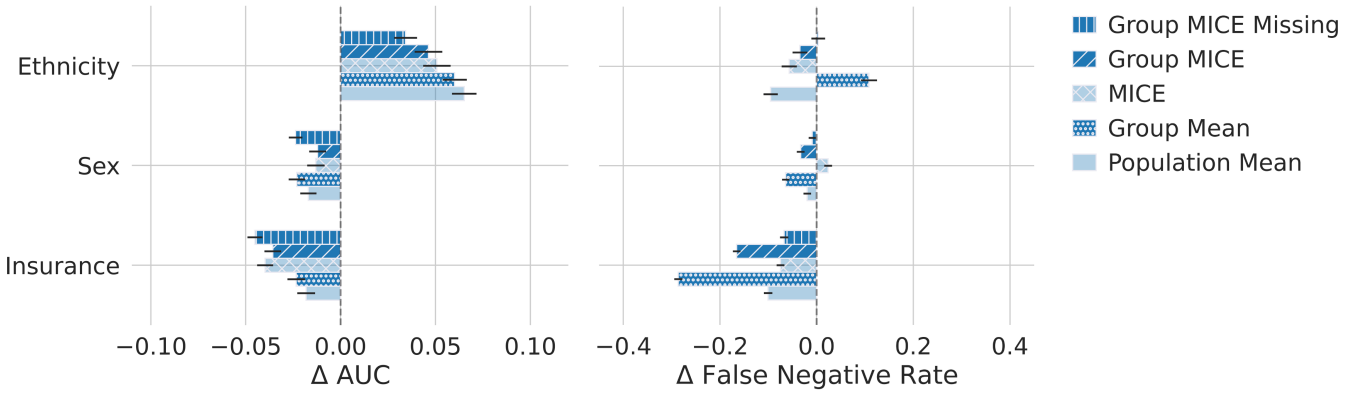


Figure 3.7: Prioritisation performance gaps Δ across marginalised groups in MIMIC III experiment, bootstrapped on the test set over 100 iterations. If $\Delta > 0$, the marginalised group has a larger value of the given metric than the rest of the population.

AUC gap when partitioning the population by insurance type. This observation shows that a given imputation strategy may be the most beneficial in terms of algorithmic fairness when considering one demographic attribute, while being the most detrimental with respect to another demographic attribute. Theorem 3.1 and Insight 2.2 demonstrate how the missingness process impacts the optimal imputation strategy for a given group. Expanding on these findings, this experiment reveals how, if different group partitions present different missingness processes, the optimal imputation strategy can be group-dependent.

Insight 3.3: Current recommendations for group-specific imputation and use of missingness indicators can increase the performance gap and yield a worse performance for the marginalised groups. Echoing Insight 2.3, using Group MICE Missing (containing the missingness indicator) results in a larger AUC gap than using Group MICE for the sex and insurance partitions, as shown in Figure 3.7's left panel. Similarly, in considering insurance subgroups, the group imputation strategies present larger absolute FNR gaps in comparison to their population alternatives. As algorithmic fairness of prediction performance is a function of both the imputation choices and the unobserved missingness processes, common practices can be suboptimal under complex patterns of clinical missingness.

3.7 Discussion

In this section, we provide a summary of our contributions, recommendations, and future research directions stemming from this work.

3.7.1 Contributions

The fairness literature has explored how ML pipelines can reinforce biases present in data. Our work demonstrates how biases may be reflected, not only in what is present, *but in what is absent from the data*. This observation repositions the overlooked imputation step as critical for improved algorithmic fairness.

In the context of ML for health, interactions between patients and the healthcare system can result in group-specific missingness patterns that may impact downstream algorithmic fairness under different imputation strategies. We reviewed historical examples from the literature and characterised three distinct mechanisms leading to group-specific missingness. Our analysis demonstrated that no imputation strategy consistently provides better performances or algorithmic fairness properties, both in terms of reconstruction error and downstream performances. An imputation strategy providing an edge under one missingness process can under-perform in another. We provided evidence that common intuitions rarely hold under complex patterns of missingness. In particular, the imputation practice of controlling for, or stratifying by, group membership may counter-intuitively hurt algorithmic fairness and group-specific performance. Crucially, the algorithmic fairness properties associated with a ML pipeline are dependent upon the choice of imputation strategy and could be reversed by using another strategy. For instance, based on the same data, one imputation strategy could result in a pipeline with similar performance between groups, while another may present strong discrepancies against or in favour of a marginalised group. The experiments conducted using the MIMIC III dataset demonstrated the relevance of the identified problem as more than a merely theoretical concern, showing that group-specific missingness patterns are present in a widely used electronic health record dataset, and the choice of imputation strategy can drastically impact algorithmic fairness properties of a downstream prediction task.

Overlooking the importance of imputation has resulted in the current practice of relying on a single strategy and does not assess whether the algorithmic fairness properties of a predictive model are sensitive to the choice of imputation. Our analysis has critical repercussions on the study of algorithmic fairness in clinical predictive analyses in which practitioners can and must carefully consider the sensitivity of algorithmic fairness to the imputation step. While imputation's reconstruction error and associated fairness gap cannot be evaluated in practice, its downstream predictive implications can be.

Critically, imputation affects both model development and evaluation. In settings where one can assume that the missingness process will remain the same during deployment, experimenting

with different imputation strategies helps in selecting the strategy that yields better fairness conclusions. Meanwhile, if efforts are made to enhance data quality and reduce the amount of missing data at deployment, then assessing the downstream algorithmic fairness properties under different imputation strategies captures the uncertainty regarding the algorithmic fairness properties of the predictive model. In these settings, this assessment is essential for quantifying the confidence in a given analysis's fairness conclusions. Our work illustrates the potential of sensitivity analysis in imputation strategies as a valuable tool when implementing data-driven prediction models in the presence of clinical missingness. We illustrate an example of this predictive group fairness sensitivity analysis in the MIMIC III experiments.

3.7.2 Recommendations

Contrary to current practice, missingness should not be considered as a disconnected problem, but rather as an integral part of our understanding of algorithmic fairness. Learning from medical data without sufficient attention to the potential disparities present in clinical missingness could reinforce and automatise inequities, and further harm historically marginalised groups. Particularly, our work calls for caution in imputation choice and its associated algorithmic fairness consequences, and shows that appropriate handling of missing data is important towards reducing health inequities. We invite model developers to:

1. *Study the missingness process*: if there is access to the upstream process generating the data, explore which assumptions best describe the missingness process. While MCAR and MAR assumptions present attractive theoretical guarantees, they may not apply in real-world clinical settings, as shown in Section 3.3.

2. *State the missingness assumptions*: report assumptions made about the missingness process in preprocessing and modelling. The importance of explicitly stating assumptions underlying imputation done during preprocessing ought to be underscored, as these are rarely reported in applications of ML to medical data. Opacity on these assumptions often leads practitioners to not account for the complexity of the observational process. Clear communication of the underlying assumptions is critical for aligning methodological choices with clinical missingness, and also for defining the analysis' scope and limitations.

3. *Consider differences in the missingness process between training and deployment*: report how the missingness process is assumed to change at deployment time. In some cases, the missingness process may remain the same between the training and the deployment phases. In other cases, the missingness process may differ between these two phases. The latter may be especially true if the deployment of the model constitutes an incentive to collect more complete, higher-quality data. Considering this difference can guide the analysis and interpretation of the impact of imputation strategies on downstream predictions, as we detail in the next point.

4. *Evaluate the impact of different imputation strategies*: measure the impact of different imputation strategies on the algorithmic fairness properties of a predictive model. Report the change in fairness conclusions one would observe under different imputation strategies. If the data quality is likely to improve at deployment, this sensitivity analysis quantifies the uncertainty in the fairness conclusions associated with the *unknown* missingness process. If the missingness process is likely to remain the same, this sensitivity analysis can guide model selection, and practitioners should select the one that achieves the most acceptable trade-off between performance and algorithmic fairness.

3.7.3 Future work

In future work, we identify three axes of potential research. First, clinical missingness is only one dimension of how clinical presence shapes the data-generating process. The interaction between patients and the healthcare system not only imprints missingness, but it may also shape aspects such as the temporality of medical time series, which may similarly convey group-specific disparities that current ML pipelines may amplify. Second, the development of a toolbox for quantification of the uncertainty of the fairness conclusions associated with the missingness process would be valuable in expressing a model's deployment risk. Third, this work focuses on the impact of imputation on algorithmic fairness. Future work can study the interplay between missingness and other stages of the ML pipeline, such as characterising how the model choice may interact with disparities in the missingness patterns.

Chapter 4

Response Heterogeneity

Associated Publications. The work presented in this chapter is based on our publications:

- Neural Survival Clustering: Non-parametric mixture of neural networks for survival clustering [145] presented at CHIL 2022.
- Identifying treatment response subgroups in observational time-to-event data [150].

Problem statement. *How can we uncover patients under-served by current medical practices?*

4.1 Motivation

A disease develops as the sum of endogenous molecular processes and exogenous environmental exposures [335]. It follows that, despite shared characteristics, any disease can encompass considerable heterogeneity in symptoms, responses to treatments, and outcomes, as exemplified in conditions like breast cancer [255], diabetes mellitus [323] and Parkinson's disease [115].

Understanding disease heterogeneity is crucial for enhancing patient outcomes by deepening our comprehension of biological pathways, advancing targeted treatment development, and refining current medical guidelines. For example, the discovery of histopathological and molecular sub-types of breast cancer [246] led to the development of novel targeted therapies that fundamentally changed the prognoses for certain patient populations [127]. Combination treatment designed for HER2-expressing breast cancers resulted in a threefold increase in life expectancy from 2001 to 2015 [208, 297].

Disease subgroup discovery often follows a novel characterisation of a condition. New technologies and medical knowledge lead to the differentiation of subgroups, which researchers then test for differences in outcomes and treatment responses through careful experimentation. This procedure remains the gold standard for exploring the causal relationships between treatments or practices and patient outcomes. However, as mentioned in the introduction,

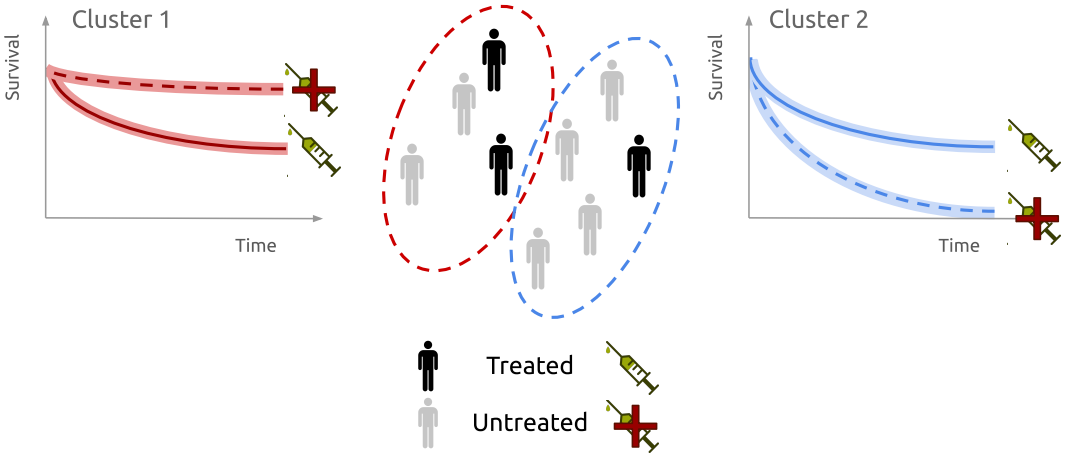


Figure 4.1: Our proposed methods identify subgroups of patients with similar survival profiles and treatment responses to guide clinical practice and design clinical trials. Our methods simultaneously model survival and clustering structure to identify subgroups that best explain the outcome of interest or, at the least, are not exposed to ineffective therapies which may have adverse side effects. In this example, our approach uncovers two clusters with different survival outcomes. The red cluster presents a better average survival but a worse treatment response than the blue one.

clinical trials are costly and time-consuming, with restricted patient cohorts unrepresentative of the real-world diversity of patients and treatment strategies [126]. In such restricted patient cohorts, subgroup analyses may fail to identify subgroups of under-represented populations. Additionally, identifying subgroups in experimental settings often relies on simple linear models on a subset of variables to mitigate the risk of false positive findings from multiple comparisons and false negatives from insufficient statistical power [41, 91], especially in small datasets.

The large amount of available observational data offers extensive and diverse cohorts to explore heterogeneity under real-world practice. However, medical practitioners do not randomise treatment in clinical settings, creating a selection bias in treatment assignments, a clinical presence challenge that we must account for to uncover the causal link between treatment and outcome. For instance, women with breast-conserving surgery are more likely to receive adjuvant radiotherapy than those with more advanced cancer following current guidelines in the United Kingdom [364]. If ignored, this non-randomisation may bias the estimate of patients' response to treatment.

When using observational data, ML approaches often focus on individual patients over groups, focusing on individualised survival and treatment effects modelling [34, 75]. Beyond individualised predictions, the volume of available observational data offers the opportunity to use neural networks and clustering algorithms for uncovering subgroups delineated by non-linear, higher-dimensional combinations of available covariates.

In this chapter, we study how observational data can guide subgroup discovery. When exploring heterogeneity in outcomes, such as death or cancer recurrence, we propose a clustering

strategy based on observed outcomes to identify subgroups with different risk profiles. With available therapeutic options, we further propose to uncover subgroups of patients for whom existing treatments do not lead to improved outcomes. Our work addresses the gap in the literature at the intersection of clustering, survival outcomes and observational studies with the following contributions:

1. We formalise the problem of uncovering heterogeneity as a latent clustering problem guided by the observed survival outcomes.

2. We propose Neural Survival Clustering, a neural network architecture with latent survival profiles, to identify subgroups of patients' outcomes. Then, we introduce its causal extension to identify subgroups of treatment responses in observational settings where treatment may not be randomised to discover subgroups that do not respond to existing treatment options.

3. We evaluate our approaches with a synthetic dataset to show the models' capacity to recover the underlying clustering structure associated with the outcome of interest. To demonstrate the models' real-world relevance, we present an analysis of breast cancer heterogeneity using the SEER dataset[1].

These contributions offer valuable insights for developing future clinical trials, using observational data to identify subgroups that could benefit from novel treatments. The proposed tools present two complementary insights into heterogeneity. Neural Survival Clustering describes the *association* between covariates — potentially including treatment — and observed outcomes, while Causal Neural Survival Clustering explores the *causal relation* between treatment and observed outcomes. Exploring associations describes how the current use of treatment may not benefit all groups equally. Examining the causal relation shows the potential outcome patients would have under a different treatment, resulting in the identification of subgroups for which treatments do not improve outcomes. These tools have the potential to inform the design of policies and guidelines with more tailored treatment recommendations, as group recommendations better align with current guidelines than individualised predictions, which could improve their adoptions.

In Section 4.2, we review the related literature on treatment effect heterogeneity and subgroup analysis before formalising the problem of subgroup discovery in Section 4.3. Then, in Section 4.4, we introduce the proposed model for identifying survival risk profiles and its extension for uncovering treatment response subgroups. Finally, we present a simulation study and applications and an observational breast cancer dataset in Sections 4.5 and 4.6 to demonstrate the relevance of our approach.

[1] `https://seer.cancer.gov/`

4.2 Related work

Survival analysis, introduced in Chapter 2, is often used to model the time to an event of interest following interventions and treatments and, consequently, inform policies and medical guidelines. In this section, we review how practitioners have used different statistical approaches to identify subgroups of patients with different survival and treatment response profiles.

4.2.1 Identifying subgroups in time-to-event data

Low discriminative performance is one of the barriers to the adoption of predictive models [129]. However, understanding the risk and the factors influencing it are even more critical in high-stakes applications [129]. While ML models for survival analysis, described in Chapter 2, may improve performance by capturing complex non-linear dependencies on covariates, this gain often comes at the cost of interpretability. Better alignment of predictive improvement with medical needs would improve medical adoption and actionability. As current guidelines frequently base treatment recommendations on identified risk groups, models identifying such subgroups would enhance actionability and interpretability [69].

Clustering survival outcomes is a natural solution to uncovering risk groups. The literature on this topic can be divided into two families of approaches: (i) post-processing and (ii) joint optimisation. First, one can consider the following steps. Practitioners fit a survival model at the population level and then use the identified predictive covariates to inform clustering. For instance, [18, 99] model survival outcomes using a Cox model and applied a K-Means clustering on the patient population using a distance weighted by the estimated coefficients. Xia et al. [357] extract embeddings through a deep learning survival model, then used to cluster the population. Nonetheless, this dissociation of outcome modelling and clustering can lead to groups inconsistent with outcomes [18, 99]. Second, one can jointly maximise clustering and survival objectives. In a non-parametric form, Mouli et al. [217] train a neural network to assign points to a cluster by maximising the divergence between clusters' Kaplan-Meier estimators. In the Bayesian profile regression [49, 196, 200], outcomes guide clustering in a mixture model. Similarly, Nagpal et al. [225] explore a mixture of Cox regression with group baselines in which individual covariates allow deviation from the group-specific Breslow estimator of the cumulative hazards. Each cluster assumes proportional hazards, and the semi-parametric approach requires an expectation-maximisation (EM) optimisation. Direct joint optimisation should be preferred as multi-stage optimisation and EM approach might lead to suboptimal solutions and slow convergence [207].

Our work is part of this last family of methods with end-to-end optimisation in a non-Bayesian framework. The proposed approach consists of a mixture of monotonic neural networks modelling subgroups' cumulative hazard functions. Each patient's survival distribution is a combination of these distributions. This method uses monotonic neural networks to obtain

clusters' survival distributions while maximising the likelihood of the observed data. This modelling choice improves the interpretability of neural network approaches while presenting more flexibility than parametric mixture models.

4.2.2 Identifying subgroups of treatment effects

In the previous section, we reviewed methods to identify groups of patients based on observed associations between covariates and outcomes. In this section, we focus on uncovering groups of causal relations between intervention and outcome. We aim to cluster *treatment effects* under time-to-event outcomes. The current literature on this topic remains sparse with a focus on estimating population or individual treatment effects [75, 153, 288, 375]. While these estimates help to understand the average response to treatment and could guide individual treatment decisions, they do not directly provide an understanding of the groups that may benefit or be harmed by treatment. Identifying such groups better aligns with current medical guidelines to target groups with improved treatment responses. This section reviews the existing literature on treatment effect estimation and subgroup identification.

Discovering intervenable subgroups is core to medical practice, particularly identifying subgroups of treatment effect as patients do not respond like average [35, 270, 278] and the average may hide potential response differences. Identification of subgroups has long been used to design Randomised Controlled Trials (RCTs). *A priori* identified groups discovered through novel medical findings were tested through trials [71, 268]. *A posteriori* analyses gained traction to uncover subgroups of patients from existing RCTs to understand the underlying variability of responses.

The first set of methodologies consists of a step-wise approach: (i) estimate the Individualised Treatment Effects (ITE) and (ii) uncover subgroups using a second model to explain the heterogeneity in ITE. Foster, Taylor, and Ruberg [91] and Qi et al. [259] describe the virtual twins approach in which one models the outcome using a decision tree for each treatment group. The difference between these decision trees results in the estimated treatment effects. A final decision tree aims to explain these estimated treatment effects to uncover subgroups. Similar approaches have been explored with different meta-learners [361], or Bayesian additive models [137], or replacing the final step with a linear predictor to uncover the feature influencing heterogeneity [55]. However, Guelman, Guillén, and Pérez-Marín [117] discuss the multiple drawbacks associated with these approaches. Notably, the two-step optimisation may not lead to recovery of the underlying subgroups of treatment effects.

Tree-based approaches have been proposed to address the limitations of step-wise approaches by jointly discovering subgroups and modelling the treatment effect. Instead of traditional splits on the observed outcomes, these causal trees aim to discover homogeneous splits regarding covariates and treatment effects. Su et al. [311] introduce a recursive population splitting based on the average difference in treatment effect between splits. Athey and Imbens [9] and

Athey and Imbens [10] improve the confidence interval estimation through the honest splitting criterion, which dissociates the splitting from the treatment effect estimation. Wager and Athey [331] agglomerate these causal trees into causal forests for improved ITE estimation. Each obtained split in the decision tree delineates two subgroups of treatment effect [188, 198]. Alternatively, McFowland III, Somanchi, and Neill [205] propose pattern detection and Wang and Rudin [339], causal rule set learning to uncover these subgroups. However, all these approaches rely on a criticised local optimisation criterion [187] and greedy split exploration. Recently, Nagpal et al. [224] addressed the local optimisation by constraining the treatment response shape to a linear form in a mixture of Cox models.

The previous approaches uncover subgroups of treatment effect but often consider RCTs with binary outcomes, not the observational setting with survival outcomes we are interested in. At the intersection with survival analysis, Zhang et al. [373] extend causal trees to survival causal trees, modifying the splitting criterion by measuring the difference in survival estimates between resulting leaves. Similarly, Hu, Ji, and Li [137] propose Bayesian additive models and Zhu and Gallego [382] propose a step-wise approach with propensity weighting to study observational data. Closest to our work, [152, 223] propose to uncover subgroups with survival outcomes. Specifically, Nagpal, Sanil, and Dubrawski [223] stratify the population into three groups: non-, positive- and negative responders to treatment. An iterative Monte Carlo optimisation is used to uncover these subgroups, characterised by a Cox model with a multiplicative treatment effect. As discussed, this step-wise optimisation may be limiting, and the assumption of RCT renders the model less relevant in observational data. Similarly, Jia et al. [152] analyse RCT with a mixture of treatment effects characterised by Weibull distributions trained in an EM framework.

The existing strategies are constrained by (i) their iterative optimisation that may lead to suboptimal survival estimation [207], (ii) the treatment response parametrisation that may not hold in medical settings [306], and (iii) the non-adjustment for treatment assignment that would bias the estimate in observational settings [75]. Our proposed methodology aims to address these challenges by accounting for the potential treatment assignment bias associated with observational settings and avoiding traditional assumptions about the survival function or treatment effect.

4.3 Latent Clustering

Between individualised and population-level modelling, we aim to uncover heterogeneity guided by observed survival outcomes. In this section, we formalise this subgroup discovery as an outcome-guided latent clustering problem and introduce the related quantities of interest to uncover survival and treatment effects profiles.

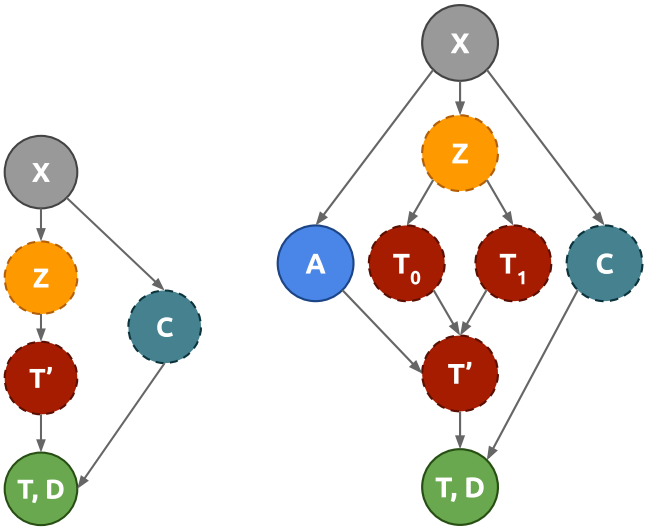
4.3.1 Outcome-guided clustering

Formalisation. Consider the random variables associated with observed covariates X, the indicator associated with the event of interest D, and the observed survival time T. Formally, the last variables are deterministically defined as $T := \min(C, T')$ and $D := \mathbb{1}(C > T')$, with T' the partially observed random variable associated with the time of the event of interest if there was no censoring and C the (right)-censoring time — as a reminder, this is the time at which the patient left the study before experiencing the event of interest. Central to this chapter is Z associated with the latent group membership. We assume that Z depends upon X and influences the survival profile T'. As our interest is to recover the survival function in the latent subgroups, we ignore the potential dependence of T' upon X, considering the observed outcome as a mixture of the different subgroups. While this assumption may hurt individual performance, it improves the interpretability of the model as it disentangles group membership from the survival profile. In this context, we aim to recover Z and the survival distribution associated with each cluster from the observed X, T and D. As a summary, the left panel in Figure 4.2 illustrates the assumed dependencies among random variables through a DAG.

Quantities of interest. Assume we aim to identify a pre-specified $K \in \mathbb{N}$ number of subgroups with similar survival outcomes. As a pre-specified number of clusters may be a limitation in a real-world setting where we do not know the underlying subgrouping structure, we explore in Section 4.5 how to select this parameter based on the likelihood of the predicted outcomes. Given K and the previously described dependencies, the expected survival (described in Section 2.2) for a given patient with covariates x becomes the marginal survival over the different clusters:

$$\begin{aligned} S(t \mid X = x) &= \sum_{k=1}^{K} \mathbb{P}(Z = k \mid X = x)\mathbb{P}(T' \geq t \mid X = x, Z = k) \\ &= \sum_{k=1}^{K} \mathbb{P}(Z = k \mid X = x)\mathbb{P}(T' \geq t \mid Z = k) \end{aligned} \tag{4.1}$$

with the following quantities of interest (i) the assignment to the different clusters k, $\alpha_k(x) := \mathbb{P}(Z = k \mid X = x)$, and (ii) $\Lambda_k(t) := -\log \mathbb{P}(T' \geq t \mid Z = k)$ extending the definition of



(a) Survival Profiles

(b) Treatment Responses Profiles

Figure 4.2: Graphical representations of the latent clustering problem between covariates (X) and outcomes (T, D). Random variables within dashed circles are unobserved, while X, A, T and D are observed. Z is the latent clustering structure variable and T', the event times if no censoring occurred. C denotes the censoring time. When considering a binary treatment, the assigned regime A depends upon the covariates X in observational data, while these two random variables are independent in RCTs. The potential outcomes are denoted as T_0 under no treatment and T_1 under treatment.

cumulative hazard from Section 2.2 to group-specific cumulative hazard, with derivative $\lambda_k(t)$, the group-specific instantaneous hazard.

Following the derivation from Chapter 2.2, one can use Equation (4.1) to obtain the likelihood associated with the observed realisation $\{x_i, t_i, d_i\}$ of the random variables X, T, D for patient i. Specifically, patients with an observed event $(d_i = 1)$ contribute the negative derivative of the survival function. Under the Assumption (4.1) of independence between the censoring time and the event of interest given the covariates (as depicted in the DAG in Figure 4.2), each censored patient $(d_i = 0)$ contributes the probability of not observing an event before t_i, i.e., $S(t_i \mid x_i)$.

Assumption 4.1 (Non-informative censoring)**.** Censoring is independent of the event of interest given the covariates. Formally, this corresponds to $T' \perp\!\!\!\perp C \mid X$.

With these notations, the log-likelihood under the assumed latent structure is equal to:

$$\begin{aligned} l &= \sum_{i,d_i=1} \log \left(-\frac{\partial S(u \mid x_i)}{\partial u}\bigg|_{u=t_i} \right) + \sum_{i,d_i=0} \log S(t_i \mid x_i) \\ &= \sum_{i,d_i=1} \log \sum_k \alpha_k(x_i)\lambda_k(t_i)e^{-\Lambda_k(t_i)} + \sum_{i,d_i=0} \log \sum_k \alpha_k(x_i)e^{-\Lambda_k(t_i)} \end{aligned} \tag{4.2}$$

In the following section, the training of the proposed architecture consists of maximising this quantity.

4.3.2 Treatment effect clustering

In the previous section, we discussed the case in which one aims to uncover different survival profiles. However, when considering treatment, the previous method would identify *associations* between treatment and cluster assignment. Practitioners may instead want to explore heterogeneity in *causal* responses to treatment given observed outcomes. In this section, we describe this problem under a binary treatment as a generalisation of the previous formalism.

Formalisation. Consider the additional *binary* treatment variable A. Patients either receive the treatment or do not. In this context, we consider the two potential outcomes: the time of events under treatment T_1 and without it T_0. The central challenge is that one can not observe both T_0 and T_1, but only T', the random variable associated with the event time under the observed treatment regime in the absence of censoring, which is equal to $A \cdot T_1 + (1 - A) \cdot T_0$ under Assumption (4.3) described below. In the causal literature, the *unobserved* potential outcome that would have occurred under the complementary treatment regime, i.e. $(1 - A) \cdot T_1 + A \cdot T_0$, is known as the *counterfactual outcome*. The right panel in Figure 4.2 describes the dependencies between the previously described random variables with the additional variables A, T_0, and T_1 compared to the previous setting. A critical assumption from the proposed setting is that the subgroups we aim to identify do not influence care, as shown by the assumed independence between treatment and group membership given the covariates, formally described as follows:

Assumption 4.2 (Unknown Latent Groups)**.** The treatment assignment is independent of the subgroup membership given the covariates. Formally, this corresponds to $A \perp\!\!\!\perp Z \mid X$, with Z being the random variable associated with the cluster membership, and A indicates treatment.

Assumption 4.3 (Consistency)**.** A patient's observed outcome is the potential outcome associated with the observed treatment. Formally, this means $T' := A \cdot T_1 + (1 - A) \cdot T_0$ with T', the observed outcome, (T_0, T_1), the potential outcomes and A, the observed treatment.

Assumption 4.4 (Ignorability)**.** The potential outcomes are independent of the treatment given the observed covariates, i.e. $A \perp\!\!\!\perp (T_0, T_1) \mid X$. Equivalently, no unobserved confounders impact both treatment and event time.

Assumption 4.5 (Overlap / Positivity)**.** Each patient has a non-zero probability of receiving the treatment, i.e. $\mathbb{P}(A \mid X) \in (0, 1)$, resulting in a non-deterministic treatment assignment.

Counterfactual Modelling. In this setting, existing works often focus on estimating the ITE τ under the additional Assumptions (4.3), (4.4), and (4.5) defined as the difference between potential outcomes given the covariates x:

$$\begin{aligned}\tau(t,x) :&= \mathbb{E}(\mathbb{1}(T_1 \geq t) - \mathbb{1}(T_0 \geq t) \mid X = x) \\ &= \mathbb{E}(\mathbb{1}(T_1 \geq t) \mid X = x) - \mathbb{E}(\mathbb{1}(T_0 \geq t) \mid X = x) \\ &= \mathbb{P}(T' \geq t \mid A = 1, X = x) - \mathbb{P}(T' \geq t \mid A = 0, X = x) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Under Asm. 4.3 and 4.4)} \\ &= S(t \mid A = 1, X = x) - S(t \mid A = 0, X = x) \qquad (4.3)\end{aligned}$$

Estimating this quantity requires accurately modelling the survival distributions under the two treatment regimes *for all* x from observed data. This estimation would be straightforward if one could access the observed survival times for all patients under both treatment regimes. One would estimate the survival using the outcomes under $A = 1$ and under $A = 0$. The observational challenge is the non-observation of the *counterfactual* survival outcome: if a patient receives the treatment, we do not observe its outcome under no treatment. In this context, we only observe a subset of the outcomes under both treatment regimes.

Under independence between treatment and covariates, as designed in RCTs, the subset of the patients receiving either treatment regime is representative of the overall population. Consequently, relying on this observed subset to estimate the survival distribution under each regime is an valid estimate for all patients. However, covariates and treatment are not independent in observational studies in which treatment recommendations depend upon the observed covariates. Formally, $\mathbb{P}(A \mid X) \neq \mathbb{P}(A)$, e.g. medical practitioners recommend more aggressive treatment to patients with more advanced conditions. This absence of randomisation results in a covariate shift [75] as the covariates for the non-treated and treated patients are drawn from different distributions [35]. Estimated observational survival probabilities are not transportable between regimes, biasing the treatment effect estimates.

Under Assumption (4.5), re-weighting [292], penalisation on the dissimilarity of learnt representations [153, 291] or a combination of these approaches [75, 122, 288] are remedies to estimate the likelihood — both factual and counterfactual — by reducing the difference between the two treatment regimes' populations when modelling the outcome of interest. Specifically, Shalit, Johansson, and Sontag [288] demonstrate that the negative log-likelihood is upper-bounded. The authors propose regularising the training objective by an Integral Probability Metric (IPM) between the distributions q_ϕ associated with an invertible embedding ϕ. This embedding is then used to estimate the survival function under both treatment regimes. By regularising ϕ, the treated and untreated patient distributions are alike. The difference between the estimated survival functions is then an accurate estimate of the treatment effect as the penalisation corrects for the shift resulting from treatment non-randomisation. The

following adapts the authors' loss function used to train the survival model and the embedding ϕ to the survival settings:

$$-l \leq -l_F^* + \gamma \cdot \text{IPM}(q_\phi^{A=0}, q_\phi^{A=1}), \tag{4.4}$$

where l is the underlying log-likelihood, equal to the sum of the factual and counterfactual log-likelihoods. As one can not compute the counterfactual likelihood due to the non-observation of the counterfactual outcomes, the authors bound this quantity by (i) l_F^*, the *weighted* factual log-likelihood with an inverse propensity of treatment weighting for each patient, and (ii) the difference between the embedding distributions under the two regimes with $q_\Phi^{A=a} = q(\Phi(X) \mid A = a)$ and γ a positive constant.

Quantities of interest. Our work differs from the existing literature in its focus on groups' average treatment effects (GATE) instead of individualised ones where Z is unobserved:

$$\begin{aligned}\tau_k(t) :&= \mathbb{E}(\tau(t, X) \mid Z = k) \\ &= \mathbb{E}(\mathbb{P}(T' \geq t \mid A = 1, Z = k, X) - \mathbb{P}(T' \geq t \mid A = 0, Z = k, X)) \\ &= S(t \mid A = 1, Z = k) - S(t \mid A = 0, Z = k)\end{aligned} \tag{4.5}$$

Note that, under the proposed setting, only the clustering assignment depends upon X. Connecting with the previously described bound on the likelihood in Equation (4.4), only the cluster assignments $\alpha := \{\alpha_k\}$ could play the role of ϕ, as the only transformation of X. However, treatment is assumed independent of latent group membership under Assumption (4.2), meaning that treatment rate does not differ between clusters. This assumption results in the regularisation term being null, i.e. $\text{IPM}(q_\alpha^{A=0}, q_\alpha^{A=1}) = 0$ (see Appendix B.1 for derivation).

Our work, therefore, focuses on the first term of the upper bound in Equation (4.4) by re-weighting the factual likelihood given the patient propensity to receive treatment [122, 288]. Using an estimator $\hat{p}_A(x)$ of the propensity of treatment $\mathbb{P}(A = 1 \mid X = x)$, we rely on a truncated propensity weighting [15] scheme to avoid unstable weights, in which the weights w_i are defined for each patient i through:

$$w_i^{-1} = \begin{cases} 0.05 & \text{if } \hat{p}_A(x_i) < 0.05 \\ 0.95 & \text{if } \hat{p}_A(x_i) > 0.95 \\ a_i \cdot \hat{p}_A(x_i) + (1 - a_i) \cdot (1 - \hat{p}_A(x_i)) & \text{otherwise} \end{cases} \tag{4.6}$$

with a_i, the realisation of A for patient i.

Using the survival expression from Equation (4.2) and the weights w_i defined in Equation (4.6), we derive the upper-bound of the negative log-likelihood in the proposed settings as

follows:

$$l_F^* = \sum_{i,d_i=1} w_i \log \left[\sum_{k \in [\![1,K]\!]} \alpha_k(x_i) \cdot \lambda_k(t_i) e^{-\Lambda_k(t_i)} \right] + \sum_{i,d_i=0} w_i \log \left[\sum_{k \in [\![1,K]\!]} \alpha_k(x_i) \cdot e^{-\Lambda_k(t_i)} \right] \tag{4.7}$$

4.4 Proposed approach

In the previous section, we identified the different components to study heterogeneity in observational settings. In this section, we propose an implementation for each element following a similar breakdown of the problem:

- We jointly uncover subgroups and maximise the likelihood of the observed survival outcomes.
- We alter this architecture to estimate the causal treatment responses.
- We correct the model training to account for the non-randomisation of treatment in observational settings.

4.4.1 Recovering underlying subgroups

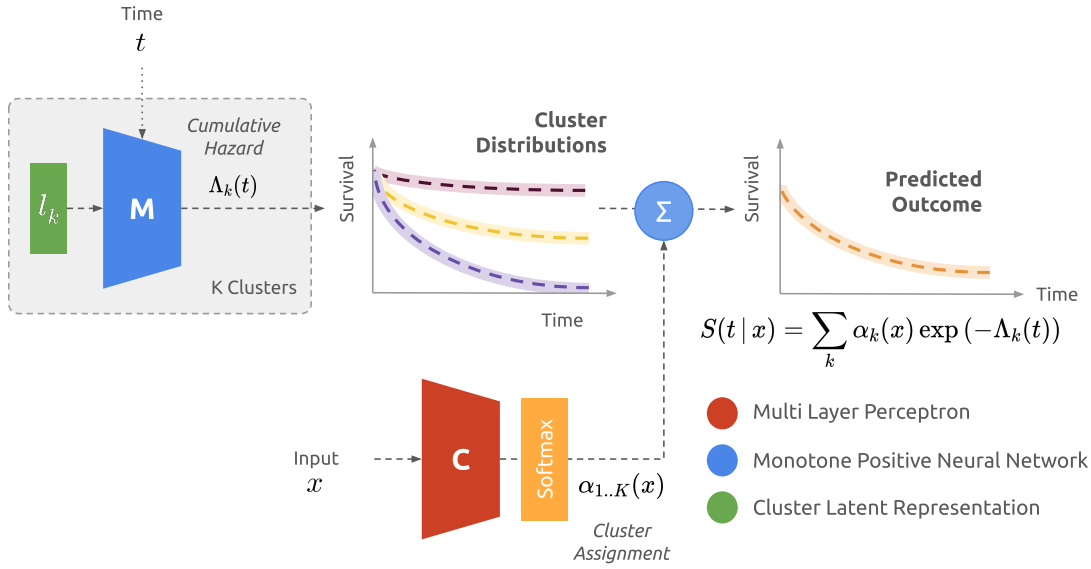


Figure 4.3: Neural Survival Clustering Architecture. A neural network assigned the covariates x to the mixture of K clusters characterised by a latent l_k used as input to a monotonic neural network M modelling the cluster-specific cumulative hazard.

In Equation (4.1), we decompose the survival as a mixture for which we need to model the cluster assignment probability α_k and the cumulative hazard Λ_k for each cluster k. While α_k is

germane to a multi-class problem in which one assigns a cluster to individuals with the covariates x, the challenge is modelling Λ_k, equivalently each cluster's survival distribution. As the integral of the hazard function λ_k, the cumulative hazard is a positive function, monotonically increasing over time, and null at time $t = 0$. Further, as we aim to maximise the observed likelihood, described in Equation (4.2), we need to obtain the hazard λ_k.

As discussed in Chapter 2, existing survival models approach this challenge by constraining Λ_k to a parametric form with a closed-form derivative or by modelling λ_k and obtaining Λ_k by numerical integration. While we could rely on any of the approaches to model this quantity and its derivative, we choose to model the survival distribution through monotonic neural networks (see Section 2.1.2.3) as in [266] to ensure the exact computation of both Λ_k and λ_k. Specifically, our modelling choices are as follows:

Cluster assignment. To model α_k, we use a multi-layer perceptron C with final Softmax layer assigning each patient characterised by x, to the different clusters through a K-dimensional normalised vector of probabilities: $[\alpha_k(x)]_{k=1}^K$.

$$C(x) := [\mathbb{P}(Z = k \mid x)]_{k=1}^K$$

with K, the selected number of clusters.

Cluster survival distributions. Each cluster k is represented by a vector $l_k \in \mathbb{R}^L$, with L its dimension, characterising the cluster's survival distribution *independently* of x. Consider the quantity l_k as the latent parametrisation of a cumulative hazard function. The vector l_k is then concatenated with t and used as inputs for a neural network M with monotonic positive outcomes to model the cumulative hazard function $\Lambda_k(t)$. Specifically, we use the following transformation to ensure that no probability is assigned to negative times — a limitation raised concerning previous monotonic neural networks [290]:

$$\Lambda_k(t) := t \cdot M(l_k, t)$$

Note that this transformation does not constrain the cumulative hazard beyond ensuring $\Lambda_k(0) = 0$.

Figure 4.3 illustrates the proposed Neural Survival Clustering. A first multi-layer perceptron C estimates the mixture weights $\{\alpha_k(x)\}$ with a Softmax to ensure that they sum up to one given a patient's covariates x. Each component of the mixture of networks takes the time t and the latent representation l_k as inputs to predict the cluster-specific cumulative hazard $\Lambda_k(t)$. Finally, the survival function estimate is obtained as the weighted sum of the components as described in equation (4.1).

Training. The model is trained by maximising the observed survival likelihood presented in Equation (4.2). Our approach uses automatic differentiation to compute the exact instantaneous hazard λ_k, necessary in estimating the log-likelihood.

4.4.2 Uncovering treatment subgroups

To uncover subgroups of treatment effects instead of subgroups of survival profiles, we propose altering the previous architecture to model the outcome under both treatment regimes while remaining independent of the covariates x.

Cluster survival distributions. As before, each cluster k is characterised by a latent parameter l_k , concatenated with t and used as inputs for a neural network M with monotonic positive output. The difference with the previous architecture is that M outputs two values: the cumulative hazard function under both treatment regimes $(\Lambda_k(t \mid A = 0), \Lambda_k(t \mid A = 1))$ — this modelling choice results in two distributions per cluster, *independent* of x. With this modelling, we can estimate the GATE, $\hat{\tau}_k(t)$, in cluster k as:

$$\hat{\tau}_k(t) := \exp(-\Lambda_k(t \mid A = 0)) - \exp(-\Lambda_k(t \mid A = 1))$$

Training. In a randomised setting, one can uncover subgroups of treatment effect by training the model using the likelihood of the observed outcomes (Equation (4.2)), as done in training the previous Neural Survival Clustering architecture.

4.4.3 Accounting for observational data

While the previous architecture allows the discovery of subgroups with similar treatment effects in randomised settings, our interest lies in the observational setting with its associated clinical presence challenge. We propose adjusting the model's training to estimate treatment effects under treatment non-randomisation.

Inverse Propensity Weighting. To correct the log-likelihood for non-randomisation, we estimate the propensity of the patient to receive treatment using a multi-layer perceptron W with a final Sigmoid transformation.

$$W(x) := \mathbb{P}(A = 1 \mid x)$$

Training. This adjustment requires a training step before the previous maximisation of the log-likelihood. We model the binary treatment assignment through W by minimising the cross-entropy of receiving treatment, defined in Equation (2.1). Then, training all other components relies on minimising the weighted factual log-likelihood introduced in Equation (4.7).

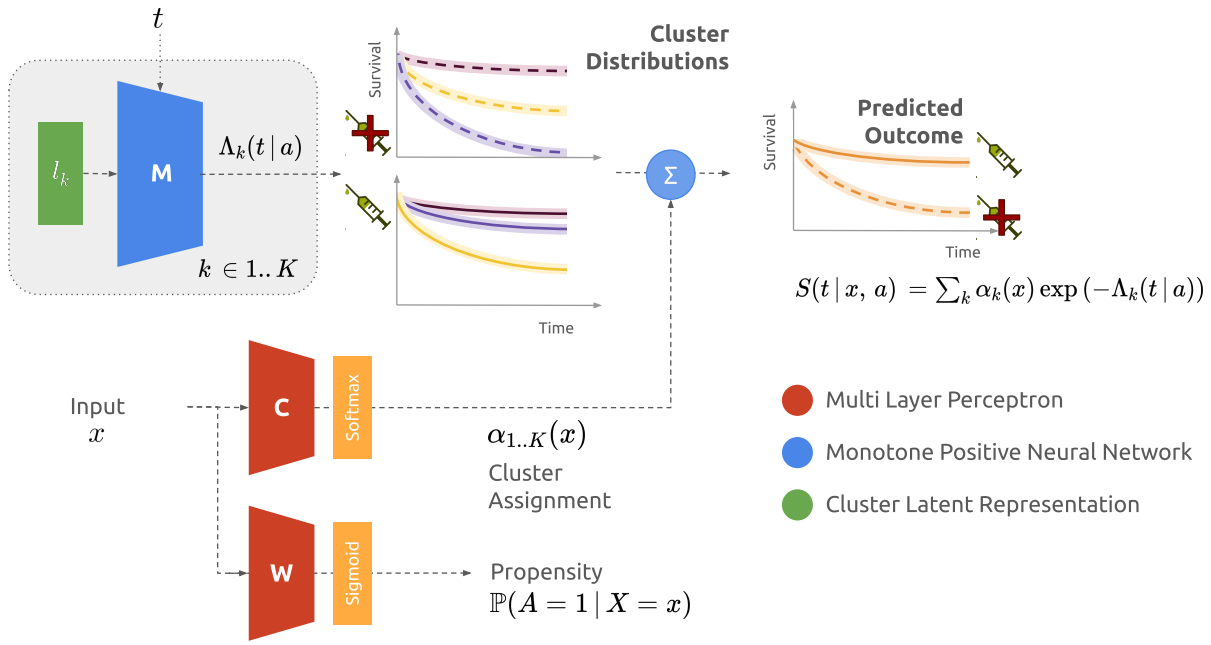


Figure 4.4: Causal Neural Survival Clustering Architecture. Latent parameters l_k characterising the cluster k are inputted in the monotonic network M to estimate the cumulative hazard $\Lambda_k(t \mid A = 0)$ and $\Lambda_k(t \mid A = 1)$. In parallel, C assigns weights α_k to each cluster given the patient's covariates x. To tackle the challenge of treatment assignment bias, we use the network W to estimate the treatment propensity and weigh the training log-likelihood.

These implementation choices result in the flexible modelling of survival and treatment effect outcomes by maximising the exact log-likelihood without constraining parametrisations of the survival function.

4.5 Synthetic evaluation

As counterfactuals are unknown in observational data, we adopt the common practice of a synthetic dataset — in which underlying survival distributions and subgroup structure are known — to evaluate our proposed methodology's capacity to uncover subgroups of treatment effect. We describe the data generation process in Section 4.5.1. Then, we introduce the empirical settings in Section 4.5.2 and compare different approaches in Sections 4.5.3 and 4.5.4. The code to generate the data and reproduce the following results is available on Github[2,3].

4.5.1 Data generation

We consider a synthetic population of $N = 30,000$ patients with 10 associated covariates $X \in \mathbb{R}^{10}$ divided into $K = 3$ clusters. The following data generation does not aim to mimic a

[2] https://github.com/Jeanselme/NeuralSurvivalClustering
[3] https://github.com/Jeanselme/CausalNeuralSurvivalClustering

particular real-world setting but follows a similar approach to [226]. The following describes our generation process:

Covariates. Each patient's membership Z is drawn from a multinomial with equal probability. Group membership informs the two first covariates through the parametrisation of the bivariate normal distribution with centres c_k equal to $(0, 2.25)$, $(-2.25, -1)$, and $(2.25, -1)$. All other covariates are drawn from standard normal distributions. Formally, this procedure is described as:

$$\begin{aligned} Z &\sim \mathsf{Mult}\left(1, \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]\right) \\ X_{[1,2]} \mid Z = k &\sim \mathsf{MVN}(c_k, I^2) \\ X_{[3:10]} &\sim \mathsf{MVN}(0, I^8) \end{aligned}$$

with MVN denoting a multivariate normal distribution, and I^n, the identity covariance matrix of dimension n.

