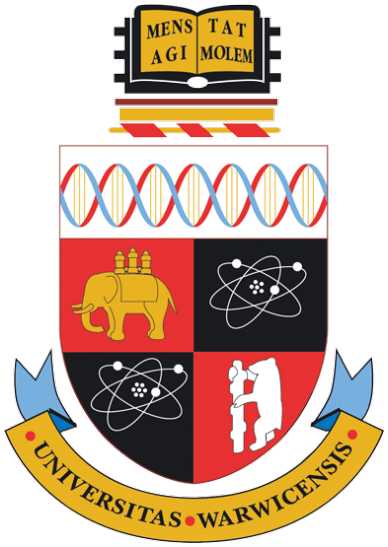
WARWICK
THE UNIVERSITY OF WARWICK



Analysing Clinical Data for Real-World Evidence Generation in Oncology

by

Carlos Serra Traynor

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

School of Engineering

December 2021

Contents

List of Tables

List of Figures

List of Algorithms

Acknowledgments

During this thesis's research and writing process, I have received a great deal of help and support.

I would like to record my thanks to all academic staff members at the University of Warwick who taught me during my research. Each of them made a contribution to my research development. In particular, I wish to thank my supervisors, Prof. Mike Chappell and Dr Neil Evans, for their continual academic support and guidance.

I want to acknowledge my colleagues from my internship at AstraZeneca for their extraordinary collaboration. I would particularly like to single out my supervisors at Da Vinci, Cambridge. Dr Tarj Sahota, Dr Ignacio Gonzalez Garcia and Miss Helen Tomkinson for your patient support and for all the opportunities to further my research.

I would also like to thank my colleagues in D229, School of Engineering, for their advice, provocative questions, and thoughtful discussion.

Finally, I would like to thank my family and friends who have always been there in every step of my career and forever supported me in the most challenging times.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been previously published by the author in the following:

[1] Carlos Traynor, Tarjinder Sahota, Helen Tomkinson, Ignacio Gonzalez-Garcia, Neil Evans, and Michael Chappell. Imputing biomarker status from rwe datasetsa comparative study. *Journal of Personalized Medicine*, 11(12):1356, 2021

[2] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. DNN multiple imputations: Improved imputation accuracy in RWE datasets. *Proceedings of Pharmacokinetics UK conference, Online*, 10-12 November 2021

[3] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian time-varying effect models- a rwe study in oncology. *To submit to Pharmacometrics and Systems Pharamcology*, 2022

[4] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian model comparison of survival models. *To submit to Proceedings of Pharmacokinetics UK conference, Canterbury*, 2-4 November 2022

Research was performed in collaboration during the development of this thesis, but does not form part of the thesis:

[5] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Multi-state model of disease progression and overall survival in nsclc. *Proceedings of the Population Approach Group in Europe, Stockholm*, 11-14 June 2019

[6] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian multistate models in pharmacometrics. *Proceedings of Pharmacokinetics UK conference, Stratford upon Avon*, 6-8 November, 2019

[7] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Elastic net applications to target sourcing in metastatic breast cancer patients. *Proceedings of Pharmacokinetics UK conference, Manchester*, 21-23 November 2018

[8] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. A model-based prediction of survival outcome for breast cancer patients stratified integrative genomics. *Proceedings of the WIN Oncology symposium, Paris*, 25-26 June 2018

Abstract

Randomised clinical trials (RCT) are the bedrock of evidence-based medicine and remain the gold standard in determining the efficacy and safety of investigational new drugs in well-defined populations. They have high internal validity and remain crucial for securing regulatory approval. However, RCTs potentially lack external validity because of the exclusion of subpopulations such as the elderly or comorbidities. Time constraints limit assessing long-term effects, and sample size may be inadequate to identify new biomarkers for personalisation. Real-world evidence (RWE) can complement RCTs' evidence by providing effectiveness and safety data in a wide range of outcomes representative of the everyday clinical setting. Similarly, the concept of real-world data (RWD) is typically associated with big datasets that advance current medical practice towards personalisation. However, if used only to predict the most beneficial treatment choice, the best-case scenario with RWE analysis could match the current medical practice. The key challenge in analysing RWD is that individualised treatment effects are never observed. Therefore, its non-randomised, observational nature is prone to biases from unrecognised factors. To properly use RWD requires finding better solutions to the unique challenges of working with clinical data: (1) a significant amount of missing data, (2) heterogeneous data, (3) seldom exist a ground truth. This dissertation addresses these specialities constructing a formal causal inference framework to enhance the statistical analysis of RWD. We focus on three problems:

- Missing data imputation
- Accurately predicting the consequences of treatment in biomarker-defined populations

- Assessing how conclusions might change in the presence of hidden factors

To appropriately tackle these problems, we propose new methodologies for RWD analysis:

1. We formalise the missing data problem to design a machine-learning algorithm to perform missing data imputation.

2. We develop Bayesian modelling techniques for treatment effects heterogeneity of survival outcomes introducing a new methodology named survival Gaussian processes, which are particularly well-suited for distributed varying treatment effects inference.

3. We extend the Bayesian approach to infer causal bounds for time-varying effects probabilistically.

To demonstrate the technique's utility, we analyse two large real-world cohorts of non-small cell lung cancer patients with epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK), kirsten rat sarcoma (KRAS), B-RAF proto-oncogene (BRAF), and immunotherapy marker programmed death-ligand 1 receptor (PD-L1) status of biomarker and treated with immune checkpoint inhibitors (ICI). The first study tackled the missing data problem by developing a new imputation algorithm for multiple imputations in synthetic and real-world examples of biomarker status missingness. The second study covered the impact of ICI in the survival time of NSCLC patients stratified by PD-L1 expression, handling missing data with the first study's results, and embracing the Bayesian approach for modelling heterogeneous treatment effects and time-varying effects. We show that the proposed methods outperform state-of-the-art methods for missing data imputation in complex datasets with non-linearities, pooling across PD-L1 per cent staining difference with Gaussian processes achieves better out-of-sample performance than conventional interaction models and estimates of causal bounds are critical for understanding the impact of unobserved confounding in analysing RWD.

Sponsorships and Grants

Grateful acknowledgements for the joint Engineering and Physical Sciences Research Council Doctoral Training Partnership Award (EPSRC) and AstraZeneca PhD Studentship funding.

Acronyms

advNSCLC advanced non-small cell lung cancer.

AIC Akaike information criterion.

ALK anaplastic lymphoma kinase.

ANN artificial neural networks.

APC antigen presentation cells.

ATE average treatment effect.

BRAF B-RAF proto-oncogene.

CATE conditional average treatment effect.

CP carboplatin, pemetrexed.

CPP carboplatin, pembrolizumab, pemetrexed.

D durvalumab.

DAG directed acyclic graph.

DNN deep neural networks.

ECOG Easter Cooperative Oncology Group.

EGFR epidermal growth factor receptor.

ELPD expected log-predictive density.

EM expectation-maximisation.

GAIN generative adversarial imputation networks.

GAN generative adversarial networks.

GLM generalised linear model.

GLU gated linear units.

GP Gaussian process.

HMC Hamiltonian Monte Carlo.

ICI immune checkpoint inhibitors.

IO immuno-oncology.

IPW inverse probability weighting.

ITE individualised treatment effect.

ITS inverse transform sampling.

KM Kaplan and Meier maximum likelihood.

KRAS kirsten rat sarcoma.

LOSO leave one subject out cross-validation.

MAR missing at random.

MCAR missing completely at random.

MCMC Markov chain Monte Carlo.

MI multiple imputations.

MICE multivariate imputations by chained equations.

MIPCA multiple imputation principal component analysis.

MIRF multiple imputation random forest.

MITABNET multiple imputations with tabular networks.

MNAR missing not at random.

MS M-splines.

MSE mean squared error.

MS-PH M-splines proportional hazards.

N nivolumab.

NOS not otherwise specified.

NPH non proportional hazard.

NSCC non squamous cell carcinoma.

NSCLC non-small cell lung cancer.

ODE ordinary differential equation.

OTS optimal treatment selection.

P pembrolizumab.

PD-1 programmed death co-receptor 1.

PD-L1 programmed death-Ligand 1.

PH proportional hazard.

PMM predictive mean matching.

PPC posterior predictive checks.

PriorPC prior predictive checks.

RCT randomised clinical trial.

RMSE root mean squared error.

RMST restricted mean survival time.

RWD real-world data.

RWE real-world evidence.

rwOS real-world overall survival.

rwTTD real-world time to treatment discontinuation.

SCC squamous cell carcinoma.

SCM structural causal model.

SIP specific informative prior distribution.

SUTVA stable unit treatment value assumption.

TABNET tabular networks.

TET treatment effect on the treatead.

TFE time fixed effect.

TMB tumour mutational burden.

TVE time varying effect.

VPC visual predictive checks.