Treatment response. For each cluster, event times under treatment and control regimes are drawn from Gompertz distributions, with parameters that are functions of group-specific coefficients (B^0 and Γ^0 for the event time when untreated and B^1 and Γ^1 when treated) and the patient's covariates.

$$\begin{aligned} B_z^0 \mid Z = z &\sim \mathsf{MVN}(0, I^{10}) \\ \Gamma_z^0 \mid Z = z &\sim \mathsf{MVN}(0, I^{10}) \\ T_0 \mid Z, X, B_z^0, \Gamma_z^0 = (z, x, \beta_z^0, \gamma_z^0) &\sim \mathsf{Gompertz}\left(w_0(\beta_z^0, x), s_0(\gamma_z^0, x)\right) \end{aligned}$$

$$\begin{aligned} B_z^1 \mid Z = z &\sim \mathsf{MVN}(0, I^{10}) \\ \Gamma_z^1 \mid Z = z &\sim \mathsf{MVN}(0, I^{10}) \\ T_1 \mid Z, X, B_z^1, \Gamma_z^1 = (z, x, \beta_z^1, \gamma_z^1) &\sim \mathsf{Gompertz}\left(w_1(\beta_z^1, x), s_1(\gamma_z^1, x)\right) \end{aligned}$$

with w_0, w_1 two functions parametrising the Gompertz distributions' shape as $w_0(\beta, x) := |\beta[0]| + (x[5:10] \cdot \beta[5:10])^2$, $w_1(\beta, x) := |\beta[0]| + (x[1:5] \cdot \beta[1:5])^2$, and the shift parameter parameterised as $s_0(\gamma, x) := |\gamma[0]| + |(x[1:5] \cdot \gamma[1:5])|$ and $s_1(\gamma, x) := |\gamma[0]| + |(x[5:10] \cdot \gamma[5:10])|$ where $v[i]$ described the i^{th} element of the vector v. These functions aim to introduce non-linear responses with discrepancies between control and treatment regimes. Note that we allow covariates to influence the survival distribution as a patient's covariates influence Gompertz's shapes and scales.

Treatment assignment. The non-randomisation of treatment is central to the problem of identifying treatment subgroups in real-world applications. After drawing the treatment assignment probability P, we assign each patient to a given treatment. We propose two treatment assignment strategies reflecting a RCT and an observational setting, denoted as "Randomised" and "Observational". "Randomised" consists of a Bernoulli draw using the realisation of P. "Observational" reflects an assignment dependent upon the observed covariates.

$$\begin{aligned} P &\sim \text{Uniform}(0.25, 0.75) \\ A_{rand} \mid P = p &\sim \text{Bernoulli}(p) \end{aligned}$$

$$A_{obs} \mid P, X = (p, x) \sim \text{Bernoulli}(F(x) \times p)$$

with $F(x) := \mathbb{P}(X[1]^2 + X[2]^2 < x[1]^2 + x[2]^2)$ the cumulative function that returns the probability that a transformation of X will take a smaller value than x following the same transformation.

Censoring. Finally, our work focuses on right-censored data. To generate censoring independent of the treatment and event, we draw censoring from another Gompertz distribution as follows:

$$\begin{aligned} B^C &\sim \text{MVN}(0, I^5) \\ C \mid X, B^C = (x, \beta) &\sim \text{Gompertz}\left(w_c(\beta, x), 0\right) \\ T' &= A \cdot T_1 + (1 - A) \cdot T_0 \\ T &= \min(C, T') \\ D &= \mathbb{1}(C > T') \end{aligned}$$

with $w_c := (x[5:10] \cdot \beta)^2$, the scale of the censoring Gompertz distribution.

4.5.2 Empirical settings

In this section, we detail the different methodologies, training procedures, and evaluations used to compare the proposed methods against existing approaches.

Survival Modelling Baselines. First, to compare the predictive quality of the proposed Neural Survival Clustering (**NSC**), we evaluate multiple baselines: a Cox Proportional Hazards model **CoxPH** [74], its deep learning extension **DeepSurv** [163], which uses a neural network to estimate the covariate effect on the hazard. Moreover, the performance of the monotone survival neural network **SuMo-net** [266] is also compared, as our work uses a similar network

for modelling each cluster's cumulative hazard. Additionally, we analyse the performance of **DeepHit** [179], which discretise the predictive horizons and Deep Survival Machine (**DSM** [221]), which uses a mixture of Weibulls parameterised through a neural network with inputs the individual covariates.

Survival Clustering Baselines. As we seek to uncover subgroups of survival outcomes and not maximise predictive performance alone, we compare our model to a mixture of Cox models known as Deep Cox Mixture (**DCM** [225]). While this method allows individual flexibility as each patient can deviate from a non-parametric cluster baseline, it relies on EM optimisation that might lead to sub-optimal modelling. Further, the proportional hazards assumption may not hold in real-world settings. We also consider a Cox-Weighted K-Means (**CWKM**) in which a K-means algorithm divides the population based on an Euclidean distance weighted by the estimated hazard ratios of a Cox regression. As a final clustering baseline, we consider a **Survival Tree** [112] in which the purity criterion used to to divide the population relies on a test on the difference in Kaplan-Meier survival estimates. This tree structure identifies, by design, subgroups of survival outcomes at each depth of the tree.

Treatment Effect Clustering Baselines The previous survival clustering baselines focus on the factual clustering of survival outcomes, and do not consider the heterogeneity in treatment responses. To study this causal relation between treatment and outcomes, we compare the proposed Causal Neural Survival Clustering (**CNSC**) against its unadjusted alternative (**CNSC Unadjusted**), which uses the unweighted factual likelihood ($\forall i, w_i = 1$) to demonstrate the necessity of controlling for the treatment assignment bias. Additionally, we compare the proposed approach to Cox Mixtures with Heterogeneous Effects (**CMHE** [226]), which uncovers treatment effect and baseline survival latent subgroups. In an EM framework, a model assigns each patient to a group for which a Cox model is fitted. The central difference with our proposed approach is that CMHE allows for direct dependencies of the survival outcome on the covariates through the Cox model. Further, the methodology dissociates treatment and survival subgrouping for more flexibility in the outcome modelling. Treatment clusters differ in treatment responses, while survival clusters have different baseline survival models. For a fair comparison, we present three alternatives: one with fixed $K = 3$ survival clusters and $L = 1$ treatment effect clusters, one with $L = 3$ and $K = 1$, and one with $K = L = 2$. Note that this last alternative allows for a total of 4 clusters. Critically, these methods assume proportional hazards for each cluster and do not account for the treatment non-randomisation. We focus on neural network approaches due to the absence of tree-based causal survival implementation with access to the counterfactual distributions and subgroup assignment, which is central to our work.

Training Procedure. All models' evaluation uses the same 5-fold cross-validation. In each fold, the development set is divided into three: 80% for training, 10% for stopping criterion and 10% for hyper-parameters selection. We use 50 draws of a random search on the following grid: learning rate (0.001 or 0.0001), batch size (100 or 250), number of layers for both assignment network and survival neural networks (1, 2, 3) with number of nodes (50 or 100), number of components for the mixture ($[\![2, 5]\!]$) and the dimensionality of the latent cluster representation l^k in (10, 50, 100). We use the TanH activation function to ensure the cumulative intensity's derivative exists. All models are optimised over 1000 epochs using an Adam optimiser [166] with early stopping.

The parameter search for all other methods uses a similar grid (when appropriate). Additionally, following [221], we optimise DSM over the type of distributions (LogNormal or Weibull) and use 10,000 warming epochs that initialise all survival distributions with parameters associated with the average survival distribution when discarding the covariates' impact on outcome. Four intervals are used for DeepHit to discretise the time horizon. These splits reflect the evaluation at 0.25, 0.5 and 0.75 quantiles. The training procedure relied on an early stopping criterion on 10% of the training split using the negative log-likelihood loss. The survival tree depth is constrained to the range $[2, 5]$, and we analyse the first two splits for clustering analysis.

Evaluation. Survival performances were measured using cumulative time-dependent C-Index [139] and Brier Score at the dataset-specific 0.25, 0.5 and 0.75 quantiles of the uncensored population event times and averaged over the 5-fold cross-validation. Means and standard deviations are reported.

As in the synthetic experiments, subgroup structure is known; we measure the adjusted[4] Rand-Index [265], which quantifies how the predicted assignment aligns with the known underlying cluster structure. In the context of treatment estimation, we additionally use the integrated squared error (ISE) between the treatment effect estimate and the ground truth, which measures how well we recover each cluster's treatment effect:

$$\text{ISE}_k(t) = \int_0^t \left(\hat{\tau}_k(s) - \tau_k(s)\right)^2 ds$$

with $\hat{\tau}_k$ the estimated treatment effect for cluster k and τ_k the ground truth. Our experiment evaluates this metric to the last observed event time. Further, we measure the ISE in modelling the average treatment effect at the population level, defined as:

$$\text{ISE}_{pop}(t) = \int_0^t \left(\mathbb{E}_x[\hat{\tau}(s, x)] - \mathbb{E}_x[\tau(s, x)]\right)^2 ds$$

[4]Random patient assignment results in an adjusted Rand-Index of 0.

4.5.3 Outcome subgrouping

Our first set of experiments aims to measure the model's capacity to uncover subgroups of survival outcomes under current medical practice. To this end, we consider treatment as one of the input covariates and model the *association* between all covariates and observed survival outcomes in the previously generated dataset.

Survival modelling. Table 4.1 displays the predictive performance of all considered survival models. Note that the proposed methodology is less predictive of survival than other approaches. This observation aligns with the assumed setting as the simulation allows each patient's survival distribution to deviate from its subgroup through dependence on the patient's covariates. By discarding variation in a given cluster, the proposed model reduces the capacity to discriminate between patients, and approaches modelling such deviation perform better. In a population with negligible deviations, this performance gap would narrow (as observed in the real-world setting of Section 4.6 and the Appendix Section B.2).

	C Index			Brier Score		
Model	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
NSC	0.832 (0.017)	0.832 (0.017)	0.823 (0.012)	0.115 (0.005)	0.114 (0.008)	0.101 (0.006)
DCM	0.844 (0.008)	0.839 (0.006)	0.822 (0.007)	0.114 (0.004)	0.115 (0.004)	0.100 (0.003)
DSM	*0.872* (0.006)	*0.848* (0.006)	0.819 (0.006)	*0.104* (0.003)	*0.107* (0.003)	*0.092* (0.001)
SuMo-net	**0.878** (0.005)	**0.849** (0.004)	*0.827* (0.004)	**0.098** (0.003)	**0.103** (0.003)	**0.089** (0.001)
DeepHit	0.841 (0.011)	0.818 (0.012)	0.792 (0.014)	0.143 (0.011)	0.192 (0.012)	0.151 (0.007)
DeepSurv	0.839 (0.006)	0.836 (0.003)	**0.828** (0.003)	0.116 (0.005)	0.118 (0.001)	0.095 (0.002)
CoxPH	0.784 (0.004)	0.746 (0.004)	0.709 (0.005)	0.135 (0.004)	0.181 (0.002)	0.187 (0.004)
Survival Tree	0.834 (0.008)	0.825 (0.007)	0.802 (0.008)	0.117 (0.005)	0.122 (0.003)	0.100 (0.003)

Table 4.1: Models' predictive performance on the synthetic dataset. Mean (std) over the 5-fold cross validation with best performance in **bold** and second best in *italic*.

Recovering *K*. Figure 4.5 presents the average negative log-likelihood obtained by cross-validation given the number of clusters K. The dotted lines represent the elbow heuristic [319], which consists of identifying a change point in the explained variability of a clustering strategy, here in the log-likelihood. From this heuristic, the number of clusters revolves around $K = 4$. Hyperparameter cross-validation further confirms this choice when considering K as a hyperparameter. This data-guided choice of the number of clusters K is a crucial strength of the proposed strategy, compared to post-analyses, such as CWKM, which cannot guide this choice through modelling. This approach offers an outcome-guided heuristic to select the number of clusters.

Subgroups recovery. Figure 4.6 displays the average clusters obtained across folds by the four considered clustering methodologies with K selected as a hyperparameter for NSC and

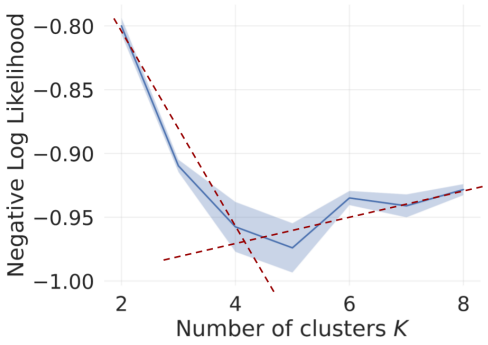


Figure 4.5: Averaged log-likelihood across 5-fold cross-validation given the number of clusters K with the shaded area representing 95% CI. The methodology presents an elbow around $K = 4$.

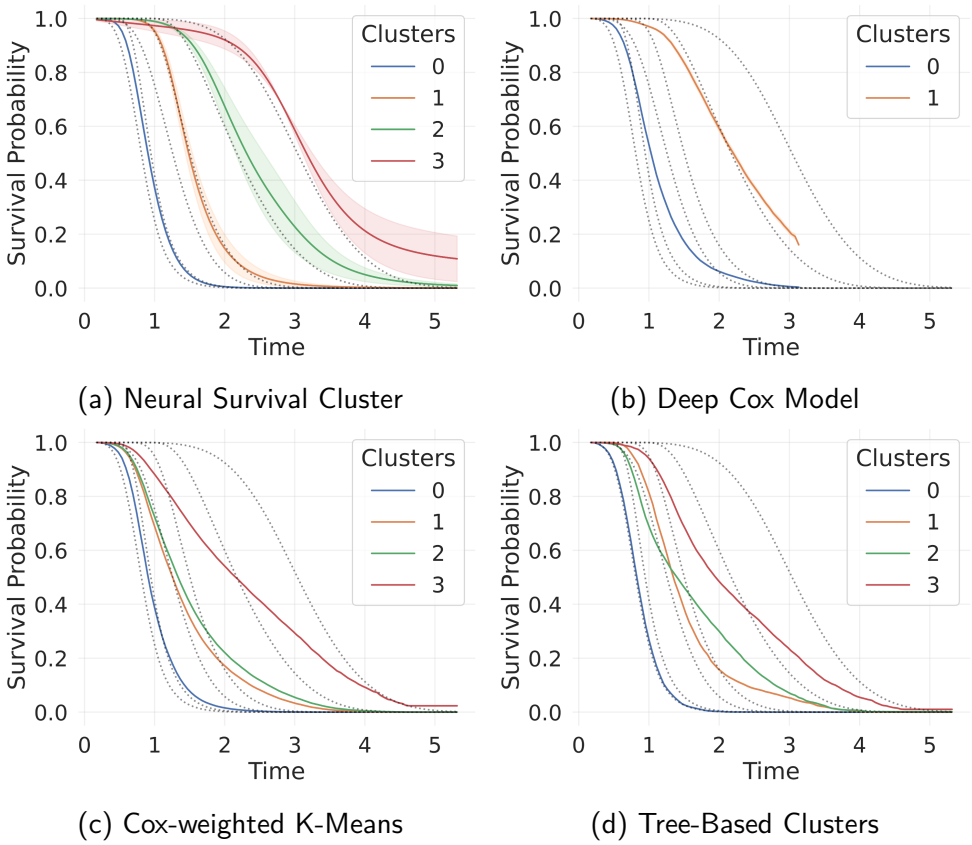
DCM. The dotted lines represent the 3×2 generated survival distributions, as each cluster presents two distributions: one under treatment and one when untreated. Note that both NSC and DCM underestimate the number of clusters as the synthetic dataset presents three similar survival distributions, explaining the discovery of 4 clusters by the proposed methodologies. This observation highlights that outcome-guided clustering only uncovers subgroups with distinct responses. Despite this limitation, the proposed strategy identifies clusters closest to the underlying clusters despite presenting less discriminative performance.

4.5.4 Treatment subgrouping

The previous experiment demonstrates the model's capacity to uncover different risk profiles. However, considering treatment as a covariate, the proposed approach identified associations between treatment and outcome, missing the underlying treatment response subgroups. This second experiment aims to uncover the causal link between treatment and outcome. As mentioned in the data generation, we explore two scenarios of treatment assignment: (i) "Randomised" and (ii) "Observational".

Recovering K. Following the previous elbow rule, we use the log-likelihood to select the number of clusters. Figure 4.7 shows a drastic change in the negative log-likelihood around $K = 3$ clusters in the observational setting, aligning with the simulation's generative process that results in 3 distinct treatment responses.

Subgroup recovery. Table 4.2 presents the performance of the different methodologies across the two studied scenarios. Recall that CMHE's parameter L describes the number of survival distributions, while K denotes the number of treatment response clusters. This disentanglement answers the question of which subgroups do not respond well to treatment, regardless of their survival distribution. Our approach, however, identifies subgroups of treatment while considering



(a) Neural Survival Cluster (b) Deep Cox Model

(c) Cox-weighted K-Means (d) Tree-Based Clusters

Figure 4.6: Averaged survival clusters across 5-fold cross-validation observed in the simulated dataset with the shaded areas representing 95% CI. Dotted lines denote the underlying survival distributions.

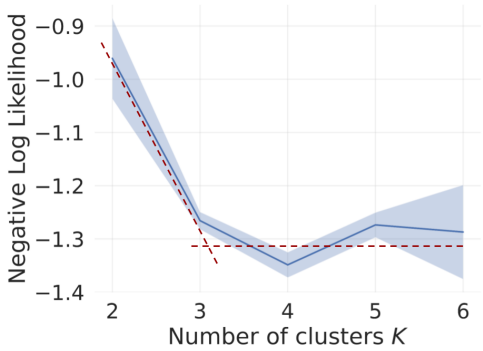


Figure 4.7: Averaged log-likelihood across 5-fold cross-validation given the number of clusters K under the "Observational" treatment assignment with the shaded area representing 95% CI. The log-likelihood presents an elbow around $K = 3$, the underlying number of clusters.

their survival distribution. While reducing flexibility in individual estimates, non-response to treatment has a more significant impact on patients with low survival than patients with better survival. Because of this difference, these patients may also present different disease pathways requiring different treatments.

Despite this reduced flexibility, the proposed method outperforms all CMHE alternatives in

	Model	Rand-Index	$\text{ISE}_{pop}(t_{max})$	$\text{ISE}_k(t_{max})$ $k=1$	$k=2$	$k=3$
Randomised	**CNSC**	**0.879** (0.016)	0.008 (0.001)	0.012 (0.002)	0.012 (0.001)	0.022 (0.001)
	CNSC Unadjusted	0.878 (0.012)	**0.007** (0.001)	**0.010** (0.002)	**0.011** (0.002)	**0.021** (0.006)
	CMHE ($K=3$)	0.489 (0.016)	0.057 (0.005)	0.123 (0.007)	0.102 (0.002)	0.286 (0.007)
	CMHE ($L=3$)	0.618 (0.237)	0.050 (0.009)	0.094 (0.013)	0.184 (0.015)	0.304 (0.014)
	CMHE ($K=L$)	0.625 (0.011)	0.020 (0.001)	0.183 (0.031)	0.358 (0.023)	0.424 (0.004)
Observational	**CNSC**	0.875 (0.005)	**0.008** (0.001)	0.011 (0.002)	**0.013** (0.002)	**0.029** (0.004)
	CNSC Unadjusted	**0.881** (0.006)	0.016 (0.002)	**0.010** (0.002)	0.019 (0.002)	0.054 (0.002)
	CMHE ($K=3$)	0.476 (0.010)	0.038 (0.010)	0.148 (0.020)	0.102 (0.002)	0.258 (0.020)
	CMHE ($L=3$)	0.468 (0.029)	0.054 (0.005)	0.103 (0.006)	0.195 (0.005)	0.295 (0.006)
	CMHE ($K=L$)	0.517 (0.086)	0.029 (0.008)	0.137 (0.046)	0.146 (0.084)	0.278 (0.071)

Table 4.2: Cross-validated performances' average (std) under the "Randomised" and "Observational" treatment simulations. ISE_{pop} quantifies the error in estimating the average treatment effect, while ISE_k reflects how the estimated treatment effects align with the ground truth. Best performances are in **bold**. The proposed CNSC best recovers the underlying treatment responses across treatment assignment scenarios.

this simulation. This gap is due to the parametrisation and assumptions made by CMHE. This latter assumes proportional hazards and a linear impact of treatment on log-hazards. Neither assumption holds in the proposed simulation. Note that increasing the number of clusters, as shown in $K = L$, improves performances in terms of clustering quality (Rand-Index) and subgroup recovery (lower ISE_{pop}). These experiments highlight the advantage of our proposed model in uncovering subgroups of treatment responses due to the flexibility in modelling complex survival distributions under both treatment regimes. Moreover, the "Observational" simulation demonstrates the importance of controlling for treatment non-randomisation. While all methodologies present larger ISE in the observational setting, we observe an increasing performance gap between CMHE and CNSC, validating the importance of accounting for the treatment assignment bias. Critically, the superiority of CNSC over its unadjusted alternative is evident in the estimation of the average treatment effect, and recovery of Cluster 3's treatment effect, in which not adjusting for treatment assignment doubles the ISE.

While this simulation setting is somewhat artificial, it underlines the proposed methodology's flexibility to recover complex treatment responses in an observational setting.

4.6 Case-study: Adjuvant radiotherapy after breast surgery and chemotherapy.

We previously demonstrated the proposed model's capacity to uncover the underlying heterogeneity of patients' outcomes and treatment responses in observational settings. In this section, we analyse how these tools can help discover subgroups in real-world observational data. The following analysis focuses on the SEER dataset, a large observational cohort of patients diagnosed with cancer across the United States. The following first introduces the subset of patients we consider. Then, we explore heterogeneity in survival outcomes and treatment responses following cancer diagnosis.

4.6.1 Dataset

The Surveillance, Epidemiology, and End Results (SEER) program in the United States monitors national cancer statistics. In our analysis, we focus on women who were diagnosed with breast cancer between the years 2010 and 2015 and followed up until 2021. We chose this diagnosis period to ensure that all the selected covariates are available and to maintain consistency in the available treatment options. We examine the disease-specific risk of death from breast cancer, considering all other causes as censored. The previous selection criteria resulted in 278,225 patients. We use 23 covariates to describe the patient demographics and disease characteristics at diagnosis, such as grade, laterality, tumour size, and observed treatments, including surgery, radiation, and chemotherapy.

4.6.2 Survival modelling

To ensure the relevance of the identified clusters in this real-world context, we must verify that the methodology recovers clusters that are predictive of the survival outcome. Table 4.3 summarises the different survival models' predictive performance averaged across the 5-fold cross-validation. The central observation is that all neural network approaches present similar performances with a slight advantage for SuMo-net, a survival monotonic neural network. This last observation further justifies our reliance on monotonic neural networks to model cluster survival distribution. The proposed NSC remains competitive with existing methodologies despite relying on survival distributions independent of the input covariates. This observation validates the presence of a latent structure explaining the observed survival outcomes. While reducing performance, this dissociation identifies survival subgroups, potentially insightful to medical practice.

Model	C Index			Brier Score		
	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
NSC	0.922 (0.009)	0.902 (0.006)	0.882 (0.006)	*0.018* (0.001)	0.034 (0.001)	0.049 (0.001)
DCM	0.918 (0.008)	0.898 (0.005)	0.879 (0.006)	*0.018* (0.000)	0.034 (0.001)	0.049 (0.001)
DSM	0.924 (0.009)	0.900 (0.006)	0.880 (0.007)	*0.018* (0.000)	0.035 (0.001)	0.052 (0.001)
SuMo-net	**0.928** (0.009)	**0.908** (0.007)	**0.889** (0.005)	**0.017** (0.001)	**0.032** (0.001)	**0.047** (0.001)
DeepHit	*0.925* (0.009)	0.903 (0.007)	0.882 (0.006)	0.019 (0.000)	0.036 (0.001)	0.052 (0.001)
DeepSurv	0.924 (0.009)	*0.905* (0.006)	*0.887* (0.005)	*0.018* (0.001)	*0.033* (0.001)	*0.048* (0.001)
Survival Tree	0.894 (0.008)	0.874 (0.003)	0.852 (0.010)	*0.018* (0.000)	0.035 (0.001)	0.050 (0.001)
CoxPH	0.904 (0.011)	0.883 (0.007)	0.862 (0.007)	0.020 (0.002)	0.039 (0.004)	0.057 (0.006)

Table 4.3: Models' performance on the SEER dataset. Mean (std) over the 5-fold cross validation with best performance in **bold** and second best in *italic*. All neural networks present similar performance, validating the existence of the subgroups identified by the proposed NSC.

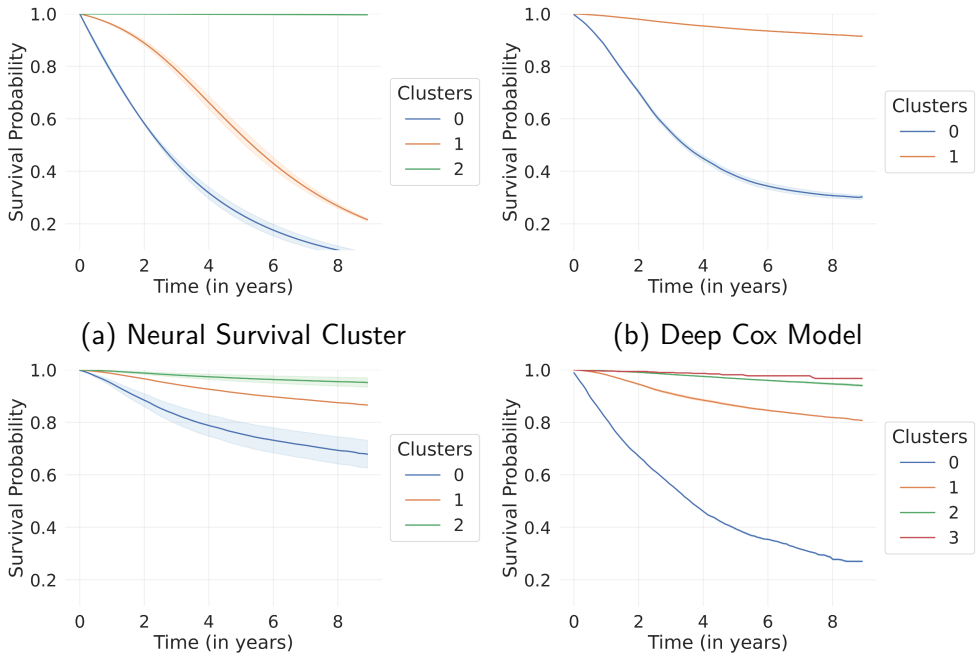
4.6.3 Survival subgrouping

Our first analysis involves exploring survival profiles after surgery and chemotherapy to identify subgroups benefiting from current medical practice or needing more medical attention. To this end, we explore survival subgroups using the proposed Neural Survival Clustering.

Figure 4.8: Averaged log-likelihood across 5-fold cross-validation given number of clusters K in the SEER dataset with the shaded area representing 95% CI. Due to the lack of a clear change in likelihood, we select the number of clusters as a hyperparameter.

Selecting number of clusters. As previously discussed, we use performance to guide the choice of the number of subgroups in the population. Figure 4.8 presents the log-likelihood as a function of K. In this real-world setting, there is no clear elbow but only a slight improvement in performance for $K = 3$. As this difference is relatively small, we select the number of clusters as a hyperparameter, chosen using a subset of the development set.

Survival profiles. Figure 4.9 displays the average survival distributions obtained with the considered methodologies. One can note how distinguishable the identified baseline distributions are. Additionally, the narrowness of the 95% confidence bands indicates NSC's consistency



(a) Neural Survival Cluster

(b) Deep Cox Model

(c) Cox-weighted K-Means

(d) Tree-Based Clusters

Figure 4.9: Averaged survival clusters across 5-fold cross-validation observed in the SEER dataset with the shaded areas representing 95% CI.

over the 5-fold cross-validation, validating the stability of these three clusters in the population contrary to the weighted K-Means alternative. A log-rank test confirmed that all survival distributions were significantly distinct at the 5% significance level.

NSC presents critical properties of interest. As evidenced by the lower DCM's performance, NSC offers more flexibility than the mixture of semi-parametric models. NSC performs similarly to DSM, which relies on patient-specific parametrisation, demonstrating that a mixture of distributions *independent* of the covariates is as representative of these real-world outcomes as more flexible individualised parametrisation. Further, NSC can assign new points to these different distributions, which means one can use NSC to assign patients to treatment in developing new trials.

Membership to a cluster and survival distributions are dissociated in the proposed modelling and can be studied separately. Notably, one may perform a permutation of the covariates [39] on the assignment network to identify which covariates most influence membership to a cluster. Following this permutation test, consisting of a random shuffle of a covariate repeated for each covariate and the measure of the change in likelihood, we identified the number of distant lymph nodes, HER2 status and ER status as the most discriminative covariates between groups. Figure 4.10 displays the ten most influential covariates, with the top 3 further described in Table 4.4. These characteristics underscore a cluster of long-term survivors observed across the different methodologies. Nearly 95% of the population presents a median survival longer than 8 years following diagnosis, reflecting the efficacy of existing procedures and treatments.

This population is characterised by a small number of metastases in distant lymph nodes, suggesting less advanced breast cancers. More medically relevant is the identification of the two smaller clusters, which present more severe cancer types with short-term survival with a more significant number of distant lymph nodes and HER2-positive cancer.

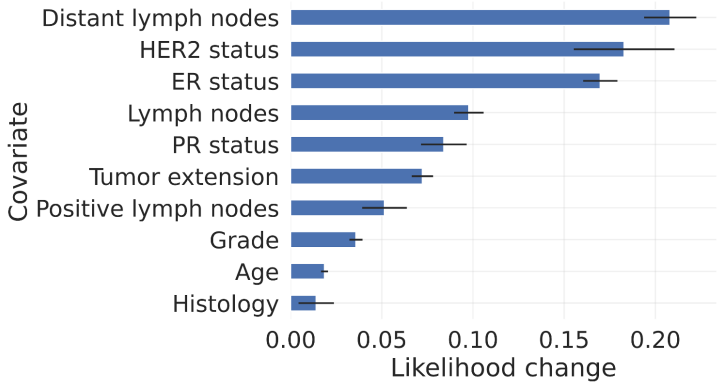


Figure 4.10: Neural Survival Clustering - Change in log-likelihood given random permutation of a given covariates. Distant lymph nodes, HER2 and ER status impact the likelihood the most.

	Median Survival (in years)	Population %	Distant Lymph Nodes	HER2 Positive	ER Positive
Cluster 0	3.07 (0.20)	4.2%	23.46 (14.32)	25.6%	49.2%
Cluster 1	6.85 (0.82)	1.1%	9.25 (14.41)	9.4%	55.5%
Cluster 2	> 8.00	94.7%	1.25 (5.96)	17.6%	46.5%

Table 4.4: NSC identified subgroups' characteristics in the SEER cohort described through percentage / mean (std).

4.6.4 Treatment response discovery

We identified three survival profiles in patients following surgery and chemotherapy. This association does not provide causal insights into who most benefits from adjuvant radiation. In this section, we examine the treatment response following adjuvant radiotherapy to identify groups of patients who received surgery and chemotherapy and may benefit from adjuvant radiotherapy. This problem is central to patients' treatment as no evidence-based guidelines for adjuvant therapy exist [177], making this setting more likely to meet the positivity assumption (Assumption (4.5)), necessary to study causality in observational data.

Uncovering treatment response. Figure 4.11 presents the identified treatment effect subgroups when using CNSC and CMHE, with the number of clusters K selected through hyperparameter tuning (see Appendix B.3 for the elbow analysis leading to the same choice). As previously mentioned, our proposed methodology presents two strengths that explain the difference in the identified clusters of treatment effects. First, the survival distribution under

(a) Neural Survival Treatment (b) Cox Mixture of Heterogenous Effects

Figure 4.11: Averaged treatment effect clusters across 5-fold cross-validation observed in the SEER dataset with the shaded areas representing 95% CI.

treatment is not constrained by the one under the control regime, resulting in more flexible, non-proportional distributions. CMHE's parametrisation, which characterises treatment as a linear shift in the log hazard, results in a proportionality assumption between treated and untreated distributions. Second, CMHE does not account for treatment non-randomisation in its average treatment effect estimation, whereas our use of inverse propensity weighting corrects for any observed ones.

Using a permutation test (see Figure B.2 in Appendix) leads to identifying the same 3 covariates most indicative of the different treatment response subgroups. Table 4.5 summarises the average value across the identified subgroups and the life expectancy gain when using adjuvant radiations measured through the Restricted Mean Survival Time (RMST) [269]. Both methodologies identify a population with limited treatment response. However, our proposed methodology identifies a second group, characterised by larger HER2 and larger distant lymph node count, with a positive treatment response, gaining more than half a year of life expectancy over the five years following diagnosis.

	RMST at 5 years	Population %	Treated %	Distant Lymph Nodes	HER2 Positive	ER Positive
Cluster 0	0.00 (0.00)	93.3%	55.5%	1.18 (5.74)	17.4%	46.6%
Cluster 1	0.62 (0.16)	6.7%	47.5%	17.75 (16.08)	23.4%	48.6%

Table 4.5: Causal Neural Survival Clustering subgroups' characteristics in the SEER cohort described through percentage / mean (std).

The proposed analysis pinpoints a group that could benefit from adjuvant radiotherapy. However, our methodologies remain hypothesis-generating tools, requiring further experimental validation, particularly due to potential confounding through hormonal therapy (not available in this dataset), the temporal nature of treatment, and the plurality of treatment options.

4.7 Discussion

We summarise this chapter's contributions, the recommendations for using the introduced strategies, and future work directions in the following.

4.7.1 Contributions

Observational data offer a valuable source for exploring disease heterogeneity. In this chapter, we formalise the problems of uncovering subgroups of outcomes and treatment effects as latent clustering problems. Following this formalisation, we introduce a neural network architecture composed of (i) monotonic neural networks to model each subgroup's survival distribution independent of patients' covariates and (ii) an assignment network that associates a cluster given patients' observed covariates. This mixture of survival distributions improves interpretability by dissociating survival distribution from individual covariates while maintaining comparable predictive performance to existing ML models, that aim for individualised predictions. Importantly, our proposed methodology avoids the parametric and proportional hazards assumptions made by existing ML mixture models [221, 225].

When treatment options exist, identifying subgroups that do not respond to these therapies is critical to inform guidelines and future clinical trials. The non-randomisation of treatment in observational data introduces a selection bias that must be corrected for treatment effect estimation. We use an inverse propensity of treatment weighting to adjust the likelihood used to train the previous architecture. Further, we alter each cluster's associated survival distribution to model two distributions: one under the treated and control regimes. This altered architecture identifies subgroups with different causal treatment responses.

The two proposed architectures offer complementary insights into observational data. The first highlights the limitations and strengths of current clinical practices by identifying subgroups of survival risk profiles. The second reveals the differences in treatment responses to identify patients that would most benefit from existing treatment and groups for which the existing options do not result in improved outcomes. We demonstrate the proposed models' capacity to uncover such subgroups and their associated survival and treatment response distributions through a synthetic dataset and further illustrate the utility of the methodology with a detailed analysis of breast cancer subgrouping in the observational SEER dataset.

This chapter presents critical implications for shaping policies and trial designs. Identifying subgroups of patients underserved by current medical practice or presenting worse outcomes holds the promise of redirecting resources towards these subgroups. Further, these methodologies can generate hypotheses to design future trials for developing targeted treatments. Beyond improved medical knowledge, subgroup identification can help combat health disparities by identifying differences in care for different subgroups [225].

4.7.2 Recommendations

Contrary to the existing literature aiming for individualised modelling when studying treatment response heterogeneity, we propose methodologies to identify subgroups in observational data directly. To uncover subgroups within observational data, we recommend the following steps to appropriately use our proposed approaches:

1. *Study heterogeneity in observed outcomes:* One can use the proposed NSC to identify different survival profiles under current practices. If the absence of medical insights, performance should guide the choice of the number of subgroups to consider, as discussed in Section 4.5.3. This tool uncovers *association* between covariates and outcomes, not causal relations. Consequently, it should not be used to guide treatment recommendations.

2. *Study heterogeneity in treatment responses:* model developers should investigate the treatment assignment strategy with medical experts when treatment options are available. Notably, one must assess the positivity assumption, ensuring that treatment is not deterministically assigned. If this assumption holds, the proposed CNSC can be used to uncover underlying subgroups of treatment responses. Once again, medical expertise is required to investigate whether the identified subgroups are not the product of unmet hypotheses.

Following these recommendations, these models can identify new subgroups from observational data. These hypothesis-generating tools have the potential to inform the design of follow-up clinical trials to validate observational results, identify alternative treatments, and ultimately shape policies.

4.7.3 Future work

This chapter opens multiple avenues for future work. First, our current approaches rely on neural networks for assigning covariates to the different clusters. Using models with fewer parameters to learn this assignment through *a posteriori* distillation [128] could lead to a more interpretable exploration of subgroups. Second, integrating other competing outcomes for unbiased survival estimates, as studied in Chapter 5, could be accounted for to model potential side effects. Finally, the current choice of the number of clusters, guided by performance, remains the practitioners' choice. To motivate this choice and inform this number when unknown, we would like to explore the integration of a Dirichlet Process prior in a Bayesian framework [316].

Chapter 5

Competing Risks

Associated Publications. The work presented in this chapter is based on our publication: Neural Fine-Gray [149] presented at CHIL 2023.

Problem statement. *Why ignoring events that preclude the outcome of interest bias predictive modelling?*

5.1 Motivation

When modelling the time to an outcome of interest such as cardiac event, other outcomes may preclude it. For instance, patients who die during the study can no longer experience a cardiac event. While these patients do not experience the event of interest, they inform its modelling as they remain event-free. However, unlike censored patients, these patients can *no longer* experience the outcome of interest. These events, known as *competing risks*, inform the modelling of the event of interest. If ignored, these competing risks bias the time estimate to the event of interest. In the previous example, discarding patients who died over the study period or treating them as censored would result in an overinflated prediction of cardiac events.

Competing risks underscore the intricate relationship between the observational process and predictive modelling. Beyond the process associated with observed covariates and treatment, clinical presence impacts observed outcomes, as events related to a patient's conditions or treatments may hinder the occurrence of the outcome of interest. As demonstrated in this chapter, failure to account for these competing risks compromises the accuracy of predictions, particularly for the patients most at risk for these events.

Despite competing risks' prevalence in medicine, these risks remain an overlooked challenge [13, 167]. In their review of 50 medical research papers investigating time-to-event outcomes, Koller et al. [167] identified competing risks among 70% of the studies, with more than half inadequately accounting for them in their analysis. Particularly, practitioners frequently ignore competing risks, considering patients experiencing them as censored [11]. As these events

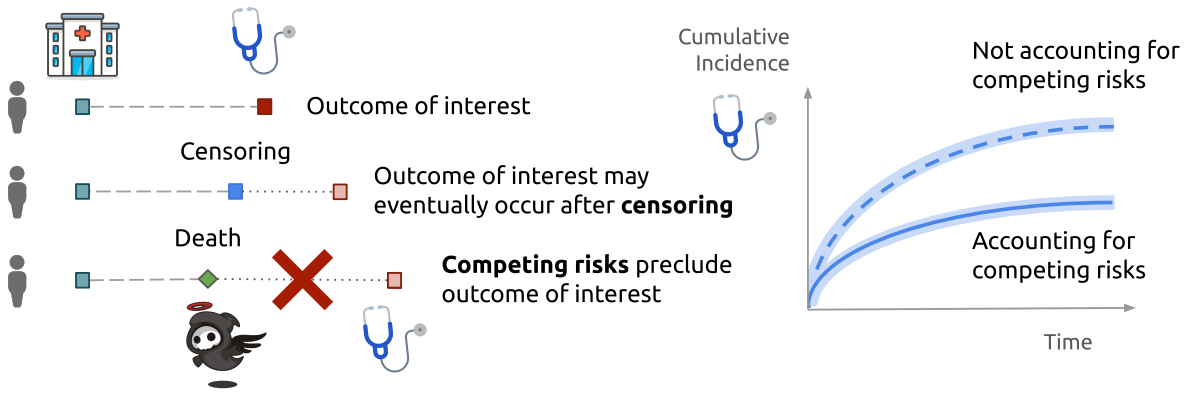


Figure 5.1: Potential biases in risk estimation emerging from competing risks. When experiencing risks precluding the event of interest, such as death, the patient can no longer experience the event of interest.

inform the observation of the event of interest, this common practice breaks the assumption of non-informative censoring. Considering competing risks as censoring, therefore, results in misestimating risk [89, 182, 279, 283] as illustrated in Figure 5.1 where not accounting for death results in an overestimation of the risk of being diagnosed with a given condition.

Risk misestimation has consequential medical repercussions [44]. At the individual level, misestimation impacts medical recommendations for treatment. At the population level, risk assessment informs guidelines and policies affecting care quality and costs. For instance, cardiac event management guidelines rely on the estimated risk to recommend treatment [201]. Overestimation of cardiac risk can lead to unnecessary treatment and prioritisation of patients at low risk. At the intersection of individual and population risk, we argue that discarding competing risks results in group-specific misestimation. For a given condition, different groups often differ in their risk profiles. This difference is not only observed for the event of interest but also for the competing risks. For instance, *at the same age*, women often present a lower risk than men for cardiac event [356], and a longer life expectancy [284]. Consequently, ignoring competing risks may impact these groups differently, increasing the error gap between groups characterised by different risk profiles.

This chapter investigates the algorithmic fairness repercussions of the common practice of discarding competing risks by quantifying the error associated with this approach. Our contribution in this regard is four-folds:

1. We provide a theoretical measure of the error in survival estimate resulting from considering competing risks as censoring. We then validate this measure in a simulation study.

2. Building upon the proposed error quantification, we analyse the algorithmic fairness consequences of ignoring competing risks. We theoretically demonstrate and empirically validate the systematic error between groups with different risk profiles.

3. Due to the demonstrated importance of accurate modelling, we introduce a novel approach to model competing risks while avoiding the approximations and assumptions made by existing models. We study the strengths and limitations of the proposed methodology with simulations and three real-world applications.

4. Finally, we underline our results' medical relevance through the analysis of the FRAMINGHAM cardiovascular risk score. In its original form, this risk score ignored competing risks. We analyse the implications for population and group performance and the resultant over-treatment.

Quantifying the error associated with discarding competing risks is critical to encourage practitioners to use models that account for competing risks and understand the impact on different patients. Our work proves that considering competing risks as censoring leads to systematic under-estimation of survival at the *individual*, *group* and *population* levels. Critically, we mathematically justify the intuition that patients most at risk of experiencing competing risks suffer the most from this practice. When groups differ in risk profile, the error is the largest for the group most at risk. Ignoring competing risks can, therefore, not only harm medical practice but also increase inequities.

This work has critical implications for predictive modelling, risk estimation and associated guidelines. Failing to account for the impact of clinical presence on the observation of competing risks diminishes models' clinical utility and exacerbates inequities between groups with different risk profiles. Our theoretical and experimental contributions highlight a simple opportunity for improvement: medical guidelines development must account for competing risks to serve all patients.

This chapter first explores the related literature on algorithmic fairness in survival analysis and the limitations of the current handling of competing risks. In Section 5.3, we formalise the problem of competing risks. We then theoretically analyse the error resulting from the common practice of disregarding competing risks on individual and group survival estimates in Section 5.4. In Section 5.5, we introduce a novel survival model to efficiently and accurately model these outcomes. We validate our theoretical results and demonstrate the benefits and limitations of our approach in a simulation study in Section 5.6, and three real-world medical datasets in Section 5.7. Finally, we further investigate the FRAMINGHAM dataset to underline the importance of considering competing risks for improved and more equitable cardiovascular disease risk estimation.

5.2 Related work

While survival analysis has received considerable attention, competing risks are less well studied [336] and even less addressed in empirical studies [214]. This section reviews the known

biases associated with the common practice of ignoring competing risks and the limitations of existing methods to tackle this problem. Finally, we discuss why survival analysis presents its own algorithmic fairness challenges and the literature at the intersection of these fields.

5.2.1 The known biases associated with ignoring competing risks

The existing literature emphasises the inadequacy of Kaplan-Meier estimates under competing risks with multiple empirical evidence [111, 301, 352]. Satagopan et al. [279] provides intuitions on the biases resulting from considering competing risks as censoring: the estimated risk for patient experiencing the event of interest is overinflated using a Kaplan-Meier estimates. Despite this evidence, Kaplan-Meier estimates continue to be misapplied [334], as demonstrated by Walraven and McAlister [334]'s analysis of 100 publications, revealing that nearly half relied on the Kaplan-Meier estimator in the presence of competing risks. Critically, a third of these studies overestimate the underlying risk by more than 10%.

Empirical evidence in the literature highlights biases introduced when ignoring competing risks beyond non-parametric estimators when comparing the Cox and Fine-Gray models [66, 283, 354]. Wolkewitz et al. [354] advocates for careful examination of assumptions underlying survival models, demonstrating that excluding competing risks in the Cox model results in overestimating risk. Wolbers et al. [351] and Berry et al. [31] echo this caution, emphasising the importance of considering competing risks, especially in frail populations who are at a higher risk for competing risks. This comment highlights the potential group-specific misestimation and its algorithmic fairness consequences studied in this work.

Despite these observations, there is a notable absence of work quantifying the theoretical errors stemming from the common practice of ignoring competing risks. Our work aims to fill this gap in the literature.

5.2.2 Limitations of existing competing risks models

Due to the biases associated with ignoring competing risks, enhancing the modelling of competing risks is crucial for informing clinical decisions. Current approaches often centre around modelling the marginal probability of each risk, represented by the Cumulative Incidence Function (CIF). However, akin to the single-risk approaches discussed in Chapter 2, these methodologies frequently rely on proportional hazards or parametric assumptions or employ numerical integration to estimate the CIF and/or the likelihood necessary for training these models. These assumptions can lead to sub-optimal target functions that misrepresent the underlying survival distribution. The following presents a review of these approaches and their associated limitations.

Proportionality assumptions. Common in the medical literature [307], the Cox model assumes proportional hazards to render tractable the covariate effect estimation. The cause-specific Cox model [256] extends the Cox model to tackle the problem of competing risks. The method independently estimates each risk-specific hazard using a Cox model. Then, it combines these models to evaluate the risk associated with the event of interest given the non-observation of all other competing risks. This approach is limited on two fronts. First, the proportional hazards assumption may not hold in real-world medical settings [307]. Invalidity of this assumption means that the ratio of hazard for any patient to the baseline hazard does not remain constant over time. Second, this step-wise approach to competing risks assumes independence between the times associated with the different risks and may, consequently, misrepresent the relative effect of these covariates across risks [13]. To address this latter, Fine-Gray [88] model the sub-hazards $h_r(t)$, i.e., the probability of observing a given event r if the patient has not experienced this event until t. Despite providing insights into the link between covariates and risk particularly suitable for prediction [12], this model still relies on an assumption of proportionality analogous to the one made in the Cox proportional hazards model to estimate the sub-hazards. Further, this approach can result in an ill-defined survival function with the sum of CIFs greater than one [14].

Parametric assumptions. Parametric models constrain the survival function form by assuming a given distribution to obtain a closed-form likelihood. When the parametric choice does not capture the complexity of the underlying survival distribution, this assumption leads to misestimating risks in both single and competing risks. [221] extend their proposed mixture of Weibull distributions to competing risks by modelling each CIF with a different mixture. Similarly, [26] introduces a Bayesian mixture of Generalised Gamma distributions to tackle competing risks.

Approximated likelihood. Contrarily to the previous methods, DeepHit [179] does not make parametric assumptions on the survival function but uses a coarse time discretisation to simplify the problem to a series of discrete classification problems. Straightforwardly, this approach extends to competing risks [179] by multiplying the output dimension by the number of risks. Similarly, [321] propose to form a hierarchy of conditional survival sub-problems. This method iteratively splits time intervals and models the probability of each competing risk happening in this split conditioned on the previous iteration's likelihood. Extrapolation of DeepHit to infinite time discretisation resembles an ordinary differential equation (ODE), as proposed in [78]. While these approaches do not make assumptions on the underlying survival distribution, the likelihood does not exhibit a closed form. Discretisation [179, 321], numerical integration [1] or pseudo-value approximation [78, 262] are then necessary to estimate this quantity, leading to modelling the underlying survival function but through an approximate loss function.

Avoiding approximations. The previous methods either optimise towards an approximated survival function or perform an approximate optimisation towards the exact underlying distribution. A third option is possible in which one does not specify a parametric survival function nor approximate the likelihood. In the single-risk context, Rindt et al. [266] propose a constrained neural network with a monotonically decreasing output over time to model the survival function, then use automatic differentiation to derive the exact likelihood. Our work generalises this work to competing risks, harnessing monotonic neural networks to model each competing risk's CIF instead of the survival function. Our proposed method tackles the limitations of existing strategies by using an exact computation of the likelihood with a lower computational cost.

For completeness, non-likelihood-based approaches such as boosted trees [27] or survival trees [280] exist but fall beyond the scope of our work, which focuses on neural networks for their flexibility in handling diverse modalities.