WAIC widely applicable information criterion.

WIP weakly informative prior distribution.

Symbols

D	observed data
A	set of treatment variables
Y	set of outcome variables
W	set of measured individual variables
X	set of measured individual variables of interest
T	time variable
T^*	observed event time
Y^a	counterfactual outcome on treatment $A = a$
$h_o(t)$	the baseline hazard function
$S(t)$	the survivor function
λ	the link function
γ	auxiliary parameters, such as Weibull's shape parameter; tunnin parameter in attentive transformer (ANN context)
d	censoring indicator function
$\perp\!\!\!\perp$	the conditional independence symbol
$\Rightarrow$	the implication symbol
$\rightarrow$	Arrow in a DAG representing causal relationships between variables.
θ	model's finite number of parameters
I	indicator function
$\sim$	relation of a random variable to its probability distribution
$\mathbb{E}$	expected value
μ	mean function
σ^2	variance.
σ	standard deviation; sigmoid function (ANN context)
df	degrees of freedom, maximum number of independent variables.
Pr	probability
$J(\theta)$	cost function.
K	covariance matrix
ρ	correlation; volatility (GP context)

α	alpha levels related to confidence intervals; Intercept parameter (Linear regression context); amplitude (GP context); the lowest possible outcome (OTS context); tunning parameter of dropout layer (ANN context)
β	log-hazard ratio (hazard regression context); the highest possible outcome (OTS context)
Δ	difference
Normal	Normal distribution
Student's t	Student's t distribution
Exponential	exponential distribution
Weibull	Weibull's distribution
$\boldsymbol{B}$	B-splines function
$\boldsymbol{M}$	M-splines function
k_0	splines knots

Chapter 1

Introduction

The rapid adoption of real-world data (RWD) presents unprecedented opportunities to advance clinical research in oncology. Since 2008, the adoption of electronic health records in the United States has grown 9-fold, from 9.4% in 2008 to over 84% in 2016 [9]. The data collected as part of the clinical management of cancerous diseases include clinical assessments, patients' general well-being, laboratory tests such as biomarker tests, proteomics and genomics data. These data represent the patient outcomes, and data scientists can analyse them to address oncology research challenges and study the treatment response to anticancer drugs.

In addition to having data, there have been notable breakthroughs in machine learning and computational statistics to quantify real-world data patterns, such as automatic differentiation [10], high-dimensional sampling algorithms for automatic inference [11], and symbolic math libraries [12, 13]. However, it is only recently that we have started to see translations of these research ideas into clinical practice [14]. To fully optimise the opportunities that RWD present, researchers need to transform machine learning into discovery tools by taking on complex problems such as conducting robust inference from incomplete, missing data, censored survival analysis [15], and heterogeneous data [16].

First, missing data is a universal problem in analysing RWE datasets. In RWE datasets, there is a clinical interest to understand which covariates best correlate with clinical outcomes, such as disease progression or survival. A substantial difficulty in real-world settings is that several covariates may be missing in the dataset, hiding meaningful information in the study. There are many practical implications when missing data are present; for example, it can lower the power, affect the precision of the confidence intervals for parameter estimates, and lead to biased estimates. Hence, excluding the underlying value of missing data may invalidate the results.

Second, analysing RWD requires reasoning about causality. Most research

in machine learning has resulted in predictive problems [12]. However, if we use the current clinical RWD to predict the most beneficial treatment choice, the best-case scenario could potentially match the current medical practice. Often the questions that we want to answer when it comes to RWD are not predictive, but causal. Clinicians expect to predict patient outcomes based on the consequences of treatment interventions and patient characteristics.

Third, clinicians place particular interest in biomarkers that modify the treatment effect. These biomarkers can be effectively used for treatment personalisation and are essential for developing new anticancer drugs [17]. RWE increasingly has the utility for biomarker sourcing and early patient stratifications by analysing large datasets to adequately identify new biomarkers for personalisation. However, individualised treatment effects are not available. Therefore, one must design a causal model to use observational RWD to tackle individualised treatment effects questions.

The work in this thesis tackles the problems introduced here in analysing RWE datasets to enhance clinical research in oncology. We separate the thesis into theoretical and applications chapters to define a framework to analyse heterogenous treatment effects from observational data. The second part advances the applications in RWD and their respective contributions.

1.1 Aims and Objectives

This thesis aims to enhance the usability of RWD by extending machine learning methods for the inference of personalised treatment effects from observational RWD. To do so, we explore how to use data-driven models with very flexible functional forms, such as neural networks [12], ensemble learning [16], and kernels [18] to model individualised treatment effects improving real-world model performance.