5.2.3 Algorithmic fairness in survival analysis

While the literature has explored novel ways to model survival outcomes from observed covariates, there is limited research at the intersection between survival analysis and algorithmic fairness. Existing works focus on the characteristic problem of survival analysis: censoring. Specifically, unobserved outcomes make the quantification and mitigation of group disparities harder. Zhang and Weiss [374] introduce a concordance impurity metric to quantify the group difference in discriminative performance with censored data. Leveraging this metric, the authors adapt survival random forests with an updated splitting procedure to maximise this metric. Hu and Chen [138] approach the problem through a distributionally robust optimisation, improving algorithmic fairness properties without access to demographics. Do et al. [82] explore a training procedure minimising the inter-group mutual information. Instead of modifying the model's training, [378] propose pre- and post-processing to maximise algorithmic fairness. More recently, Rahman and Purushotham [261] proposed a pseudo-value model to reduce the bias resulting from ignoring the censoring mechanisms, particularly when assumptions of non-informativeness do not hold.

While enforcing fairness has been studied in the context of survival analysis with a focus on censoring, to our knowledge, no prior study has analysed the fairness gap resulting from the common practice of ignoring competing risks.

5.3 Competing risks

Competing risks are mutually exclusive outcomes: observing one precludes the others. When analysing survival data with competing risks, one is interested in modelling the time to one of the events, but any could occur. Critically, the occurrence of any event informs the others

as they can no longer happen. This section introduces a formalisation of this problem to understand how this problem differs from single-risk survival analysis and the related quantities of interest.

5.3.1 Formalisation

Consider the observed random variables: X, the observed covariates, T, the time of the first observed event, and D, its associated event type. Formally, the latter variables are deterministically defined as: $T := \min(C, T_1, ..., T_R)$ and D is the associated index, i.e. 0 if censoring occurs first and r if T_r is the minimum realised times, with T_r the random variable associated with the time of event $r \in [\![1, R]\!]$ and C the (right-)censoring time. Similarly to the non-competing setting, censoring describes exit from the study before experiencing *any of the considered events*. From these variables, one aims to model the event time associated with the risk r in the presence of the different potential outcomes. Critically, T_r is not this quantity as it does not account for the competing events. To describe the dependence between competing events, we introduce $T_{r|\neg r}$, the time associated with event r *considering the occurrence of any competing risk*. This variable is the key quantity of interest as it quantifies the event process while considering the other events. Formally, this random variable is defined as:

$$T_{r|\neg r} := \begin{cases} \infty & \text{if any competing risk occurred before event r, i.e., } \exists r' \neq r, T_{r'} < T_r \\ T_r & \text{otherwise} \end{cases}$$

As one cannot observe the event r if any other precedes it, its associated time is set to

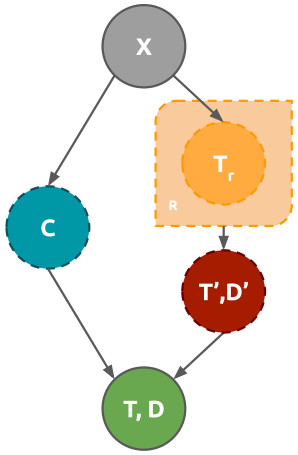


Figure 5.2: Competing risks DAG. *Each node represents a random variable — a dotted contour indicates that the variable is unobserved, a solid one, observed. The square around the random variable T_r denotes the repetition of the variable for $r \in [\![1, R]\!]$. Arrows between nodes represent dependencies.*

infinity. While $T_{r|\neg r}$ calls attention to the quantity of interest, it does not lend itself to easy mathematical manipulations. To tackle this issue, we decompose this quantity into the

equivalent following notations: the time of the first event T' of any type D' (independently of the censoring time). Note that the joint distribution of T' and $D' = r$ is equivalent to $T_{r|\neg r}$. Further, in the context of a single risk, $D' = 1$ is the only possible outcome, resulting in $T' = T_r = T_{r|\neg r}$. To limit the complexity of the following mathematical notations, we do not use the notation $T_{r|\neg r}$, focusing on T' and D'.

Figure 5.2 illustrates the dependence structure between the previously introduced variables through a DAG. X, D and T are observed, while the variables of interest T' and D' are not. The plate notation denotes repeated variables across the R competing risks. Importantly, this representation underlines the challenge associated with competing risks through the dependence of T', D' on *all* T_r.

5.3.2 Quantities of interest

The proposed formalism allows us to describe the quantities modelled when considering competing risks. First, the overall survival function $S(t)$ becomes the probability of observing none of the competing risks before time t:

$$S(t) := \mathbb{P}(T' \geq t) = 1 - \sum_{r \in [\![1,R]\!]} F_r(t)$$

where F_r, the CIF for the event r, denotes the probability of observing the event r before time t without prior occurrence of any competing event(s), i.e.:

$$F_r(t) := \mathbb{P}(T' < t, D' = r) \tag{5.1}$$

Existing approaches to competing risks often estimate this quantity by extending single-risk models. For instance, the cause-specific Cox model [256] consists of fitting independent Cox models on each risk while treating all other outcomes as censoring. Then, one computes the CIF as the integral of observing the event r in an infinitesimal interval given no other prior event:

$$F_r(t) = \int_0^t \lambda_r(u) e^{-\int_0^u \sum_k \lambda_k(s) ds} du \tag{5.2}$$

with $\lambda_r(t)$ the cause-specific hazard, i.e., the instantaneous risk of observing the event r in the instant following t, in the absence of prior occurrence of competing events, defined as:

$$\lambda_r(t) := \lim_{\delta t \to 0} \frac{\mathbb{P}(t < T' < t + \delta t, D' = r \mid T' \geq t)}{\delta t}$$

This two-step approach can extend any survival model introduced in Chapter 2. However, in the absence of a closed-form expression, Equation (5.2) necessitates numerical integration, resulting in an approximation of these quantities of interest. Further, this staged modelling may

misestimate the covariate effects [325] as it assumes independence of the different outcomes. Fine-Gray [88] overcomes this issue by modelling another related quantity: the sub-distribution hazards h_r defined as:

$$h_r(t) := \lim_{\delta t \to 0} \frac{\mathbb{P}(t < T' < t + \delta t, D' = r | [T' \geq t, D' = r] \cup [D' \neq r])}{\delta t}$$

Particularly, [88] decomposes this quantity as:

$$h_r(t \mid x) = \bar{h}_r(t \mid x) e^{\eta(x)}$$

with η a function parameterising the deviation from the population's non-parametric sub-hazard estimate $\bar{h}_r$ given the covariates x.

Note, when there are no competing risks ($R = 1$), cause-specific hazards and sub-distribution hazards are equal, and all the previous quantities simplify to the definitions introduced in Chapter 2. In this case, the Fine-Gray model is equivalent to the Cox model.

5.3.3 Likelihood

To fit a model of the previous quantities, one often aims to maximise the likelihood of an observed set of points. Consider a realisation of the previous observed random variables of the form $\{x_i, t_i, d_i\}$ with x_i the covariates for patient i, $t_i \in \mathbb{R}^+$ the time of the first observed outcome (including whether censored) from the appropriately defined time origin and $d_i \in [\![0, R]\!]$ its associated cause. If $d_i \in [\![1, R]\!]$, the patient left the study due to one of the R considered risks. Otherwise, $d_i = 0$, the patient is right-censored. Specifically, we assume non-informative censoring *once controlled* on all identified competing risks, meaning that the patient left the study for a reason unrelated to any of the competing risks. Under this assumption, one can express the likelihood analogously to (2.2): patients with observed events contribute to the likelihood as the probability of observing the event d_i at t_i with no other prior events, i.e., $\lambda_{d_i}(t_i|x_i)S(t_i|x_i)$ using the previously introduced notations. Note that this quantity is the partial derivative of F_r with respect to t evaluated at t_i. Remaining censored patients influence the likelihood as the probability of observing no event until t_i, i.e., $S(t_i|x_i)$. The *competing risks log-likelihood* is, therefore, expressed as:

$$l = \sum_{r \in [\![1,R]\!]} \sum_{i, d_i = r} \log \frac{\partial F_r(u|x_i)}{\partial u}\bigg|_{u=t_i} + \sum_{i, d_i = 0} \log \left[1 - \sum_r F_r(t_i|x_i)\right] \tag{5.3}$$

5.4 The impact of ignoring competing risks

Equipped with the previous notations, this section mathematically formalises how ignoring competing risks biases the CIF estimate F_r both at an individual and group level.

5.4.1 Impact on cumulative incidence estimate

As previously described, a common practice is approaching competing risks as censoring, effectively discarding outcomes potentially informative of the quantity of interest S_r. This strategy effectively estimates the distribution of T_r, instead of the joint distribution of T' and $D' = r$. Adequately modelling competing risks consists of evaluating the CIF described in equation (5.2) — denoted as F_r^C in this section, whereas ignoring them means estimating the conditional distribution:

$$F_r^{NC} := \mathbb{P}(T' < t \mid D' = r) = \mathbb{P}(T_r < t)$$

We aim to quantify the overestimation resulting from this common practice and highlight its associated medical consequences. To this end, we define the relative cumulative incidence discrepancy when not adequately accounting for competing risks. This quantity corresponds to the difference in estimates divided by the quantity modelled, formally:

Definition 5.1 (Relative cumulative incidence discrepancy)**.** The relative cumulative incidence discrepancy for risk r, denoted by L^r, from considering competing risks as censoring is the relative difference between the CIF curves ignoring competing risks and considering them:

$$L^r(t, x) := \frac{F_r^{NC}(t \mid x) - F_r^C(t \mid x)}{\max(F_r^{NC}(t \mid x), F_r^C(t \mid x))} \tag{5.4}$$

Intuition. The magnitude of L^r quantifies the disagreement in the probability estimates of experiencing r from the two approaches considered for modelling competing risks. The normalisation ensures that the quantity is in the range $[\![-1, 1]\!]$, where 0 indicates perfect alignment between the modelling strategies, positive values denote an overestimation of the risk from discarding competing risks, and negative values indicate an overestimation of the risk from the strategy modelling competing risks.

Direct application of the previous definition and Bayes' theorem lead to the following result expressing the relative discrepancy as the probability of observing any competing risk (See proof in Appendix C.1).

Theorem 5.1 (Relative cumulative incidence discrepancy induced by ignoring competing risks)**.** *The relative cumulative incidence discrepancy for a given outcome r is the probability of not observing this outcome given the covariates x:*

$$L^r(t, x) = \mathbb{P}(D' \neq r \mid x)$$

with D', the random variable associated with the first observed competing risk if there was no censoring.

Intuition. This expression underlines how ignoring competing risks biases the cumulative incidence estimate proportionally to the individual likelihood of experiencing other competing outcomes. The theorem encapsulates two key properties: (i) independence of time as dependent on which event occurs first, irrespective of when and (ii) positivity. While the independence on time demonstrates a constant discrepancy, the latter property substantiates the empirical evidence that treating competing risks as censoring *always* overestimates the CIF [89, 182, 279, 283]. When competing risks are independent of individual covariates (i.e., $\forall s \neq r, T_s \perp\!\!\!\perp X$), ignoring competing risks results in a uniform shift in the survival estimate for all patients, resulting in no difference in the two methods discriminative performance. However, this result highlights a potential correction of models ignoring competing risks by estimating the relative risk of event r. Note that as D' is unobserved, this requires a model adjusting for the censoring mechanism. Crucially, this theorem demonstrates that a patient who is more likely to observe competing risks derives more benefits from accounting for competing risks.

5.4.2 Impact on group-specific estimate

Theorem 5.1 underlines the relative discrepancy emerging from ignoring competing risks at the *individual level*. The following describes the differential group impact and the resulting algorithmic fairness consequences. Consider g_i, the group membership for patient i. Following the previous definitions of algorithmic fairness introduced in Section 3.4, we define the group-specific discrepancy and the associated algorithmic fairness gap.

Definition 5.2 (Group-specific discrepancy)**.** The group-specific discrepancy from considering competing risks as censoring is the expected relative cumulative incidence discrepancy across all members of a group g for the event of interest r:

$$L_g^r := \mathbb{E}_{x_i|g_i=g}\left[L^r(x_i)\right]$$

Definition 5.3 (Discrepancy gap)**.** The gap from considering competing risks as censoring between members of a group g and the rest of the population is the difference in group-specific discrepancies:

$$\Delta_g^r := L_g^r - L_{\neg g}^r$$

Intuition. The gap in CIF discrepancy quantifies how groups are differently impacted by the common practice of ignoring competing risks. The group-specific discrepancies are in the range $[\![0, 1]\!]$, resulting in a difference that spans the range $[\![-1, 1]\!]$. A value of 0 indicates no difference in the impact of ignoring competing risks at the group level. Positive values signify that group g exhibits a larger discrepancy, indicating a greater error from ignoring competing risks than the rest of the population. Conversely, negative values suggest the opposite scenario, denoting a decreased impact for group g compared to the overall population.

These definitions and Theorem 5.1 directly lead to the following expression of the inter-group error gap.

Theorem 5.2 (Fairness gap resulting from ignoring competing risks)**.** *The gap in survival estimate errors when ignoring competing risks is the difference in the probabilities of observing any of the competing risks other than r:*

$$\Delta_g^r = \mathbb{P}(D' \neq r \mid g) - \mathbb{P}(D' \neq r \mid \neg g)$$

Intuition. Modelling competing risks reduces the algorithmic fairness gap proportionally to the difference in the probabilities of observing competing risks between the considered groups. Critically, accurate survival modelling is not enough to reduce this gap if competing risks are ignored, nor is the common practice of controlling for group membership in the hope of reducing disparities. As long as groups present with different competing risk profiles, one must account for competing risks to improve the model's algorithmic fairness.

In conclusion, this theoretical analysis quantifies the error arising from ignoring competing risks and the associated group disparity. These results invite practitioners to account for competing risks and complement these recommendations with a novel insight into how discarding competing risks has algorithmic fairness consequences.

5.5 Proposed approach

The previous theoretical analysis of the error reduction associated with modelling competing risks relies on accurate survival estimates. Approaches to tackle the problem of competing risks often make simplifying assumptions or approximations that diminish their capacity to capture the complexity of the survival distributions. This section introduces Neural Fine-Gray to tackle existing models' limitations. Our model is named after the original Fine-Gray model [88] due to its joint approach to model competing risks. The following equation highlights the link between sub-distribution hazards h_r and CIFs F_r, i.e., between the quantities jointly modelled by the original and Neural Fine-Gray models:

$$h_r(t) = -\frac{\partial \log(1 - F_r(u))}{\partial u}\bigg|_{u=t}$$

5.5.1 Architecture

The core challenge in accurately modelling competing risks is to maximise towards the *exact* likelihood. To compute this quantity, one must avoid approximations of this quantity and ensure that the set of distributions allowed by the proposed model contains the underlying survival distribution. Existing approaches satisfy one or the other requirement, often constraining

the survival function form to obtain a closed-form likelihood, or allowing flexibility on the distribution but approximating likelihood or CIF. These techniques aim to render tractable the integration in Equation (5.2). Integration is computationally expensive and inexact, whereas differentiation is exact through one backward pass by automatic differentiation — available in most ML libraries used for training neural networks.

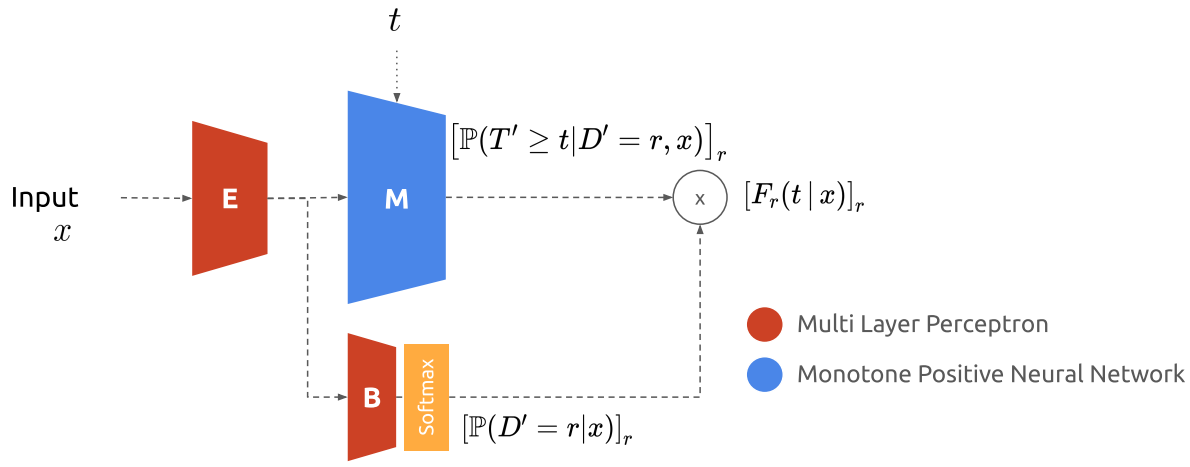


Figure 5.3: Neural Fine-Gray Architecture. *E embeds the covariate(s) x, which are then inputted in the monotonic networks M and balancing network B to estimate the CIFs.*

Our proposed approach, Neural Fine-Gray (NeuralFG) illustrated in Figure 5.3, aims to take advantage of automatic differentiation by considering competing risks as a differentiation problem instead of an integration one. Specifically, our method outputs the CIFs and computes the likelihood through differentiation. By using a universal approximator, the proposed approach avoids parametric assumptions upon the CIFs. Further, automatic differentiation results in the exact likelihood computation (Equation (5.3)), while reducing the computational cost compared to inexact numerical integration techniques.

To approach competing risks as a differentiation problem, we must ensure that the model's outputs satisfy all CIFs' properties: monotonicity, constrained between 0 and 1, and summation across risks is upper-bounded by 1. To this end, we decompose F_r as follows:

$$\begin{aligned} F_r(t|x) &= \mathbb{P}(D' = r|x) \cdot \mathbb{P}(T' \leq t | D' = r, x) && \text{(Bayes' Theorem)} \\ &= B(E(x))_r \cdot [1 - \exp(-t \times M_r(t, E(x)))], \end{aligned}$$

with three neural networks: E, M and B.

Embedding network (E). A first multi-layer perceptron E with inter-layer dropout extracts an embedding $\tilde{x}$ shared between the different cause-specific networks from the covariates x.

Sub-distribution networks ($[M_r]_{r\in[\![1,R]\!]}$). The embedding $\tilde{x}$ is inputted in R positive monotonic networks[1] $[M_r]_{r\in[\![1,R]\!]}$ representing a lifetime distribution conditioned on one risk r, through the relation $\mathbb{P}(T' \leq t|\tilde{x}, D' = r) = 1 - \exp(-t \times M_r(t, \tilde{x}))$. Recall, a *positive monotonic neural network* is a network constrained to have its outcome monotonic and positive given its inputs. We enforce these constraints by enforcing all weights to be positive through a square function and use a final *SoftPlus* layer for positivity. In this context, the embedding network E ensures the outcome to be monotonic in terms of $\tilde{x}$ and t, but not directly x - which would be an unrealistic assumption. Finally, the chosen transformation with the multiplication by t is to address the limitation raised by Shchur, Biloš, and Günnemann [290] that monotonic neural networks can attribute non-null density to non-positive times, i.e., $F_r(t = 0) \neq 0$. Our proposed transformation addresses this issue and avoids an unnecessary logarithmic operation in the log-likelihood computation.

Balancing network (B). A multi-layer perceptron B with a final *SoftMax* layer leverages $\tilde{x}$ to balance the probability of observing each risk $B(\tilde{x}) := [\mathbb{P}(D' = r|\tilde{x})]_r$. This weighting ensures that the survival function is correctly constrained, i.e., $\sum_{r\in[\![1,R]\!]} F_r(t|x) \leq 1$, contrarily to the standard Fine-Gray model [14].

The proposed approach directly models F_r by multiplying the distribution's outputs and balancing networks. Automatic differentiation of the model's output results in the derivative $\frac{\partial F_r(u|x_i)}{\partial u}\Big|_{u=t_i}$ then used for computing the likelihood (Equation (5.3)). Monotonic neural networks are central to ensuring the properties of CIFs are satisfied. However, any architecture or optimisation ensuring the monotonicity given time and allowing automatic differentiation are possible alternatives.

Remark 5.1. As previously mentioned, the proposed methodology is a generalisation of the survival model Sumo-Net [266] that estimates S in the single-risk setting. If $R = 1$, then $F_r = 1 - S$ and $B_r = \mathbb{1}$, resulting in the equivalence of the proposed approach with Sumo-Net. Moreover, the proposed decomposition echoes DeSurv [78] while avoiding the proposed numerical integration necessary to estimate the lifetime distributions.

5.5.2 Computational complexity

The proposed method does not restrict the CIFs to parametric forms and uses the exact likelihood computation. However, another architecture, DeSurv [78], converges in the limit towards these quantities, i.e., when increasing the number of point estimates for the numerical integral used

[1]Note that for model flexibility, we used R neural networks. As an alternative, we explore in Appendix C.3.1 how using one network with R outcomes impacts performance.

for the likelihood computation. This section demonstrates NeuralFG's computational gain at training and prediction times compared with DeSurv.

To this end, one must understand how DeSurv [78] approaches the modelling of the CIFs. DeSurv's architecture outputs the cause-specific hazards with only positivity constraint. Then, the CIFs is the solution of an ODE, one aims to obtain the integral of the cause-specific hazard with an initial condition. Specifically, Danks and Yau [78] assumes the following form:

$$F_r(t|x) = h(x)\text{TanH}(v(x,t))$$

with v being the solution to the ODE defined as $\frac{\partial v(x,u)}{\partial u}\Big|_{u=t} = g(x,t) > 0$ and $v(x,0) = 0$ with g, a neural network with positive range, and h, a second neural network modelling the probability of observing event r. This constraint, achieved through a final *SoftPlus* transformation, ensures the monotonicity of v and, consequently, of F_r as the integral of a positive function with a monotonic transformation: TanH. However, the proposed approach requires solving the ODE. For efficiency, the authors propose a Gauss-Legendre quadrature to obtain v, instead of relying on a complex solver. This approximation consists of estimating the neural network's output at n time points $[t_j(t)]_{j\in[\![1,n]\!]}$ weighted by the associated $[w_j]_{j\in[\![1,n]\!]}$ (see [257] for a detailed description of Gauss-Legendre quadrature). To estimate the CIF, one must compute $\frac{\partial v(x,u)}{\partial u}\Big|_{u=t_j(t)}$ at the n points and then evaluate:

$$\hat{F}_r(t|x) = h(x)\text{TanH}\left(\frac{t}{2}\sum_{j\in[\![1,n]\!]} w_j g\left(x, \frac{t}{2}\cdot t_j(t)\right)\right)$$

DeSurv's training cost. Similar to our approach, DeSurv relies on backpropagation of the competing risks' log-likelihood for training as described in Equation (5.3). Estimation of $\hat{F}_r(t_i \mid x_i)$ and $\frac{\partial \hat{F}_r(u|x_i)}{\partial u}\Big|_{u=t_i} = h(x)g(x,t_i)(1-\text{TanH}(\hat{F}_r(t_i|x))^2)$ are therefore necessary to estimate the likelihood at each epoch. As previously described, $\hat{F}_r(t|x)$ necessitates n forward passes with n, the number of points used for the numerical integration. $\frac{\partial \hat{F}_r(u|x_i)}{\partial u}\Big|_{u=t_i}$ requires $n+1$ forward passes in total, but only one additional once $\hat{F}_r(t|x)$ is estimated. As a result, one epoch requires $\mathcal{O}((n+1)N)$ passes with N the number of patients in the study, and n the number of points used to approximate the CIF.

NeuralFG's training cost. For our proposed model, F_r is estimated in one forward pass, and $\frac{\partial \hat{F}_r(u|x_i)}{\partial u}\Big|_{u=t_i}$ in one backward pass, through automatic differentiation. Assuming the *same computational cost for forward and backward passes*, the likelihood estimation has an $\mathcal{O}(2N)$ complexity. Our proposed methodology, therefore, presents a theoretical $(n+1)/2$

computational gain compared to DeSurv in estimating the likelihood used at each training epoch.

Prediction cost. Once trained, DeSurv still requires n forward passes to estimate the CIF, whereas our method needs one forward pass. This results in an $O(n)$ gain when deployed.

Our proposed methodology offers a significant computational gain for training and testing compared with DeSurv while avoiding approximations of the quantities of interest. NeuralFG, therefore, presents a theoretical advantage for model development and deployment for accurate survival modelling under competing risks.

5.6 Empirical evidence of the impact of different competing risks handling strategies

In Section 5.4.2, we quantified the gap in considering competing risks as censoring on both individual and group-specific errors. Because of the potential biases resulting from inappropriate modelling, we introduced a novel approach to tackle the limitations of existing methodologies in Section 5.5. We present in this section a simulation study to empirically (i) compare the different models and (ii) validate the theoretical biases associated with modelling competing risks. First, in Section 5.6.1, we describe the data generation process used in the proposed simulation study. In Section 5.6.2, we then detail the empirical setting with the considered modelling strategies. To demonstrate our approach's edge over existing methodologies, we analyse in Sections 5.6.3 and 5.6.4 the performance and training times of the different models. Finally, we validate in Section 5.6.5 the theoretical individual and group gain from modelling competing risks previously introduced.

5.6.1 Data generation

Our experiments rely on a synthetic population of $N = 30,000$ patients with 10 associated covariates $X \in \mathbb{R}^{10}$, group membership $G \in \{0, 1\}$ and associated time and cause of event T, D. The following data generation does not aim to reflect a specific real-world setting but highlights the model's flexibility in modelling complex survival distributions. To generate the data, we followed the next three steps combining the one proposed in [179, 226] and resembling the one proposed in Chapter 4:

Covariates. To generate the covariates, we first draw the group membership following a Bernoulli distribution. Then, we generated the two independent covariates following two normal distributions with group-specific centres, and unit variance, specifically $c_1 = (1.5, 1.5)$ and $c_2 = -c_1$. All other covariates were independently drawn from standard normal distributions.

Formally, the group and covariates were modelled using the following procedure:

$$\begin{aligned} G &\sim \text{Bernoulli}(0.5) \\ X_{[1,2]} \mid G = g &\sim \text{MVN}(c_g, I^2) \\ X_{[3:10]} &\sim \text{MVN}(0, I^8) \end{aligned}$$

with MVN denoting a multivariate normal distribution, and I^n, the identity covariance matrix of dimension n.

Competing risks. From the generated covariates, we then draw two competing risks $R = 2$ following the procedure introduced in [33]. This procedure consists of the following three steps:

- Define the cause specific hazards λ_r associated with each competing risk r.
- Simulate the time of the first observed event T' following the all-cause hazard equal to the sum of the cause-specific hazards: $\sum_r \lambda_r$.
- For each simulated time $T' = t$, draw the associated event type d from a Bernoulli with probability the relative hazard for event r: $\frac{\lambda_r(t)}{\sum_s \lambda_s(t)}$.

For our simulation, we choose each cause-specific hazard distribution to follow a Gompertz distribution with a shared scale parameter across events. This parametrisation results in the all-cause hazard to follow a Gompertz distribution. To ensure each individual and each group have different survival profiles, we draw group-specific coefficients K and Φ from normal distributions, used to parametrise transformations of the covariates defining the Gompertz's scale and shape. Formally, each risk r has for cause-specific hazard λ_r, the form:

$$\lambda_r(t \mid x, g) = w_r(\kappa_r^g, x) \cdot \exp(w_s(\phi^g, x) \times t)$$

with κ_r^g and ϕ^g, realisations of the K and Φ for group g, and w_r and w_s transformations of the covariates x used to ensure the positivity and non-linearity of the Gompertz's scale and shape defined, similarly to [179], as:

$$\begin{aligned} w_1(\kappa_1^g, x) &= \left|(\kappa_1^g[5:10] \cdot x[5:10])^2 + \kappa_1^g[1:5] \cdot x[1:5]\right| && \text{(Shape Cause 1)} \\ w_2(\kappa_2^g, x) &= \left|(\kappa_2^g[1:5] \cdot x[1:5])^2 + \kappa_2^g[5:10] \cdot x[5:10]\right| && \text{(Shape Cause 2)} \\ w_s(\phi^g, x) &= \left|\phi^g \cdot x[5:10]\right| && \text{(Shift)} \end{aligned}$$

with the notation $v[a:b]$ corresponding at the selection of the dimension a to b of the vector v, and $|\cdot|$, the absolute value. Following the previously described procedure from [33], the overall hazard associated with the first observed event (T') is the sum of the cause-specific hazards

with associated event type (D') drawn from a trial following the relative risk of each event.

$$
\begin{aligned}
K_1^g &\sim \text{MVN}(0, \sigma_K^2 I^{10}) && \text{(Group-specific shape for } r = 1\text{)} \\
K_2^g &\sim \text{MVN}(0, \sigma_K^2 I^{10}) && \text{(Group-specific shape for } r = 2\text{)} \\
\Phi^g &\sim \text{MVN}(0, \sigma_\Phi^2 I^5) && \text{(Shared scale)}
\end{aligned}
$$

$$
\begin{aligned}
T' \mid X, G, K_1^g, K_2^g, \Phi^g &= (x, g, \kappa_1^g, \kappa_2^g, \phi^g) \\
&\sim \text{Gompertz}\left(w_1(\kappa_1^g, x) + w_2(\kappa_2^g, x), w_s(\phi^g, x)\right) \\
D' = 1 \mid X, G, K_1^g, K_2^g &= (x, g, \kappa_1^g, \kappa_2^g) \\
&\sim \text{Bernoulli}\left(\frac{w_1(\kappa_1^g, x)}{w_1(\kappa_1^g, x) + w_2(\kappa_2^g, x)}\right)
\end{aligned}
$$

Censoring. To mimic the real-world settings, we generate censoring independent from the two competing risks. We draw censoring times following a Gompertz hazard distribution[2] with shape: $w_c(\zeta, x) = \left|\zeta \cdot x[5:10]\right|$ with ζ, the realisation of Z a multivariate normal random variable [3] with variance σ_Z. Observed event times and types (T, D) are the minimum between the first observed competing risks time T' and censoring C.

$$
\begin{aligned}
Z &\sim \text{MVN}(0, \sigma_Z^2 I^6) \\
C \mid X = x, Z = \zeta &\sim \text{Gompertz}(w_c(\zeta, x)) \\
T &= \min(C, T') \\
D &= \mathbb{1}(C > T')
\end{aligned}
$$

with $\mathbb{1}$, the indicator function. Using this procedure, we generate 25 simulated datasets. For each, we aim to model the survival distribution associated with cause 1 (T_1') relying only on the observed T, D, X, G. While the number of simulations may appear limited due to the prohibitive computational cost of training each of these methods, this number improves the quantification of performance compared to the common reliance on a unique synthetic dataset in the ML literature. The code to generate the data and reproduce the following results is available on Github[4].

5.6.2 Empirical settings

Following the previously described data generation process, the proposed pipeline compares approaches to handle competing risks. This section introduces these methodologies and their associated training procedures.

[2] This differs from [179, 226] to ensure non-informative censoring.
[3] In our simulations, we choose $\sigma_K = \sigma_Z = \sigma_\Phi = 1$ for all experiments.
[4] `https://github.com/Jeanselme/NeuralFineGray`

Baselines. We compare the proposed Neural Fine-Gray (**NeuralFG**) against multiple alternatives. First, we considered the well-established cause-specific Cox model (**CS Cox** [256]) and **Fine-Gray** model [88] with a linear parametric form for the covariate effect. The cause-specific Cox model models each cause independently using a Cox proportional hazards model, while Fine-Gray models the sub-hazard functions assuming proportional sub-hazards[5]. Thereafter, we compare with the existing competing risks neural networks introduced in Section 5.2: Deep Survival Machine (**DSM**, [221]), **DeepHit** [179] and, closer to our work, **DeSurv** [78] [6].

Additionally, we considered each model's **Non-Competing** alternative. To train these models while considering the competing risks as censoring, we encode the event type as $\mathbb{1}(d = 1)$.

These non-competing models' architectures differ as they aim to model only one outcome. Using the same number of parameters for one or multiple outcomes may mislead us to believe that competing risks' modelling does not improve performance because of the larger number of parameters used for the outcome of interest. To ensure the same allocation of parameters for the outcome of interest, we propose an additional non-competing NeuralFG*. This model relies on the same architecture as NeuralFG, and, consequently, the same number of parameters. Only the methodology's training differs by backpropagating the sum of the cause-specific losses, i.e.:

$$l = \sum_r \left[\sum_{i,d_i=r} \log \lambda_r(t_i|\tilde{x}_i) - \sum_i \Lambda_r(t_i|\tilde{x}_i) \right]$$

Each monotonic network, therefore, models the cumulative hazard function for risk r, Λ_r, by maximising the likelihood of one cause whilst considering the rest of the population as censored, relying on a shared embedding $\tilde{x}$. Automatic differentiation outputs $[\lambda_r]_{r\in[\![1,R]\!]}$. This log-likelihood differs from Equation (5.3) as it does not depend upon the CIFs but the cause-specific cumulative hazards.

Training procedure. We analyse the performance of these different strategies over the 25 repetitions of the previously described simulations. For each simulated dataset, we split the data into two: 80% for development and the rest for testing. We further divide the development set into 3: 10% for hyperparameter tuning, 10% for early stopping and the rest for training. We use random search on the following grid over 100 iterations: learning rate (10^{-3} or 10^{-4}), batch size (100, 250), dropout rate (0, 0.25, 0.5 or 0.75), number of layers ($[\![1, 4]\!]$) and nodes (25 or 50). All activation functions are TanH to ensure proper differentiation — note that any $\mathcal{C}^1$ activation function would work. All models are optimised using an Adam optimiser [166] over $1,000$ epochs with early stopping.

[5]Expert knowledge could provide more relevant parametric forms in real-world settings than the linear relation used in these experiments. However, this highlights one of the advantages of using neural network approaches to automatically discover this relation between covariates and outcomes from the data.

[6]This architecture was modified to match both DeepHit and NeuralFG with a shared embedding before the risk-specific models.

All other methods are optimised over the same grid (if applicable). Additionally, we explore both Log-Normal and Weibull distributions for DSM and use $10,000$ initial iterations to estimate the parametric form closest to the average survival as proposed in the original paper [221]. These iterations ignore the assignment network and focus on estimating the distributions' parameters that best describe the population survival for faster convergence. For DeSurv, we followed the original paper's recommendation of a 15-point Gauss-Legendre quadrature to estimate the CIFs. Similarly, DeepHit uses a regular discretisation of the time horizon in 15 bins. Finally, for a fair comparison, we ensure an equal maximum number of layers for all models in the grid search.

5.6.3 Performance comparison

The first aim of the proposed simulations is to compare different approaches to competing risks in terms of the quality of survival estimates. Following the recommendation of accounting for competing risks in evaluation metrics [326], we use extensions of the metrics introduced in Chapter 2: the time-dependent Brier score [281] and truncated C-index [353] to measure calibration and discrimination at the dataset-specific 0.25, 0.5 and 0.75 quantiles of the event times. Table 5.1 presents these metrics' means and standard deviations over the 25 simulations (Appendix C.2.2 presents the complementary performances on the competing risk). While this evaluation quantifies the models' predictive performances, the knowledge of the underlying survival distributions allows us to compute the exact error in modelling the survival distribution. Specifically, Table 5.2 presents the mean squared error between the estimated CIF and the predicted one.

Insight 1: Modelling competing risks improves survival estimates. The mean squared error between the estimated CIF for risk 1 and the underlying distribution, presented in Table 5.2, illustrates the impact of different strategies on survival estimates. With the least constraints, NeuralFG and DeSurv present the lowest error, equivalently the best overall recovery of the underlying CIFs. DeepHit follows with a coarser recovery of the survival distribution, as shown by the second-best integrated MSE. Fine-Gray and CS Cox lead to an accurate survival estimate because of their non-parametric survival baseline, resulting in a smaller average error than some more flexible strategies. Finally, DSM's parametric assumption reduces its capacity to estimate the underlying distribution. DSM's implementation relies on Weibull for each competing risk, not resulting in closed-form CIFs or likelihood computation. This observation explains DSM's absence of improvement when considering competing risks. Note that all other methodologies benefit from the modelling of competing risks. This edge increases at longer time horizons as the competing risk is more likely to occur. Note that the improvement between non-competing and competing modelling is even stronger when the number of parameters is equal, as demonstrated by NeuralFG*. For the same number of parameters, modelling

competing risks improves survival estimates.

	Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
Competing	**NeuralFG**	*0.752* (0.044)	*0.674* (0.046)	*0.634* (0.044)	**0.064** (0.019)	**0.125** (0.029)	**0.179** (0.029)
	DeSurv	**0.754** (0.044)	**0.677** (0.046)	**0.638** (0.044)	**0.064** (0.019)	**0.125** (0.029)	**0.179** (0.030)
	DeepHit	0.716 (0.050)	0.637 (0.054)	0.606 (0.049)	0.069 (0.022)	0.135 (0.033)	0.189 (0.031)
	DSM	0.726 (0.047)	0.656 (0.049)	0.610 (0.050)	*0.065* (0.020)	*0.128* (0.030)	0.200 (0.036)
	Fine-Gray	0.562 (0.099)	0.568 (0.067)	0.573 (0.051)	0.068 (0.021)	0.132 (0.032)	0.188 (0.033)
	CS Cox	0.615 (0.068)	0.591 (0.054)	0.581 (0.047)	0.067 (0.021)	0.131 (0.032)	*0.187* (0.033)
Non-Comp.	NeuralFG	0.747 (0.045)	0.672 (0.048)	0.623 (0.051)	**0.064** (0.019)	*0.128* (0.029)	0.196 (0.031)
	NeuralFG*	0.746 (0.045)	0.670 (0.049)	0.627 (0.052)	0.075 (0.027)	0.158 (0.047)	0.247 (0.055)
	DeSurv	0.742 (0.044)	0.666 (0.046)	0.616 (0.049)	0.068 (0.024)	0.141 (0.047)	0.222 (0.077)
	DeepHit	0.705 (0.055)	0.630 (0.054)	0.587 (0.048)	0.073 (0.024)	0.151 (0.038)	0.233 (0.035)
	DSM	0.718 (0.047)	0.655 (0.050)	0.614 (0.049)	*0.065* (0.020)	*0.128* (0.030)	0.198 (0.036)
	Cox	0.616 (0.068)	0.590 (0.053)	0.569 (0.055)	0.067 (0.021)	0.132 (0.033)	0.199 (0.037)

Table 5.1: Comparison of model performance by means (std) across 25 simulations on event $r = 1$. Best performances are in **bold**, second best in *italics*.

	Model	MSE *(Smaller is better)*			
		Integrated	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
Competing	**NeuralFG**	*0.086* (0.008)	*0.016* (0.004)	*0.036* (0.007)	*0.059* (0.008)
	DeSurv	**0.085** (0.010)	**0.015** (0.005)	**0.035** (0.009)	**0.058** (0.012)
	DeepHit	0.134 (0.015)	0.053 (0.016)	0.084 (0.017)	0.111 (0.018)
	DSM	0.410 (0.063)	0.022 (0.005)	0.044 (0.010)	0.086 (0.018)
	Fine-Gray	0.121 (0.018)	0.023 (0.008)	0.047 (0.013)	0.073 (0.018)
	CS Cox	0.118 (0.017)	0.023 (0.008)	0.046 (0.013)	0.073 (0.017)
Non-Comp.	NeuralFG	0.362 (0.051)	0.019 (0.004)	0.047 (0.010)	0.093 (0.014)
	NeuralFG*	0.428 (0.051)	0.039 (0.017)	0.093 (0.033)	0.165 (0.043)
	DeSurv	0.383 (0.077)	0.025 (0.021)	0.059 (0.045)	0.115 (0.073)
	DeepHit	0.426 (0.057)	0.059 (0.019)	0.106 (0.022)	0.166 (0.023)
	DSM	0.404 (0.062)	0.022 (0.005)	0.043 (0.010)	0.083 (0.018)
	Cox	0.358 (0.050)	0.024 (0.008)	0.054 (0.014)	0.098 (0.020)

Table 5.2: Comparison of survival estimate error by means (std) across 25 simulations on the risk $r = 1$. Best performances are in **bold**, second best in *italics*.

Insight 2: Modelling competing risks improves predictive performance. When comparing the mean square error between models accounting for competing risks and their non-competing alternatives, Table 5.2 shows a consistent error reduction. The accompanying Table 5.1 presents a similar conclusion regarding discrimination and calibration. Performances show a consistent gain between non-competing and competing approaches. Note that this gain is larger with longer time horizons, as the competing risk is more likely to occur.

Insight 3: Inadequate assumptions upon the survival distribution lower performances. Table 5.1 highlights that the fewer assumptions and approximations one makes upon the survival distribution, the better the performances are. Specifically, the two models without constraints upon the survival distributions, NeuralFG and DeSurv, present the best performance. DSM and DeepHit present poorer performance due to their parametric assumption and coarse temporal approximations. Finally, the standard assumption of linearity reduces the performances of the Cox and Fine-Gray approaches, highlighting the advantage of the automatic discovery of non-linearities between covariates and outcomes when this relation is unknown.

5.6.4 Training time comparison

The previous section shows the importance of modelling competing risks for improved survival estimates and performance. We empirically validated the benefit gained from avoiding assumptions upon the survival distribution. However, these simulations show little difference between DeSurv and NeuralFG. As discussed in Section 5.5.2, our approach seeks to reduce the computational complexity. This section presents the training time for all neural network methodologies given a fixed architecture[7,8]. Figure 5.4 displays the different strategies' average relative training time compared with NeuralFG. Note that only DeSurv and NeuralFG present the same architecture and training procedure, whereas we relied on the original paper's implementation and training strategies for the other approaches. Further, we trained three versions of DeSurv with different numbers of point estimates[9] for numerical integration $n = \{1, 5, 15\}$. We include DSM and DeepHit as comparators to show the computational advantage of their associated assumptions and approximations. For these models, we vary the number n of time discretisation for DeepHit and the number k of Weibull distributions for DSM.

Insight 4: NeuralFG presents faster exact survival modelling. NeuralFG presents longer or equal training time than DeepHit and DSM. These two approaches' approximations reduce the computational complexity to a single forward pass per epoch. DSM further reduces its time to convergence with its pretraining procedure, estimating the Weibull parameters closest to the observed events' distribution without consideration of the covariates, i.e. without propagation. However, under applications in which the underlying distributions are unknown, NeuralFG and DeSurv hold an edge. In this context, NeuralFG is, on average, 4.2 times faster than the recommended $n = 15$ DeSurv and 1.5 times faster than $n = 1$. Despite not observing the theoretical $(n + 1)/2$ gain, NeuralFG results in a faster average training time. Notably, the discrepancy with the theory introduced in Section 5.5.2 results from (i) considering one

[7]We left aside the faster non-neural network approaches due to the non-comparable training procedure and architecture.

[8]This set of experiments is performed on a i7-10810U CPU with 15 GB of memory.

[9]Danks and Yau [78] recommends $n = 15$, but, if a lower degree Taylor expansion results in an accurate estimate of the CIFs' derivative, smaller n may present similar performance.

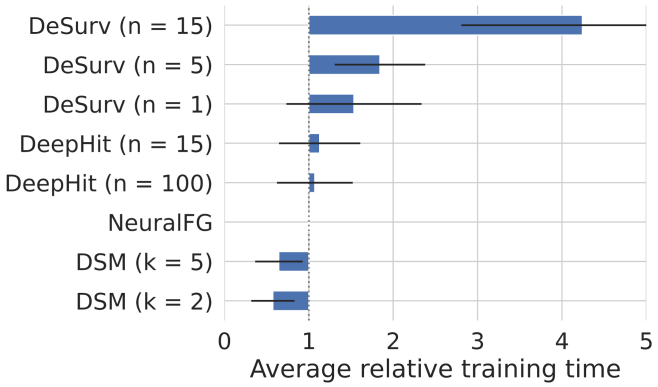


Figure 5.4: Average training gain over NeuralFG across 25 simulations with 95% CI marked in black.

iteration in terms of passes, not the total number of operations and iterations necessary for convergence in terms of execution time, and (ii) assuming equal computational cost for the forward and backward passes. While the training average time until convergence is more susceptible to hyperparameters' choice, initialisation and datasets' characteristics, this metric is more relevant to measure the real-world edge of the proposed strategy. Key to the discrepancy with theory is the number of iterations necessary to converge that constrained neural networks and approximated likelihood may impact. Despite these caveats, NeuralFG presents the fastest exact estimation of the underlying survival distributions.

5.6.5 Impact on population and group-specific discrepancy

These simulations present an opportunity to measure the alignment between theoretical and empirical discrepancies. The theoretical analysis in Section 5.4.2 considers the underlying distributions, not the imperfect modelling of these quantities. In the following, we aim to validate these results in more realistic settings. To this end, we measure the gain observed between the methodologies considering competing risks and their non-competing alternatives over the 25 simulations. Figure 5.5 displays the observed relative discrepancy on the y-axis evaluated using Equation (5.4) and the theoretical one on the x-axis at the 0.5 and 1 quantiles of the observed event times ($q_{0.5}$ and q_1). Due to the nature of the simulation, one can derive from the cause-specific risks the average theoretical relative discrepancy as:

$$\mathbb{E}_x[L^r(x)] = \mathbb{E}_x[\mathbb{P}(D' \neq 1 \mid x)] = \mathbb{E}_x\left(\frac{w_2(\kappa_2^g, x)}{w_1(\kappa_1^g, x) + w_2(\kappa_2^g, x)}\right) \tag{5.5}$$

In this figure, each point represents a simulation, and the line is the associated linear regression for a given method. Table 5.3 presents the slope of the linear regression and associated concordance correlation coefficient (CCC [175]). This last metric quantifies the empirical alignment with theory; 0 corresponds to no alignment, while 1 represents a perfect match

Model	CCC(L_r)		Slope(L_r)		CCC(Δ_g)		Slope(Δ_g)	
	$q_{0.5}$	q_1	$q_{0.5}$	q_1	$q_{0.5}$	q_1	$q_{0.5}$	q_1
NeuralFG	0.045	0.980	0.498	0.925	0.517	0.984	0.388	**0.899**
DSM	-0.004	-0.000	-0.093	-0.009	-0.069	-0.015	-0.046	-0.009
DeSurv	0.060	0.951	**0.608**	0.972	**0.675**	**0.986**	0.491	0.896
DeepHit	**0.070**	**0.990**	0.497	**1.020**	0.198	0.643	0.117	0.390
Fine-Gray	0.048	0.985	0.591	0.974	0.377	0.897	**0.686**	0.674

Table 5.3: Concordance correlation coefficient and slope between empirical and theoretical discrepancies with the most aligned performance between experiments and theory highlighted in **bold**.

between empirical and theoretical values. Contrary to Pearson's correlation, this metric accounts for the potential difference between the considered values.

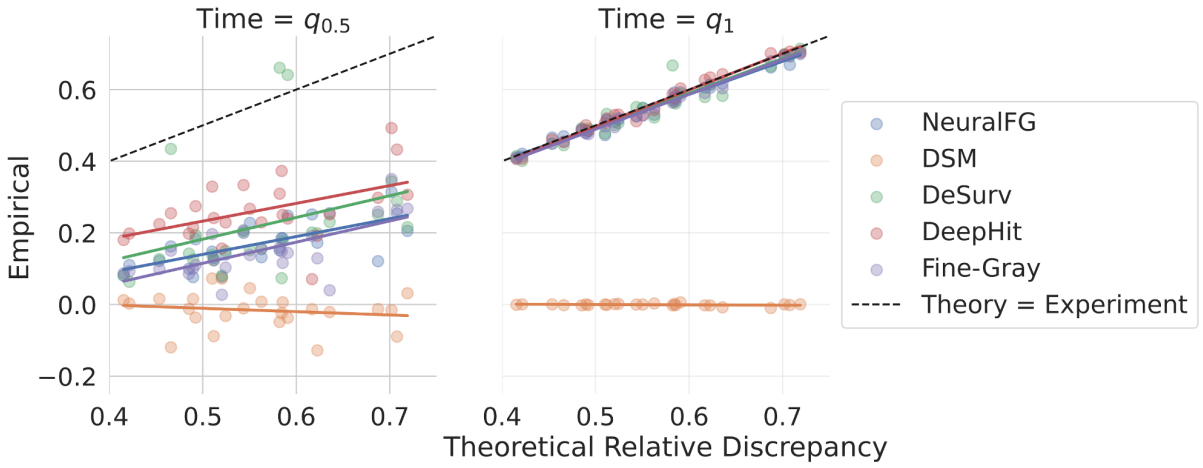


Figure 5.5: Theoretical and empirical relative discrepancy computed for each simulation and method at evaluation times $q_{0.5}$ and q_1. Each line represents the linear regression fit associated with a method's discrepancy across simulations. The closer to the dashed theoretical line, the more aligned the methodology is with the theory.

Additionally, the knowledge of the group structure and survival distributions allows us to verify the theoretical inter-group discrepancy gap introduced in Equation (5.5). Similarly to the overall gain, Figure 5.6 presents the measured and theoretical gaps in performance between groups defined by G at the same two horizons. Similarly to Equation (5.5), the x-axis displays the theoretical gap between group g and the rest of the population computed as:

$$\Delta_g = \mathbb{E}_{x|g}\left(\frac{w_2(\kappa_2^g, x)}{w_1(\kappa_1^g, x) + w_2(\kappa_2^g, x)}\right) - \mathbb{E}_{x|\neg g}\left(\frac{w_2(\kappa_2^{\neg g}, x)}{w_1(\kappa_1^{\neg g}, x) + w_2(\kappa_2^{\neg g}, x)}\right)$$

Insight 5: The gain associated with modelling competing risks aligns with the theory in the limit. In Figure 5.6, all methods, except DSM — echoing the previously described shortcomings in modelling competing risks — present a high correlation between the theoretical

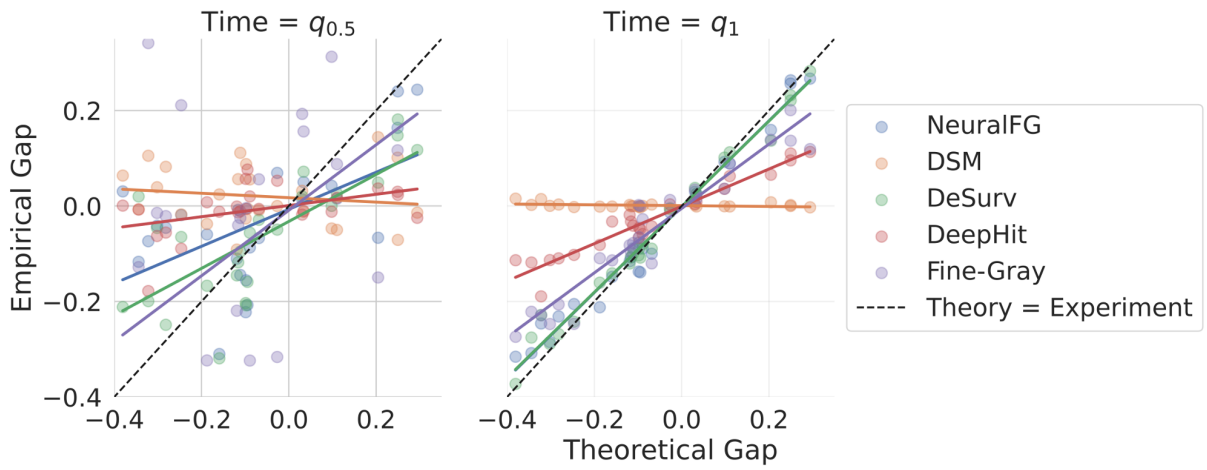


Figure 5.6: Theoretical and empirical gap in relative cumulative incidence discrepancy computed for each simulation and method at evaluation times $q_{0.5}$ and q_1. The closer to the dashed theoretical line, the more aligned the methodology is with the theory.

and empirical gain at $q_{1.0}$ as shown in Table 5.3. Despite imperfect modelling, ignoring competing risks results in a discrepancy proportional to the probability of observing the competing risks. Note that, at $q_{0.5}$, the empirical relative discrepancy consistently underestimates the theoretical error. This observation means that the difference between the model accounting for and the one ignoring competing risks is smaller than theoretically predicted. This phenomenon may be explained by a more flexible single-risk model than competing risk one *for the same number of parameters*, reducing the discrepancy.

The comparison of the predicted fairness gap and the observed one evidences the direct impact of competing risks on algorithmic fairness. In Figure 5.6, the stronger the difference between groups — as shown by extreme values on the x-axis, the larger the gain from modelling competing risks. Table 5.3 echoes this conclusion with positive slopes for all methodologies at both horizons. This observation emphasises the importance of accurate and flexible modelling of the underlying survival distribution for fairer modelling.

5.7 Real-world analysis

We previously introduced simulations to test the proposed methodology's limitations and strengths, and verify the theoretical gain associated with modelling competing risks. However, simulations' distributional choices may not capture the complexity of medical applications. This section demonstrates the real-world improvement observed when modelling competing risks.

5.7.1 Datasets

For this analysis, we considered three datasets with competing risks commonly used in the literature:

- PBC [318] consists of the data collected at the start of a 10-year RCT led by Mayo Clinic between 1974 and 1984 to measure the impact of D-penicillamine on Primary Biliary Cholangitis (PBC). This progressive auto-immune disease of the liver involves the inflammation of the bile ducts which eventually collapse. Without a liver transplant, this condition leads to the patient's death. In this context, the considered models rely on the 312 patients and their associated 25 covariates related to their initial blood test and demographics to model their outcome. Death on the waiting list is the primary outcome, and a liver transplant is the competing risk.

- FRAMINGHAM [159]: the epidemiological Framingham Heart Study aimed to quantify the impact of multiple risk factors on the propensity of observing cardiovascular disease (CVD). Patients were followed for over 20 years, and recurrent questionnaires and clinical measurements such as blood tests, lung function, and treatment were regularly collected. Our analysis focuses on the subset of covariates obtained at the first medical appointment for 4,434 patients to model CVD risk. As multiple patients did not observe the primary event of interest and died from other causes over the study period, we consider this outcome as a competing risk.

- SEER[10]: As introduced in the previous chapter, the Surveillance, Epidemiology, and End Results gathers American cancer statistics. Following [78, 179], we focus on 658,354 women diagnosed with breast cancer between 1992 and 2017, with 23 covariates describing the patient demographics and disease characteristics at diagnosis, such as grade, laterality and tumour size. Instead of focusing on the risk of death from breast cancer studied in Chapter 4, we select patients who may also died from CVD. In this context, death from breast cancer (BC) is the primary event of interest, and death from CVD is a competing risk. Note that this population is larger and more heterogeneous than the previously studied population.[11]

Table 5.4 summarises the datasets' characteristics with the respective proportion of outcome and censoring. These datasets present a large range of sample sizes and censoring rates to evaluate the models' performance under different real-world clinical settings.

Dataset	Sample size	Covariates	Primary Outcome	Competing risk	Censored
PBC	312	25	Death (44.87 %)	Transplant (9.29 %)	45.83 %
FRAMINGHAM	4,434	18	CVD (26.09 %)	Death (17.75 %)	56.16 %
SEER	658,354	23	BC Death (16.51 %)	CVD Death (5.69 %)	77.80 %

Table 5.4: Considered datasets' characteristics.

[10] `https://seer.cancer.gov/`

[11] Due to the dataset's large dimensionality, we consider larger batch sizes: $1,000$ or $5,000$ to reduce memory access for each epoch.

To train and evaluate the models, we use a 5-fold cross-validation using the same training procedure as in Section 5.6.2. We measure performance employing the same calibration and discrimination metrics computed at the dataset-specific quantiles of the uncensored population event times (See Appendix C.2.1 for the datasets' characteristics at these time points and Appendix C.2.2 for the performances on the competing risk)). A difference with simulation is the presence of missing data. As the impact of missingness on performance is out of the scope of this analysis, we mean imputed all missing values.

5.7.2 Performance

Analysis of these datasets complements the simulations with new insights into the models' limitations and strengths under real-world settings. This section focuses on performance across different datasets. Table 5.5 summarises the calibration and discriminative performance of the analysed models on the primary outcome averaged across the 5-fold cross-validation.

Insight 6: Flexible competing risks modelling can improve real-world predictive performance... Focusing on the FRAMINGHAM and SEER datasets, modelling competing risks consistently improves performance compared to the models' non-competing alternative. Contrarily to Insight 1, real-world competing risks modelling can improve predictive performance. Further, Cox and Fine-Gray's linearity and proportional hazards assumptions and DSM's parametric form reduce performance, echoing Insight 3. Critically, these medical datasets illustrate the shortcomings associated with approximations. DeepHit and DeSurv present lower calibration and discrimination than NeuralFG across all time horizons. While DeepHit presents a different architecture than the proposed model, DeSurv differs solely on its likelihood computation, demonstrating that numerical approximation of the loss does not only slow computation but can reduce performance.

Insight 7: ... but hurts small-data regimes' performance. The PBC dataset provides a more nuanced conclusion: more flexible approaches require large sample sizes. PBC's limited number of patients results in more complex models' overfitting. Practitioners should prefer models with fewer parameters in small-data settings to avoid this problem. For instance, the linear Fine-Gray and CS Cox models result in competitive performances in this setting, outperforming most of the neural network approaches. However, their underlying assumptions diminish performance under more complex covariate effects as in the SEER dataset. As in Insight 3, DSM's parametric assumption results in the best discrimination in this setting. A smaller number of parameters captures the underlying distribution with less data.

		Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
			$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Competing	**NeuralFG**	0.809 (0.079)	0.791 (0.119)	0.759 (0.119)	0.099 (0.028)	0.140 (0.017)	0.172 (0.044)
		DeSurv	0.799 (0.104)	0.813 (0.041)	0.772 (0.045)	0.111 (0.016)	0.157 (0.023)	0.225 (0.029)
		DeepHit	0.822 (0.098)	0.839 (0.042)	0.780 (0.028)	**0.090** (0.030)	0.131 (0.015)	0.179 (0.016)
		DSM	**0.864** (0.062)	**0.863** (0.035)	**0.826** (0.051)	**0.090** (0.039)	0.124 (0.015)	0.162 (0.013)
		Fine-Gray	0.825 (0.130)	0.852 (0.046)	0.816 (0.055)	**0.090** (0.041)	*0.101* (0.008)	0.145 (0.033)
		CS Cox	0.826 (0.119)	0.850 (0.042)	0.811 (0.060)	*0.091* (0.038)	*0.101* (0.007)	*0.144* (0.032)
	Non-Comp.	NeuralFG*	0.815 (0.088)	0.847 (0.044)	*0.819* (0.054)	0.116 (0.031)	0.128 (0.023)	0.162 (0.057)
		DeSurv	0.814 (0.051)	0.800 (0.080)	0.766 (0.098)	0.123 (0.014)	0.193 (0.038)	0.280 (0.093)
		DeepHit	0.808 (0.088)	0.816 (0.036)	0.758 (0.052)	0.109 (0.025)	0.161 (0.029)	0.213 (0.027)
		DSM	**0.864** (0.062)	**0.863** (0.035)	**0.826** (0.051)	**0.090** (0.039)	0.124 (0.015)	0.162 (0.013)
		Cox	*0.827* (0.119)	*0.854* (0.037)	0.816 (0.054)	*0.091* (0.038)	**0.099** (0.009)	**0.141** (0.033)
FRAMINGHAM	Competing	**NeuralFG**	**0.871** (0.025)	**0.809** (0.030)	**0.775** (0.018)	**0.050** (0.003)	**0.096** (0.009)	**0.130** (0.004)
		DeSurv	0.855 (0.028)	0.774 (0.023)	0.727 (0.008)	0.054 (0.006)	0.126 (0.010)	0.211 (0.019)
		DeepHit	0.854 (0.026)	0.778 (0.026)	0.737 (0.015)	*0.053* (0.003)	0.100 (0.006)	0.138 (0.002)
		DSM	*0.866* (0.023)	*0.803* (0.023)	*0.771* (0.014)	0.057 (0.004)	0.103 (0.006)	0.138 (0.002)
		Fine-Gray	0.841 (0.025)	0.791 (0.025)	0.764 (0.016)	0.057 (0.006)	0.099 (0.005)	*0.132* (0.002)
		CS Cox	0.844 (0.020)	0.794 (0.023)	0.765 (0.016)	0.056 (0.006)	*0.098* (0.005)	*0.132* (0.002)
	Non-Comp.	NeuralFG*	0.861 (0.029)	*0.803* (0.032)	*0.771* (0.021)	*0.053* (0.004)	0.104 (0.010)	0.143 (0.005)
		DeSurv	0.840 (0.034)	0.764 (0.033)	0.725 (0.020)	0.067 (0.008)	0.186 (0.028)	0.359 (0.047)
		DeepHit	0.842 (0.035)	0.752 (0.032)	0.721 (0.028)	0.056 (0.004)	0.104 (0.008)	0.140 (0.004)
		DSM	*0.866* (0.022)	0.800 (0.024)	0.766 (0.014)	0.057 (0.004)	0.102 (0.006)	0.138 (0.002)
		Cox	0.844 (0.020)	0.794 (0.023)	0.766 (0.016)	0.055 (0.005)	*0.098* (0.005)	0.133 (0.001)
SEER	Competing	**NeuralFG**	**0.899** (0.002)	**0.863** (0.001)	**0.824** (0.000)	**0.037** (0.000)	**0.068** (0.000)	**0.100** (0.001)
		DeSurv	0.839 (0.008)	0.767 (0.009)	0.699 (0.009)	0.054 (0.002)	0.148 (0.004)	0.294 (0.008)
		DeepHit	**0.899** (0.001)	0.847 (0.010)	0.794 (0.012)	0.039 (0.001)	0.075 (0.001)	0.110 (0.001)
		DSM	0.890 (0.001)	0.850 (0.002)	0.812 (0.002)	0.039 (0.000)	0.075 (0.001)	0.110 (0.001)
		Fine-Gray	0.852 (0.003)	0.811 (0.002)	0.769 (0.001)	0.042 (0.001)	0.080 (0.000)	0.117 (0.001)
		CS Cox	0.854 (0.003)	0.811 (0.002)	0.768 (0.001)	0.042 (0.001)	0.080 (0.000)	0.116 (0.001)
	Non-Comp.	NeuralFG*	0.894 (0.002)	*0.859* (0.002)	*0.823* (0.002)	**0.037** (0.000)	*0.070* (0.000)	*0.102* (0.001)
		DeSurv	0.833 (0.008)	0.747 (0.013)	0.638 (0.013)	0.062 (0.004)	0.182 (0.009)	0.376 (0.012)
		DeepHit	*0.898* (0.002)	0.844 (0.004)	0.774 (0.015)	*0.038* (0.000)	0.075 (0.000)	0.109 (0.001)
		DSM	0.890 (0.002)	0.851 (0.000)	0.814 (0.002)	0.039 (0.000)	0.075 (0.000)	0.110 (0.001)
		Cox	0.854 (0.003)	0.812 (0.002)	0.769 (0.001)	0.042 (0.001)	0.079 (0.000)	0.117 (0.001)

Table 5.5: Comparison of model performance by means (std) across 5-fold cross-validation on the primary outcome. Best performances are in **bold**, second best in *italics*.

5.7.3 Case study: The impact of ignoring competing risks on cardiovascular risk management.

This chapter has demonstrated the importance of modelling competing risks for improved survival modelling, predictive performance and reducing group inequities. This final section aims to further connect these findings with medical practice by exploring how the way practitioners handle competing risks may have practical repercussions on patients' care. We propose a

further analysis of the FRAMINGHAM dataset as this landmark study has been seminal to cardiovascular risk scores guiding medical practice. First, we present the performance differences between the proposed model in comparison to the same architecture maximising the cause-specific likelihoods. Then, we explore which population subgroups most benefit from this modelling. Finally, we study how guidelines would differ under the proposed NeuralFG and its non-competing alternative.

Why account for competing risks? To measure how modelling competing risks impacts performance, while ensuring the *same number of parameters*, we use the same non-competing architecture NeuralFG* presented in Section 5.6.1 that maximises the sum of the cause-specific likelihoods. Table 5.6 summarises the discrimination and calibration differences for the event of interest. Modelling competing risks improves performance for CVD, the primary outcome of interest, without significant differences for death, the competing risk. Since patients who die from other causes during the study period can no longer suffer from CVD, not accounting for all-cause mortality results in an upward-biased estimate of CVD risk.

Death	Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
CVD	**Comp.**	**0.871** (0.025)	**0.809** (0.030)	**0.775** (0.018)	**0.050** (0.003)	**0.096** (0.009)	**0.130** (0.004)
	Non-Comp.	0.861 (0.029)	0.803 (0.032)	0.771 (0.021)	0.053 (0.004)	0.104 (0.010)	0.143 (0.005)
Death	**Comp.**	**0.730** (0.056)	**0.698** (0.037)	**0.687** (0.023)	**0.025** (0.002)	**0.065** (0.004)	**0.103** (0.004)
	Non-Comp.	0.712 (0.052)	0.680 (0.043)	0.671 (0.025)	0.041 (0.004)	0.096 (0.007)	0.141 (0.009)

Table 5.6: Modelling competing risk - means (std) across the 5-fold cross-validation. Modelling competing risks improves the predictive performance of the outcome of interest.

Who may benefit? This comparison allows us to explore the group-specific impact from modelling competing risks. As shown in Theorem 5.1, patients who are the most likely to experience competing risks benefit the most from this modelling. As risk profiles change with age and sex, we propose to explore the impact of modelling competing risks on these groups. Table 5.7 summarises the calibration differences for the different age group. This highlights that older patients benefit the most from modelling death as a competing risk, as they are most at risk.

Similarly, we study sex differences. The average age between men and women is similar (50.0 for women and 49.8 for men, t-test p-value 0.35). However, women in the US generally have longer life expectancy [284], rendering them less likely to die over a fixed term study in comparison to men of the same age. We observe this phenomenon in the dataset with a sex-specific risk of death from other causes (16.8% for women, against 18.9% for men, Fisher's exact test p-value 0.07). While non-significant, this difference in the competing risk mechanism could result in differential gains when modelling competing risks. Table 5.8 confirms this

Age	Brier Score Difference		
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
<40	-0.000 (0.000)	-0.001 (0.002)	0.000 (0.005)
40-50	-0.001 (0.001)	-0.003 (0.002)	-0.005 (0.002)
50-60	*-0.004* (0.006)	*-0.008* (0.003)	*-0.016* (0.006)
60+	**-0.011** (0.012)	**-0.031** (0.015)	**-0.041** (0.023)

Table 5.7: Calibration differences - Means and standard deviations over 5-fold cross-validation. Larger negative values correspond to better calibration for the competing risk model.

intuition with the difference in Brier score between the competing and non-competing models doubling for men. Importantly, we compute the algorithmic fairness gap between the group gains in the last row. This metric extends the theoretical results presented in Theorem 5.2 to Brier score, confirming that modelling competing risks can reduce the algorithmic fairness gap when the competing risk's mechanism is group-dependent.

Sex	Brier Score Difference		
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
Male	-0.005 (0.004)	-0.012 (0.004)	-0.018 (0.004)
Female	-0.002 (0.002)	-0.005 (0.002)	-0.009 (0.003)
Δ_{M-F}	-0.002 (0.005)	-0.007 (0.005)	-0.009 (0.005)

Table 5.8: Calibration differences between sex - Means and standard deviations over 5-fold cross-validation. Larger negative values correspond to better calibration for the competing risk model. Δ_{M-F} corresponds to the average gain between men and women from modelling competing risks.

What is the impact on medical practice? The FRAMINGHAM dataset was used to derive the eponymous 10-year cardiovascular disease (CVD) risk score [350]. This score guides clinical practice in preventatively treating patients, usually with a combination of cholesterol-lowering therapy, e.g., statins, and holistic treatment of other CVD risk factors [38]. To minimise overtreatment and adverse side effects, accurate risk estimates are critical for targeting the population most at risk so as to maximise the benefit-risk ratio [201]. However, the original FRAMINGHAM score relies on a non-competing risk model [164, 201].

Clinical treatment often relies on a discretisation of this risk [38]: low, intermediate and high risk, at $< 10\%$, $10 - 20\%$ and $> 20\%$ chance, respectively, of observing a CVD event in the following 10 years. Recent guidelines in the United States suggest placing all patients with $\geq 7.5\%$ risk on cholesterol-lowering drugs [201]. Furthermore, in the US alone, several million patients are on these medications [333]. Therefore, even modest shifts in patient risk classification could, at scale, amount to considerable numbers either inappropriately receiving preventative treatment or inappropriately receiving none. To demonstrate how considering competing risks can fundamentally alter such risk profiling, we present in Table 5.9 the

	Risk	Non - Competing Low	Inter.	High	Total
Comp.	Low	502	228	23	753
	Intermediate	2	189	229	420
	High	0	9	314	323
	Total	504	426	566	1496

(a) Patients *with no event* in the 10-year follow-up.

	Risk	Non - Competing Low	Inter.	High	Total
Comp.	Low	23	28	5	56
	Intermediate	1	37	41	79
	High	2	4	248	254
	Total	26	69	294	389

(b) Patients *with* an observed event during the 10-year follow-up.

Table 5.9: Reclassification matrices between competing and non-competing risk scores for patients older than 50. Red (resp. blue) shows when the competing risks score is less aligned with the 10-year observed outcome than the non-competing model (resp. more aligned). Note that censored patients are ignored.

reclassification matrices of risk levels given competing and non-competing NeuralFG stratified by observed outcomes for patients aged 50 or over. For instance, note that 251 of those who did not experience an event deemed intermediate-to-high risk by the non-competing risks model are reclassified as lower risk by the competing risks model, who, in turn, could have avoided the initiation of therapy. However, 33 patients who experienced an event wrongly received a low risk estimate by the competing risk model. In summary, using a non-competing risk score would have important clinical consequences of over- and under-treatment [283]. More predictive models accounting for competing risks must be preferred to ensure better care.

These findings align with existing literature highlighting the misestimation of risk when competing risks are overlooked [164, 197]. While some updated risk assessment tools in Europe, such as SCORE2 [67], account for competing risks, this practice remains uncommon. Our study not only echoes the importance of modelling competing risks for improved predictive performance but also underscores its broader fairness implications.

5.8 Discussion

Observation of an outcome is the product of the patient's condition and its interaction with the healthcare system. Clinical presence, therefore, goes beyond observed *covariates* as observed *outcomes* can reflect this process. Understanding why a particular outcome is observed, and its potential influence on the outcome of interest is critical for adequate modelling. This chapter explored the observation of outcomes, known as competing risks, that preclude the observation of the event of interest. The following summarises our contributions, recommendations and future work directions regarding this challenge.

5.8.1 Contributions

The statistical literature has long argued for modelling competing risks for improved survival estimates [31]. Despite the advocacy for such modelling, the lack of formalisation and quantification of errors has often led medical practitioners to treat competing risks as censoring, neglecting their potential informativeness. This chapter complements the literature by theoretically quantifying the error associated with this practice and empirically validating these results through simulations and medical datasets. Critically, we show that patients present errors in survival estimates proportional to their risk of experiencing competing events.

The theoretical insight provided in this chapter quantifies how ignoring competing risks biases individual predictions. Moreover, our work demonstrates that this practice can have implications for the algorithmic fairness properties of survival models. Specifically, the gap in relative survival between two groups equals the risk difference for the competing risks. Through simulations, we show that these results are more than theoretical findings, validating that ignoring competing risks impacts inter-group performance gaps.

Recognising the critical importance of accurate competing risk modelling, we introduce a novel methodology, Neural Fine-Gray. This approach estimates the underlying survival distribution through monotonic neural networks, avoiding simplifying assumptions or numerical approximations made by state-of-the-art competing risk neural networks. Our choice not only results in improved survival modelling but also comes at a lower computational cost, providing a valuable tool for accurately and efficiently modelling survival data with competing risks.

To further connect our work to medical practice, we evaluate different approaches to competing risks and their impact on real-world predictive performance and treatment recommendations. Our analysis of the FRAMINGHAM dataset and its eponymous risk score for cardiovascular disease invites practitioners to use competing risk modelling in risk score development. This recommendation not only aligns with previous calls for improved care [2, 16, 197, 283] but also holds promise for reducing disparities in healthcare outcomes.

5.8.2 Recommendations

Our work underscores the statistical imperative of incorporating competing risks into survival modelling, a recommendation grounded in our theoretical and empirical results. To enhance survival modelling accuracy and algorithmic fairness, we invite model developers to conscientiously consider the outcome process(es) and take the following steps:

1. *Study the different possible outcomes*: when modelling survival data, thoroughly examine all recorded events that may have influenced patients' risk or the observation of the event of interest. One should be wary of discarding available outcomes or data, as they may inform the process one aims to model. If these outcomes are associated with protected attributes, precluding, or correlated with the outcome of interest, one should model them.

2. *Verify censoring assumption*: considering competing risks as censoring breaks the assumption of uninformative censoring. However, accounting for competing risks does not guarantee the independence of censoring. While beyond the extent of this chapter, one should carefully assess and test that the uninformative censoring assumption holds in their study [142] as invalidity of this assumption could similarly bias the survival estimate.

5.8.3 Future work

This chapter opens multiple axes of research for future work. From a *theoretical point of view*, the problem's formalisation and error quantification is germane to multi-class problems, presenting a potential avenue to compare multi-class with binary approaches. From an *empirical point of view*, the promise of using neural architectures for survival modelling is their flexibility in handling various modalities. Exploration of survival modelling based on images, text and time series as in [144, 178, 220] could be an avenue to improve existing risk score models. Further, our proposed methodology adopts a simple approach to model monotonic functions. Alternative architectures and optimisation strategies ensuring the same behaviour are worth exploring for improved convergence. From a *medical point of view*, our paper demonstrates the potential impact on health equity. Historical evidence, outcomes' and financial impact quantification would strengthen this argument, particularly at the intersection of survival estimation and treatment recommendation policies.

Chapter 6

Clinical Presence Shift

Associated Publication. This chapter is based on preliminary results presented at NeurIPS Workshop - Learning from Time Series for Health (2022) [147], IBC (2022), ISCB (2022) and CHIL (2023).

Problem statement. *How can we improve transportability while leveraging clinical presence informativeness?*

6.1 Motivation

Predictive models' performances often drop when applied to different regions, hospitals or over time [118, 176, 230, 264, 294, 363] due to the distribution shift between these settings. We argue that a fundamental distribution shift exists in the observational processes across hospitals [154], regions, and countries, as medical training and practice may differ and may be subject to policies [312], insurance incentives, and the evolution of medical knowledge [103]. This *clinical presence shift* highlights a crucial limitation of predictive models using observational data as models may embed clinical presence patterns that do not transport under shift as practices and policies evolve [298]. Given the risk associated with the decisions these models inform, improving models' transportability under these naturally occurring shifts is critical to ML applications in healthcare.

Under the assumption of stable clinical presence between development and deployment, practitioners often use observational patterns to inform modelling and improve predictive performance. Specifically, missing tests, number of visits, and time since the last visit have been used as proxies for the patient's condition, as temporality and missingness patterns often reflect the severity of a patient's condition. While using these patterns results in improved predictive performance, no work has explored their transportability under shifts. Our work aims to fill this gap by examining how handling irregularities in longitudinal data may impact models' transportability under shifts.

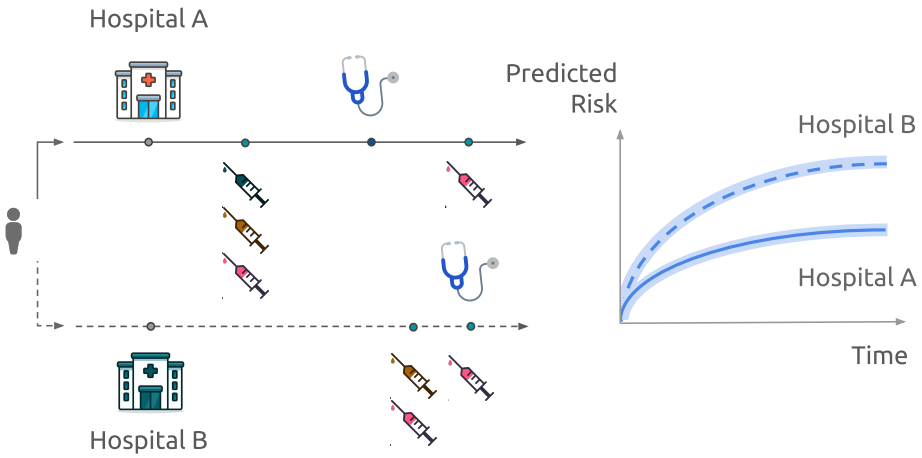


Figure 6.1: Our proposed predictive model aims to improve transportability under clinical presence shifts, e.g. between hospitals A and B, which may present different patterns of observations for the same patient. We introduce a joint modelling architecture to represent both clinical presence and the outcome of interest.

When considering potential shifts, current medical pipelines discard clinical presence to improve transportability between observational settings. Although practitioners justify this practice to avoid leveraging a changeable and unpredictable process [298, 359] in the hope of improved transportability, no evidence has demonstrated this behaviour in the clinical setting. On the contrary, we show that predictive performances decrease under shift when ignoring longitudinal irregularities. Further, multiple works [118, 302] empirically prove how existing strategies from the literature on distribution shift are ill-suited to tackle distribution shifts observed in medical practice due to unrealistic assumptions. In this work, we connect statistical joint models and multitask neural networks to offer a new solution to this problem.

Clinical presence may inform the modelling of the outcome of interest, resulting in tension between using its associated informativeness and accounting for its potential impact on transportability due to its inevitable evolution. This chapter explores the intersection of clinical presence and transportability under changes in longitudinal observational patterns and complements the literature through the following contributions:

1. We formalise the problem of shifts in the observational process, described as *clinical presence shift*, as a subtype of distribution shift when dissociating observed from observable covariates in the traditional distribution shift framework.

2. To improve transportability while taking advantage of informative longitudinal irregularities, we propose a multitask neural network [70] to jointly model the survival outcome and the longitudinal missingness and temporal patterns associated with clinical presence. We motivate this modelling choice through the statistical literature on joint modelling and a theoretical result from the adversarial ML literature.

3. We empirically validate our proposed strategy's superior transportability and measure the limitations of existing strategies to handle longitudinal irregularities under a clinical presence shift: *the weekend effect*, naturally occurring in the MIMIC III dataset.

The challenge of transportability is at the core of clinical presence as observational patterns evolve under the influence of multiple factors. Our work provides evidence that counter-intuitively discarding clinical presence not only reduces predictive performance but also reduces models' transportability. We highlight the importance of carefully considering the observational process in developing predictive models. In this chapter, we adapt joint models proposed in the statistical literature to multitask ML models and study their transportability. The proposed model relies on neural networks to model two dimensions of clinical presence: temporality and missingness. These added components improve the model's transportability by regularising the shared embedding. By considering clinical presence as *one of the model's outputs*, our modelling approach takes a step towards developing predictive models in real-world settings where the observational process is likely to evolve. Note that our contribution is not in specific implementation choices but in modelling clinical presence as a model's output to improve predictive modelling under changing observational patterns.

In Section 6.2, we review the literature on modelling medical time series, handling their irregularities and missingness patterns, and the literature on distribution shifts. Then, we formalise the concept of clinical presence shift as an extension of the existing distribution shift framework in Section 6.3. In Section 6.4, we introduce a multitask architecture jointly modelling longitudinal irregularities and the outcome of interest. In Section 6.5, we present a natural experiment in the MIMIC III dataset by considering insured and uninsured patients and measuring how a model built on one population may not perform on the other.

6.2 Related work

In the literature, multiple approaches have been introduced to model medical longitudinal time series and to tackle distribution shift. In the following, we review these works.

6.2.1 Irregularities modelling

A central challenge in longitudinal medical data is handling their irregularities. Observations rarely occur following a precise schedule, and practitioners only order a subset of the medical tests at each encounter. These temporal and missingness dimensions are active research questions in the statistical literature (see [296] for a detailed review). In this work, we divide the methodologies used in the literature into three categories: (i) Preprocessing, (ii) Featured, and (iii) Jointly modelled.

Preprocessing. Practitioners deal with irregularities as a nuisance dealt with at preprocessing time. First, if practitioners assume that the observed covariate distribution $q(X^*)$ is a representative sample of the underlying observable covariate distribution $q(X)$, with X the covariates and X^*, the observed ones, one can discard these irregularities. For example, regular re-sampling and MCAR imputation reflect this assumption. If inaccurate, this can lead to biased estimates and potentially misleading conclusions. Second, if practitioners recognise these irregularities are informative, they aim to recover $q(X)$ from the observed $q(X^*)$. Lipton, Kale, and Wetzel [190] underline how difficult it is to adequately reverse this observational process and question these methodologies' utility for recovering the underlying distribution.

Featurised. In this setting, practitioners assume that irregularities are not nuisances to address but proxies to the patient's condition. One can extract measures of clinical presence and use them as inputs to their models. For instance, the use of missing indicators [190], observation times [4, 50, 300], inter-observation time [43, 58, 59, 216, 377], or frequency [252] has improved models' performance.

Jointly modelled Instead of the two-step approaches described above, practitioners can directly model these irregularities. Joint models [98, 303, 310], marked point process [141] and Markov models [6] jointly model the visit process and the outcome of interest. Through shared effects in joint models, the observational process informs the modelling of the outcome of interest. However, these methods rely on parametric assumptions or do not scale to large datasets. In the ML literature, approaches similarly model missingness [324] or irregularities in time series [341]. These last methods aim to integrate the temporal information by slowly converging to a "stable" latent state as time passes through a decay parameter [24, 50, 250] or multi-level memory [218]. For instance, GRU-D [50] has shown promising results by extending the Gated Recurrent Unit (GRU) with an exponential decay on the hidden state, incorporating the temporal dimension in the embedding.

While these approaches tackle the problem of irregularities, they rely on different assumptions regarding clinical presence informativeness, usually focusing on one dimension of the process. To our knowledge, no study has evaluated their sensitivity to distribution shifts.

6.2.2 Distribution shifts

Predictive performances often degrade at deployment. This phenomenon is explained by a drift in the joint distribution $q(X, Y)$ of covariates X and outcomes Y between training and deployment. This phenomenon, known as distribution shift [260], has extensively been studied in the literature [215, 372] with two subtypes receiving particular attention as potential remedies exist:

1. *Covariate shift.* The covariate distribution varies $q_{train}(X) \neq q_{test}(X)$ while the conditional relation Y given X remains the same between training and deployment. Depending on the assumed generative process ($X \leftarrow Y$ or $X \rightarrow Y$), the issue of target shift has also been similarly investigated in the literature.

2. *Concept shift.* The covariate distribution remains stable but the conditional distribution evolves $q_{train}(Y \mid X) \neq q_{test}(Y \mid X)$.

Extensive literature exists on detection and mitigation strategies to tackle these shifts. Beyond the scope of this thesis, we invite the reader to refer to existing reviews [199, 227], and we focus in the following on the contributions that shaped our proposed approach and study.

Detecting shift. A first step towards addressing distribution shifts is to alert of a model's unreliability. The similarity between the covariate distributions in the training and deployment datasets [242] may provide evidence of a distribution shift. However, when only a sample is available from the deployment distribution, practitioners must rely on anomaly detection [48] to identify how a given point differs from the original distribution or favour models with outputs that capture this uncertainty [124, 181].

Domain adaptation. When samples from the deployment distribution are available, one may alter the model's training to better model the deployment distribution. For instance, Zhang et al. [372] describe an inverse weighting strategy to improve modelling under the two previously described shifts. Similarly, Fang et al. [87] further adapt inverse weighting to improve the transportability of deep learning models. Lipton, Wang, and Smola [189] propose a strategy when one does not have access labels in the target domain. However, all these strategies assume access to samples from the deployment distribution.

Domain generalisation. In opposition to domain adaptation, domain generalisation aims to create models transportable under shifts *without access to samples from the deployment distribution*. Zhou et al. [380] review the literature on this topic. Closest to our work, regularisation and self-supervised learning techniques [213] may improve the transportability of a model to a new domain. For instance, Mohseni et al. [213] propose image reconstruction as self-pretraining, resulting in detecting out-of-distribution images. Our work resembles this self-supervision as we aim to train the architecture to model the observational process.

Domain shifts in medicine. Despite the extensive literature on this topic, existing methods assume specific shifts or access to samples from the deployment distribution, not reflecting the nature of real-world medical shifts [302]. Spathis and Hyland [302] criticise this lack of real-world shift studies and demonstrate that multiple state-of-the-art strategies do not improve out-of-distribution prediction across hospitals of the eICU dataset [253] on in-hospital mortality

prediction. Similarly, Guo et al. [118] show that domain generalisation or adaption techniques fail in improving transportability in the MIMIC dataset across years. In this same setting, Nestor et al. [230] propose extracting clinically-relevant features to improve transportability. The authors demonstrate these features reduce the necessary modelling complexity, with logistic regression presenting one of the best performances and transportability. Note that while these works study mortality prediction, only [247] explore the transportability of survival models and show limited gain from distribution shift mitigation strategies.

A final challenge in clinical distribution shifts is the lack of publicly available benchmarks with limited data across hospitals — with eICU being a rare exception with multiple institutions — and limited common measurements to evaluate shifts. In this chapter, we introduce a novel way to assess transportability through natural experiments, such as the weekend effect, present in medical datasets.

6.3 Clinical presence shift

At the core of our work is the challenge of leveraging the signal contained in clinical presence through the observed irregularities while ensuring transportability under potential shifts in observational processes. This section formalises the challenge of longitudinal irregularities and the possible clinical presence shift in the observational process.

6.3.1 Irregularities

Consider the longitudinal setting where a subset of medical tests occur irregularly. Formally, we can define a set of random variables $\{X_j^*, E_j, O_j\}$, with X_j^*, the covariate vector associated with the observed values at the j^{th} encounter; E_j, the time since the last encounter and O_j, the indicator vector of covariates observed: one dimension per covariate. Consider the unobserved vector X_j of all covariates that could be observed at encounter j, and the patient outcome Y, considered as a final univariate outcome in our setting. In this context, we formalise the problem of missingness similarly to Chapter 3 and extend it to the longitudinal setting in which we model the next O_j given previous observations.

Missingness. At each encounter j, a physician only orders a subset of all possible covariates. Following [212]'s notations to express missingness, we denote the observed covariates as:

$$\forall k, X_{j,k}^* = \begin{cases} \varnothing & \text{if } O_{j,k} = 0 \\ X_{j,k} & \text{otherwise} \end{cases}$$

with k describing the different covariates.

As data are longitudinal, this formalisation extends Rubin [271]'s categorisation conditioned on *all observed variables* $\mathcal{H}_j = \{X_l^*, E_l, O_l\}_{l<j}$ until encounter j excluded:

- **Missing Completely At Random** (MCAR) in which the missingness process is independent of all variables,
$$\mathbb{P}(O_j \mid \mathcal{A}) = \mathbb{P}(O_j)$$
where $\mathcal{A}$ denotes the set of all potential random variables.
- **Missing at Random** (MAR) in which the missingness process depends only on observed covariates,
$$\mathbb{P}(O_j \mid \mathcal{A}) = \mathbb{P}(O_j \mid \mathcal{H}_j)$$
- **Missing Not At Random** (MNAR) in which the missingness process relies on unobserved covariates or the missing value itself.

Time of observation. To characterise the temporal irregularities of medical data, we adopt a temporal point process formalisation, similarly to [98, 258, 328]. However, we focus on the inter-encounter times, equivalent to a time to event analysis as introduced in Chapter 2. We express the time to the encounter j from the current encounter $j-1$ through the intensity function λ_j, which characterises the instantaneous risk of observing an event in the interval $[\zeta, \zeta+\delta\zeta]$ *following the* $j-1$ *encounter*. Following the same distinction in possible observational patterns as for missingness, we characterise this process as:

- **Sampled Completely At Random** (SCAR) in which the time to the next event is independent of all other variables,
$$\lambda_j(\epsilon \mid \mathcal{A}) = \lambda_j(\epsilon) = \lim_{\delta\epsilon \to 0} \frac{\mathbb{P}(\epsilon < E_j < \epsilon + \delta\epsilon)}{\delta\epsilon}$$
- **Sampled at Random** (SAR) in which the time to the next event depends only on previously observed variables,
$$\lambda_j(\epsilon \mid \mathcal{A}) = \lambda_j(\epsilon \mid \mathcal{H}_j) = \lim_{\delta\epsilon \to 0} \frac{\mathbb{P}(\epsilon < E_j < \epsilon + \delta\epsilon \mid \mathcal{H}_j)}{\delta\epsilon}$$
- **Sampled Not At Random** (SNAR) in all other settings.

Many existing longitudinal models in ML ignore the stochasticity of this process, with a deterministic assumption of regular observation that rarely applies in medical settings. When explicitly studied, previous works often focus on one of these two dimensions; we reconcile both by modelling the joint process.

Remark 6.1. Equivalently, one could jointly model each covariate point process, i.e. the time to the next measurement of each covariate, instead of the the joint inter-encounter and missingness process. However, implementing this approach with the proposed architecture requires a larger number of parameters.

Notations. In the following, we consider the problem of a population of N patients with a series of longitudinal realisations of $\{X^*_{i,j}, E_{i,j}, O_{i,j}\}$ observed during the first day after admission to an Intensive Care Unit (ICU) denoted by $x^*_{i,j}$ for the vector of laboratory tests — with missing values — where the additional subscript i denotes patient i. Note that $j \in [\![0, l_i]\!]$, meaning that different patients may have different numbers of encounters with l_i, the last encounter in the 24 hours following admission for patient i. We denote the time since the last encounter as $\epsilon_{i,j}$ and $o_{i,j}$, the indicator vector of observed covariates at encounter j. Additionally, we consider the survival outcome after the first 24 hours post-admission, each patient has an associated observed follow-up time t_i, and an event indicator, d_i, with $d_i = 0$ denoting discharge and $d_i = 1$ associated with death in the hospital.

6.3.2 Shift

While practitioners often models informative irregularities, a shift in clinical presence may endanger the resulting model's transportability. Despite the previously mentioned works evidencing distribution shift, formalisation of the observational process shift has yet to be proposed. Note that closest to our work, Zhou, Balakrishnan, and Lipton [379] describe the problem of missingness shift with proposed correction under MCAR patterns, we describe the more general longitudinal *clinical presence shift*.

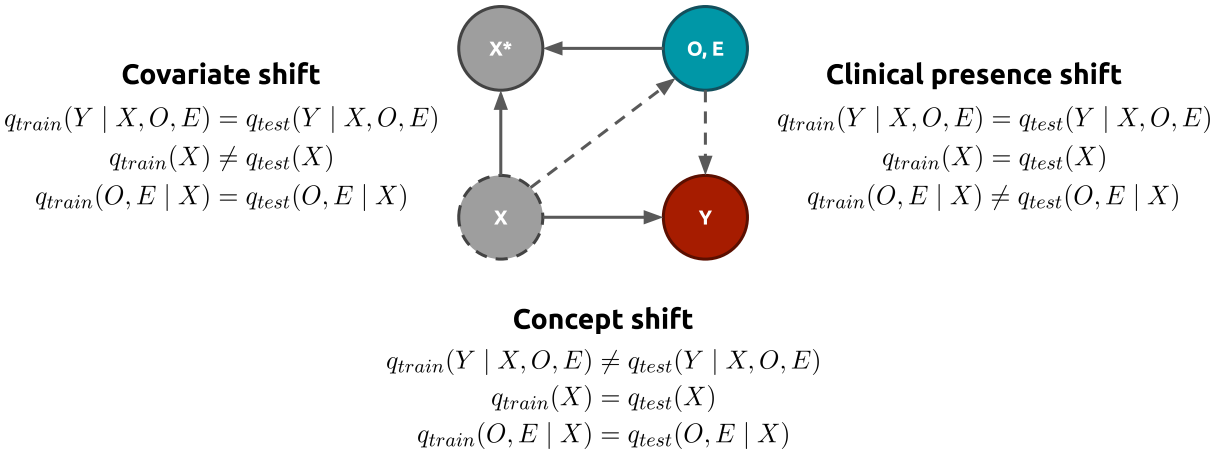


Figure 6.2: Clinical Presence shift. Full circled covariates are observed, and dashed ones are unobserved. Dashed arrows indicate potential dependencies, while the solid ones show assumed dependencies. Distinct from covariate and concept shift, clinical presence shift is a shift in the joint distribution O, E while all other distributions remain the same.

As discussed in Section 6.2.2, different characterisations of distribution shift and associated mitigation strategies have been proposed but largely discard the critical impact of the observational process. Figure 6.2 redefines the problem of distribution shifts by adding the observational process characterised by (O, E). In this figure, we dissociate observed X^* and observable X covariates and introduce their dependencies. Additionally, we display potential dependencies between unobserved covariates and the observational process that would reflect SNAR and/or MNAR patterns, as well as one between the observational process and observed outcomes, indicating the possible impact of the observation on the outcome itself. This representation highlights a new type of shift: a *clinical presence shift*. Assuming an unchanging concept distribution, $q(Y \mid X, E, O)$ and covariate distribution, $q(X)$, a clinical presence shift is a change in the distribution $q(E, O)$ resulting in shifts in the observed covariate distribution $q(X^*)$ and the modelled $q(Y \mid X^*)$. Generally, the problem of clinical presence shift can be seen as both a concept and covariate shift as both joint distributions change. In the clinical setting, this means that even when faced with the same underlying population with the same associated outcome, a model trained under a given observational process may not be adapted to a new one.

Remark 6.2. In this formalisation, we dropped the subscript j for clarity. However, to account for the longitudinal dependencies, Figure 6.2 would be repeated for each encounter j with temporal dependencies upon $\mathcal{H}_j$ between these DAGs as described in the previous section.

6.4 Joint modelling of clinical presence and survival outcome

To tackle the problem of both performance under clinical presence and transportability under shift, we propose a recurrent neural network that *jointly* models the outcome of interest and the assumed SAR and MAR observational processes. Specifically, we model the survival outcome with DeepSurv [163], and, in parallel, the time to the following observation and missingness. The proposed model, referred to as **DeepJoint**, allows scalable joint modelling by an end-to-end gradient descent maximisation of the full likelihood.

6.4.1 Motivation

The proposed methodology aims to tackle two challenges emerging from clinical presence: performance and transportability.

DeepJoint for performance. The proposed method is at the intersection of two fields of research: joint modelling and multitask learning. First, joint models have been proposed in the statistical literature to incorporate informative processes into the outcome model through a

shared effect [98, 303, 310]. These methods often consider a regression for *one dimension* of the observational process and a second for the outcome of interest. The shared effect consists of a parameter shared between both regressions to embed the informative observational process in the outcome model. These models suffer from poor scalability to the number of covariates. Our proposed deep learning architecture tackles this issue and relaxes the parametric assumptions often made by these models. Then, this method is also inspired by the multitask learning literature in which models aim to predict multiple outcomes to improve performance [45, 46]. Theoretical foundations are still lacking [274] and rely on the intuition of regularisation of the shared embedding [277]. Our work bridges the gap between these two domains by using multitask learning to model the informative observational process, forcing the shared embedding to contain a representation informative of both the observational process and outcome of interest.

DeepJoint for transportability. Under small shifts in inputs, multitask learning improves transportability. Specifically, Mao et al. [202] show the proportionality of the error to the inverse of the number of tasks modelled under adversarial attacks. In the following section, we adapt their theorem to the problem of clinical presence shift and discuss its limitations. This theorem describes that, under *similar medical practices*, one would improve transportability by modelling multiple uncorrelated outcomes. The theory, therefore, motivates the use of multitask learning to tackle the transportability under shifts.

6.4.2 Theoretical transportability

In this section, we adapt the definition and theorem proposed in [202] to formalise the improved transportability of a multitask neural network under clinical presence shift (refer to original work for the detailed proof). This theorem justifies how modelling clinical presence as an outcome should render the model more transportable.

First, we introduce the *clinical presence vulnerability* defined as the expected change in loss under two different observational processes. With the observed time series at development $\mathcal{H} = \{x_l^*, t_l, o_l\}_l$, and at deployment $\mathcal{H}'$, the target outcomes y, a loss $\mathcal{L}$, such that the difference between the marked point processes resulting from the two observational processes lies in a p-norm ($||\cdot||_p$) bounded ball with radius r, i.e. $||\mathcal{H} - \mathcal{H}'||_p < r$. Note that we assume the existence of a p-normed time series space, which includes an operator $+$, where $+\delta$ describes a perturbation of the time series. Under this assumption, the clinical presence vulnerability over the target dataset is:

$$\Delta\mathbb{E} := \mathbb{E}\Big[|\mathcal{L}(\mathcal{H}, y) - \mathcal{L}(\mathcal{H}', y)|\Big] \leq \mathbb{E}\Big[\max_{||\delta||_p < r} |\mathcal{L}(\mathcal{H}, y) - \mathcal{L}(\mathcal{H} + \delta, y)|\Big]$$

Assuming *infinitesimal perturbations* in the observational processes, i.e. $r \to 0$, one can develop

the upper-bound using Taylor expansion:

$$|\mathcal{L}(\mathcal{H}, y) - \mathcal{L}(\mathcal{H} + \delta, y)\ | = |\partial_H \mathcal{L}(\mathcal{H}, y)\delta + \mathcal{O}(\delta)|$$

with $\partial_H \mathcal{L}$ the partial derivative of the loss given the input time series H. Using the property of the dual norm d, one can show that:

$$\Delta\mathbb{E} \propto \partial_H \mathbb{E}\Big[||\mathcal{L}(\mathcal{H}, y)||_d\Big]$$

Theorem 6.1 (Multitask robustness from [202])**.** *If the tasks are correlated with each other such that the covariance between the vector of gradients of the loss for task u and task v is $Cov(g_u, g_v)$, and these gradients are i.i.d. with zero for average, then:*

$$\Delta\mathbb{E} \propto \sqrt{\frac{1 + \frac{2}{M}\sum_{u=1}^{M}\sum_{v=1}^{u-1}\frac{Cov(g_u, g_v)}{Cov(g_u, g_u)}}{M}}$$

where M is the number of output tasks and g_u, the vector of gradient defined as $\forall \mathcal{H}, \partial_H \mathcal{L}_u(\mathcal{H}, y)$ associated with the task-specific loss associated with the time series $\mathcal{H}$.

Intuition. The theorem quantifies the error resulting from a shift between two time series. When the difference between these time series is negligible, the theorem states that the more tasks a multitask neural network models, the smaller the error in the predicted outcome. Additionally, the nominator shows that the more uncorrelated the tasks are, the more robust the model is to perturbations. Note that this result relies on three assumptions: (i) the gradients are i.i.d., which requires the times series at deployment to be i.i.d, (ii) the average gradient is null, corresponding to the training having converged, and (iii) the perturbation is infinitesimal. The distance between time series is challenging to measure and unlikely to be negligible between patients in different medical settings, this last assumption is less likely to be met. In our future work, we aim to relax this assumption to improve this result's clinical relevance. However, this theorem aims to provide an intuition of why the proposed modelling may improve transportability.

6.4.3 Implementation

The proposed architecture models two dimensions of clinical presence: inter-encounter times and missingness, modelled through two different neural networks. Each relies on the embedding h_j outputted by a Recurrent Neural Network (RNN) with input (x_j^*, o_j, ϵ_j) at encounter j. Note that, by training the RNN, we embed in h_j all observed observations up to j *included*, corresponding to $\mathcal{H}_{j+1}$. In our implementation, we adopt an LSTM [130] that models the sequential nature of the observed data and captures temporal dependencies. Note that the

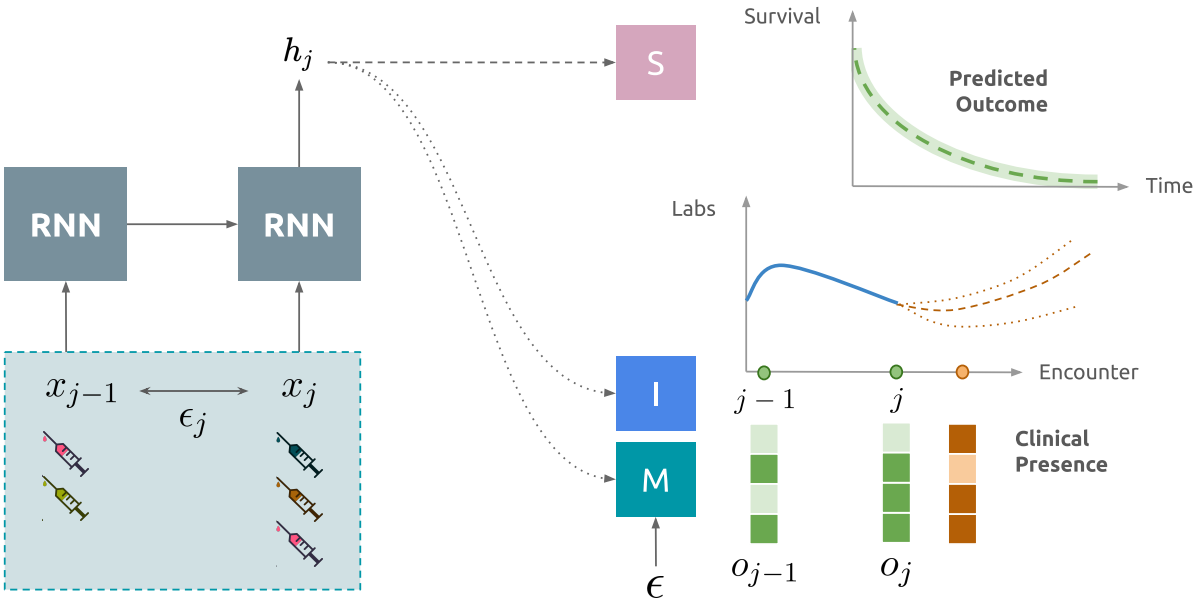


Figure 6.3: Deep Joint Model - Multitask modelling of clinical presence and survival outcome. A RNN extracts an embedding h_j then used to model clinical presence through the networks **M** — to model the missingness patterns, **I** — to model the inter-observation times, and the survival outcome through the network **S**

.

LSTM's inputs are the series of observations at *irregular times*. The resulting shared embedding at the end of the 24-hour observation period is inputted to the survival model. The following describes the implementation choice for each of these components.

Temporal process: When will be the subsequent measurement? A monotonic positive neural network $\mathbf{I}$ models the cumulative hazard associated to the time to the next event. As previously described, we enforce (i) monotonicity by using the square of all weights and (ii) postivity through a final Softplus layer. Formally, the network models:

$$\mathbb{P}(E_{j+1} < \epsilon \mid \mathcal{H}_{j+1}) = 1 - \exp(-\mathbf{I}(h_j, \epsilon))$$

Missingness process: Which laboratory tests will be performed next? A multi-layer perceptron $\mathbf{M}$ uses h_j as inputs to model the probability of observing the different covariates at encounter $j+1$, i.e. which tests are likely to be performed by ϵ.

$$\mathbb{P}(O_{j+1} = o \mid E_{j+1} < \epsilon, \mathcal{H}_{j+1}) = \mathbf{M}(h_j, \epsilon)$$

Survival outcome: How long will the patient survive? Finally, to model the hazard function $\lambda(t \mid \mathcal{H}_{l_i+1})$ at the last encounter l_i in the observation period, we use the DeepSurv model [163] that relies on a multi-layer perceptron $\mathbf{S}$ to extract a non-linear covariates shift

used in a standard multiplicative proportional hazards Cox model. Given λ_0, the baseline hazard, the hazard of observing an event at horizon t is expressed as:

$$\lambda(t \mid \mathcal{H}_{l_i+1}) = \lambda_0(t) \exp(\mathbf{S}(h_{l_i}))$$

The final model, therefore, combines existing architectures in a novel way to model clinical presence and survival outcomes jointly. This results in a latent representation h_j, which embeds both the observational process and survival outcome.

6.4.4 Training

Training consists of maximising the longitudinal irregularities and survival likelihoods. Each objective is backpropagated simultaneously by averaging the loss of the different tasks: survival, inter-encounter temporality and missingness processes. Following the multitask literature, we ensure that inter-encounter temporality and missingness are balanced using a dynamic weighting average scheme [194], as imbalances can significantly impact performance [109]. Specifically, we weights these two tasks using the relative change at iteration s as follows:

$$\forall\, task \in \{I, M\}, w_{task}(s) = \frac{L_{task}(s)}{L_{task}(s-1) \cdot \theta}$$

which is then normalised between the two clinical presence dimensions using a Softmax. L_{task} is the average training likelihood for the given task, and θ is a temperature hyperparameter that controls softness, i.e. larger values would lead to equal weights. In this work, we follow the empirical recommendation of the original paper and set θ to 2.

Temporal process. Our choice of monotonic networks results in the exact computation of likelihood associated with the temporal process. The model outputs the cumulative intensity at horizon ϵ and automatic differentiation results in the instantaneous intensity.

$$l_I = \sum_i \frac{1}{l_i - 1} \sum_{j \in [\![1, l_i - 1]\!]} \left(\mathbf{I}(h_{i,j}, \epsilon_{i,j+1}) - \log \left. \frac{\partial \mathbf{I}(h_{i,j}, \epsilon)}{\partial \epsilon} \right|_{\epsilon = \epsilon_{i,j+1}} \right)$$

Missingness process. Similarly, the likelihood for the associated mark for the missingness process results in the binary cross entropy loss:

$$l_M = -\sum_i \frac{1}{l_i - 1} \sum_{j \in [\![1, l_i - 1]\!]} \Big((1 - o_{i,j+1}) \cdot \log[\mathbf{M}(h_{i,j}, \epsilon_{i,j+1})] + o_{i,j+1} \cdot \log[1 - \mathbf{M}(h_{i,j}, \epsilon_{i,j+1})] \Big)$$

Survival outcome. As described in Chapter 2, DeepSurv [163] uses a neural network to express the log-hazard in a Cox model, following the same optimisation of the partial

log-likelihood:

$$l_S = \frac{1}{\sum_{i,d_i=1} 1} \sum_{i,d_i=1} \left(\mathbf{S}(h_{i,l_i}) - \log \sum_{k,t_k>t_i} \exp(\mathbf{S}(h_{i,l_i})) \right)$$

Finally, the population baseline hazard, $\lambda_0(t)$, is estimated with a Breslow estimator.

The final loss l is defined at iteration s as

$$l(s) = (1-\alpha) \cdot l_S + \alpha \sum_{task \in \{I,M\}} w_{task}(s) \cdot l_{task}$$

with $\alpha \in [0,1]$, a hyperparameter balancing the survival and observational process losses. This loss is computed on the training set and backpropagated throughout the architecture. We also compute it on a validation set for early stopping of the multitask training with $w_{task} = 1$. Then, we fine-tune each network head, not backpropagating through the shared RNN.

6.5 Case study: weekend effect

Using laboratory tests performed during the 24-hour observation period after admission to the ICU in the Medical Information Mart for Intensive Care III (MIMIC III) [155] dataset, we model survival. As in the publicly available version of the dataset, dates have been shuffled for anonymisation, and the data is gathered in a single hospital, it is not possible to study transportability across time and place. However, as admission days have been maintained, we investigate how models' performance would change under the natural experiment of weekend effect [17, 244], where a distribution shift occurs between patients and clinical care over weekends and weekdays. Specifically, we evaluate how a model trained using patients admitted on weekdays would perform on the population admitted on weekends. For reproducibility, the proposed models' and experiments' implementations are available on Github[1].

6.5.1 Dataset

The MIMIC III dataset gathers laboratory tests, vital signs and diagnoses of 38,597 anonymised patients admitted to the Beth Israel Deaconess Medical Centre between 2001 and 2012. This analysis focused on laboratory tests as they are more likely to embed practitioners' expertise, and consequently inform outcome and other informative modalities might present different clinical presence patterns, e.g. semi-automatic vital signs collection. After following the pre-processing from [338], we selected a set of adults with shared laboratory tests using an ECLAT algorithm [368] such that each laboratory test is performed at least once during the 24-hour

[1] `https://github.com/Jeanselme/MultitaskTransportability`

post-admission, and patients survived this period. The resultant subset consists of 31,692 patients with 21 different laboratory tests (See Tables D.1 and D.2 in Appendix for description).

6.5.2 Empirical setting

Baselines. We compare **DeepJoint** against different strategies for handling longitudinal irregularities. All methods rely on the same normalised data imputed using the last observations carried forward with patient-mean imputation, i.e. most recent laboratory tests replace missing data; remaining missing tests use the lab overall mean. The compared approaches extract a representation of the longitudinal laboratory tests. Based on this embedding, DeepSurv, a multi-layer perceptron, estimates the hazard shift from the population hazard baseline.

First, we compare against two non-longitudinal approaches:

- **Last encounter (Last)**: Extract the last encounter l_i as representation for each patient i. This approach assumes no informativeness in the evolution of the laboratory results.
- **Summarising (Count)**: In addition to the last observed laboratory results, add the count of each test performed in the first 24 hours. This common practice utilises the counting process to reflect the severity of the patient's condition. Still, it ignores whether the patient is improving or worsening as the temporal evolution is left aside.

Second, we use RNN-based approaches to take advantage of the longitudinal evolution of the laboratory tests:

- **Ignoring clinical presence (Ignore)**: An LSTM is trained on the imputed data modelling the inputs' temporal order but ignoring their irregularity and missingness patterns.
- **Resampling (Resample)**: Imputed data are re-sampled every hour to satisfy the LSTM regularity assumption. This resampling presumes the non-informativeness of the observational process.
- **Modelling (GRU-D)**: The imputed data concatenated with missingness indicators serve as inputs to a GRU-D model [50], which models inter-observation times with a decaying function, following the intuition that the hidden state stabilises with time.
- **Featurization (Feature)**: Missingness indicators and time elapsed since previous observation [190] are two informative proxies of clinical presence. An LSTM uses for inputs both the laboratory tests and these features to model survival.

The proposed **DeepJoint** architecture relies on the same input as Feature to meet the assumption of SAR. However, the proposed architecture differs in how the latent embedding at encounter j is used to model the time to and missingness at the subsequent encounter.

Remark 6.3. Our approach is not dependent on a specific choice of RNN; any alternative to the proposed LSTM — used for a fair comparison with Feature — could be used for modelling longitudinal data.

Training procedure. All methods use the same 80%-20% train-test *patients split* to train the different models. Their training relies on gradient backpropagation with an Adam optimiser [166] over 1000 epochs with early stopping on 10% of the training set. The entire network is optimised for 500 epochs. The remaining iterations are for fine-tuning the survival network. We perform hyperparameter tuning using a 10% left-aside set of patients from the training set on 50 random draws from the grid presented in Appendix Table D.3.

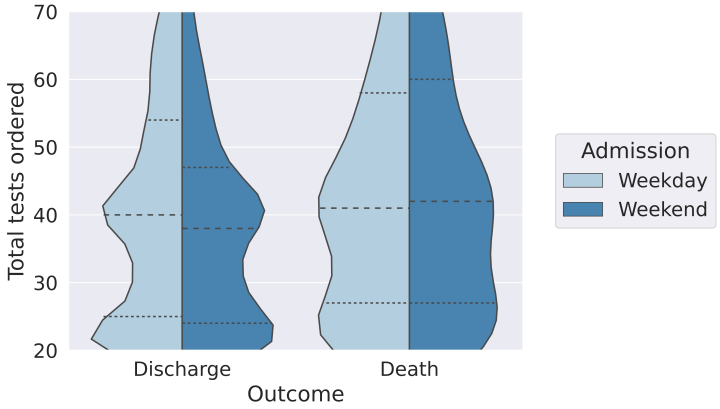


Figure 6.4: Total number of distinct tests ordered in the 24 hours after admission to the ICU. Patients admitted on weekends with favourable outcomes are less often tested than similar patients admitted on weekdays.

The previous experiment explores how handling irregularities can impact predictive performance. In a second set of experiments, we focus on how the day of admission may impact the observational process. Motivated by the difference in the counts of tests performed on patients admitted on weekends and weekdays, shown in Figure 6.4, we hypothesise that practice might differ between weekends and weekdays due to physicians and laboratory availability.

The weekend effect is a well-known phenomenon in public health: patients admitted on weekends have worse outcomes than patients admitted on weekdays [244]. While providing an opportunity to study survival models' transportability, this outcome gap represents a target shift under which we cannot directly compare performance. To make this comparison possible, we adopt the following evaluation. We split patients given days of admission — we define weekends admissions from Friday 8 pm to Sunday 8 pm and weekdays as any other. Each group is further divided between training and testing. As described in Figure 6.5, a first model uses the train set of patients admitted on weekdays and tested on the test set of weekends-admitted patients. Then, a second model uses the train set of patients admitted on weekends and tested on the test set of weekends-admitted patients. The two resulting models are comparable as

evaluated on the same test set. However, due to the difference in population size between patients admitted on weekdays compared to weekends, the quantity of data could confound the estimated transportability. To address this challenge, we follow [302] 's subsampling approach on the weekdays-admitted population to match the size of the weekend one. In the following, we focus on the performance for patients admitted on weekends; we present the converse analysis on the weekdays-admitted patient set and further ablation studies demonstrating the transportability of the proposed architecture in Appendix D.1.3.

Figure 6.5: Transportability evaluation between patients admitted on weekdays and weekends. To ensure a fair comparison, all models are evaluated on the same population subset.

Evaluation. We compute survival prediction at the last observation l_i in the 24-hour post-admission period. Models are compared using a time-dependent C-index and Brier score evaluated at time horizons 7 and 30 days after the observation period. Additionally, we adopt an Integrated C-Index to quantify the ranking quality of patients' risks, as risk at a given time horizon in the ICU setting may be less relevant than patients' prioritisation. Standard deviations were obtained using 100 bootstrapped iterations on the test set predicted values.

6.5.3 Predictive performance

Figure 6.6 describes the models' discriminative performance at different evaluation horizons on the test set (see Appendix D.1.3 for Brier score and tabular results). Note that all models rely on a DeepSurv network to model the survival outcome and only differ in their inputs and how they handle longitudinal irregularities.

Insight 1: Clinical presence informativeness fades on longer term. All methodologies have decreasing performance at 30 days compared to the 7-day horizon, and all converge towards the same discriminative performance. In the ICU context, modelling long-term survival is challenging as patients are in critically unstable states. This decrease stresses how survival is a complex outcome to model in the ICU. Critically, the lesser performance improvement associated with methodologies modelling clinical presence shows that the longitudinal irregularities reflect short-term instability and are less relevant in the long term.

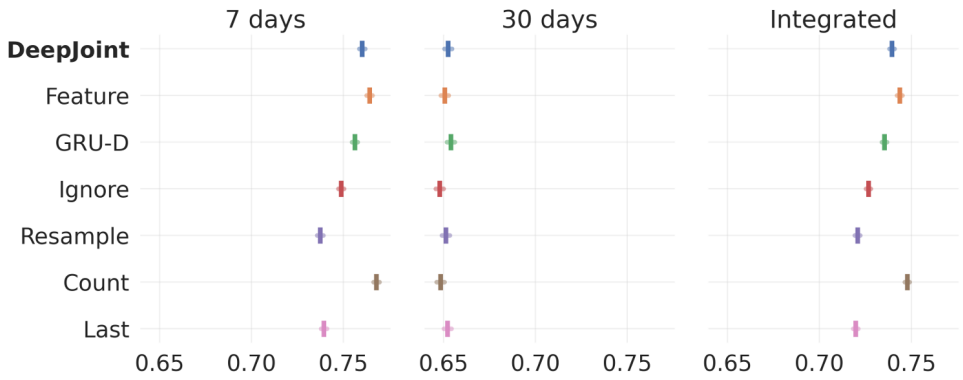


Figure 6.6: Performance comparison through C-Index on a random split of the MIMIC III population with 95% CI. The proposed Deepjoint (in bold) performs similarly to Feature, which uses the same inputs. All methodologies considering clinical presence present improved performance compared to their associated alternatives discarding it.

Insight 2: Observational patterns inform predictive modelling. Integrating clinical presence features improves models' performance: Count improves over Last, and Feature over Ignore in each of the three considered metrics. Note that these models differ in the input information with Count using the laboratory test count in addition to their last observed values, and Feature including the mask and time since the last observation in addition to the longitudinal tests modelled by Ignore. The central observation from the integrated metric is that modelling longitudinal irregularities improves performance, as shown by the increased performance of GRU-D, Feature and DeepJoint over Ignore. Focusing on the proposed Deepjoint, its performances are comparable with Feature, showing that the shared embedding regularisation does not significantly reduce performance. Finally, the simple Count baseline presents one of the best performances in this setting, echoing [230]'s remark on the superiority of medically relevant features over more complex modelling for improved predictive performances. In the following, we show that these features may be more sensitive to the deployment setting than the proposed methodology.

This first experiment demonstrates the informativeness of longitudinal irregularities to model survival. Analyses in the literature often stop at this stage. In the following, we propose to study how an internal shift may impact these strategies.

6.5.4 Transportability

Figure 6.7 presents the performance when a model is transferred from the weekday to the weekend setting (y-axis), and when a model is trained and tested in the same setting (x-axis). A model transportable under shift performs similarly when transferred from another setting as when trained under the same setting, i.e. models close to the diagonal. This diagonal delimits underfitting (above) from overfitting (under) to the training observational patterns. Practitioners should, therefore, select a model close to the diagonal with the best discriminative

performance (upper right corner). As a measure of transportability, we introduce the absolute difference in C-index performance between the model trained and tested in the same setting and the transferred one, referred to as *transfer loss*. The smaller the transfer loss, the more transportable the model is. Table 6.1 describes each model's transfer loss at the three considered time horizons used for evaluation.

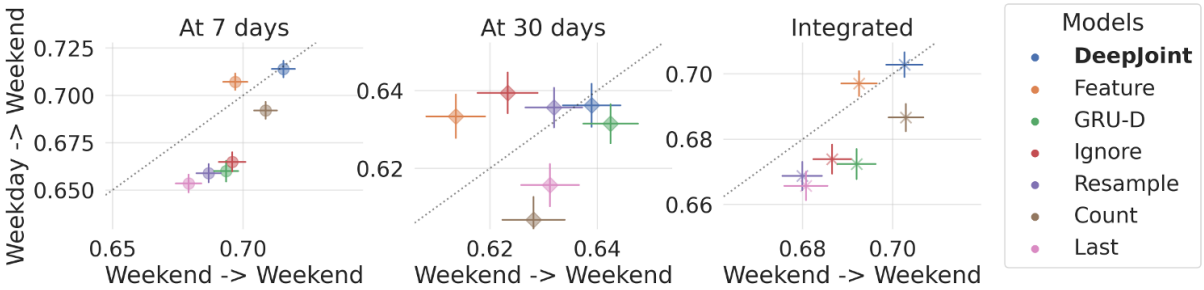


Figure 6.7: Discriminative performance evaluated on patients admitted on weekends for a transferred model (*y-axis*) and a model trained on weekends-admitted patients and tested on the test set of this same group (*x-axis*) (with 95% CI). Models closer to the diagonal are more transportable. DeepJoint presents improved transportability under the weekend effect shift.

Insight 3: Current practices under-perform under shift. In this population, performance differs from the previous random split, as shown on the x-axis performance. Specifically, laboratory test counts are less informative, resulting in lower performance. However, leveraging longitudinal irregularities improves performance, with DeepJoint, Count and Feature presenting the best integrated C-Index. Conversely, models based on the last observation and using resampled data show the worst discriminative performance in this population, aligning with the intuition that ignoring the observational patterns reduces performance. Focusing on the models' transportability, methods ignoring irregularities or using simple features tend to underperform under shifts. This counter-intuitive observation underlines that ignoring irregularities does not improve transportability. Notably, we would like to echo the remark on the risk of using informative observational patterns raised by [190, 298] in which the authors underline how taking advantage of observational patterns, such as missingness, might lead to a mismatch between the development and deployment settings, but that ignoring this information might not be possible. Finally, note that existing strategies considering the longitudinal irregularities also drift away from the diagonal, demonstrating their sensitivity to shift. Consequently, ignoring or using longitudinal irregularities is not enough to improve transportability.

Insight 4: Joint model improves transportability. DeepJoint achieves state-of-the-art discrimination performance in both settings as shown in Figure 6.7. Particularly, when compared with Feature, which uses the same input, our proposed model presents improved performance

Model	Evaluated on weekends		
	7 days	30 days	Overall
DeepJoint	**0.014** (0.012)	**0.013** (0.010)	**0.011** (0.009)
Feature	*0.015* (0.011)	0.021 (0.012)	**0.011** (0.009)
GRU-D	0.033 (0.015)	*0.015* (0.010)	0.020 (0.012)
Ignore	0.032 (0.017)	0.020 (0.014)	0.016 (0.012)
Resample	0.029 (0.015)	*0.015* (0.010)	*0.015* (0.011)
Count	0.020 (0.013)	0.022 (0.012)	0.017 (0.010)
Last	0.026 (0.012)	0.017 (0.012)	*0.015* (0.009)

Table 6.1: Transfer loss - Mean (std). The smaller the loss, the more transportable the model is. Best performances are in **bold**, second best in *italics*. DeepJoint presents the best transportability properties.

and transportability with lower transfer loss. This underlines how modelling *jointly* survival and clinical presence regularises the embedding against shifts.

This experiment demonstrates that traditional strategies for handling longitudinal irregularities appear to improve performance in internal validation but are prone to overfitting when transferred to a different clinical presence setting. The proposed joint model tackles this challenge under the weekend effect shift. Connecting these results with the previous performance on the random split, a trade-off exists between same setting performance and performance under shift, as hypothesised in [312]. In our experiments, models such as Count and Feature perform best in the random split but suffer from clinical presence shifts.

6.5.5 Algorithmic fairness and clinical presence shift

As argued throughout this thesis, clinical presence patterns can vary across populations. Consequently, a clinical presence shift may impact these groups differently. A growing literature at the intersection of fairness and distribution shifts [282, 295] explores how to maintain algorithmic fairness properties under distribution shifts. In this section, we empirically assess the impact of the weekend effect on the considered approaches to handle longitudinal irregularities. Specifically, we measure the integrated C-Index for male and female patients. Figure 6.8 illustrates how the different models behave under shifts in each subgroup.

Insight 5: Longitudinal irregularities handling impact subgroup transportability. Figure 6.8 shows how the considered models have varying discriminative performance across groups. Note that the proposed DeepJoint presents the best performance in both groups for both transferred and internal models, suggesting a potential algorithmic fairness consequence of Theorem 6.1 that we will explore as future work. When focusing on transportability, GRU-D, Count, and Ignore present a larger performance drift for women than men, while Feature and DeepJoint present smaller change. This observation indicates that distribution shifts may

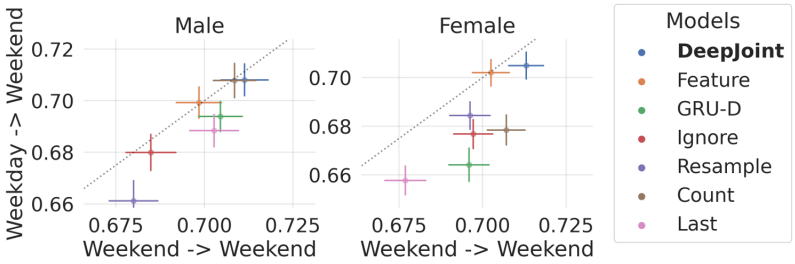


Figure 6.8: Integrated C-Index evaluated on patients admitted on weekends for a transferred model (*y-axis*) and a model trained on weekends-admitted patients and tested on the test set of this same group (*x-axis*) stratified by sex (with 95% CI).

impact groups differently and that the choice of irregularities handling strategy may play a critical role in improving algorithmic fairness properties under shifts.

6.6 Discussion

In this chapter, we have explored the impact of clinical presence shifts on predictive models. We discuss the contributions, recommendations, and future works in the following.

6.6.1 Contributions

The ML community has studied and developed mitigation strategies for a subset of distribution shifts that do not capture the complexity of the shifts observed in medical data. In this chapter, we decouple observed covariates from observable ones. This distinction underscores a new type of shift: *clinical presence shift*. This formalisation offers a new framework to understand the problem of medical distribution shift as the concept and the underlying covariate distributions may remain stable, but the observational process evolves. This contribution is central in developing predictive models as their adoption in medical practice will alter the observational process.

The central challenge with clinical presence shifts is that one would like to take advantage of informative dimensions of the outcome of interest while ensuring the model's transportability under an evolving process. At the intersection of adversarial ML and statistical joint model, we introduce a multitask neural network that models the survival outcome and the longitudinal irregularities associated with clinical presence, namely temporality and missingness. The motivation behind this joint modelling is the mounting evidence of improved performance and adversarial robustness when considering multiple tasks. Our analysis of the MIMIC III dataset demonstrates that the practice of ignoring the observational process lowers predictive performances. Not only does it reduce performance in a given setting, but counterintuitively, ignoring the observational process reduces transportability under shift. On the contrary, the proposed method presents improved transportability properties.

Publicly available datasets may not allow the study of temporal or inter-hospital distribution shifts. Following the public policy literature, we use natural experiments to measure models' transportability. Specifically, our proposed transportability evaluation relies on the distribution shift observed between weekends and weekday admissions.

6.6.2 Recommendations

In this chapter, we demonstrate the critical importance of considering clinical presence for improving predictive models' performance and transportability. Specifically, we invite model developers to:

1. *Model clinical presence.* Our work introduces a novel way to not only featurise clinical presence but model it as an output. As shown in this chapter, this approach presents a step towards transportability. Nonetheless, one should remain wary in using clinical presence for predictions and carefully assess whether patterns are indicative of the outcome or reflective of potential biases.

2. *Leverage natural experiments.* While the necessity of transportability across hospitals is sometimes questioned [94], we argue that models should be transportable under naturally occurring shifts *in a given hospital*. We invite model developers to take advantage of natural experiments to study such transportability properties, as proposed in the proposed weekend effect analysis in Section 6.5.

6.6.3 Future work

The proposed analysis empirically shows the importance of modelling clinical presence, which opens multiple avenues for future work. First, the proposed formalisation invites further theoretical work and alternative mitigation strategies with deeper mathematical grounding as multitasking remains empirically superior, but the theory still needs to be developed. Additionally, we would like to model additional tasks, such as the longitudinal values, as they should theoretically improve transportability further. Then, on the empirical side, we would like to use longer observational periods to measure the gain from longitudinal modelling compared to simple counts. To confirm the results presented in this chapter, we will also explore additional natural experiments such as the July effect [367], grouping practitioners by expertise, and other real-world datasets. Further, synthetic data would be a valuable tool to quantify the shifts under which the proposed approach fails, and evaluate the model's capacity to *detect* distribution shifts.

Chapter 7

Conclusion

Wald's pioneering investigation into reinforcing aircraft [332] underscores the pervasive issue of the observational process. Constrained to analyse planes that returned from the battlefield, Wald was tasked with identifying parts to fortify on new aircraft. Instead of reinforcing the returning planes' most damaged parts, Wald recommended to strengthen those seemingly untouched. His rationale was that planes damaged in these specific areas likely did not return from the field. This widely recognised survivorship bias underscores the critical importance of understanding the observational process for accurate data interpretation.

A similar challenge arises in the realm of clinical observation. Neglecting the observational process can lead to flawed interpretations of data. This thesis highlighted and proposed mitigation strategies for some of the challenges associated with this process. Each chapter delves into a distinct aspect of clinical presence and its ramifications on predictive modelling and algorithmic fairness. The following summarises this thesis' contributions, associated recommendations, limitations, and future axes of research.

7.1 Contributions

Predictive modelling often consists of learning from historical decisions to assist practitioners. In this context, the algorithmic fairness literature has identified potential biases in expert decisions and offered multiple mitigation strategies to tackle this issue. This thesis demonstrates that *not only do observed decisions present biases, but so too does the decision to observe*. Our opening question: "How should we handle clinical presence for fair predictive modelling?" has a nuanced answer as clinical presence encompasses both signal and biases. The following highlights some of these nuances by reviewing our contributions regarding some clinical presence challenges.

7.1.1 Chapter 3: Missingness

Challenge. Missingness in medical domains is rarely random but reflects multiple confounders such as the patient's health, medical expertise and socio-medical factors. Predictive pipelines often treat missingness as a preprocessing issue without considering its downstream impact on the algorithmic fairness of predictive models built on these data. We question this practice by identifying systematic patterns in clinical missingness and studying the impact of imputation on predictive performance and their associated algorithmic fairness.

Contributions. We provide historical evidence of group-specific missingness patterns and propose a theoretical and empirical analysis of the impact of different imputation strategies on algorithmic fairness. We show that more accurate imputation strategies, such as group-specific imputation, can increase downstream performance differences. Critically, we discuss how the choice of the imputation strategy that most reduces the gap in performance cannot be determined beforehand. In this chapter, we emphasise the importance of careful imputation, consideration of the missingness assumptions, their impact on algorithmic fairness, and their temporal evolution to avoid reinforcing inequities.

Implications. Algorithmic fairness is not only affected by what is present in the data but by what is absent from it. Predictive models developed and deployed on medical data with missingness require more careful use of imputation in their development and monitoring of the evolution of the missingness patterns in their deployment to ensure that they benefit all.

7.1.2 Chapter 4: Response Heterogeneity

Challenge. Exploring the heterogeneity of response to disease and treatment is central to guiding medical research and directing resources to improve patients' outcomes. Analyses to uncover subgroups often rely on RCTs, which are costly and limited in sample size and diversity. Further, methodological advances frequently focus on individual treatment effects and defer the problem of subgrouping as postprocessing. This chapter explores how to identify subgroups and jointly model observed survival outcomes under treatment non-randomisation.

Contributions. The chapter's contributions include formalising the subgroup identification problem as a latent clustering problem, proposing two neural network architectures for subgroup discovery in observational settings, and evaluating these approaches through simulations and a medical dataset.

Implications. By exploring heterogeneity in outcomes and treatment responses under observational settings, medical practitioners can design improved treatment guidelines and develop

clinical trials for groups under-served by current medical practice, ultimately advancing care for all.

7.1.3 Chapter 5: Competing Risks

Challenge. Clinical presence encompasses not only the process associated with covariates and interventions but also the observed outcomes. Critically, outcomes could preclude the event of interest. The observation of these competing risks is informative of the outcome of interest. Despite their prevalence in medical data, these competing risks are often overlooked, with many studies treating them as censored events, thereby introducing biases into risk estimation.

Contributions. This chapter examines the algorithmic fairness implications of ignoring competing risks, demonstrating that this common practice leads to systematic errors in survival estimates at individual, group, and population levels and exacerbates healthcare inequities, particularly affecting patients most at risk. We quantify the error associated with ignoring competing risks, empirically validate this result, and introduce a novel survival model, Neural Fine-Gray, to accurately account for competing risks. Our analysis emphasises the importance of competing risks in predictive modelling to ensure equitable healthcare outcomes for all patients.

Implications. Accounting for competing risks is crucial not only for improved predictive performance — as already known in the literature — but also to combat health disparities.

7.1.4 Chapter 6: Clinical Presence Shift

Challenge. A central challenge associated with leveraging clinical presence to inform predictive models is its propensity to change. When developing predictive models using observational data and advanced modelling strategies, practitioners risk embedding the associated observational process to model the outcome of interest. As these patterns may differ between places, across time and medical practitioners, it is critical to ensure predictive models are transportable under potential shifts.

Contributions. By differentiating observed from observable covariates in the distribution shift framework, we formalise the problem of clinical presence shift as a distribution shift. We propose DeepJoint, a multitask neural network that jointly models the longitudinal irregularities reflective of the observational process and the outcome of interest. Using a theoretical result from the adversarial ML literature, we provide an intuition as to why joint models may improve transportability. Through a natural experiment, we validate this result and show that the current handling of longitudinal irregularities may suffer under shifts occurring in clinical practice.

Implications. Ignoring the observational process may not be possible, and current approaches to do so underperform under shifts. Jointly modelling the observational process with the outcome of interest may enhance performance and transportability under the distributional shifts observed in the healthcare system.

7.2 Handling Clinical Presence

This thesis shows that one's handling of clinical presence can impact the performance and algorithmic fairness of ML pipelines. This section gathers recommendations on how to handle clinical presence at different stages of the ML pipeline: (i) collection, (ii) development and (iii) deployment.

Collection. While beyond the scope of this thesis, improved collection is central to improved modelling. As multiple confounders may influence clinical presence, controlling these factors through random or systematic sampling would help combat this bias. For instance, in psychology, it is common to sample participants' inputs — through a questionnaire — either regularly or at random times as described in [79, 81], where researchers use randomness to remove the impact of expectation on measurements. However, in the clinical setting, randomisation of measurements is unethical, and regular and thorough measurements are impracticable and expensive and may not even be necessary if confounders are carefully measured. We encourage the detailed recording of the reasons for collecting data, e.g. scheduled measurement versus medical initiative, medical practitioners' characteristics, and reasons for follow-up. This information could provide critical insights into the condition evolution and the observational process to better account for its associated biases.

Development. As previously discussed, we highlighted the critical importance of clinical presence in (i) observed covariates, (ii) counterfactual outcomes, and (iii) competing risks. While we proposed methodologies to address some of the biases associated with clinical presence, our main recommendation is to consider the observational process carefully. One must question why data is observed and when and if clinical presence is an informative signal in modelling the outcome of interest or a potential source of biases.

Deployment. A central risk of leveraging clinical presence is its potential evolution. Medical practice evolves, and patients interact differently with the healthcare system. ML practitioners must closely monitor the observational shift and ensure the proposed models can detect or, even better, be transportable under this shift. Critically, we want to reiterate the importance of monitoring these changes at patient and group levels, as population change may not reflect differences observed by patients under-served by current clinical practice. Further to naturally

occurring clinical presence shifts, one must monitor the observational process due to the potential impact of *predictive models on clinical presence*.

7.3 Limitations and future directions

This thesis has some limitations that raise exciting avenues for future work. This section highlights some key directions we would like to explore.

Balancing bias correction and analysis power. While this thesis underscores the necessity of addressing various observational biases, we acknowledge a limitation in our work: examining these biases in isolation. A key question emerges: how should we control for multiple biases simultaneously, and if so, without compromising the analytical power of our study? A nuanced comprehension of clinical practices and study design is essential to pinpoint the biases most likely present in the dataset.

Combining randomised and observational data. Our work focuses on improving observational data modelling. However, we will investigate how insights from randomised trials can enrich observational data modelling. Integrating informed priors on overlapping populations could improve observational data modelling while aligning with medical knowledge obtained under a more controlled setting.

Independence assumption. As commonly done in the literature, all proposed methodologies assume independence between patients and, consequently, their observational processes. However, real-world practices often deviate from this assumption. For instance, due to limited medical resources, practitioners in the same ICU unit prioritise patients based on severity. These phenomena potentially lead to correlated observational processes and outcomes. Developing clinical presence correction accounting for this potential correlation is an important avenue to improve the real-world relevance of predictive models.

Modalities. Our analyses deal with structured covariates in static and dynamic settings. Unstructured data, such as medical notes, may be marked by similar observational biases. For instance, patients with less severe conditions may present shorter notes, and the quality of these notes may reflect the practitioner's fatigue [136]. Exploring modality-specific biases and mitigating them is an avenue for future work.

Model interaction. Recognising the pivotal role of clinical presence in model development, we identify the critical need to explore how these models may influence this observational process. Kwong et al. [170] discuss the risk of induced belief revision in which assisting decisions

may alter medical behaviour. Understanding the interplay between model predictions and the intricate dynamics of clinical practice is imperative for refining our predictive models and, consequently, improving patient outcomes.

Beyond modelling. Our work highlights the importance of the observational process in modelling. However, a predictive pipeline is more than modelling alone. Many design choices [53], e.g., optimised target [240], influence the algorithmic fairness of a model and should be carefully considered to translate algorithmic fairness into fair medical practice [168].

Bibliography

[1] P. H. Aastha and Y. Liu. "DeepCompete: A deep learning approach to competing risks in continuous time domain". In: *AMIA Annual Symposium Proceedings*. Vol. 2020. 2020 (Cited on page 107).

[2] H. Abdel-Qadir et al. "Importance of considering competing risks in time-to-event analyses: application to stroke risk in a retrospective cohort study of elderly patients with atrial fibrillation". In: *Circulation: Cardiovascular Quality and Outcomes* 11.7 (2018) (Cited on page 134).

[3] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Vol. 55. US Government printing office, 1948 (Cited on page 24).

[4] D. Agniel, I. S. Kohane, and G. M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". In: *BMJ* 361 (2018) (Cited on pages 19, 50, 140).

[5] M. A. Ahmad, C. Eckert, and A. Teredesai. "The challenge of imputation in explainable artificial intelligence models". In: *arXiv preprint arXiv:1907.12669* (2019) (Cited on page 47).

[6] A. M. Alaa, S. Hu, and M. Schaar. "Learning from clinical judgments: Semi-markov-modulated marked hawkes processes for risk prognosis". In: *International Conference on Machine Learning*. 2017 (Cited on page 140).

[7] L. Antolini, P. Boracchi, and E. Biganzoli. "A time-dependent discrimination index for survival data". In: *Statistics in medicine* 24.24 (2005) (Cited on page 41).

[8] N. Arnould et al. "Breast cancer in men: are there similarities with breast cancer in women?" In: *Gynecologie, Obstetrique & Fertilite* 34.5 (2006) (Cited on page 49).

[9] S. Athey and G. Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27 (2016) (Cited on page 77).

[10] S. Athey and G. W. Imbens. "Machine learning methods for estimating heterogeneous causal effects". In: *stat* 1050.5 (2015) (Cited on page 77).

[11] P. C. Austin and J. P. Fine. "Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement". In: *Statistics in medicine* 36.8 (2017) (Cited on page 103).

[12] P. C. Austin and J. P. Fine. "Practical recommendations for reporting Fine-Gray model analyses for competing risk data". In: *Statistics in medicine* 36.27 (2017) (Cited on page 107).

[13] P. C. Austin, D. S. Lee, and J. P. Fine. "Introduction to the analysis of survival data in the presence of competing risks". In: *Circulation* 133.6 (2016) (Cited on pages 103, 107).

[14] P. C. Austin, E. W. Steyerberg, and H. Putter. "Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1". In: *Statistics in Medicine* 40.19 (2021) (Cited on pages 107, 116).

[15] P. C. Austin and E. A. Stuart. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". In: *Statistics in medicine* 34.28 (2015) (Cited on page 83).

[16] P. C. Austin et al. "Developing points-based risk-scoring systems in the presence of competing risks". In: *Statistics in medicine* 35.22 (2016) (Cited on page 134).

[17] P. Aylin. *Making sense of the evidence for the "weekend effect"*. 2015 (Cited on page 150).

[18] E. Bair, R. Tibshirani, and T. Golub. "Semi-supervised methods to predict patient survival from gene expression data". In: *PLoS biology* 2.4 (2004) (Cited on page 76).

[19] D. Barik and A. Thorat. "Issues of unequal access to public health in India". In: *Frontiers in public health* 3 (2015) (Cited on page 49).

[20] S. Barocas and A. D. Selbst. "Big data's disparate impact". In: *California law review* (2016) (Cited on page 47).

[21] S. Basu, A. Meghani, and A. Siddiqi. "Evaluating the health impact of large-scale public policy changes: classical and novel approaches". In: *Annual review of public health* 38 (2017) (Cited on page 18).

[22] G. E. Batista, M. C. Monard, et al. "A study of K-nearest neighbour as an imputation method." In: *His* 87.251-260 (2002) (Cited on page 46).

[23] A. G. Baydin et al. "Automatic differentiation in machine learning: a survey". In: *Journal of Machine Learning Research* 18 (2018) (Cited on page 31).

[24] I. M. Baytas et al. "Patient subtyping via time-aware lstm networks". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017 (Cited on page 140).

[25] B. K. Beaulieu-Jones et al. "Characterizing and managing missing structured data in electronic health records: data analysis". In: *JMIR medical informatics* 6.1 (2018) (Cited on page 46).

[26] A. Bellot and M. Schaar. "Tree-based bayesian mixture model for competing risks". In: *International Conference on Artificial Intelligence and Statistics*. 2018 (Cited on page 107).

[27] A. Bellot and M. van der Schaar. "Multitask boosting for survival analysis with competing risks". In: *Advances in Neural Information Processing Systems* 31 (2018) (Cited on page 108).

[28] D. A. Bennett. "How can I deal with missing data in my study?" In: *Australian and New Zealand journal of public health* 25.5 (2001) (Cited on page 46).

[29] O. Benveniste et al. "Long-term observational study of sporadic inclusion body myositis". In: *Brain* 134.11 (2011) (Cited on page 18).

[30] J. Bergstra and Y. Bengio. "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (2012) (Cited on page 32).

[31] S. D. Berry et al. "Competing risk of death: an important consideration in studies of older adults". In: *Journal of the American Geriatrics Society* 58.4 (2010) (Cited on pages 106, 134).

[32] D. Bertsimas, A. Orfanoudaki, and C. Pawlowski. "Imputation of clinical covariates in time series". In: *Machine Learning* 110.1 (2021) (Cited on page 46).

[33] J. Beyersmann et al. "Simulating competing risks data in survival analysis". In: *Statistics in medicine* 28.6 (2009) (Cited on page 119).

[34] I. Bica, J. Jordon, and M. van der Schaar. "Estimating the effects of continuous-valued interventions using generative adversarial networks". In: *Advances in Neural Information Processing Systems* 33 (2020) (Cited on page 74).

[35] I. Bica et al. "From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges". In: *Clinical Pharmacology & Therapeutics* 109.1 (2021) (Cited on pages 77, 82).

[36] J. Bishara et al. "Appropriateness of antibiotic therapy on weekends versus weekdays". In: *Journal of antimicrobial chemotherapy* 60.3 (2007) (Cited on page 19).

[37] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. Springer, 2006 (Cited on page 23).

[38] N. J. Bosomworth. "Practical use of the Framingham risk score in primary prevention: Canadian perspective". In: *Canadian Family Physician* 57.4 (2011) (Cited on page 132).

[39] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001) (Cited on page 98).

[40] N. E. Breslow. "Contribution to discussion of paper by DR Cox". In: *Journal of the Royal Statistical Society, Series B* 34 (1972) (Cited on page 36).

[41] J. F. Burke et al. "Three simple rules to ensure reasonably credible subgroup analyses". In: *BMJ* 351 (2015) (Cited on page 74).

[42] K. B. Burt et al. "Structural brain correlates of adolescent resilience". In: *Journal of Child Psychology and Psychiatry* 57.11 (2016) (Cited on page 52).

[43] X. Cai et al. "Medical concept embedding with time-aware attention". In: *arXiv preprint arXiv:1806.02873* (2018) (Cited on page 140).

[44] M. Cainzos-Achirica and M. J. Blaha. "Cardiovascular risk perception in women: true unawareness or risk miscalculation?" In: *BMC medicine* 13 (2015) (Cited on page 104).

[45] R. Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997) (Cited on page 146).

[46] R. Caruana and J. O'Sullivan. "Multitask pattern recognition for autonomous robots". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications (Cat. No. 98CH36190)*. Vol. 1. 1998 (Cited on page 146).

[47] S. Caton, S. Malisetty, and C. Haas. "Impact of Imputation Strategies on Fairness in Machine Learning". In: *Journal of Artificial Intelligence Research* 74 (2022) (Cited on page 47).

[48] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009) (Cited on page 141).

[49] P. Chapfuwa et al. "Survival cluster analysis". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020 (Cited on page 76).

[50] Z. Che et al. "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1 (2018) (Cited on pages 140, 151).

[51] I. Chen, F. D. Johansson, and D. Sontag. "Why is my classifier discriminatory?" In: *Advances in Neural Information Processing Systems* 31 (2018) (Cited on page 41).

[52] I. Y. Chen, P. Szolovits, and M. Ghassemi. "Can AI help reduce disparities in general medical and mental health care?" In: *AMA journal of ethics* 21.2 (2019) (Cited on page 41).

[53] I. Y. Chen et al. "Ethical Machine Learning in Healthcare". In: *Annual Review of Biomedical Data Science* 4 (2020) (Cited on pages 20, 164).

[54] R. J. Chen et al. "Algorithmic fairness in artificial intelligence for medicine and healthcare". In: *Nature biomedical engineering* 7.6 (2023) (Cited on page 44).

[55] V. Chernozhukov et al. *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*. Tech. rep. National Bureau of Economic Research, 2018 (Cited on page 77).

[56] P. Chilinski and R. Silva. "Neural likelihoods via cumulative distribution functions". In: *Conference on Uncertainty in Artificial Intelligence*. 2020 (Cited on page 28).

[57] K. Cho et al. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259* (2014) (Cited on pages 26, 27).

[58] E. Choi et al. "Doctor ai: Predicting clinical events via recurrent neural networks". In: *Machine learning for healthcare conference*. 2016 (Cited on page 140).

[59] E. Choi et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in Neural Information Processing Systems*. 2016 (Cited on page 140).

[60] A. Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2 (2017) (Cited on page 41).

[61] A. Chouldechova and A. Roth. "A snapshot of the frontiers of fairness in machine learning". In: *Communications of the ACM* 63.5 (2020) (Cited on pages 20, 47).

[62] A. Chouldechova et al. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In: *Conference on Fairness, Accountability and Transparency*. 2018 (Cited on page 41).

[63] J. A. Christen and F. J. Rubio. "Dynamic survival analysis: modelling the hazard function via ordinary differential equations". In: *arXiv preprint arXiv:2308.05205* (2023) (Cited on page 39).

[64] M. Chun et al. "Utility of single versus sequential measurements of risk factors for prediction of stroke in Chinese adults". In: *Scientific reports* 11.1 (2021) (Cited on page 52).

[65] F. Cismondi et al. "Missing data in medical databases: Impute, delete or classify?" In: *Artificial intelligence in medicine* 58.1 (2013) (Cited on page 20).

[66] M. Coemans et al. "Bias by censoring for competing events in survival analysis". In: *BMJ* 378 (2022) (Cited on page 106).

[67] E. C. R. Collaboration, S. W. Group, et al. "SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe". In: *European Heart Journal* 42.25 (2021) (Cited on page 133).

[68] D. Collett. *Modelling survival data in medical research*. CRC press, 2023 (Cited on page 33).

[69] F. S. Collins and H. Varmus. "A new initiative on precision medicine". In: *New England journal of medicine* 372.9 (2015) (Cited on page 76).

[70] R. Collobert and J. Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning". In: *Proceedings of the 25th international conference on Machine learning*. 2008 (Cited on page 138).

[71] D. I. Cook, V. J. Gebski, and A. C. Keech. "Subgroup analysis in clinical trials". In: *Medical Journal of Australia* 180.6 (2004) (Cited on page 77).

[72] C. Cox. "The generalized F distribution: an umbrella for parametric survival analysis". In: *Statistics in medicine* 27.21 (2008) (Cited on page 35).

[73] C. Cox et al. "Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution". In: *Statistics in medicine* 26.23 (2007) (Cited on page 35).

[74] D. R. Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972) (Cited on pages 36, 89).

[75] A. Curth, C. Lee, and M. van der Schaar. "Survite: Learning heterogeneous treatment effects from time-to-event data". In: *Advances in Neural Information Processing Systems* 34 (2021) (Cited on pages 74, 77, 78, 82).

[76] C. Curtis et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". In: *Nature* 486.7403 (2012) (Cited on page 214).

[77] H. Daniels and M. Velikova. "Monotone and partially monotone neural networks". In: *IEEE Transactions on Neural Networks* 21.6 (2010) (Cited on page 28).

[78] D. Danks and C. Yau. "Derivative-Based Neural Modelling of Cumulative Distribution Functions for Survival Analysis". In: *International Conference on Artificial Intelligence and Statistics*. 2022 (Cited on pages 39, 107, 116, 117, 121, 124, 128).

[79] P. A. Delespaul. *Assessing schizophrenia in daily life: The experience sampling method*. 1995 (Cited on page 162).

[80] K. M. Detre et al. "Observer agreement in evaluating coronary angiograms." In: *Circulation* 52.6 (1975) (Cited on page 19).

[81] M. W. DeVries. *The experience of psychopathology: Investigating mental disorders in their natural settings*. Cambridge University Press, 1992 (Cited on page 162).

[82] H. Do et al. "When More is Less: Incorporating Additional Datasets Can Hurt Performance By Introducing Spurious Correlations". In: *Machine Learning for Healthcare Conference*. 2023 (Cited on page 108).

[83] S. Du et al. "Gradient descent finds global minima of deep neural networks". In: *International conference on machine learning*. 2019 (Cited on page 32).

[84] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011) (Cited on page 30).

[85] C. Dugas et al. "Incorporating second-order functional knowledge for better option pricing". In: *Advances in neural information processing systems* 13 (2000) (Cited on page 24).

[86] C. Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012 (Cited on page 41).

[87] T. Fang et al. "Rethinking importance weighting for deep learning under distribution shift". In: *Advances in neural information processing systems* 33 (2020) (Cited on page 141).

[88] J. P. Fine and R. J. Gray. "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American statistical association* 94.446 (1999) (Cited on pages 107, 111, 114, 121).

[89] L. Fisher and P. Kanarek. "Presenting censored survival data when censoring and survival times may not be independent". In: *Reliability and Biometry* (1974) (Cited on pages 104, 113).

[90] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". In: *Fed. Probation* 80 (2016) (Cited on page 41).

[91] J. C. Foster, J. M. Taylor, and S. J. Ruberg. "Subgroup identification from randomized clinical trial data". In: *Statistics in medicine* 30.24 (2011) (Cited on pages 74, 77).

[92] H. P. Freeman and R. Payne. *Racial injustice in health care*. 2000 (Cited on page 44).

[93] C. Fricke et al. *Missing Fairness: The Discriminatory Effect of Missing Values in Datasets on Fairness in Machine Learning*. 2020 (Cited on page 47).

[94] J. Futoma et al. "The myth of generalisability in clinical research and machine learning in health care". In: *The Lancet Digital Health* 2.9 (2020) (Cited on page 158).

[95] J. J. Gagne et al. "Innovative research methods for studying treatments for rare diseases: methodological review". In: *BMJ* 349 (2014) (Cited on page 18).

[96] K. K. Ganju et al. "The role of decision support systems in attenuating racial biases in healthcare delivery". In: *Management science* 66.11 (2020) (Cited on page 47).

[97] M. Garcia et al. "Cardiovascular disease in women: clinical perspectives". In: *Circulation research* 118.8 (2016) (Cited on page 49).

[98] A. Gasparini et al. "Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study". In: *Statistica Neerlandica* 74.1 (2020) (Cited on pages 140, 143, 146).

[99] S. Gaynor and E. Bair. "Identification of relevant subtypes via preweighted sparse clustering". In: *Computational statistics & data analysis* 116 (2017) (Cited on page 76).

[100] T. A. Gerds and M. Schumacher. "Consistent estimation of the expected Brier score in general survival models with right-censored event times". In: *Biometrical Journal* 48.6 (2006) (Cited on page 41).

[101] T. A. Gerds et al. "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring". In: *Statistics in medicine* 32.13 (2013), pp. 2173–2184 (Cited on page 40).

[102] E. Getzen et al. "Mining for equitable health: Assessing the impact of missing data in electronic health records". In: *Journal of Biomedical Informatics* 139 (2023) (Cited on page 47).

[103] M. Ghassemi et al. "Practical guidance on artificial intelligence for health-care data". In: *The Lancet Digital Health* 1.4 (2019) (Cited on page 137).

[104] M. Ghassemi et al. "A review of challenges and opportunities in machine learning for health". In: *AMIA Summits on Translational Science Proceedings* 2020 (2020) (Cited on page 47).

[105] M. A. Gianfrancesco et al. "Potential biases in machine learning algorithms using electronic health record data". In: *JAMA internal medicine* 178.11 (2018) (Cited on pages 47, 49).

[106] S. H. Giordano. "Breast cancer in men". In: *New England Journal of Medicine* 378.24 (2018) (Cited on page 49).

[107] H. M. Gloster Jr and K. Neal. "Skin cancer in skin of color". In: *Journal of the American Academy of Dermatology* 55.5 (2006) (Cited on page 49).

[108] B. A. Goldstein et al. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* 24.1 (2017) (Cited on pages 19, 66).

[109] T. Gong et al. "A comparison of loss weighting strategies for multi task learning in deep neural networks". In: *IEEE Access* 7 (2019) (Cited on page 149).

[110] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016 (Cited on pages 23, 24).

[111] T. A. Gooley et al. "Estimation of failure probabilities in the presence of competing risks: new representations of old estimators". In: *Statistics in medicine* 18.6 (1999) (Cited on page 106).

[112] L. Gordon and R. A. Olshen. "Tree-structured survival analysis." In: *Cancer treatment reports* 69.10 (1985) (Cited on pages 39, 90).

[113] J. Gould and J. Ashton-Smith. "Missed diagnosis or misdiagnosis? Girls and women on the autism spectrum". In: *Good Autism Practice (GAP)* 12.1 (2011) (Cited on page 49).

[114] E. Graf et al. "Assessment and comparison of prognostic classification schemes for survival data". In: *Statistics in medicine* 18.17-18 (1999) (Cited on page 40).

[115] J. C. Greenland, C. H. Williams-Gray, and R. A. Barker. "The clinical heterogeneity of Parkinson's disease and its therapeutic implications". In: *European Journal of Neuroscience* 49.3 (2019) (Cited on page 73).

[116] R. H. Groenwold. "Informative missingness in electronic health record systems: the curse of knowing". In: *Diagnostic and prognostic research* 4.1 (2020) (Cited on pages 19, 61).

[117] L. Guelman, M. Guillén, and A. M. Pérez-Marín. "Uplift random forests". In: *Cybernetics and Systems* 46.3-4 (2015) (Cited on page 77).

[118] L. L. Guo et al. "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine". In: *Scientific reports* 12.1 (2022) (Cited on pages 137, 138, 142).

[119] X. Han, M. Goldstein, and R. Ranganath. "Survival mixture density networks". In: *Machine Learning for Healthcare Conference*. 2022 (Cited on page 39).

[120] M. Hardt, E. Price, and N. Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016) (Cited on pages 41, 47).

[121] E. Hariton and J. J. Locascio. "Randomised controlled trials–the gold standard for effectiveness research". In: *BJOG: an international journal of obstetrics and gynaecology* 125.13 (2018) (Cited on page 18).

[122] N. Hassanpour and R. Greiner. "CounterFactual Regression with Importance Sampling Weights." In: *IJCAI*. 2019 (Cited on pages 82, 83).

[123] J. S. Haukoos and C. D. Newgard. "Advanced statistics: missing data in clinical research–part 1: an introduction and conceptual framework". In: *Academic Emergency Medicine* 14.7 (2007) (Cited on pages 45, 46, 61).

[124] D. Hendrycks and K. Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016) (Cited on page 141).

[125] S. Heo et al. "Sex differences in heart failure symptoms and factors associated with heart failure symptoms". In: *Journal of Cardiovascular Nursing* 34.4 (2019) (Cited on pages 50, 51).

[126] M. A. Hernán and J. M. Robins. "Using big data to emulate a target trial when a randomized trial is not available". In: *American journal of epidemiology* 183.8 (2016) (Cited on pages 18, 74).

[127] M. J. Higgins, J. Baselga, et al. "Targeted therapies for breast cancer". In: *The Journal of clinical investigation* 121.10 (2011) (Cited on page 73).

[128] G. Hinton, O. Vinyals, and J. Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015) (Cited on page 102).

[129] F. Hobbs et al. "Barriers to cardiovascular disease risk scoring and primary prevention in Europe". In: *QJM: An International Journal of Medicine* 103.10 (2010) (Cited on page 76).

[130] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997) (Cited on pages 26, 27, 147).

[131] A. E. Hoerl and R. W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970) (Cited on page 33).

[132] C. Hoffman and J. Paradise. "Health insurance and access to health care in the United States". In: *Annals of the New York Academy of Sciences* 1136.1 (2008) (Cited on page 49).

[133] T. R. Holford. "The analysis of rates and of survivorship using log-linear models". In: *Biometrics* (1980) (Cited on page 35).

[134] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989) (Cited on page 25).

[135] V. J. Howard et al. "Disparities in stroke incidence contributing to disparities in stroke mortality". In: *Annals of neurology* 69.4 (2011) (Cited on pages 45, 52).

[136] C.-C. Hsu, Z. Obermeyer, and C. Tan. "Clinical Notes Reveal Physician Fatigue". In: *arXiv preprint arXiv:2312.03077* (2023) (Cited on page 163).

[137] L. Hu, J. Ji, and F. Li. "Estimating heterogeneous survival treatment effect in observational data using machine learning". In: *Statistics in medicine* 40.21 (2021) (Cited on pages 77, 78).

[138] S. Hu and G. H. Chen. "Distributionally robust survival analysis: A novel fairness loss without demographics". In: *Machine Learning for Health*. 2022 (Cited on page 108).

[139] H. Hung and C.-T. Chiang. "Estimation methods for time-dependent AUC models with survival data". In: *Canadian Journal of Statistics* 38.1 (2010) (Cited on page 91).

[140] H. Ishwaran et al. *Random survival forests*. 2008 (Cited on page 39).

[141] K. T. Islam et al. "Marked point process for severity of illness assessment". In: *Machine Learning for Healthcare Conference*. 2017 (Cited on page 140).

[142] D. Jackson et al. "Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation". In: *Statistics in medicine* 33.27 (2014) (Cited on page 135).

[143] K. J. Janssen et al. "Missing covariate data in medical research: to impute is better than to ignore". In: *Journal of clinical epidemiology* 63.7 (2010) (Cited on page 61).

[144] V. Jeanselme, N. Agarwal, and C. Wang. "Review of Language Models for Survival Analysis". In: *AAAI 2024 Spring Symposium on Clinical Foundation Models*. 2024 (Cited on page 135).

[145] V. Jeanselme, B. Tom, and J. Barrett. "Neural Survival Clustering: Non-parametric mixture of neural networks for survival clustering". In: *Conference on Health, Inference, and Learning*. 2022 (Cited on page 73).

[146] V. Jeanselme et al. "Sex differences in post cardiac arrest discharge locations". In: *Resuscitation plus* 8 (2021) (Cited on page 44).

[147] V. Jeanselme et al. "DeepJoint: Robust Survival Modelling Under Clinical Presence Shift". In: *arXiv preprint arXiv:2205.13481* (2022) (Cited on page 137).

[148] V. Jeanselme et al. "Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness". In: *Machine Learning for Health*. 2022 (Cited on page 43).

[149] V. Jeanselme et al. "Neural Fine-Gray: Monotonic neural networks for competing risks". In: *Proceedings of the Conference on Health, Inference, and Learning*. Vol. 209. Proceedings of Machine Learning Research. PMLR, 2023 (Cited on page 103).

[150] V. Jeanselme et al. "Identifying treatment response subgroups in observational time-to-event data". In: *arXiv preprint arXiv:2408.03463* (2024) (Cited on page 73).

[151] H. Jeong, H. Wang, and F. P. Calmon. "Fairness without imputation: A decision tree approach for fair prediction with missing values". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022 (Cited on pages 47, 52–54).

[152] B. Jia et al. "Inferring latent heterogeneity using many feature variables supervised by survival outcome". In: *Statistics in medicine* 40.13 (2021) (Cited on page 78).

[153] F. Johansson, U. Shalit, and D. Sontag. "Learning representations for counterfactual inference". In: *International conference on machine learning*. 2016 (Cited on pages 77, 82).

[154] A. E. Johnson, T. J. Pollard, and T. Naumann. "Generalizability of predictive models for intensive care unit patients". In: *arXiv preprint arXiv:1812.02275* (2018) (Cited on page 137).

[155] A. E. Johnson et al. "MIMIC-III , a freely accessible critical care database". In: *Scientific data* 3.1 (2016) (Cited on pages 66, 150).

[156] D. Kahneman, O. Sibony, and C. R. Sunstein. "Bias Is a Big Problem. But So Is' Noise.'." In: *International New York Times* (2021) (Cited on page 19).

[157] F. Kamiran and T. Calders. "Classification with no discrimination by preferential sampling". In: *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. 6. 2010 (Cited on page 47).

[158] T. Kamishima, S. Akaho, and J. Sakuma. "Fairness-aware learning through regularization approach". In: *2011 IEEE 11th International Conference on Data Mining Workshops*. 2011 (Cited on page 47).

[159] W. B. Kannel and D. L. McGee. "Diabetes and cardiovascular disease: the Framingham study". In: *Jama* 241.19 (1979) (Cited on pages 128, 214).

[160] E. L. Kaplan and P. Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958) (Cited on page 36).

[161] U. Kartoun et al. "Prediction performance and fairness heterogeneity in cardiovascular risk models". In: *Scientific Reports* 12.1 (2022) (Cited on page 20).

[162] C. Kartsonaki. "Survival analysis". In: *Diagnostic Histopathology* 22.7 (2016) (Cited on page 35).

[163] J. L. Katzman et al. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC medical research methodology* 18.1 (2018) (Cited on pages 39, 89, 145, 148, 149).

[164] B. J. van Kempen et al. "Performance of Framingham cardiovascular disease (CVD) predictions in the Rotterdam Study taking into account competing risks and disentangling CVD into coronary heart disease (CHD) and stroke". In: *International journal of cardiology* 171.3 (2014) (Cited on pages 132, 133).

[165] L. K. Kim et al. "Sex-based disparities in incidence, treatment, and outcomes of cardiac arrest in the United States, 2003–2012". In: *Journal of the American Heart Association* 5.6 (2016) (Cited on page 44).

[166] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015* (2015) (Cited on pages 30, 91, 121, 152).

[167] M. T. Koller et al. "Competing risks and the clinical community: irrelevance or ignorance?" In: *Statistics in medicine* 31.11-12 (2012) (Cited on page 103).

[168] N. Kordzadeh and M. Ghasemaghaei. "Algorithmic bias: review, synthesis, and future research directions". In: *European Journal of Information Systems* 31.3 (2022) (Cited on page 164).

[169] M. J. Kusner et al. "Counterfactual fairness". In: *Advances in neural information processing systems* 30 (2017) (Cited on page 41).

[170] J. C. Kwong et al. "When the Model Trains You: Induced Belief Revision and Its Implications on Artificial Intelligence Research and Patient Care–A Case Study on Predicting Obstructive Hydronephrosis in Children". In: *NEJM AI* 1.2 (2024) (Cited on page 163).

[171] N. Laird and D. Olivier. "Covariance analysis of censored survival data using log-linear analysis techniques". In: *Journal of the American Statistical Association* 76.374 (1981) (Cited on page 35).

[172] C. S. Lam et al. "Sex differences in heart failure". In: *European heart journal* 40.47 (2019) (Cited on page 51).

[173] B. Lang. "Monotonic multi-layer perceptron networks as universal approximators". In: *International conference on artificial neural networks*. 2005 (Cited on page 28).

[174] A. J. Larrazabal et al. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". In: *Proceedings of the National Academy of Sciences* 117.23 (2020) (Cited on page 64).

[175] I. Lawrence and K. Lin. "A concordance correlation coefficient to evaluate reproducibility". In: *Biometrics* (1989) (Cited on page 125).

[176] D. Lazer et al. "The parable of Google Flu: traps in big data analysis". In: *science* 343.6176 (2014) (Cited on page 137).

[177] G. Lazzari et al. "Current Trends and Challenges in Real-World Breast Cancer Adjuvant Radiotherapy: A Practical Review.: New trends in adjuvant radiotherapy in BC". In: *Archives of Breast Cancer* 10.1 (2023) (Cited on page 99).

[178] C. Lee, J. Yoon, and M. Van Der Schaar. "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data". In: *IEEE Transactions on Biomedical Engineering* 67.1 (2019) (Cited on page 135).

[179] C. Lee et al. "Deephit: A deep learning approach to survival analysis with competing risks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018 (Cited on pages 37, 90, 107, 118–121, 128).

[180] C. Lee et al. "Temporal quilting for survival analysis". In: *The 22nd international conference on artificial intelligence and statistics*. 2019 (Cited on page 39).

[181] K. Lee et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks". In: *Advances in neural information processing systems* 31 (2018) (Cited on page 141).

[182] K.-M. Leung, R. M. Elashoff, and A. A. Afifi. "Censoring issues in survival analysis". In: *Annual review of public health* 18.1 (1997) (Cited on pages 104, 113).

[183] H. Li et al. "Deep convolutional neural networks for imaging data based survival analysis of rectal cancer". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019 (Cited on page 39).

[184] J. Li et al. "Imputation of missing values for electronic health record laboratory data". In: *NPJ digital medicine* 4.1 (2021) (Cited on page 45).

[185] J. Lin et al. "Racial differences in family health history knowledge of type 2 diabetes: exploring the role of interpersonal mechanisms". In: *Translational Behavioral Medicine* 8.4 (2018) (Cited on page 44).

[186] Y.-K. Lin, M. Lin, and H. Chen. "Do electronic health records affect quality of care? Evidence from the HITECH Act". In: *Information Systems Research* 30.1 (2019) (Cited on page 49).

[187] I. Lipkovich and A. Dmitrienko. "Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES". In: *Journal of biopharmaceutical statistics* 24.1 (2014) (Cited on page 78).

[188] I. Lipkovich et al. "Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations". In: *Statistics in medicine* 30.21 (2011) (Cited on page 78).

[189] Z. Lipton, Y.-X. Wang, and A. Smola. "Detecting and correcting for label shift with black box predictors". In: *International conference on machine learning*. 2018 (Cited on page 141).

[190] Z. C. Lipton, D. Kale, and R. Wetzel. "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series". In: *Machine Learning for Healthcare Conference*. 2016 (Cited on pages 19, 61, 140, 151, 155, 225).

[191] R. J. Little et al. "The prevention and treatment of missing data in clinical trials". In: *New England Journal of Medicine* 367.14 (2012) (Cited on page 43).

[192] R. J. Little and D. B. Rubin. "The analysis of social science data with missing values". In: *Sociological Methods & Research* 18.2-3 (1989) (Cited on page 46).

[193] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019 (Cited on page 46).

[194] S. Liu, E. Johns, and A. J. Davison. "End-to-end multi-task learning with attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019 (Cited on pages 149, 224).

[195] X. Liu et al. "Certified monotonic neural networks". In: *Advances in Neural Information Processing Systems* 33 (2020) (Cited on page 28).

[196] S. Liverani et al. "Clustering method for censored and collinear survival data". In: *Computational Statistics* 36.1 (2021) (Cited on page 76).

[197] D. M. Lloyd-Jones et al. "Framingham risk score and prediction of lifetime risk for coronary heart disease". In: *The American journal of cardiology* 94.1 (2004) (Cited on pages 17, 133, 134).

[198] W.-Y. Loh, X. He, and M. Man. "A regression tree approach to identifying subgroups with differential treatment effects". In: *Statistics in medicine* 34.11 (2015) (Cited on page 78).

[199] J. Lu et al. "Learning under concept drift: A review". In: *IEEE Transactions on Knowledge and Data Engineering* 31.12 (2018) (Cited on page 141).

[200] L. Manduchi et al. "A Deep Variational Approach to Clustering Survival Data". In: *International Conference on Learning Representations*. 2022 (Cited on page 76).

[201] C. M. Mangione et al. "Statin use for the primary prevention of cardiovascular disease in adults: US preventive services task force recommendation statement". In: *JAMA* 328.8 (2022) (Cited on pages 104, 132).

[202] C. Mao et al. "Multitask learning strengthens adversarial robustness". In: *European Conference on Computer Vision*. 2020 (Cited on pages 146, 147).

[203] F. Martínez-Plumed et al. "Fairness and missing values". In: *arXiv preprint arXiv:1905.12728* (2019) (Cited on page 47).

[204] F. Mauvais-Jarvis et al. "Sex and gender: modifiers of health, disease, and medicine". In: *The Lancet* 396.10250 (2020) (Cited on page 49).

[205] E. McFowland III, S. Somanchi, and D. B. Neill. "Efficient Discovery of Heterogeneous Quantile Treatment Effects in Randomized Experiments via Anomalous Pattern Detection". In: *arXiv preprint arXiv:1803.09159* (2018) (Cited on page 78).

[206] G. J. McLachlan and D. C. McGiffin. "On the role of finite mixture models in survival analysis". In: *Statistical methods in medical research* 3.3 (1994) (Cited on page 35).

[207] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007 (Cited on pages 76, 78).

[208] D. Mendes et al. "The benefit of HER2 -targeted therapies on overall survival of patients with metastatic HER2 -positive breast cancer–a systematic review". In: *Breast Cancer Research* 17 (2015) (Cited on page 73).

[209] R. B. Merkatz et al. "Women in Clinical Trials of New Drugs–A Change in Food and Drug Administration Policy". In: *New England Journal of Medicine* 329.4 (1993) (Cited on page 18).

[210] S. Mitchell et al. "Algorithmic fairness: Choices, assumptions, and definitions". In: *Annual Review of Statistics and Its Application* 8 (2021) (Cited on pages 41, 47).

[211] R. Mitra et al. "Learning from data with structured missingness". In: *Nature Machine Intelligence* 5.1 (2023) (Cited on pages 47, 48).

[212] K. Mohan and J. Pearl. "Graphical models for processing missing data". In: *Journal of the American Statistical Association* 116.534 (2021) (Cited on pages 50, 142).

[213] S. Mohseni et al. "Self-supervised learning for generalizable out-of-distribution detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020 (Cited on page 141).

[214] K. Monterrubio-Gómez, N. Constantine-Cooke, and C. A. Vallejos. "A review on statistical and machine learning competing risks methods". In: *Biometrical Journal* 66.2 (2024) (Cited on page 105).

[215] J. G. Moreno-Torres et al. "A unifying view on dataset shift in classification". In: *Pattern recognition* 45.1 (2012) (Cited on page 140).

[216] R. Moskovitch et al. "Outcomes prediction via time intervals related patterns". In: *2015 IEEE international conference on data mining*. 2015 (Cited on page 140).

[217] S. C. Mouli et al. "Deep lifetime clustering". In: *arXiv preprint arXiv:1910.00547* (2019) (Cited on page 76).

[218] M. C. Mozer, D. Kazakov, and R. V. Lindsey. "Discrete event, continuous time RNNs". In: *arXiv preprint arXiv:1710.04110* (2017) (Cited on page 140).

[219] R. Nabi and I. Shpitser. "Fair inference on outcomes". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018 (Cited on page 41).

[220] C. Nagpal, V. Jeanselme, and A. Dubrawski. "Deep Parametric Time-to-Event Regression with Time-Varying Covariates". In: *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*. Vol. 146. Proceedings of Machine Learning Research. PMLR, 2021 (Cited on pages 39, 66, 135).

[221] C. Nagpal, X. Li, and A. Dubrawski. "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks". In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021) (Cited on pages 38, 90, 91, 101, 107, 121, 122, 214).

[222] C. Nagpal, W. Potosnak, and A. Dubrawski. "auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data". In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. Vol. 182. Proceedings of Machine Learning Research. PMLR, 2022 (Cited on page 220).

[223] C. Nagpal, V. Sanil, and A. Dubrawski. "Recovering Sparse and Interpretable Subgroups with Heterogeneous Treatment Effects with Censored Time-to-Event Outcomes". In: *Proceedings of Machine Learning Research vol TBD* 1 (2023) (Cited on page 78).

[224] C. Nagpal et al. "Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020 (Cited on page 78).

[225] C. Nagpal et al. "Deep Cox mixtures for survival regression". In: *Machine Learning for Healthcare Conference*. 2021 (Cited on pages 76, 90, 101).

[226] C. Nagpal et al. "Counterfactual Phenotyping with Censored Time-to-Events". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. 2022 (Cited on pages 88, 90, 118, 120).

[227] N. G. Nair, P. Satpathy, J. Christopher, et al. "Covariate shift: A review and analysis on classifiers". In: *2019 Global Conference for Advancement in Technology (GCAT)*. 2019 (Cited on page 141).

[228] V. Nair and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010 (Cited on page 24).

[229] J. A. Nelder and R. W. Wedderburn. "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972) (Cited on page 25).

[230] B. Nestor et al. "Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks". In: *Machine Learning for Healthcare Conference*. 2019 (Cited on pages 137, 142, 154).

[231] C. D. Newgard and J. S. Haukoos. "Advanced statistics: missing data in clinical research–part 2: multiple imputation". In: *Academic Emergency Medicine* 14.7 (2007) (Cited on page 61).

[232] C. D. Newgard and R. J. Lewis. "Missing data: how to best account for what is not known". In: *Jama* 314.9 (2015) (Cited on page 46).

[233] B. Nicholson et al. "Determinants and extent of weight recording in UK primary care: an analysis of 5 million adults' electronic health records from 2000 to 2017". In: *BMC medicine* 17.1 (2019) (Cited on page 50).

[234] T. G. Nick and K. M. Campbell. "Logistic Regression". In: *Topics in Biostatistics*. Totowa, NJ: Humana Press, 2007 (Cited on page 66).

[235] M. S. Nielsen et al. "Predictors of weight loss after bariatric surgery–a cross-disciplinary approach combining physiological, social, and psychological measures". In: *International Journal of Obesity* 44.11 (2020) (Cited on page 52).

[236] S. Nijman et al. "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review". In: *Journal of Clinical Epidemiology* 142 (2022) (Cited on pages 43, 52).

[237] A. Noriega-Campero et al. "Active fairness in algorithmic decision making". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019 (Cited on page 41).

[238] K. Norris and A. R. Nissenson. "Race, gender, and socioeconomic disparities in CKD in the United States". In: *Journal of the American Society of Nephrology* 19.7 (2008) (Cited on page 44).

[239] C. Nwankpa et al. "Activation functions: Comparison of trends in practice and research for deep learning". In: *arXiv preprint arXiv:1811.03378* (2018) (Cited on page 24).

[240] Z. Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019) (Cited on page 164).

[241] T. Omi, K. Aihara, et al. "Fully neural network based model for general temporal point processes". In: *Advances in Neural Information Processing Systems*. 2019 (Cited on page 28).

[242] C. Park et al. "Reliable and trustworthy machine learning for health using dataset shift detection". In: *Advances in Neural Information Processing Systems* 34 (2021) (Cited on page 141).

[243] A. Paszke et al. "Automatic differentiation in pytorch". In: (2017) (Cited on page 31).

[244] L. A. Pauls et al. "The weekend effect in hospitalized patients: a meta-analysis." In: *Journal of hospital medicine* 12.9 (2017) (Cited on pages 150, 152).

[245] A. B. Pedersen et al. "Missing data and multiple imputation in clinical epidemiological research". In: *Clinical epidemiology* 9 (2017) (Cited on page 20).

[246] C. M. Perou et al. "Molecular portraits of human breast tumours". In: *Nature* 406.6797 (2000) (Cited on page 73).

[247] F. Pfisterer et al. "Evaluating domain generalization for survival analysis in clinical studies". In: *Conference on Health, Inference, and Learning*. 2022 (Cited on page 142).

[248] S. Pfohl et al. "Creating fair models of atherosclerotic cardiovascular disease risk". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019 (Cited on page 41).

[249] D. Pham and D. Karaboga. *Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks*. Springer Science & Business Media, 2012 (Cited on page 29).

[250] T. Pham et al. "Deepcare: A deep dynamic memory model for predictive medicine". In: *Pacific-Asia conference on knowledge discovery and data mining*. 2016 (Cited on page 140).

[251] M. Phelan, N. A. Bhavsar, and B. A. Goldstein. "Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference". In: *eGEMs* 5.1 (2017) (Cited on page 20).

[252] R. Pivovarov et al. "Identifying and mitigating biases in EHR laboratory tests". In: *Journal of biomedical informatics* 51 (2014) (Cited on page 140).

[253] T. J. Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific data* 5.1 (2018) (Cited on page 141).

[254] S. Pölsterl. "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn". In: *The Journal of Machine Learning Research* 21.1 (2020) (Cited on page 220).

[255] K. Polyak et al. "Heterogeneity in breast cancer". In: *The Journal of clinical investigation* 121.10 (2011) (Cited on page 73).

[256] R. L. Prentice et al. "The analysis of failure times in the presence of competing risks". In: *Biometrics* (1978) (Cited on pages 107, 110, 121).

[257] W. H. Press et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007 (Cited on page 117).

[258] E. M. Pullenayegum and L. S. Lim. "Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design". In: *Statistical methods in medical research* 25.6 (2016) (Cited on page 143).

[259] W. Qi et al. "Explaining heterogeneity of individual treatment causal effects by subgroup discovery: an observational case study in antibiotics treatment of acute rhino-sinusitis". In: *Artificial Intelligence in Medicine* 116 (2021) (Cited on page 77).

[260] J. Quiñonero-Candela et al. *Dataset shift in machine learning*. Mit Press, 2008 (Cited on page 140).

[261] M. M. Rahman and S. Purushotham. "Fair and interpretable models for survival analysis". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022 (Cited on page 108).

[262] M. M. Rahman et al. "DeepPseudo: pseudo value based deep learning models for competing risk analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021 (Cited on page 107).

[263] A. Rajkomar et al. "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12 (2018) (Cited on pages 41, 47, 63).

[264] A. Rajkomar et al. "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1.1 (2018) (Cited on page 137).

[265] W. M. Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971) (Cited on page 91).

[266] D. Rindt et al. "Survival regression with proper scoring rules and monotonic neural networks". In: *International Conference on Artificial Intelligence and Statistics*. 2022 (Cited on pages 28, 39, 85, 89, 108, 116).

[267] E. Röösli, S. Bozkurt, and T. Hernandez-Boussard. "Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model". In: *Scientific Data* 9.1 (2022) (Cited on page 64).

[268] P. M. Rothwell. "Subgroup analysis in randomised controlled trials: importance, indications, and interpretation". In: *The Lancet* 365.9454 (2005) (Cited on page 77).

[269] P. Royston and M. K. Parmar. "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome". In: *BMC medical research methodology* 13.1 (2013) (Cited on page 100).

[270] S. J. Ruberg, L. Chen, and Y. Wang. "The mean does not mean as much anymore: finding sub-groups for tailored therapeutics". In: *Clinical trials* 7.5 (2010) (Cited on page 77).

[271] D. B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976) (Cited on page 143).

[272] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004 (Cited on page 46).

[273] D. B. Rubin. *Multiple imputation*. Chapman and Hall/CRC, 2018 (Cited on page 61).

[274] S. Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017) (Cited on page 146).

[275] A. Rusanov et al. "Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research". In: *BMC medical informatics and decision making* 14.1 (2014) (Cited on page 50).

[276] M. Saar-Tsechansky and F. Provost. "Handling missing values when applying classification models". In: *Journal of Machine Learning Research* (2007) (Cited on page 61).

[277] S. Sagawa et al. "Distributionally robust neural networks". In: *International Conference on Learning Representations*. 2019 (Cited on page 146).

[278] P. Sanchez et al. "Causal machine learning for healthcare and precision medicine". In: *Royal Society Open Science* 9.8 (2022) (Cited on page 77).

[279] J. Satagopan et al. "A note on competing risks in survival data analysis". In: *British journal of cancer* 91.7 (2004) (Cited on pages 104, 106, 113).

[280] M. Schmid and M. Berger. "Competing risks analysis for discrete time-to-event data". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.5 (2021) (Cited on page 108).

[281] R. Schoop et al. "Quantifying the predictive accuracy of time-to-event models in the presence of competing risks". In: *Biometrical Journal* 53.1 (2011), pp. 88–112 (Cited on page 122).

[282] J. Schrouff et al. "Diagnosing failures of fairness transfer across distribution shift in real-world medical settings". In: *Advances in Neural Information Processing Systems* 35 (2022) (Cited on page 156).

[283] N. A. Schuster et al. "Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis". In: *Journal of clinical epidemiology* 122 (2020) (Cited on pages 104, 106, 113, 133, 134).

[284] J. E. Seifarth, C. L. McGowan, and K. J. Milne. "Sex and life expectancy". In: *Gender medicine* 9.6 (2012) (Cited on pages 104, 131).

[285] S. Selvin. *Survival analysis for epidemiologic and medical research*. Cambridge University Press, 2008 (Cited on page 33).

[286] C. W. Seymour et al. "Precision medicine for all? Challenges and opportunities for a precision medicine approach to critical illness". In: *Critical Care* 21 (2017) (Cited on page 17).

[287] L. Seyyed-Kalantari et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. 2020 (Cited on page 41).

[288] U. Shalit, F. D. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. 2017 (Cited on pages 77, 82, 83).

[289] A. Sharafoddini et al. "A new insight into missing data in intensive care unit patient profiles: observational study". In: *JMIR medical informatics* 7.1 (2019) (Cited on pages 43, 50).

[290] O. Shchur, M. Biloš, and S. Günnemann. "Intensity-Free Learning of Temporal Point Processes". In: *International Conference on Learning Representations*. 2020 (Cited on pages 85, 116).

[291] C. Shi, D. Blei, and V. Veitch. "Adapting neural networks for the estimation of treatment effects". In: *Advances in neural information processing systems* 32 (2019) (Cited on page 82).

[292] H. Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of statistical planning and inference* 90.2 (2000) (Cited on page 82).

[293] J. D. Singer and J. B. Willett. "It's about time: Using discrete-time survival analysis to study duration and the timing of events". In: *Journal of educational statistics* 18.2 (1993) (Cited on pages 35, 37).

[294] H. Singh, V. Mhasawade, and R. Chunara. "Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database". In: *PLOS Digital Health* 1.4 (2022) (Cited on page 137).

[295] H. Singh et al. "Fair predictors under distribution shift". In: *NeurIPS Workshop on Fair ML for Health*. 2019 (Cited on page 156).

[296] R. Sisk et al. "Informative presence and observation in routine health data: A review of methodology for clinical risk prediction". In: *Journal of the American Medical Informatics Association* (2020) (Cited on pages 50, 139).

[297] D. J. Slamon et al. "Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2". In: *New England journal of medicine* 344.11 (2001) (Cited on page 73).

[298] M. van Smeden, R. H. Groenwold, and K. G. Moons. "A cautionary note on the use of the missing indicator method for handling missing data in prediction research". In: *Journal of clinical epidemiology* 125 (2020) (Cited on pages 137, 138, 155).

[299] K. T. Smith et al. "Access is necessary but not sufficient: factors influencing delay and avoidance of health care services". In: *MDM Policy & Practice* 3.1 (2018) (Cited on page 49).

[300] R. T. Sousa, L. A. Pereira, and A. S. Soares. "Improving Irregularly Sampled Time Series Learning with Dense Descriptors of Time". In: *arXiv preprint arXiv:2003.09291* (2020) (Cited on page 140).

[301] D. A. Southern et al. "Kaplan–Meier methods yielded misleading results in competing risk scenarios". In: *Journal of clinical epidemiology* 59.10 (2006) (Cited on page 106).

[302] D. Spathis and S. L. Hyland. "Looking for Out-of-Distribution Environments in Multi-center Critical Care Data". In: *arXiv preprint arXiv:2205.13398* (2022) (Cited on pages 138, 141, 153).

[303] M. Sperrin, E. Petherick, and E. Badrick. "Informative observation in health data: association of past level and trend with time to next measurement". In: *Stud Health Technol Inform* 235 (2017) (Cited on pages 140, 146).

[304] M. Sperrin et al. "Missing data should be handled differently for prediction than for description or causal explanation". In: *Journal of Clinical Epidemiology* 125 (2020) (Cited on page 61).

[305] N. Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014) (Cited on page 33).

[306] D. M. Stablein, W. H. Carter Jr, and J. W. Novak. "Analysis of survival data with nonproportional hazard functions". In: *Controlled clinical trials* 2.2 (1981) (Cited on pages 37, 78).

[307] M. J. Stensrud and M. A. Hernán. "Why test for proportional hazards?" In: *Jama* 323.14 (2020) (Cited on page 107).

[308] J. A. Sterne et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls". In: *BMJ* 338 (2009) (Cited on page 20).

[309] A. Strömberg and J. Mårtensson. "Gender differences in patients with heart failure". In: *European Journal of Cardiovascular Nursing* 2.1 (2003) (Cited on pages 20, 50).

[310] L. Su et al. "A sensitivity analysis approach for informative dropout using shared parameter models". In: *Biometrics* 75.3 (2019) (Cited on pages 140, 146).

[311] X. Su et al. "Subgroup analysis via recursive partitioning." In: *Journal of Machine Learning Research* 10.2 (2009) (Cited on page 77).

[312] A. Subbaswamy and S. Saria. "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics* 21.2 (2020) (Cited on pages 137, 156).

[313] T. Szandała. "Review and comparison of commonly used activation functions for deep neural networks". In: *Bio-inspired neurocomputing* (2021) (Cited on page 24).

[314] A. Szczepura. "Access to health care for ethnic minority populations". In: *Postgraduate medical journal* 81.953 (2005) (Cited on page 49).

[315] W. Tang et al. "SODEN: A Scalable Continuous-Time Survival Model through Ordinary Differential Equation Networks". In: *Journal of Machine Learning Research* 23.34 (2022) (Cited on page 39).

[316] Y. W. Teh et al. "Dirichlet Process." In: *Encyclopedia of machine learning* 1063 (2010) (Cited on page 102).

[317] L. Thabane et al. "A tutorial on sensitivity analyses in clinical trials: the what, why, when and how". In: *BMC medical research methodology* 13.1 (2013) (Cited on page 47).

[318] T. M. Therneau et al. *The Cox model*. Springer, 2000 (Cited on page 128).

[319] R. L. Thorndike. "Who belongs in the family?" In: *Psychometrika* 18.4 (1953) (Cited on page 92).

[320] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996) (Cited on page 33).

[321] D. Tjandra, Y. He, and J. Wiens. "A hierarchical approach to multi-event survival analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021 (Cited on page 107).

[322] E. J. Tsiklidis, T. Sinno, and S. L. Diamond. "Predicting risk for trauma patients using static and dynamic information from the MIMIC III database". In: *Plos one* 17.1 (2022) (Cited on page 66).

[323] T. Tuomi et al. "The many faces of diabetes: a disease with increasing heterogeneity". In: *The Lancet* 383.9922 (2014) (Cited on page 73).

[324] B. Twala, M. Jones, and D. J. Hand. "Good methods for coping with missing data in decision trees". In: *Pattern Recognition Letters* 29.7 (2008) (Cited on page 140).

[325] S. Van Der Pas, R. Nelissen, and M. Fiocco. "Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events: illustrated with 138,234 hip (124,560 patients) and 139,070 knee (125,213 patients) replacements from the Dutch Arthroplasty Register". In: *Acta Orthopaedica* 89.2 (2018) (Cited on page 111).

[326] N. Van Geloven et al. "Validation of prediction models in the presence of competing risks: a guide through modern methods". In: *BMJ* 377 (2022) (Cited on page 122).

[327] M. Van Ryn. "Research on the provider contribution to race/ethnicity disparities in medical care". In: *Medical care* (2002) (Cited on page 20).

[328] T. Vanderschueren et al. "Accounting For Informative Sampling When Learning to Forecast Treatment Outcomes Over Time". In: (2023) (Cited on page 143).

[329] S. Verma and J. Rubin. "Fairness definitions explained". In: *2018 ieee/acm international workshop on software fairness (fairware)*. 2018 (Cited on page 41).

[330] B. Vogel et al. "The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030". In: *The Lancet* 397.10292 (2021) (Cited on pages 20, 49).

[331] S. Wager and S. Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018) (Cited on page 78).

[332] A. Wald. "A method of estimating plane vulnerability based on damage of survivors". In: *Statistical Research Group, Columbia University. CRC* 432 (1943) (Cited on page 159).

[333] H. K. Wall et al. "Vital signs: prevalence of key cardiovascular disease risk factors for Million Hearts 2022–United States, 2011–2016". In: *Morbidity and Mortality Weekly Report* 67.35 (2018) (Cited on page 132).

[334] C. van Walraven and F. A. McAlister. "Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals". In: *Journal of clinical epidemiology* 69 (2016) (Cited on page 106).

[335] M. Wang et al. "Statistical methods for studying disease subtype heterogeneity". In: *Statistics in medicine* 35.5 (2016) (Cited on page 73).

[336] P. Wang, Y. Li, and C. K. Reddy. "Machine learning for survival analysis: A survey". In: *ACM Computing Surveys (CSUR)* 51.6 (2019) (Cited on page 105).

[337] Q. Wang et al. "A comprehensive survey of loss functions in machine learning". In: *Annals of Data Science* (2020) (Cited on page 29).

[338] S. Wang et al. "Mimic-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020 (Cited on pages 66, 150, 207).

[339] T. Wang and C. Rudin. "Causal rule sets for identifying subgroups with enhanced treatment effects". In: *INFORMS Journal on Computing* 34.3 (2022) (Cited on page 78).

[340] Y. Wang and L. Singh. "Analyzing the impact of missing values and selection bias on fairness". In: *International Journal of Data Science and Analytics* 12.2 (2021) (Cited on page 47).

[341] P. B. Weerakody et al. "A review of irregular time series data handling with gated recurrent neural networks". In: *Neurocomputing* (2021) (Cited on page 140).

[342] N. G. Weiskopf, A. Rusanov, and C. Weng. "Sick patients have more data: the non-random completeness of electronic health records". In: *AMIA Annual Symposium Proceedings*. Vol. 2013. 2013 (Cited on page 50).

[343] N. G. Weiskopf et al. "Defining and measuring completeness of electronic health records for secondary use". In: *Journal of biomedical informatics* 46.5 (2013) (Cited on page 43).

[344] N. G. Weiskopf and C. Weng. "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research". In: *Journal of the American Medical Informatics Association* 20.1 (2013) (Cited on page 43).

[345] J. S. Weissman et al. *Delayed access to health care: risk factors, reasons, and consequences*. 1991 (Cited on page 49).

[346] B. J. Wells et al. "Strategies for handling missing data in electronic health record derived data". In: *Egems* 1.3 (2013) (Cited on page 50).

[347] P. J. Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10 (1990) (Cited on page 31).

[348] I. R. White and J. B. Carlin. "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values". In: *Statistics in medicine* 29.28 (2010), pp. 2920–2931 (Cited on page 46).

[349] I. R. White, P. Royston, and A. M. Wood. "Multiple imputation using chained equations: issues and guidance for practice". In: *Statistics in medicine* 30.4 (2011) (Cited on pages 46, 61).

[350] P. W. Wilson et al. "Prediction of coronary heart disease using risk factor categories". In: *Circulation* 97.18 (1998) (Cited on page 132).

[351] M. Wolbers et al. "Prognostic models with competing risks: methods and application to coronary risk prediction". In: *Epidemiology* (2009) (Cited on page 106).

[352] M. Wolbers et al. "Competing risks analyses: objectives and approaches". In: *European heart journal* 35.42 (2014) (Cited on page 106).

[353] M. Wolbers et al. "Concordance for prognostic models with competing risks". In: *Biostatistics* 15.3 (2014), pp. 526–539 (Cited on page 122).

[354] M. Wolkewitz et al. "Interpreting and comparing risks in the presence of competing events". In: *BMJ* 349 (2014) (Cited on page 106).

[355] A. M. Wood, I. R. White, and S. G. Thompson. "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals". In: *Clinical trials* 1.4 (2004) (Cited on pages 47, 61).

[356] M. Woodward. "Cardiovascular disease and the female disadvantage". In: *International journal of environmental research and public health* 16.7 (2019) (Cited on page 104).

[357] E. Xia et al. "Outcome-Driven Clustering of Acute Coronary Syndrome Patients Using Multi-Task Neural Network with Attention." In: *MedInfo*. 2019 (Cited on page 76).

[358] Y. Xia and J. Wang. "A general projection neural network for solving monotone variational inequalities and related optimization problems". In: *IEEE Transactions on Neural Networks* 15.2 (2004) (Cited on page 28).

[359] C. Xiao, E. Choi, and J. Sun. "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* 25.10 (2018) (Cited on page 138).

[360] J. Xu et al. "Association of sex with clinical outcome in critically ill sepsis patients: a retrospective analysis of the large clinical database MIMIC-III". In: *Shock (Augusta, Ga.)* 52.2 (2019) (Cited on page 66).

[361] Y. Xu et al. "Treatment heterogeneity with survival outcomes". In: *Handbook of Matching and Weighting Adjustments for Causal Inference*. Chapman and Hall/CRC, 2023 (Cited on page 77).

[362] M. Yalaza, A. İnan, and M. Bozer. "Male breast cancer". In: *The journal of breast health* 12.1 (2016) (Cited on page 49).

[363] H. Yao et al. "Wild-time: A benchmark of in-the-wild distribution shift over time". In: *Advances in Neural Information Processing Systems* 35 (2022) (Cited on page 137).

[364] J. Yarnold. "Early and locally advanced breast cancer: diagnosis and treatment National Institute for Health and Clinical Excellence guideline 2009". In: *Clinical Oncology* 21.3 (2009) (Cited on page 74).

[365] J. Yoon, J. Jordon, and M. Van Der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets". In: *International conference on learning representations*. 2018 (Cited on page 17).

[366] J. H. Yoon et al. "Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit". In: *Critical Care* 24.1 (2020) (Cited on page 17).

[367] J. Q. Young et al. "'July effect': impact of the academic year-end changeover on patient outcomes: a systematic review". In: *Annals of internal medicine* 155.5 (2011) (Cited on page 158).

[368] M. J. Zaki et al. "Parallel algorithms for discovery of association rules". In: *Data mining and knowledge discovery* 1.4 (1997) (Cited on page 150).

[369] J. Zeidan et al. "Global prevalence of autism: A systematic review update". In: *Autism Research* 15.5 (2022) (Cited on page 49).

[370] H. Zhang et al. "Hurtful words: quantifying biases in clinical contextual word embeddings". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020 (Cited on page 41).

[371] H. Zhang et al. "Improving the Fairness of Chest X-ray Classifiers". In: *Proceedings of the Conference on Health, Inference, and Learning*. Vol. 174. Proceedings of Machine Learning Research. PMLR, 2022 (Cited on page 64).

[372] K. Zhang et al. "Domain adaptation under target and conditional shift". In: *International Conference on Machine Learning*. 2013 (Cited on pages 140, 141).

[373] W. Zhang et al. "Mining heterogeneous causal effects for personalized cancer treatment". In: *Bioinformatics* 33.15 (2017) (Cited on page 78).

[374] W. Zhang and J. C. Weiss. "Longitudinal fairness with censorship". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 2022 (Cited on page 108).

[375] Y. Zhang, A. Bellot, and M. Schaar. "Learning overlapping representations for the estimation of individualized treatment effects". In: *International Conference on Artificial Intelligence and Statistics*. 2020 (Cited on page 77).

[376] Y. Zhang and Q. Long. "Fairness-aware Missing Data Imputation". In: *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. 2022 (Cited on page 47).

[377] Y. Zhang. "ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling." In: *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019), pp. 4369-4375, Macao, China.* 2019 (Cited on page 140).

[378] Z. Zhao and T. L. J. Ng. "Fairness-Aware Processing Techniques in Survival Analysis: Promoting Equitable Predictions". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2023 (Cited on page 108).

[379] H. Zhou, S. Balakrishnan, and Z. Lipton. "Domain adaptation under missingness shift". In: *International Conference on Artificial Intelligence and Statistics*. 2023 (Cited on page 144).

[380] K. Zhou et al. "Domain generalization: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022) (Cited on page 141).

[381] X.-H. Zhou, G. J. Eckert, and W. M. Tierney. "Multiple imputation in public health research". In: *Statistics in medicine* 20.9-10 (2001) (Cited on page 61).

[382] J. Zhu and B. Gallego. "Targeted estimation of heterogeneous treatment effect in observational survival analysis". In: *Journal of Biomedical Informatics* 107 (2020) (Cited on page 78).

Appendix A

Supplemental material Chapter 3: Missingness

A.1 Proofs

In this section, we demonstrate the theoretical results introduced in Section 3.4.

Notations. Consider an imputation strategy $\mathcal{I}$ that replaces missing data with a constant value $c_g^{\mathcal{I}}$ for the group g. Our work analyses $L_g^{\mathcal{I}}$, the group-specific reconstruction error for imputation $\mathcal{I}$, and $\Delta_g^{\mathcal{I}}$, the gap in reconstruction error between group g and the rest of the population.

All capital letters stand for random variables: G for group membership, O for the observation process and X for the covariates. Throughout the proofs, we characterise the missingness process with $\rho_g = \text{Corr}_{P_g}(O, X)$, the unobserved correlation between the observation indicator and the ground truth covariate values, and $\alpha_g = \mathbb{E}[O \mid G = g]$, the observation rate in the group g. The covariate X is described with μ for its mean, σ for its variance. Exponent O expresses an observed quantity. Subscript g indicates subgroup characteristics.

Finally, the population is defined by the ratio of members in group g in comparison to the rest of the population, i.e., $r_g = \mathbb{E}[G = g]$.

A.1.1 Theorem 3.1

Theorem. Assuming i.i.d. data points $\{x_i\}$, one can express the reconstruction error in group g resulting from group mean imputation as:

$$L_g^{group} = \left(\overbrace{-\frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} \cdot \rho_g \cdot \sigma_{X|G=g}}^{B_g^{group}} \right)^2 + \sigma^2_{X|O=0,G=g} \tag{A.1}$$

Variance of unobserved data

Standard deviation

Missingness process

where the missingness process is represented through (i) $\rho_g = \text{Corr}(O, X \mid G = g)$, the unobserved correlation between the observation indicator and the ground truth covariate values and (ii) $\alpha_g = \mathbb{E}[O \mid G = g]$, the observation rate in group g, which is observable. Other values impacting the reconstruction error are reflective of the underlying covariate distribution. This includes $\sigma^2_{X|G=g} = \text{Var}(X \mid G = g)$, the ground truth variance of the covariate in the group g; and $\sigma^2_{X|O=0,G=g}$, the variance of the unobserved values of this same group.

Under the same assumptions, one can compute the reconstruction error in a group g using population mean imputation as a function of the previously-defined quantity B_g^{group}:

$$L_g^{pop} = \left(B_g^{group} + \mu_g^O - \mu^O \right)^2 + \sigma^2_{X|O=0,G=g} \tag{A.2}$$

Group imputation bias

Difference between group- and population- observed means

Proof. First, we express the reconstruction error of a constant imputation strategy considering each covariate independently. For clarity, we denote by $\neg O$ if $O = 0$, $\neg G$ if $G \neq g$.

$$\begin{aligned}
L_g^{\mathcal{I}} :=& \mathbb{E}\left[||c_g^{\mathcal{I}} - X||_2^2 \,\middle|\, \neg O, G \right] && \textit{(Using Definition 3.1)} \\
=& \mathbb{E}\left[(c_g^{\mathcal{I}} - X)^2 \,\middle|\, \neg O, G \right] \\
=& \mathbb{E}\left[X^2 \,\middle|\, \neg O, G \right] + c^{\mathcal{I}^2} - 2c_g^{\mathcal{I}} \mathbb{E}\left[X \,\middle|\, \neg O, G \right] \\
=& \mathbb{E}\left[X \,\middle|\, \neg O, G \right]^2 + \sigma^2_{X|\neg O,G} + c^{\mathcal{I}^2} - 2c_g^{\mathcal{I}} \mathbb{E}\left[X \,\middle|\, \neg O, G \right] && \textit{(By definition of variance)} \\
=& \left(\mathbb{E}\left[X \mid \neg O, G \right] - c_g^{\mathcal{I}} \right)^2 + \sigma^2_{X|\neg O,G} && \text{(A.3)}
\end{aligned}$$

Remark A.1. This first expression demonstrates that the reconstruction error of *any* constant imputation is lower bounded by the variance of the unobserved data: these constant strategies do not capture any of this variance.

Using this decomposition, we further explore how group mean imputation impacts the reconstruction error. As a reminder, group mean imputation replaces missing values in group g with the observed group mean, i.e., $c_g^{group} = \mathbb{E}[X \mid O, G]$. The square root of the first term in Equation (A.3), that we refer as B_g^{group}, therefore simplifies to:

$$\begin{aligned}
B_g^{group} :&= \mathbb{E}[X \mid \neg O, G] - \mathbb{E}[X \mid O, G] \\
&= \frac{\mathbb{E}[(1-O)X \mid G]}{\mathbb{E}[(1-O) \mid G]} - \frac{\mathbb{E}[OX \mid G]}{\mathbb{E}[O \mid G]} \\
&= \frac{\mathbb{E}[X \mid G] - \mathbb{E}[OX \mid G]}{1 - \mathbb{E}[O \mid G]} - \frac{\mathbb{E}[OX \mid G]}{\mathbb{E}[O \mid G]} \\
&= \frac{-\mathsf{Corr}(O, X \mid G) \cdot \sigma_{O|G} \cdot \sigma_{X|G}}{(1 - \mathbb{E}[O \mid G])\mathbb{E}[O \mid G]} && \textit{(By definition of covariance)} \\
&= -\rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \cdot \sigma_{X|G}
\end{aligned}$$

Similarly, in the context of population mean imputation, which replaces the missing values with the observed mean, i.e., $c^{pop} = \mathbb{E}[X \mid O]$, results in the simplification of the first term as:

$$\begin{aligned}
B_g^{pop} :=& \mathbb{E}[X \mid \neg O, G] - \mathbb{E}[X \mid O] \\
=& \mathbb{E}[X \mid \neg O, G] - \mathbb{E}[X \mid O, G] + \mathbb{E}[X \mid O, G] - \mathbb{E}[X \mid O] \\
=& B_g^{group} + \mu_g^O - \mu^O
\end{aligned}$$

One can further decompose this equality to disentangle distributions' properties from missingness processes:

$$\begin{aligned}
B_g^{pop} =& B_g^{group} + \mathbb{E}[X \mid O, G] - \mathbb{E}[G \mid O]\mathbb{E}[X \mid O, G] - \mathbb{E}[\neg G \mid O]\mathbb{E}[X \mid O, \neg G] \\
=& B_g^{group} + (1 - \frac{\alpha_g r_g}{\alpha})\mu_g^O - \frac{\alpha_{\neg g}(1-r_g)}{\alpha}\mu_{\neg g}^O && \textit{(By Bayes' theorem)} \\
=& B_g^{group} + \frac{\alpha_{\neg g}(1-r_g)}{\alpha}[\mu_g^O - \mu_g + \mu_g - \mu_{\neg g} + \mu_{\neg g} - \mu_{\neg g}^O] \\
=& B_g^{group} + \frac{\alpha_{\neg g}(1-r_g)}{\alpha}\left[\rho_g \cdot \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot \sigma_{X|G} + \mu_g - \mu_{\neg g} - \rho_{\neg g} \cdot \sqrt{\frac{1-\alpha_{\neg g}}{\alpha_{\neg g}}} \cdot \sigma_{X|\neg G}\right]
\end{aligned} \tag{A.4}$$

with $r_g = \frac{|P_g|}{|P|}$, the proportion of patients member of group g, and α, the overall observation rate, i.e. $\alpha = \alpha_g r_g + \alpha_{\neg g}(1-r_g)$. This last expression is obtained by a decomposition of $\mu_g^O - \mu_g$ similar to the one proposed for the computation of B_g^{group}. □

A.1.2 Theorem 3.2

Theorem. The group reconstruction error resulting from group mean imputation is larger than the one resulting from population mean imputation, i.e. $L_g^{group} > L_g^{pop}$, iff one of the following conditions holds:

$$\rho_g \cdot \frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} < \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} < 0 \text{ or } 0 < \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} < \rho_g \cdot \frac{1}{\sqrt{\alpha_g(1-\alpha_g)}} \tag{A.5}$$

Missingness process

Distribution characteristics

Proof. This inequality can be rewritten as:

$$L_g^{group} > L_g^{pop} \Leftrightarrow |B_g^{group}| > \left| B_g^{group} + \mu_g^O - \mu^O \right|$$

To demonstrate this inequality, we explore the four different cases.

Case 1: $B_g^{group} > 0$ and $B_g^{group} + \mu_g^O - \mu^O > 0$

$$\begin{cases} B_g^{group} > 0 \\ B_g^{group} + \mu_g^O - \mu^O > 0 \\ B_g^{group} > B_g^{group} + \mu_g^O - \mu^O \end{cases} \Leftrightarrow \begin{cases} \rho_g < 0 \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} < \frac{\mu_g^O - \mu^O}{\sigma_{X|G}} \\ \mu^O > \mu_g^O \end{cases}$$

Case 2: $B_g^{group} > 0$ and $B_g^{group} + \mu_g^O - \mu^O < 0$

$$\begin{cases} B_g^{group} > 0 \\ B_g^{group} + \mu_g^O - \mu^O < 0 \\ B_g^{group} > -B_g^{group} - \mu_g^O + \mu^O \end{cases} \Leftrightarrow \begin{cases} \rho_g < 0 \\ \mu^O > \mu_g^O \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \in [\frac{\mu_g^O - \mu^O}{\sigma_{X|G}}, \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}}] \end{cases}$$

Case 3: $B_g^{group} < 0$ and $B_g^{group} + \mu_g^O - \mu^O > 0$

$$\begin{cases} B_g^{group} < 0 \\ B_g^{group} + \mu_g^O - \mu^O > 0 \\ -B_g^{group} > B_g^{group} + \mu_g^O - \mu^O \end{cases} \Leftrightarrow \begin{cases} \rho_g > 0 \\ \mu^O < \mu_g^O \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \in [\frac{\mu_g^O - \mu^O}{2\sigma_{X|G}}, \frac{\mu_g^O - \mu^O}{\sigma_{X|G}}] \end{cases}$$

Case 4: $B_g^{group} < 0$ and $B_g^{group} + \mu_g^O - \mu^O < 0$

$$\begin{cases} B_g^{group} < 0 \\ B_g^{group} + \mu_g^O - \mu^O < 0 \\ -B_g^{group} > -B_g^{group} - \mu_g^O + \mu^O \end{cases} \Leftrightarrow \begin{cases} \rho_g > 0 \\ \mu^O < \mu_g^O \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} > \frac{\mu_g^O - \mu^O}{\sigma_{X|G}} \end{cases}$$

Combining cases 1 and 2, and cases 3 and 4 lead to:

$$\begin{cases} \rho_g < 0 \\ \mu^O > \mu_g^O \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} < \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} \end{cases} \cup \begin{cases} \rho_g > 0 \\ \mu^O < \mu_g^O \\ \rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} > \frac{\mu_g^O - \mu^O}{2\sigma_{X|G}} \end{cases}$$

□

A.1.3 Theorem 3.3

Theorem. Under the simplifying assumptions $\sigma^2_{X|\neg O,G} = \sigma^2_{X|\neg O,\neg G}$, and $\mu_g^O > \mu^O$, both imputation strategies penalise the marginalised group and the reconstruction gap is larger for the group imputation than the population one (i.e., $\Delta_g^{group} > \Delta_g^{pop} > 0$) iff:

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) > \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot d(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot d(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

or

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) < \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot d(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot d(\alpha_{\neg g}, 1 - r_g, \alpha_g) < ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

with $r_g = \mathbb{E}[G = g]$, the ratio of the population belonging to group g, $f(\alpha_g, r_g, \alpha_{\neg g}) = \frac{2\alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1-r_g) - \alpha_g r_g)$, $e(\alpha_g) = \sqrt{\frac{\alpha_g}{1-\alpha_g}}$, and $d(\alpha_g, r_g, \alpha_{\neg g}) = \frac{\alpha_g r_g + \alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1-r_g) - \alpha_g r_g)$.

Proof. Let consider these following two expressions $\Delta_g^{group} > \Delta_g^{pop}$ and $\Delta_g^{pop} > 0$ separately.

First, consider the expression $\Delta_g^{group} > \Delta_g^{pop}$, this can further be expressed as follows:

$$\begin{aligned}\Delta_g^{group} > \Delta_g^{pop} \Leftrightarrow & B_g^{group^2} - B_{\neg g}^{group^2} > B_g^{pop^2} - B_{\neg g}^{pop^2} \\ \Leftrightarrow & \left(B_{\neg g}^{pop} - B_{\neg g}^{group}\right)\left(B_{\neg g}^{pop} + B_{\neg g}^{group}\right) > \left(B_g^{pop} - B_g^{group}\right)\left(B_g^{pop} + B_g^{group}\right) \\ \Leftrightarrow & \left(\mu_{\neg g}^{O} - \mu^{O}\right)\left(B_{\neg g}^{pop} + B_{\neg g}^{group}\right) > \left(\mu_g^{O} - \mu^{O}\right)\left(B_g^{pop} + B_g^{group}\right)\end{aligned}$$

Using Equation (A.4), this corresponds to:

$$\alpha_g r_g \gamma\left(2\rho_{\neg g} \cdot \sqrt{\frac{1}{\alpha_{\neg g}(1-\alpha_{\neg g})}} \cdot \sigma_{X|\neg G} + \frac{\alpha_g r_g}{\alpha} \cdot \gamma\right) > \alpha_{\neg g}(1-r_g)\gamma\left(-2\rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \cdot \sigma_{X|G} + \frac{\alpha_{\neg g}(1-r_g)}{\alpha} \cdot \gamma\right)$$

Assuming $\mu_g^O > \mu^O$ results in $\gamma > 0$:

$$\begin{aligned}&\alpha_g r_g 2\rho_{\neg g} \cdot \sqrt{\frac{1}{\alpha_{\neg g}(1-\alpha_{\neg g})}} \cdot \sigma_{X|\neg G} + \alpha_{\neg g}(1-r_g)2\rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \cdot \sigma_{X|G} \\ &> \frac{-(\alpha_g r_g)^2 + (\alpha_{\neg g}(1-r_g))^2}{\alpha_g r_g + \alpha_{\neg g}(1-r_g)}\gamma \\ \Leftrightarrow & \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1-r_g \quad, \alpha_g) > ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g})\end{aligned}$$

with $\gamma = \rho_g \cdot \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot \sigma_{X|G} + \mu_g - \mu_{\neg g} - \rho_{\neg g} \cdot \sqrt{\frac{1-\alpha_{\neg g}}{\alpha_{\neg g}}} \cdot \sigma_{X|\neg G}$,
and $f(\alpha_g, r_g, \alpha_{\neg g}) = \frac{2\alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1-r_g) - \alpha_g r_g)$, $f : [0,1]^3 \longrightarrow \mathbb{R}^+$.

Second, consider $\Delta_g^{pop} > 0$, we focus on the case in which both components are positive. The equivalence is obtained by considering the complementary case in which both components are negative.

$$\begin{cases} B_g^{group} - B_{\neg g}^{group} + \mu_g^O - \mu_{\neg g}^O > 0 \\ B_g^{group} + B_{\neg g}^{group} + \mu_g^O + \mu_{\neg g}^O - 2\mu^O > 0 \end{cases} \quad \textit{(Assuming } \sigma^2_{X|\neg O,G} = \sigma^2_{X|\neg O,\neg G}\textit{)}$$

$$\Leftrightarrow \begin{cases} -\rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \cdot \sigma_{X|G} + \rho_{\neg g} \cdot \sqrt{\frac{1}{\alpha_{\neg g}(1-\alpha_{\neg g})}} \cdot \sigma_{X|\neg G} > -\gamma \\ -\rho_g \cdot \sqrt{\frac{1}{\alpha_g(1-\alpha_g)}} \cdot \sigma_{X|G} - \rho_{\neg g} \cdot \sqrt{\frac{1}{\alpha_{\neg g}(1-\alpha_{\neg g})}} \cdot \sigma_{X|\neg G} > -\frac{\alpha_{\neg g}(1-r_g) - \alpha_g r_g}{\alpha_g r_g + \alpha_{\neg g}(1-r_g)}\gamma \end{cases}$$

$$\Leftrightarrow \begin{cases} \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) < \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot h(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot h(\alpha_{\neg g}, 1-r_g, \alpha_g) < ((1-r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

with $e(\alpha_g) = \sqrt{\frac{\alpha_g}{1-\alpha_g}}$ and $h(\alpha_g, r_g, \alpha_{\neg g}) = \frac{\alpha_g r_g + \alpha_{\neg g}(1-r_g)}{\sqrt{\alpha_g(1-\alpha_g)}} - \sqrt{\frac{1-\alpha_g}{\alpha_g}} \cdot (\alpha_{\neg g}(1-r_g) - \alpha_g r_g)$.

Therefore $\Delta_g^{group} > \Delta_g^{pop} > 0$ is equivalent to satisfy the following set of equations:

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) > \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot h(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot h(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

or

$$\begin{cases} \rho_g \cdot \sigma_{X|G} \cdot f(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot f(\alpha_{\neg g}, 1 - r_g, \alpha_g) > ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \\ \rho_g \cdot \sigma_{X|G} \cdot e(\alpha_g) - \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot e(\alpha_{\neg g}) < \mu_g - \mu_{\neg g} \\ \rho_g \cdot \sigma_{X|G} \cdot h(\alpha_g, r_g, \alpha_{\neg g}) + \rho_{\neg g} \cdot \sigma_{X|\neg G} \cdot h(\alpha_{\neg g}, 1 - r_g, \alpha_g) < ((1 - r_g)\alpha_{\neg g} - r_g\alpha_g)(\mu_g - \mu_{\neg g}) \end{cases}$$

□

Example. Consider a dataset with the following observed characteristics: observed means $\mu_g^O = 0.5$ and $\mu_{\neg g}^O = 0$, the marginalised group ratio $r_g = 25\%$, and the observation rates: $\alpha_g = 0.7$ and $\alpha_{\neg g} = 0.8$. Further, we assume the underlying data characteristic $\sigma_{X|G} = \sigma_{X|\neg G} = 0.5$, and $\sigma_{X|\neg O,G} = \sigma_{X|\neg O,\neg G}$. Figure A.1 illustrates the theoretical fairness gap difference and the area satisfying the previous theorems under varying missingness characteristics ρ_g and $\rho_{\neg g}$.

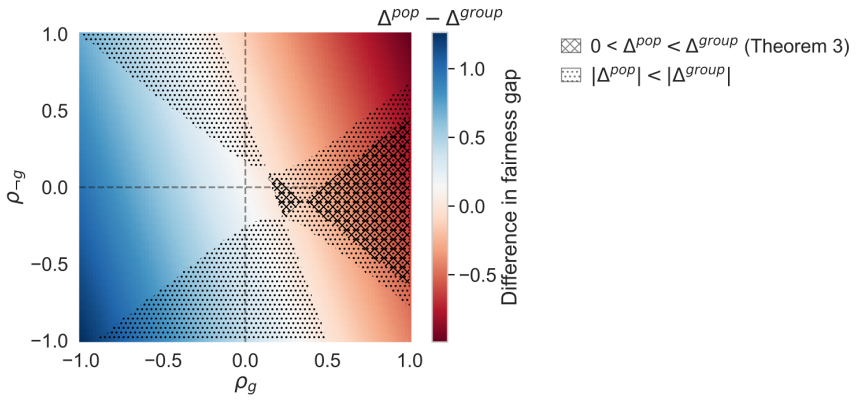


Figure A.1: Difference in fairness gap between population imputation and group imputation reconstruction errors. In red, the fairness gap is larger for the group imputation strategy than the population one. In blue, the opposite is true. The crossed area describes settings satisfying Theorem 3.3, i.e. when both strategies result in larger reconstruction errors for group g but population mean imputation reduces the fairness gap compared to its group imputation alternative. The dotted area presents the generalisation of the Theorem when population mean imputation reduces the absolute fairness gap.

This example provides evidence that for a set of observed characteristics, the problem of the optimal imputation strategy from a reconstruction error point of view is under-determined.

Specifically, two missingness processes could lead to the same observed data characteristics but impact which imputation to choose.

A.2 Experiments

This section provides additional details on the experimental design.

A.2.1 Simulation study

Data Generation. The proposed synthetic population consists of 100,000 points for the majority group and 1,000 for the marginalised group resulting in a sample size of $N = 101,000$ with a ratio of 100:1. Each individual is represented in this dataset as a pair of covariates, i.e., $X \in \mathbb{R}^2$. For each group, 2/3 presents the condition, i.e., $\mathbb{P}(Y_i = 1) = 0.66$. Negatives are drawn from the normal distribution $\mathcal{N}((0, 0), 0.25)$. The condition characterisation, i.e., the boundary between positive and negatives, differs between groups with positive from the majority (resp. the marginalised group) sampled from $\mathcal{N}((1, 0), 0.25)$ (resp. $\mathcal{N}((0, 1), 0.25)$). Figure A.2 shows the density distribution of the generated population.

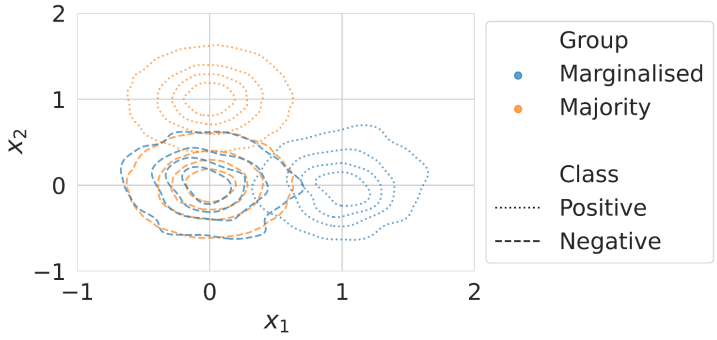


Figure A.2: Density distributions of the generated population.

Missingness. In this synthetic population, 50% of the dimension X_2 is removed in a given subgroup to enforce the three clinical presence scenarios. We enforce the following clinical missingness:

- Limited access to quality care (S1): $O_2 \mid [G = 1] \sim \text{Bernoulli}(0.5)$
- (Mis)-informed collection (S2): $O_2 \mid [X_1 > 0.5] \sim \text{Bernoulli}(0.5)$
- Confirmation bias (S3): $O_2 \mid [X_2 > 0.5] \sim \text{Bernoulli}(0.5)$

with O_2, the observation indicator associated with X_2 and G, the group membership ($G = 1$ indicates a member of the marginalised group).

Modelling. We generate 100 datasets and enforce the different missingness patterns before running a logistic regression with an l2 penalty ($\lambda = 1$). Results are computed on the 20% test set and averaged over the 100 iterations with 95% confidence bounds reported.

Tabular results. Table A.1 presents the reconstruction error differentiated by groups, scenarios and imputation strategies averaged over 100 simulations presented in Section 3.5.3. Similarly, Table A.2 presents the AUC averaged over the different imputation strategies as presented in Section 3.5.4.

		Imputation strategy	
Scenario	Group	Population Mean	Group Mean
(S1)	Overall	0.493 (0.016)	0.062 (0.004)
	Majority	0.000 (0.000)	0.000 (0.000)
	Marginalised	0.493 (0.016)	0.062 (0.004)
	$\Delta^{\mathcal{I}}_{Marginalised}$	0.493 (0.016)	**0.062** (0.004)
(S2)	Overall	0.332 (0.007)	0.236 (0.006)
	Majority	0.286 (0.007)	0.286 (0.007)
	Marginalised	0.490 (0.019)	0.062 (0.005)
	$\Delta^{\mathcal{I}}_{Marginalised}$	**0.204** (0.021)	-0.224 (0.009)
(S3)	Overall	0.333 (0.002)	0.325 (0.002)
	Majority	0.333 (0.002)	0.325 (0.002)
	Marginalised	0.020 (0.010)	0.370 (0.035)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.313 (0.010)	**0.045** (0.035)

Table A.1: Group-specific reconstruction errors and error gap Δ mean (std) across scenarios on 100 synthetic experiments. If $\Delta < 0$, the marginalised group has a smaller reconstruction error than the majority. Lower reconstruction error is better.

Missingness Indicators. In Section 3.5.4, we only considered the addition of the missingness indicators to the Group MICE. In this Appendix, we present how this technique impacts each imputation strategy performance. Table A.3 presents the AUC for each pipeline relying on the imputed data resulting from the different imputation strategies and their concatenation with the missingness indicator (as described in Section 3.5.2). These results echo the same insights:

- Different imputation strategies may result in equal downstream performance at the population level while having different group performance gaps
- No strategy consistently outperforms the others across clinical presence scenarios
- Current recommendations for group-specific imputation can increase the performance gap and yield a worse performance for the marginalised group.

When comparing Table A.2 and Table A.3, these experiments underline that predictive performances are often improved by the addition of the missingness indicators. However, this gain is not equally distributed across groups and may increase the fairness gaps.

		Imputation strategy			
	Group	Population Mean	Group Mean	MICE	Group MICE
(S1)	Overall	0.995 (0.000)	0.995 (0.000)	0.996 (0.000)	0.995 (0.000)
	Majority	0.997 (0.000)	0.997 (0.000)	0.997 (0.000)	0.997 (0.000)
	Marginalised	0.679 (0.037)	0.872 (0.026)	0.651 (0.039)	0.795 (0.033)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.318 (0.037)	**-0.125** (0.026)	-0.347 (0.040)	-0.201 (0.033)
(S2)	Overall	0.994 (0.000)	0.993 (0.000)	0.995 (0.000)	0.994 (0.000)
	Majority	0.996 (0.000)	0.995 (0.000)	0.997 (0.000)	0.996 (0.000)
	Marginalised	0.785 (0.031)	0.753 (0.040)	0.815 (0.028)	0.804 (0.032)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.212 (0.031)	-0.242 (0.040)	**-0.182** (0.028)	-0.192 (0.032)
(S3)	Overall	0.983 (0.001)	0.984 (0.001)	0.967 (0.001)	0.968 (0.001)
	Majority	0.987 (0.001)	0.987 (0.001)	0.970 (0.001)	0.972 (0.001)
	Marginalised	0.641 (0.037)	0.650 (0.037)	0.652 (0.037)	0.630 (0.038)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.346 (0.037)	-0.337 (0.037)	**-0.318** (0.037)	-0.341 (0.038)

Table A.2: Group-specific AUCs and gaps Δ mean performance (std) across scenarios on 100 synthetic experiments. If $\Delta < 0$, the marginalised group has lower AUC than the majority. Higher AUC is better.

		Imputation strategy with **missingness indicator**			
	Group	Population Mean	Group Mean	MICE	Group MICE
(S1)	Overall	0.996 (0.000)	0.995 (0.000)	0.996 (0.000)	0.996 (0.000)
	Majority	0.997 (0.000)	0.997 (0.000)	0.997 (0.000)	0.997 (0.000)
	Marginalised	0.725 (0.034)	0.872 (0.026)	0.684 (0.037)	0.694 (0.036)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.272 (0.034)	**-0.125** (0.026)	-0.313 (0.037)	-0.304 (0.036)
(S2)	Overall	0.995 (0.000)	0.993 (0.000)	0.995 (0.000)	0.994 (0.000)
	Majority	0.997 (0.000)	0.995 (0.000)	0.997 (0.000)	0.996 (0.000)
	Marginalised	0.835 (0.027)	0.753 (0.040)	0.831 (0.027)	0.804 (0.032)
	$\Delta^{\mathcal{I}}_{Marginalised}$	**-0.161** (0.027)	-0.242 (0.040)	-0.166 (0.027)	-0.192 (0.032)
(S3)	Overall	0.991 (0.001)	0.984 (0.001)	0.990 (0.001)	0.990 (0.001)
	Majority	0.993 (0.001)	0.987 (0.001)	0.993 (0.001)	0.993 (0.001)
	Marginalised	0.773 (0.032)	0.650 (0.037)	0.776 (0.032)	0.774 (0.032)
	$\Delta^{\mathcal{I}}_{Marginalised}$	-0.220 (0.032)	-0.337 (0.037)	**-0.216** (0.032)	-0.219 (0.032)

Table A.3: Group-specific AUCs and gaps Δ mean performance (std) across scenarios on 100 synthetic experiments when using missingness indicators as a predictor. If $\Delta < 0$, the marginalised group has lower AUC than the majority. Higher AUC is better.

Sensitivity to distribution. The proposed simulations have focused on a given underlying distribution of data in which the marginalised group presents the condition differently than the majority. In this section, we propose to study when the condition manifestation is the same across groups but the condition *prevalence* differs. As discussed in Section 3.3, this group difference may result in different group-specific missingness processes.

Consider a population of $N = 101,000$ patients with a ratio of 100:1 for the majority. Patients without the condition are drawn from the normal distribution $\mathcal{N}((0,0), 0.25)$ and positives are sampled from $\mathcal{N}((1,1), 0.25)$. Contrarily to the previous simulations, the marginalised group has a prevalence of 50% while the rest of the population, 10%. We then enforce the three previously described missingness processes. Figure A.3 illustrates how the same three proposed missingness processes would be expressed in this population. Importantly, due to differences in the prevalence, the missingness processes still differentially affect the two groups. For instance, the proposed (S2) affects the positive cases of both groups, representing 10% resp. 50% of these groups.

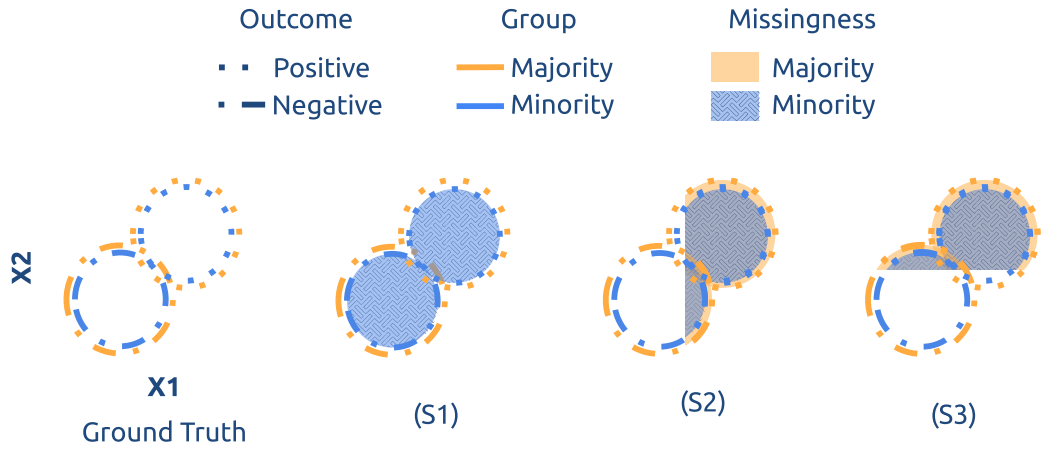


Figure A.3: Graphical summary of clinical missingness in the simulation experiments with identical condition manifestation but different prevalence across groups. Missingness is enforced on X_2, affecting 50% of the shaded regions for the indicated group.

Figures A.4 and A.5 present the associated reconstruction error and AUC performance differentiated by scenario and imputation strategies. This set of experiments shows that group-specific condition prevalence would lead to similar discrepancies in missingness process, reconstruction errors and performance gaps. Importantly, no imputation strategy would consistently lead to a better or fairer performance in the studied settings. This result stands despite no difference in condition manifestation, highlighting the need for a thorough evaluation of different imputation strategies.

Correlated covariates. The introduced distributions do not present correlations between the covariates at the group level. While the missingness process may introduce informative

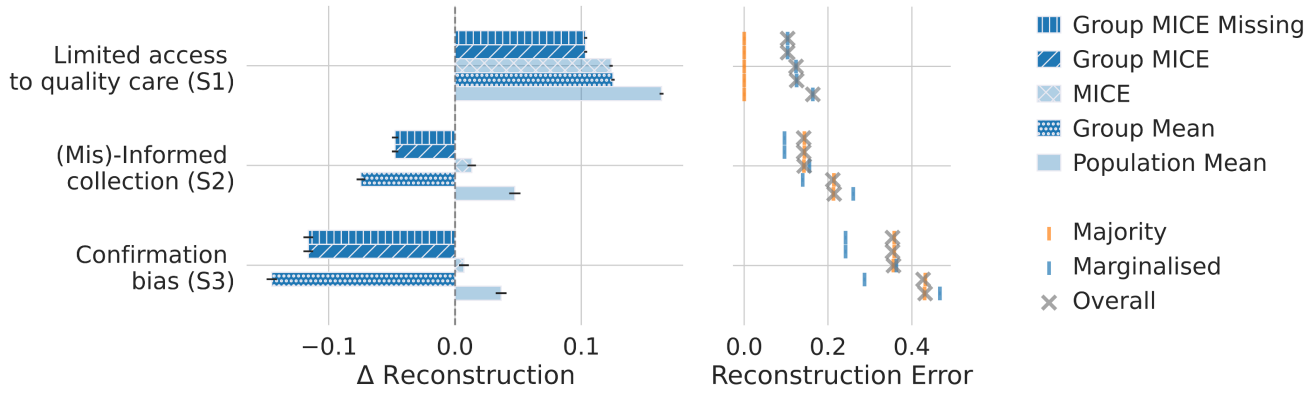


Figure A.4: Reconstruction error gap (on the left) and group-specific reconstruction errors (on the right) across scenarios on 100 synthetic experiments with the same condition manifestation across groups. If $\Delta < 0$, the marginalised group has a smaller reconstruction error than the majority. Lower reconstruction error is better.

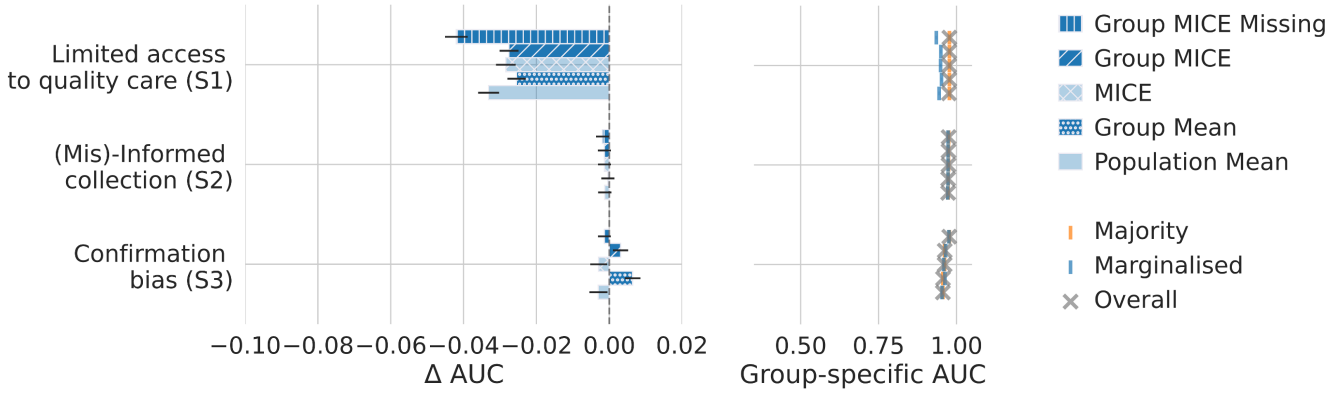


Figure A.5: AUC performance gaps Δ (on the left) and group-specific AUCs (on the right) across scenarios on 100 synthetic experiments with the same condition manifestation across groups. If $\Delta < 0$, the marginalised group has lower AUC than the majority. Higher AUC is better.

correlation that MICE imputation may leverage for improved imputation, this setting may not present the full strength of MICE imputation strategies. In this final set of experiments, we enforce correlation between X_1 and X_2 using the same simulations than in Section 3.5.1 except that the value of X_1 is added to X_2. This results in the distribution with correlated covariates as schematised in Figure A.6, in which we enforce the same three missing scenarios.

Figures A.4 and A.5 present the associated reconstruction error and AUC performance differentiated by scenario and imputation strategies. These figures echo the results presented in Section 3.5. Importantly, MICE and its group alternative do not consistently present superior performance. For instance, Group Mean presents the smallest reconstruction gap in (S3) and Population Mean reduced predictive difference in both (S2) and (S3).

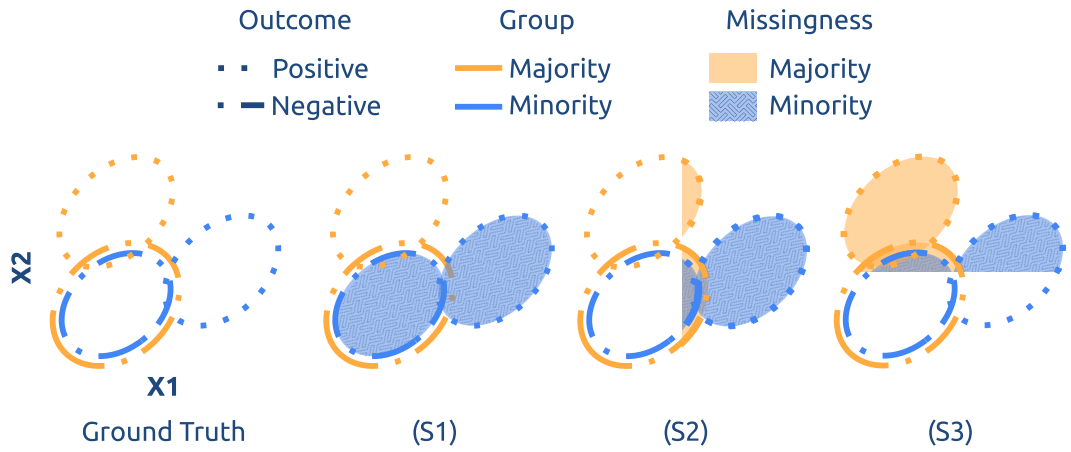


Figure A.6: Graphical summary of clinical missingness in the simulation experiments with correlated covariates. Missingness is enforced on X_2, affecting 50% of the shaded regions for the indicated group.

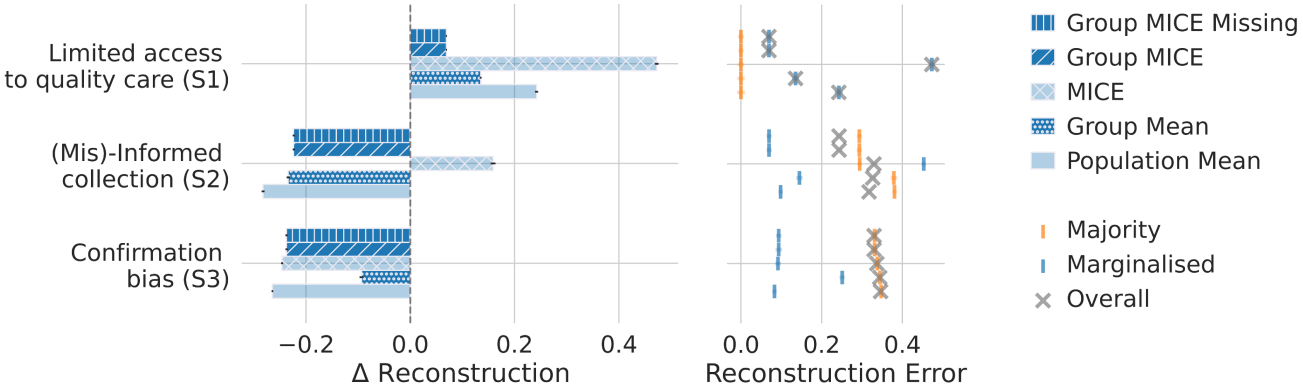


Figure A.7: Reconstruction error gap (on the left) and group-specific reconstruction errors (on the right) across scenarios on 100 synthetic experiments with correlated covariates. If $\Delta < 0$, the marginalised group has a smaller reconstruction error than the majority. Lower reconstruction error is better.

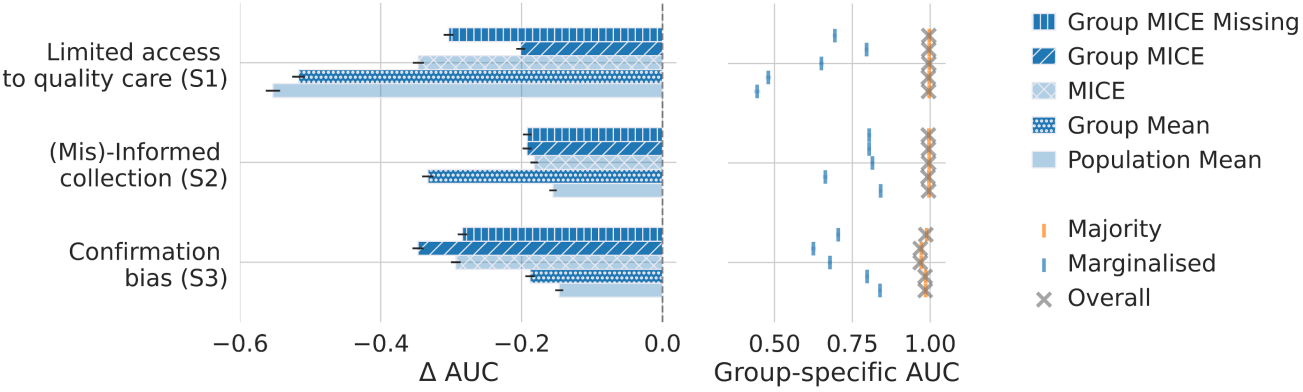


Figure A.8: AUC performance gaps Δ (on the left) and group-specific AUCs (on the right) across scenarios on 100 synthetic experiments with correlated covariates. If $\Delta < 0$, the marginalised group has lower AUC than the majority. Higher AUC is better.

A.2.2 Mimic III

Dataset. After preprocessing [338] and standardisation, the MIMIC III dataset consists of $36,296$ patients with 67 different laboratory tests. Focusing on the three marginalised groups of interest, the population can be further divided into marginalised subgroups as presented in Figure A.9. This representation underlines the importance of identifying subgroups at risk in the studied population.

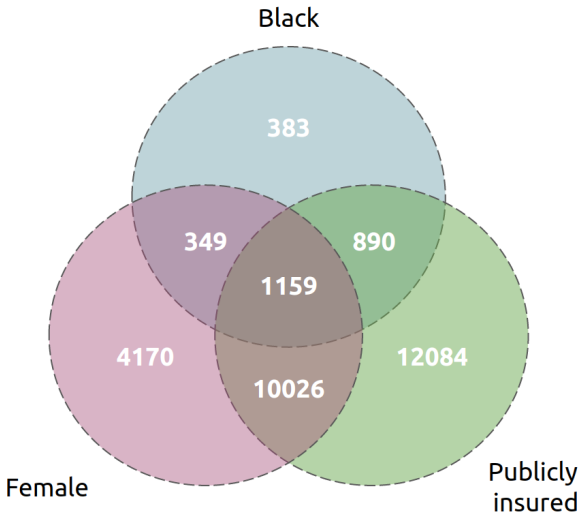


Figure A.9: Venn diagram of the population distribution in the three marginalised groups.

Experimental design. For this real-world dataset, patients are split into three groups: 80% for training, 10% for validation and 10% for hyper-parameters selection. The hyper-parameter search consisted of the l2 penalty selection for the logistic regression among $\lambda \in [0.1, 1., 10., 100.]$.

We bootstrapped the test set 100 times and report the mean and 95% confidence bounds.

Tabular results. Table A.4 presents the AUC for each group and imputation strategy. Similarly, Table A.5 details the false negative rate performances.

Missingness Indicators. Similarly, Table A.6 presents the AUC for each group and imputation strategy when missingness indicators are added to the regression model. Table A.7 shows the false negative rate for the same pipelines. This analysis shows that the missingness patterns are informative of the outcome of interest as adding the missingness indicators as regressors improves performance. Note, however, that MICE Group presents lower AUC performance than MICE in Table A.6. This observation echoes the criticism that group imputation may lead to suboptimal performance. Finally, these tables underline how no strategy consistently outperforms the others across groups.

	Imputation strategy			
Group	Population Mean	Group Mean	MICE	Group MICE
Black	**0.807** (0.030)	0.802 (0.030)	0.785 (0.035)	0.782 (0.036)
Non Black	**0.742** (0.010)	**0.742** (0.010)	0.735 (0.010)	0.735 (0.010)
Female	**0.736** (0.016)	0.733 (0.016)	0.729 (0.018)	0.726 (0.018)
Male	0.753 (0.016)	**0.756** (0.016)	0.742 (0.015)	0.738 (0.014)
Public	**0.735** (0.011)	0.732 (0.012)	0.722 (0.012)	0.715 (0.013)
Private	0.753 (0.021)	0.755 (0.021)	**0.762** (0.019)	0.751 (0.021)

Table A.4: AUC performance divided by group and imputation strategy - Bootstrapped mean (std). Bold indicates the highest AUC.

	Imputation strategy			
Group	Population Mean	Group Mean	MICE	Group MICE
Black	**0.271** (0.074)	0.458 (0.082)	0.314 (0.078)	0.336 (0.076)
Non Black	0.366 (0.017)	**0.350** (0.017)	0.371 (0.018)	0.371 (0.018)
Female	0.348 (0.030)	**0.330** (0.028)	0.381 (0.030)	0.361 (0.032)
Male	0.367 (0.027)	0.394 (0.027)	**0.357** (0.026)	0.394 (0.026)
Public	0.336 (0.019)	**0.310** (0.018)	0.348 (0.019)	0.344 (0.021)
Private	0.437 (0.041)	0.596 (0.041)	**0.423** (0.039)	0.510 (0.036)

Table A.5: False Negative rate divided by group and imputation strategy - Bootstrapped mean (std). Bold indicates the smallest FNR.

	Imputation strategy with **missingness indicators**			
Group	Population Mean	Group Mean	MICE	Group MICE
Black	**0.827** (0.027)	0.825 (0.027)	0.818 (0.028)	0.817 (0.029)
Non Black	**0.786** (0.010)	0.785 (0.010)	0.781 (0.010)	0.783 (0.010)
Female	0.770 (0.013)	0.770 (0.013)	**0.772** (0.014)	**0.772** (0.014)
Male	**0.801**(0.013)	0.800 (0.013)	0.793 (0.013)	0.796 (0.013)
Public	**0.773** (0.010)	**0.773** (0.011)	0.767 (0.011)	0.766 (0.010)
Private	0.816 (0.016)	0.810 (0.017)	**0.819** (0.017)	0.811 (0.018)

Table A.6: AUC performance divided by group and imputation strategy - Bootstrapped mean (std). Bold indicates the highest AUC.

	Imputation strategy with **missingness indicators**			
Group	Population Mean	Group Mean	MICE	Group MICE
Black	**0.224** (0.063)	0.459 (0.077)	0.229 (0.063)	0.294 (0.070)
Non Black	0.299 (0.017)	**0.273** (0.017)	0.294 (0.017)	0.290 (0.017)
Female	0.295 (0.024)	**0.255** (0.025)	0.292 (0.025)	0.278 (0.026)
Male	0.296 (0.025)	0.314 (0.024)	**0.285** (0.025)	0.287 (0.026)
Public	0.290 (0.020)	**0.250** (0.020)	0.286 (0.019)	0.280 (0.019)
Private	0.302 (0.038)	0.448 (0.043)	**0.295** (0.037)	0.348 (0.040)

Table A.7: False Negative rate divided by group and imputation strategy - Bootstrapped mean (std). Bold indicates the smallest FNR.

Controlling for all groups. In Section 3.6, the group-specific results correspond to the imputation associated with each respective group. As practitioners aim to reduce the performance gap across all presented groups simultaneously, we propose alternative group imputation strategies in which we control over all groups. Specifically, group mean imputation associate to each patient the group mean associated with its Ethnicity, Sex and Insurance. Similarly, Group MICE regresses on the three groups simultaneously.

The use of these imputation strategies result in the updated Figure A.10. These experiments leads to identical conclusions.

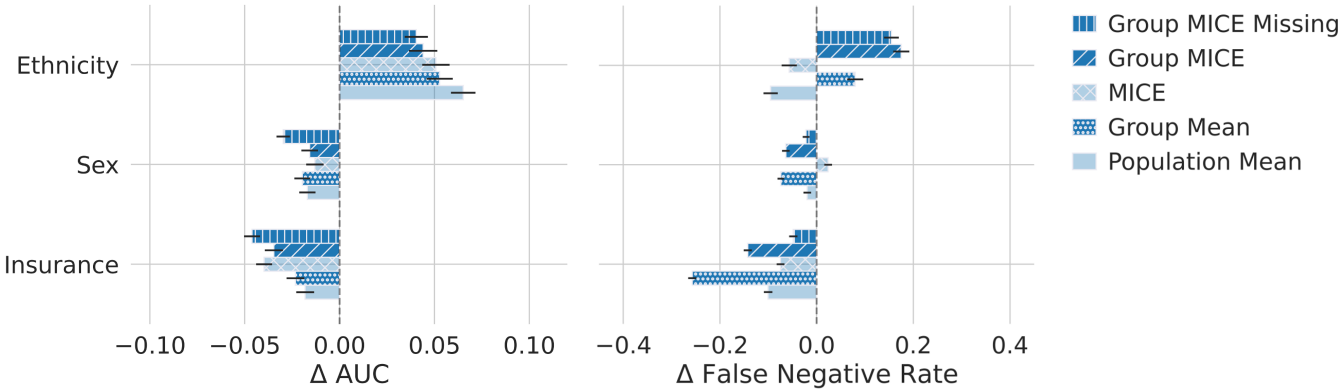


Figure A.10: Prioritisation performance gaps Δ across marginalised groups in MIMIC III experiment when controlling on inter-sectional groups. If $\Delta > 0$, the marginalised group has a larger value of the given metric than the rest of the population.

Threshold sensitivity. In Section 3.6, we present results for a policy of 30% additional care. As we arbitrarily chose this threshold, we propose to measure how the results vary under two different thresholds: 5% and 50%. Additionally, we introduce an additional metric to quantify equity in care delivery. When considering a fixed resource threshold (30%), we measure each group's prioritisation — the proportion of all patients who would receive care under this policy, regardless of their ground-truth condition. A gap in prioritisation quantifies how care is delivered differently across groups while a gap in FNR between groups illustrates how patients would be incorrectly de-prioritised across groups.

Figures A.11, A.12 and A.13 present the results at 5%, 30% and 50% thresholds. First, note that the magnitude of the Δ in prioritisation increases with larger thresholds, but similar trends are observed. This indicates that members of the same group have similar risk scores. Increasing the threshold, therefore, further penalises this whole group. Second, the Δ in false positive rates demonstrates how the choice of imputation is sensitive to the target task. In addition to validating the insights from Section 3.6, this set of experiments demonstrates that the target task may also affect whether an imputation strategy favours or penalises a given group.

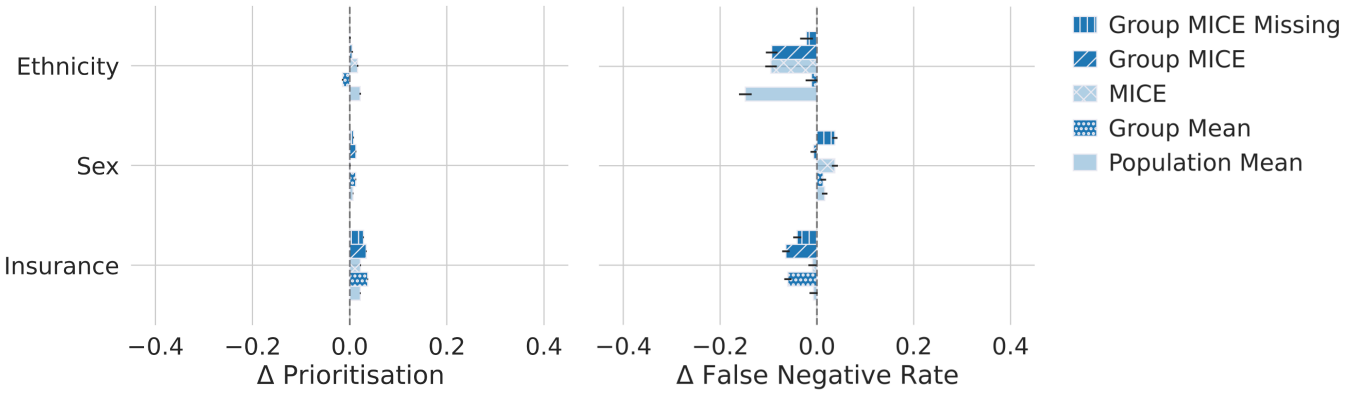


Figure A.11: Prioritisation performance gaps Δ across marginalised groups in MIMIC III experiment for 5% additional care. If $\Delta > 0$, the marginalised group has a larger value of the given metric than the rest of the population.

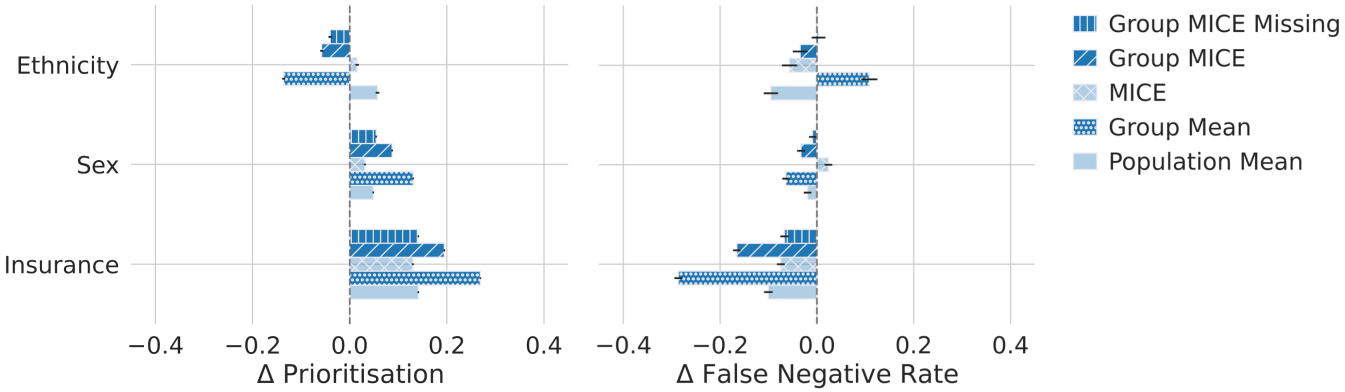


Figure A.12: Prioritisation performance gaps Δ across marginalised groups in MIMIC III experiment for 30% additional care. If $\Delta > 0$, the marginalised group has a larger value of the given metric than the rest of the population.

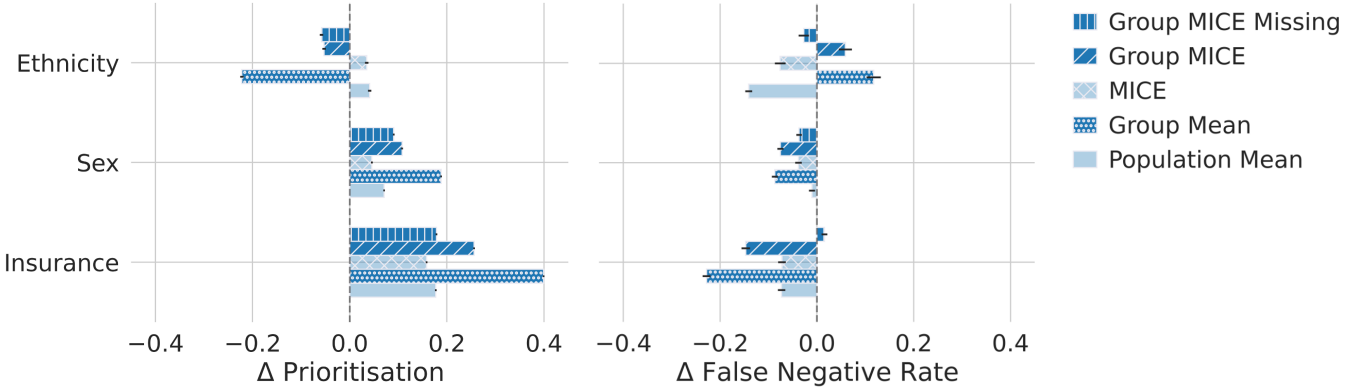


Figure A.13: Prioritisation performance gaps Δ across marginalised groups in MIMIC III experiment for 50% additional care. If $\Delta > 0$, the marginalised group has a larger value of the given metric than the rest of the population.

Appendix B

Supplemental material Chapter 4: Response Heterogeneity

B.1 Proof

In Section 4.3, we claim that under the considered DAGs,

$$\text{IPM}(q_\alpha^{A=0}, q_\alpha^{A=1}) = 0$$

with A, being the treatment assignment and $\alpha(x) = \{\mathbb{P}(Z = k \mid X = x)\}$, the probability vector of cluster assignment given the covariates x. The following derive this result:

Proof.

$$\begin{aligned} q_\alpha^{A=a} &:= q(\{\mathbb{P}(Z = k \mid X)\} \mid A = a) \\ &= q(\{\mathbb{P}(Z = k \mid X, A = a)\}) \\ &= q(\{\mathbb{P}(Z = k \mid X)\}) \qquad \text{(Under Asm. 4.2)} \end{aligned}$$

From this expression, $q_\alpha^{A=0} = q_\alpha^{A=1}$, which results in the distance between these distributions being null. □

B.2 Neural survival clustering in real-world setting

In this section, we introduce two additional real-world datasets to measure the capacity of Neural Survival Clustering to predict survival outcomes in medical data using the same evaluation pipeline as presented in Section 4.6.

B.2.1 Datasets description

Following a similar experiment setting and pre-processing as in [221], we present results on the following single-event and single-risk datasets:

- METABRIC [76] with 1,904 patients presenting 9 genetics and clinical covariates. 57.9% of the population died from breast cancer.
- FRAMINGHAM [159] with 4,434 patients followed over 20 years. The outcome of interest is the observation of CVD, affecting 34.9% of the population.

B.2.2 Performances

Table B.1 echoes Chapter 4's results with the performance of the proposed methodology similar to existing survival models. This observation validates the hypothesis that the considered population's outcome is well explained through a discrete mixture of survival functions.

	Model	C Index			Brier Score		
		$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
METABRIC	**NSC**	**0.703** (0.043)	**0.670** (0.013)	**0.651** (0.019)	**0.117** (0.012)	*0.190* (0.007)	*0.222* (0.007)
	DCM	0.575 (0.085)	0.565 (0.092)	0.579 (0.076)	0.125 (0.010)	0.207 (0.009)	0.243 (0.014)
	DSM	*0.688* (0.025)	0.652 (0.019)	0.627 (0.023)	0.119 (0.009)	0.195 (0.004)	0.232 (0.019)
	SuMo-net	*0.688* (0.038)	*0.669* (0.008)	0.644 (0.022)	*0.118* (0.013)	**0.189** (0.005)	*0.222* (0.010)
	DeepHit	0.687 (0.025)	0.620 (0.037)	0.599 (0.036)	0.122 (0.008)	0.205 (0.008)	0.239 (0.004)
	DeepSurv	0.658 (0.032)	0.648 (0.022)	0.648 (0.022)	0.120 (0.010)	0.192 (0.008)	**0.221** (0.008)
	CoxPH	0.632 (0.018)	0.628 (0.019)	0.636 (0.020)	0.121 (0.009)	0.195 (0.008)	*0.222* (0.007)
	Survival Tree	0.626 (0.047)	0.618 (0.038)	0.611 (0.029)	0.127 (0.011)	0.202 (0.011)	0.233 (0.010)
FRAMINGHAM	**NSC**	**0.779** (0.015)	**0.760** (0.021)	**0.761** (0.012)	*0.070* (0.008)	*0.116* (0.006)	*0.143* (0.006)
	DCM	0.753 (0.027)	0.739 (0.021)	0.737 (0.016)	0.073 (0.009)	0.124 (0.009)	0.156 (0.012)
	DSM	0.776 (0.015)	*0.759* (0.019)	0.756 (0.011)	0.071 (0.008)	0.120 (0.005)	0.150 (0.006)
	SuMo-net	0.766 (0.019)	0.757 (0.019)	0.757 (0.010)	0.071 (0.009)	*0.116* (0.008)	0.144 (0.006)
	DeepHit	0.764 (0.014)	0.739 (0.022)	0.736 (0.017)	0.073 (0.009)	0.130 (0.007)	0.167 (0.005)
	DeepSurv	*0.777* (0.018)	0.759 (0.021)	*0.759* (0.011)	**0.069** (0.008)	*0.116* (0.008)	0.144 (0.005)
	CoxPH	0.774 (0.014)	**0.760** (0.022)	**0.761** (0.012)	*0.070* (0.008)	**0.114** (0.006)	**0.142** (0.005)
	Survival Tree	0.730 (0.011)	0.715 (0.016)	0.715 (0.009)	0.077 (0.009)	0.127 (0.008)	0.158 (0.007)

Table B.1: Comparison of model performance by means (standard deviations) across the METABRIC and FRAMINGHAM dataset. Best performances are in **bold**, second best in *italics*.

B.3 Further SEER's analysis

Selecting number of treatment clusters. Similarly to the simulation, we computed the likelihood given an increasing number of clusters K and observed no change in likelihood after $K = 2$ as shown in Figure B.1. Note that one should not expect the same number of clusters as with NSC; the number of treatment responses may be more or less diverse than the number of survival profiles.

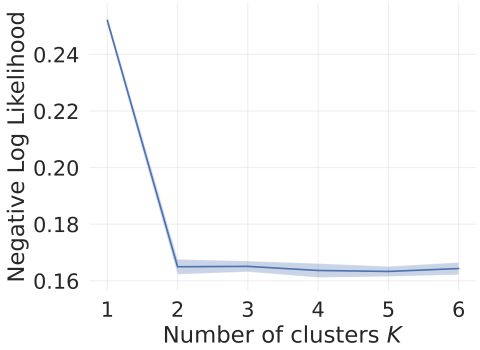


Figure B.1: Causal Neural Survival Clustering - Averaged log-likelihood across 5-fold cross-validation given the number of clusters K in the SEER dataset. The absence of change indicates $K = 2$.

Discriminative covariates. As for the Neural Survival Clustering, we performed a permutation test on the covariates to identify the most predictive of the treatment effect subgroups. Figure B.2 displays the ten most predictive covariates. Note that the set of identified covariates overlaps with the NSC one.

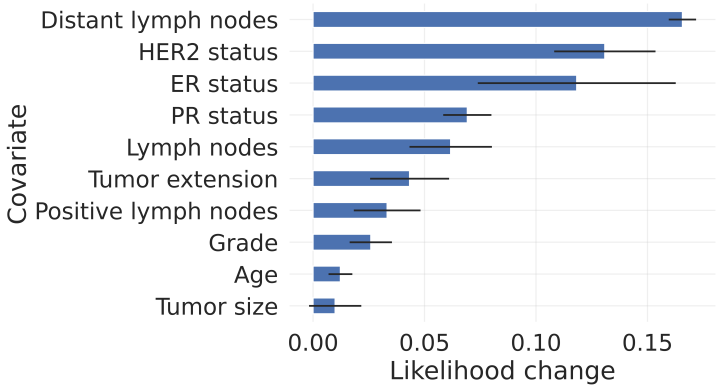


Figure B.2: Causal Neural Survival Clustering - Change in log-likelihood given random permutation of a given covariates.

Appendix C

Supplemental material Chapter 5: Competing Risks

C.1 Theoretical analysis of ignoring competing risks

Notations. Remind that F_{NC} is the cumulative incidence function when considering competing risks as censoring while F_C is the same quantity when considering competing risks. In this context, T_r is the unobserved event time of risk r and T' and D', the first event time and type if there was no censoring.

Theorem. One can express the survival estimate error for a given outcome r as the probability of not observing this outcome given the covariates x:

$$L^r(x) = \mathbb{P}(D' \neq r \mid x)$$

with D' the random variable associated with the type of observed risk in the absence of censoring.

Proof. First, we focus on the denominator of the expression $L^r(x)$:

$$\begin{aligned}
\max(F_r^{NC}(t \mid x), F_r^C(t \mid x)) &:= \max(\mathbb{P}(T_r > t \mid x), \mathbb{P}(T' > t, D' = r \mid x)) \\
&= \max(\mathbb{P}(T_r > t \mid x), \mathbb{P}(T_r > t, D' = r \mid x)) \\
&\qquad \text{(By Definition of } T') \\
&= \mathbb{P}(T_r > t \mid x)
\end{aligned}$$

Then, we can simplify the equality as follows:

$$
\begin{aligned}
L^r(t,x) &:= \frac{F_r^{NC}(t \mid x) - F_r^{C}(t \mid x)}{F_r^{NC}(t \mid x)} \\
&= \frac{\mathbb{P}(T_r > t \mid x) - \mathbb{P}(T' > t, D' = r \mid x)}{\mathbb{P}(T_r > t \mid x)} \\
&= \frac{\mathbb{P}(T_r > t \mid x) - \mathbb{P}(T_r > t, D' = r \mid x)}{\mathbb{P}(T_r > t \mid x)} && \text{(By Definition of } T'\text{)} \\
&= \mathbb{P}(D' \neq r \mid x) && \text{(By Bayes' Theorem)}
\end{aligned}
$$

□

C.2 Experiments

C.2.1 Datasets characteristics

Table C.1 presents the times and observed outcomes corresponding to the different quantiles of the uncensored population used for evaluation, differentiated by datasets.

		Quantiles		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Time (years)	3.19	4.95	7.45
	Censoring	0.00 %	0.00%	11.54%
	Death	12.82%	23.72%	32.69%
	Transplant	0.64%	3.21%	7.69%
FRAM.	Time (years)	5.90	12.57	18.14
	Censoring	0.00 %	0.00%	0.00%
	Death	2.66%	7.40%	12.56%
	CVD	8.30%	14.52%	20.32%
SEER	Time (years)	1.67	4.00	8.08
	Censoring	10.34%	22.20 %	39.59%
	BC	4.53%	9.32%	13.37%
	CVD	0.80%	1.76%	3.23%

Table C.1: Observed outcomes of interest at the different evaluation horizons.

C.2.2 Competing Risk Performance

To complement the performance on the primary outcome of interest presented in Section 5.6.2, this section presents the C-index and Brier score evaluated at the dataset-specific 0.25, 0.5 and 0.75 quantiles on the *competing risk* for all competing models.

C.2.2.1 Simulations

Table C.2 summarises the performances on the second synthetic outcome $r = 2$ across the 25 simulations. These results echo the conclusion made in the main text.

	Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
Competing	**NeuralFG**	**0.802** (0.055)	*0.735* (0.060)	*0.691* (0.055)	**0.083** (0.014)	**0.148** (0.018)	**0.194** (0.016)
	DeSurv	**0.802** (0.055)	**0.737** (0.059)	**0.693** (0.052)	*0.084* (0.014)	**0.148** (0.018)	*0.195* (0.018)
	DeepHit	0.777 (0.062)	0.709 (0.062)	0.670 (0.053)	0.094 (0.019)	0.169 (0.023)	0.214 (0.015)
	DSM	*0.782* (0.059)	0.724 (0.060)	0.685 (0.053)	0.086 (0.014)	*0.151* (0.018)	0.209 (0.017)
	Fine-Gray	0.614 (0.133)	0.613 (0.099)	0.609 (0.075)	0.096 (0.019)	0.167 (0.024)	0.213 (0.021)
	CS Cox	0.635 (0.115)	0.620 (0.092)	0.607 (0.077)	0.095 (0.019)	0.166 (0.025)	0.213 (0.022)

Table C.2: Comparison of model performance by means (standard deviations) across 25 simulations on event $r = 2$. Best performances are in **bold**, second best in *italics*.

C.2.2.2 Real-world analysis

Table C.3 summarises the performances on the competing risk for all three real-world datasets.

		Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
			$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Competing	**NeuralFG**	0.772 (0.141)	0.661 (0.286)	0.700 (0.142)	*0.017* (0.002)	0.035 (0.012)	0.079 (0.015)
		DeSurv	*0.854* (0.113)	0.695 (0.270)	0.722 (0.121)	0.023 (0.006)	*0.033* (0.013)	0.075 (0.012)
		DeepHit	0.780 (0.014)	0.630 (0.168)	0.643 (0.098)	*0.017* (0.001)	*0.033* (0.017)	0.078 (0.008)
		DSM	0.708 (0.181)	*0.713* (0.262)	0.715 (0.148)	**0.016** (0.000)	**0.029** (0.015)	0.074 (0.011)
		Fine-Gray	**0.879** (0.079)	0.704 (0.332)	*0.752* (0.120)	**0.016** (0.000)	0.035 (0.014)	**0.072** (0.014)
		CS Cox	0.846 (0.010)	**0.742** (0.277)	**0.792** (0.109)	**0.016** (0.001)	0.038 (0.015)	*0.073* (0.013)
FRAMINGHAM	Competing	**NeuralFG**	**0.730** (0.056)	**0.698** (0.037)	0.687 (0.023)	**0.025** (0.002)	**0.065** (0.004)	**0.103** (0.004)
		DeSurv	0.667 (0.074)	0.633 (0.054)	0.601 (0.038)	*0.026* (0.003)	0.072 (0.006)	0.124 (0.007)
		DeepHit	0.703 (0.030)	0.675 (0.028)	0.670 (0.015)	*0.026* (0.002)	0.067 (0.004)	*0.106* (0.004)
		DSM	0.701 (0.059)	0.681 (0.035)	0.670 (0.023)	*0.026* (0.002)	0.068 (0.004)	0.109 (0.002)
		Fine-Gray	0.723 (0.036)	*0.697* (0.037)	*0.690* (0.023)	**0.025** (0.002)	**0.065** (0.005)	**0.103** (0.004)
		CS Cox	*0.727* (0.050)	0.694 (0.042)	**0.691** (0.024)	**0.025** (0.002)	*0.066* (0.004)	**0.103** (0.004)
SEER	Competing	**NeuralFG**	0.849 (0.003)	*0.849* (0.003)	**0.846** (0.002)	**0.009** (0.000)	**0.019** (0.000)	**0.036** (0.000)
		DeSurv	0.771 (0.008)	0.767 (0.006)	0.777 (0.005)	**0.009** (0.000)	*0.020* (0.000)	0.039 (0.001)
		DeepHit	**0.862** (0.004)	**0.855** (0.003)	0.842 (0.004)	**0.009** (0.000)	*0.020* (0.000)	*0.037* (0.000)
		DSM	*0.858* (0.002)	0.847 (0.003)	0.835 (0.003)	**0.009** (0.000)	*0.020* (0.000)	0.038 (0.000)
		Fine-Gray	0.835 (0.003)	0.842 (0.003)	0.843 (0.002)	**0.009** (0.000)	**0.019** (0.000)	**0.036** (0.000)
		CS Cox	0.849 (0.003)	*0.849* (0.003)	*0.845* (0.003)	**0.009** (0.000)	**0.019** (0.000)	**0.036** (0.000)

Table C.3: Comparison of model performance by means (standard deviations) across 5-fold cross-validation on the primary outcome. Best performances are in **bold**, second best in *italics*.

C.2.3 Implementation details

The proposed experiments rely on the `scikit-survival` [254][1] and `pycox`[2] libraries for evaluation. For baselines' implementations, we used the R library `riskRegression`[3] for CS Cox and Fine-Gray, `pycox` for DeepHit and `auton-survival` [222][4] for Deep Survival Machines.

C.3 Neural Fine-Gray

C.3.1 Using *R* outputs vs. *R* networks

In this section, we investigate the impact of using multiple networks — one for each competing risk, as proposed in Chapter 5 — instead of one network with multiple outcomes. The model **MonoFG** consists of the same architecture presented in Figure 5.3 with only one monotonic network with R outputs. Table C.4 shows non-significant differences between the two architectures. While multiple networks result in more flexibility in the survival distribution of different competing risks, a single neural network requires a smaller amount of data as it has fewer parameters, reduces the risk of overfitting and is faster to train.

	Risk	Model	C-Index *(Larger is better)*			Brier Score *(Smaller is better)*		
			$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Dea.	**NeuralFG**	0.809 (0.079)	0.791 (0.119)	0.759 (0.119)	0.099 (0.028)	0.140 (0.017)	0.172 (0.044)
		MonoFG	**0.813** (0.086)	**0.792** (0.102)	**0.771** (0.112)	**0.095** (0.025)	**0.135** (0.024)	**0.160** (0.055)
	Tra.	**NeuralFG**	**0.772** (0.141)	**0.661** (0.286)	**0.700** (0.142)	**0.017** (0.002)	**0.035** (0.012)	**0.079** (0.015)
		MonoFG	0.674 (0.119)	0.593 (0.255)	0.629 (0.068)	**0.017** (0.001)	0.038 (0.016)	0.081 (0.016)
FRAM.	CVD	**NeuralFG**	**0.871** (0.025)	**0.809** (0.030)	**0.775** (0.018)	**0.050** (0.003)	**0.096** (0.009)	**0.130** (0.004)
		MonoFG	0.869 (0.025)	0.803 (0.029)	0.770 (0.020)	**0.050** (0.003)	**0.096** (0.007)	**0.130** (0.004)
	Dea.	**NeuralFG**	**0.730** (0.056)	**0.698** (0.037)	**0.687** (0.023)	**0.025** (0.002)	**0.065** (0.004)	**0.103** (0.004)
		MonoFG	0.715 (0.047)	0.694 (0.035)	0.686 (0.019)	0.026 (0.003)	0.066 (0.005)	0.104 (0.004)
SEER	BC	**NeuralFG**	0.899 (0.002)	**0.863** (0.001)	**0.824** (0.000)	**0.037** (0.000)	**0.068** (0.000)	**0.100** (0.001)
		MonoFG	**0.901** (0.003)	**0.863** (0.002)	**0.824** (0.002)	**0.037** (0.000)	**0.068** (0.000)	**0.100** (0.001)
	CVD	**NeuralFG**	0.849 (0.003)	0.849 (0.003)	0.846 (0.002)	**0.009** (0.000)	**0.019** (0.000)	**0.036** (0.000)
		MonoFG	**0.850** (0.003)	**0.850** (0.003)	**0.847** (0.002)	**0.009** (0.000)	**0.019** (0.000)	**0.036** (0.000)

Table C.4: Comparison of multi-head and mono-head NFG's performance by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**.

[1] `https://github.com/sebp/scikit-survival`
[2] `https://github.com/havakv/pycox`
[3] `https://github.com/tagteam/riskRegression`
[4] `https://github.com/autonlab/auton-survival`

Appendix D

Supplemental material Chapter 6: Clinical Presence Shift

D.1 Mimic III - Experiments

D.1.1 Data characteristics

Table D.1 presents the demographic characteristics of the studied population, and Table D.2 summarises the set of tests selected with the mean number of tests performed during the 24 hours post-admission and their mean values. The results are presented at the population level and differentiated by the subgroups used to study the impact of clinical presence shift. Finally, Figure D.1 displays the Kaplan-Meier survival estimates following 24 hours of observation.

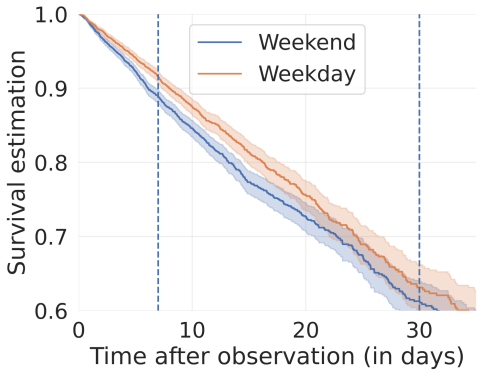


Figure D.1: Kaplan-Meier survival estimates following 24 hours of observation stratified by the admission group. Dotted lines denote evaluation horizons.

			Population		
			Overall	Weekday	Weekend
	Number of patients		31,692	24,865	6,827
Outcome	Length of stay (in days*)		10.00	9.97	10.11
	Death	Overall (%)	13.23	12.69	15.19
		7 days^{+} (%)	6.90	6.45	8.48
		30 days^{+} (%)	12.28	11.73	14.30
Demographics	Gender	Male (%)	57.12	57.36	56.22
		Female (%)	42.88	42.64	43.78
	Ethnicity	Asian (%)	2.34	2.34	2.34
		Black (%)	7.39	7.28	7.78
		Hispanic (%)	3.16	3.06	3.52
		White (%)	71.34	72.02	68.87
		Other (%)	15.77	15.30	17.49
	Insurance	Public (%)	65.42	65.07	66.69
		Private (%)	33.45	33.97	31.55
		Self Pay (%)	1.13	0.96	1.76

* Mean (std)
$^{+}$ After first day of observation

Table D.1: MIMIC III - Population characteristics between patients admitted on weekdays and weekends.

Laboratory test	Lab Value			Number Test		
	Weekday		Weekend	Weekday		Weekend
Anion gap	13.73 (3.24)	*	14.15 (3.28)	1.78 (1.03)	*	1.92 (1.09)
Bicarbonate	24.17 (4.23)	*	23.79 (4.56)	1.87 (1.00)	*	1.94 (1.09)
Blood urea nitrogen	24.51 (19.83)	*	25.96 (21.50)	1.90 (0.99)	*	1.95 (1.08)
Calcium (total)	8.37 (0.80)	*	8.31 (0.77)	1.45 (1.10)	*	1.69 (1.10)
Chloride	105.30 (5.70)		105.25 (5.98)	1.92 (1.03)	*	1.98 (1.12)
Creatinine	1.30 (1.34)	*	1.36 (1.36)	1.91 (0.99)	*	1.96 (1.08)
Glucose	137.85 (48.08)		137.75 (53.83)	3.32 (2.99)	*	2.46 (2.00)
Hematocrit	32.63 (5.37)	*	32.88 (5.39)	3.30 (2.62)	*	2.66 (2.03)
Hemoglobin	11.00 (1.92)	*	11.10 (1.93)	2.73 (2.31)	*	2.09 (1.71)
Magnesium	1.99 (0.39)	*	1.97 (0.34)	1.68 (1.07)	*	1.85 (1.10)
MCH[1]	30.31 (2.48)		30.30 (2.58)	1.79 (1.01)	*	1.72 (1.01)
MCH[1] conc.	33.90 (1.51)	*	33.80 (1.56)	1.79 (1.01)	*	1.72 (1.01)
MCV[2]	89.49 (6.54)	*	89.71 (6.77)	1.79 (1.01)	*	1.72 (1.01)
PTH[3]	38.02 (20.04)	*	38.94 (21.37)	1.55 (1.30)		1.56 (1.37)
Phosphate	3.59 (1.16)	*	3.49 (1.17)	1.44 (1.11)	*	1.69 (1.10)
Platelets	219.89 (106.52)	*	224.16 (114.58)	1.94 (1.13)	*	1.83 (1.15)
Potassium	4.16 (0.55)	*	4.10 (0.56)	1.98 (1.10)	*	2.11 (1.16)
Red blood cell	5.26 (15.58)		5.66 (18.21)	1.91 (1.06)	*	1.84 (1.08)
Sodium	138.84 (4.33)	*	139.03 (4.69)	1.88 (1.09)	*	2.01 (1.16)
White blood cell	11.98 (13.90)		12.18 (15.34)	1.92 (1.06)	*	1.86 (1.09)
pH	6.93 (0.76)	*	6.85 (0.80)	3.16 (3.74)	*	2.30 (2.94)

[1] Mean corpuscular hemoglobin.
[2] Mean corpuscular volume.
[3] Partial thromboplastin time.
* signifies that a 2-sided T-test has a p-value < 0.05

Table D.2: List of laboratory tests used with the associated mean number of tests and values (and standard deviations) differentiated by admission populations.

D.1.2 Hyperparameters tuning

All models' hyper-parameters were selected over the following grid of hyperparameters (if appropriate) following 100 iterations of random search. All experiments were ran on a A100 GPU over 100 hours for each experiment.

	Hyperparameter	Values
Training	Learning rate	10^{-3}, 10^{-4}
	Batch size	512 , 1024
	α	0.1, 0.3
	θ	2^*
RNN	Layers	1, 2
	Hidden nodes	10, 25
Survival	Layers	0, 1, 2, 3
	Nodes	50
Clinical Presence	Temporal (I)	Same parameters explored as survival
	Missingnesss (M)	Same parameters explored as survival

* Following the results from [194].

Table D.3: Grid used for hyperparameters search

D.1.3 Modelling clinical presence

Table D.4 presents the discriminative results on the test set of the first set of experiments for which patients are randomly split into train and test sets.

Model	C Index			Brier Score		
	7	30	Overall	7	30	Overall
DeepJoint	0.760 (0.008)	*0.652* (0.010)	0.739 (0.007)	0.076 (0.002)	*0.267* (0.011)	**0.127** (0.084)
Feature	*0.764* (0.008)	0.651 (0.011)	*0.744* (0.007)	*0.075* (0.002)	0.272 (0.012)	-
GRU-D	0.756 (0.008)	**0.654** (0.011)	0.735 (0.007)	**0.074** (0.002)	**0.266** (0.011)	0.160 (0.049)
Resample	0.737 (0.009)	0.651 (0.011)	0.721 (0.008)	0.076 (0.002)	0.270 (0.011)	-
Ignore	0.749 (0.008)	0.648 (0.011)	0.727 (0.007)	*0.075* (0.002)	0.273 (0.011)	-
Count	**0.768** (0.008)	0.648 (0.011)	**0.748** (0.006)	*0.075* (0.002)	0.278 (0.012)	*0.144* (0.081)
Last	0.739 (0.008)	*0.652* (0.011)	0.720 (0.007)	*0.075* (0.002)	0.269 (0.011)	0.145 (0.066)

Table D.4: Comparison of model performance by means (standard deviations) in the random split experiment of MIMIC III dataset. Best performances are in **bold**, second best in *italics*. '-' denotes the divergence of the Brier score.

D.1.4 Transfer performances

Weekday Performance. Figure D.2 shows the performance on patients admitted on weekdays. Performance resembles the one presented in Section 6.5.4 with DeepJoint presenting better transportability than Feature, its alternative that does not model clinical presence as a model's output, except at 30 days where the model underfit. Count presents the best internal performance in this setting but suffers when transferred as counts during weekends and weekdays may reflect different information. An important observation is that the proposed method does not overfit in the presented setting, as the multitask learning regularises the shared embedding.

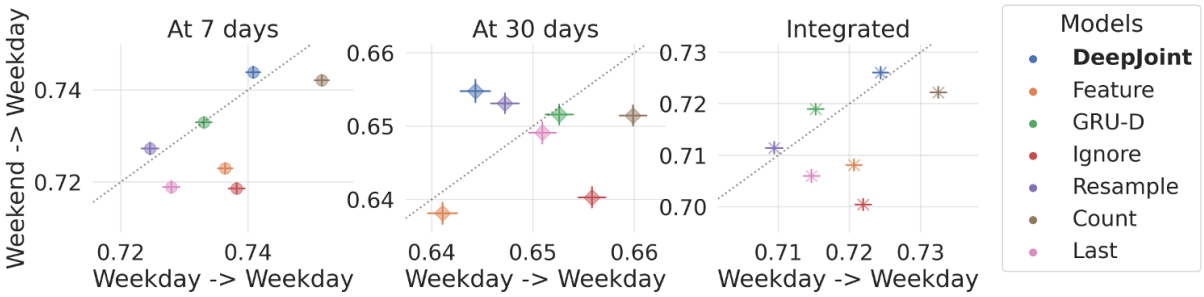


Figure D.2: Transportability evaluation for patients admitted on weekdays.

Tabular Performance. Tables D.5, D.6 and D.7 present the discriminative difference on the weekends-admissions test set between the model trained on weekends-admissions and the model transferred from weekdays to weekends (and the opposite scenario).

These results emphasise that the proposed joint modelling **DeepJoint** is more transportable than state-of-the-art approaches based on the same inputs. Interestingly, ignoring the clinical process (**Resample** or **Ignore**) seems to be less transportable under shifts in the observational process, echoing the remark made by [190] about the difficulty of ignoring it. Finally, note how simple features such as in **Count** might be detrimental under shift — particularly pronounced when applied to weekends-admissions.

	Horizon: 7 days after observation					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	**0.714** (0.022)	**0.715** (0.023)	**0.014** (0.012)	**0.744** (0.006)	*0.741* (0.006)	*0.005* (0.003)
Feature	*0.707* (0.022)	0.697 (0.024)	*0.015* (0.011)	0.722 (0.006)	0.737 (0.006)	0.014 (0.005)
GRU-D	0.660 (0.028)	0.693 (0.024)	0.033 (0.015)	0.733 (0.006)	0.732 (0.007)	**0.003** (0.002)
Ignore	0.665 (0.026)	0.696 (0.026)	0.032 (0.017)	0.718 (0.006)	0.738 (0.007)	0.019 (0.005)
Resample	0.659 (0.025)	0.687 (0.024)	0.029 (0.015)	0.727 (0.007)	0.724 (0.007)	*0.005* (0.003)
Count	0.692 (0.023)	*0.709* (0.022)	0.020 (0.013)	*0.742* (0.006)	**0.751** (0.006)	0.010 (0.004)
Last	0.653 (0.024)	0.679 (0.025)	0.026 (0.012)	0.718 (0.007)	0.727 (0.007)	0.009 (0.004)

Table D.5: C-Index performance at 7 days for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std). Best performances are in **bold**, second best in *italics*.

	Horizon: 30 days after observation					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	*0.636* (0.028)	*0.639* (0.027)	**0.013** (0.010)	**0.655** (0.008)	0.644 (0.007)	0.011 (0.005)
Feature	0.633 (0.028)	0.614 (0.028)	0.021 (0.012)	0.638 (0.007)	0.641 (0.007)	*0.004* (0.004)
GRU-D	0.631 (0.026)	**0.642** (0.026)	*0.015* (0.010)	0.652 (0.007)	0.653 (0.007)	**0.003** (0.002)
Ignore	**0.639** (0.027)	0.623 (0.029)	0.020 (0.014)	0.640 (0.007)	*0.656* (0.007)	0.016 (0.006)
Resample	*0.636* (0.026)	0.632 (0.027)	*0.015* (0.010)	*0.653* (0.007)	0.647 (0.007)	0.007 (0.004)
Count	0.607 (0.030)	0.628 (0.030)	0.022 (0.012)	0.651 (0.007)	**0.660** (0.007)	0.009 (0.005)
Last	0.616 (0.028)	0.631 (0.028)	0.017 (0.012)	0.649 (0.007)	0.651 (0.007)	*0.004* (0.003)

Table D.6: C-Index performance at 30 days for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std). Best performances are in **bold**, second best in *italics*.

	Integrated					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	**0.703** (0.019)	**0.703** (0.020)	**0.011** (0.009)	**0.726** (0.006)	*0.724* (0.006)	**0.004** (0.003)
Feature	*0.697* (0.019)	*0.693* (0.020)	**0.011** (0.009)	0.708 (0.006)	0.721 (0.006)	0.013 (0.004)
GRU-D	0.672 (0.024)	0.692 (0.022)	0.020 (0.012)	0.719 (0.005)	0.715 (0.006)	**0.004** (0.003)
Ignore	0.674 (0.023)	0.687 (0.021)	0.016 (0.012)	0.700 (0.006)	0.722 (0.006)	0.021 (0.005)
Resample	0.669 (0.022)	0.680 (0.022)	*0.015* (0.011)	0.711 (0.006)	0.709 (0.006)	**0.004** (0.003)
Count	0.687 (0.021)	**0.703** (0.020)	0.017 (0.010)	*0.722* (0.006)	**0.732** (0.006)	0.010 (0.004)
Last	0.666 (0.022)	0.681 (0.024)	*0.015* (0.009)	0.706 (0.006)	0.715 (0.006)	*0.009* (0.004)

Table D.7: Integrated C-Index performance for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std). Best performances are in **bold**, second best in *italics*.

D.1.5 Algorithmic fairness

This section presents the complementary results to Section 6.5.5.Figure D.3 displays the transportability of the considered models using integrated C-index for these groups. These results echo the main text conclusions, with DeepJoint presenting one of the smallest performance drifts in both groups.

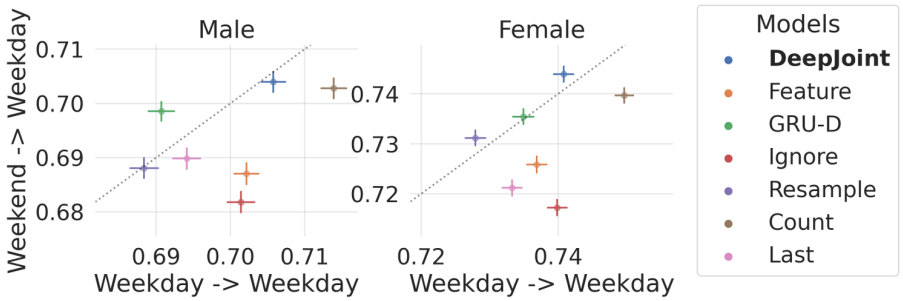


Figure D.3: Discriminative performance evaluated on patients admitted on weekdays for a transferred model (*y-axis*) and a model trained on weekdays-admitted patients and tested on the test set of this same group (*x-axis*) (with 95% CI) stratified by sex.

D.2 Ablation studies

D.2.1 Impact of *I* and *M* networks on performance

In this section, we explore how the different dimensions of clinical presence impact performance. Patients were randomly assigned to train and test sets as presented in the first set of experiments. Based on the laboratory values, mask and time of observations as inputs of the network, we compared all possible combinations of components as Table D.8 summarises.

Model	S	I	M
DeepJoint	•	•	•
DeepJoint$_M$	•		•
DeepJoint$_I$	•	•	
Feature	•		

Table D.8: Components associated to each architecture.

Table D.9 presents the discriminative performance of these different models. Note how each individual model performs similarly to **Feature** baseline based on the same inputs.

Model	C Index			Brier Score		
	7	30	Integrated	7	30	Integrated
DeepJoint	0.760 (0.008)	**0.652** (0.010)	0.739 (0.007)	*0.076* (0.002)	**0.267** (0.011)	**0.127** (0.084)
DeepJoint$_M$	**0.764** (0.008)	**0.652** (0.011)	**0.744** (0.007)	**0.075** (0.002)	**0.267** (0.011)	0.145 (0.068)
DeepJoint$_I$	*0.762* (0.008)	0.650 (0.011)	*0.741* (0.007)	*0.076* (0.002)	0.279 (0.013)	-
Feature	**0.764** (0.008)	*0.651* (0.011)	**0.744** (0.007)	**0.075** (0.002)	0.272 (0.012)	-

Table D.9: Comparison of model performance by means (standard deviations) in the random split experiment of MIMIC III dataset. Best performances are in **bold**, second best in *italics*.

D.2.2 Impact of *I* and *M* networks on transportability

From Theorem (6.1), more uncorrelated tasks are modelled, more robust would be the proposed architecture. Therefore, we investigated the transportability of the previous models. The smaller difference between settings shows that DeepJoint$_I$ and DeepJoint$_M$ improve above the Feature baseline, but their combination outperforms each individual.

	Horizon: 7 days after observation					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	*0.714* (0.022)	**0.715** (0.023)	**0.014** (0.012)	**0.744** (0.006)	*0.741* (0.006)	**0.005** (0.003)
DeepJoint$_M$	**0.715** (0.023)	*0.713* (0.026)	0.019 (0.014)	0.732 (0.007)	**0.743** (0.006)	0.011 (0.006)
DeepJoint$_I$	0.705 (0.022)	**0.715** (0.022)	0.018 (0.013)	*0.739* (0.006)	0.735 (0.006)	*0.006* (0.004)
Feature	0.707 (0.022)	0.697 (0.024)	*0.015* (0.011)	0.722 (0.006)	0.737 (0.006)	0.014 (0.005)

Table D.10: C-Index performance at 7 days for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std). Best performances are in **bold**, second best in *italics*.

	Horizon: 30 days after observation					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	*0.636* (0.028)	*0.639* (0.027)	**0.013** (0.010)	**0.655** (0.008)	*0.644* (0.007)	0.011 (0.005)
DeepJoint$_M$	**0.648** (0.028)	0.637 (0.027)	*0.018* (0.013)	*0.648* (0.008)	**0.648** (0.007)	*0.005* (0.004)
DeepJoint$_I$	0.631 (0.029)	**0.647** (0.028)	0.019 (0.013)	*0.648* (0.008)	0.639 (0.007)	0.009 (0.006)
Feature	0.633 (0.028)	0.614 (0.028)	0.021 (0.012)	0.638 (0.007)	0.641 (0.007)	**0.004** (0.004)

Table D.11: C-Index performance at 30 days for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std).

	Integrated					
	Evaluated on weekends			Evaluated on weekdays		
Model	Transfer	Internal	Difference	Transfer	Internal	Difference
DeepJoint	*0.703* (0.019)	*0.703* (0.020)	**0.011** (0.009)	**0.726** (0.006)	*0.724* (0.006)	**0.004** (0.003)
DeepJoint$_M$	**0.706** (0.019)	**0.705** (0.022)	0.015 (0.011)	0.716 (0.006)	**0.726** (0.006)	*0.010* (0.004)
DeepJoint$_I$	0.696 (0.019)	0.702 (0.020)	0.013 (0.011)	*0.720* (0.006)	0.719 (0.006)	**0.004** (0.003)
Feature	0.697 (0.019)	0.693 (0.020)	**0.011** (0.009)	0.708 (0.006)	0.721 (0.006)	0.013 (0.004)

Table D.12: Integrated C-Index performance for models trained and tested on the same type of patients (Internal), transferred from the other setting (Transfer) and their difference after the first day of observation - Mean (std). Best performances are in **bold**, second best in *italics*.