Currently, several limitations exist for the adequate application of RWE, especially when there are missing values, many interaction effects or unobserved confounders. The optimal use of RWD needs the development of methods to handle the following challenges:

- Missing data that appear as gaps in the dataset that hide meaningful values for analysis.
- Patients may respond differently to treatments, and better methodologies are necessary to model heterogeneous data.
- Unrecognised factors potentially overestimate bias of treatment effects.

Therefore, one must develop a bottom-up approach to tackling these challenges by defining a fundamental causal framework with causal assumptions in a form

that explicitly invokes the systems properties. Furthermore, the framework must assess the sensitivity to violations of the assumptions, such as the impact of unmeasured factors or confounders. To do so, we have the following **objectives**:

- Address the missing data imputation problem and propose a systematic approach for the comparison of imputation methods on RWE datasets.
- Present a step-by-step strategy for developing a new imputation algorithm by adapting state-of-the-art machine learning tools.
- Recognise the impact of heterogeneous treatment effects in new therapeutic modalities such as immunotherapy and develop new methods for predicting treatment response making efficient use of the available data with sparse biomarker measurements.
- Expand methods for modelling and simulation survival outcomes for typical RWE datasets where non-linear treatment effects are present, suggesting that treatment acts as a selection force.
- Develop a framework for conducting sensitivity to unrecognised factors in survival analysis with censored data.

1.2 Summary and Contributions

The following sections summarises this thesis' chapters and our contributions therein:

Chapter 2 : Provides an introduction to causal inference by reviewing a selection of papers in biomedical science. This thesis's novelties include using causal inference tools to evaluate new imputation algorithms and perform survival analyses that are general for RWE studies. In chapter 2, the author introduces the topics of causal inference, missing data and survival analyses, conducting a background and literature review of a list of critical publications on those topics.

Chapter 3 : Defines the building blocks of causal inference: causal estimands with causal models, statistical estimands with data and corresponding parameter estimation and identification. The foundations of causal inference presented allow us to discuss average treatment effects, treatment effects modification, and individualised treatment effects, which will apply in the RWE studies developed by the author. Further, we discuss the frequentist and Bayesian paradigms to statistical inference with two pinnacle examples, the

Hamiltonian Monte Carlo algorithm [11] and the backpropagation algorithm [12, 13]. Finally, we discuss the topic of sensitivity analysis and causal bounds.

Chapter 4 : Develops a new imputation algorithm named MITABNET, which extends recent advances in tabular learning with ANNs [19] for performing multiple imputations with mixed data types. The author contributes to developing MITABNET and finding real-world applications in analysing a large RWE dataset of more than 35,000 NSCLC patients with partially measured biomarker statuses for epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK), kirsten rat sarcoma (KRAS), B-RAF proto-oncogene (BRAF), and immunotherapy marker programmed death-Ligand 1 (PD-L1). Using the RWE dataset, we compare the imputation performance of MITABNET with state-of-the-art methods on missing data: predictive mean matching, expectation-maximisation, factorial analysis, random forest and generative adversarial networks. We also conduct extensive synthetic data experiments with structural causal models.

Chapter 5 : Illustrates a causal survival analysis using a Bayesian general formula to estimate treatment effects with censored outcomes. The author investigates the concept of non-linear covariate effects modelling with the survival Gaussian process (GP). Furthermore, it discusses the pitfalls of conventional proportional hazard models in analysing RWD examples and proposes solutions by using flexible parametric hazard models and time-varying effects models from recent literature. Finally, the author expands methods to assess the impact of unobserved factors on RWE studies and contributes by extending the optimal treatment selection causal bound in survival analyses. The author applies the new techniques to an RWE dataset cohort of advanced non-small cell lung cancer (advNSCLC) patients treated with double-platinum chemotherapy or immunotherapy.

Chapter 6 : Provides conclusions for this thesis by reviewing our contributions in advancing the concept of personalised treatments in oncology by enhancing the use of RWD. It provides a review of the most significant results in our applications of new techniques for missing data and causal survival analysis in RWE datasets and assesses the extent to which this work accomplishes its aims and objectives. Finally, it discusses potential applications of the work presented and future research directions.

Chapter 2

Background and Literature Review

The rapid adoption of RWD from the biomedical industry, research institutions and regulatory bodies has enhanced cancer drug development by transforming the generation of evidence, augmenting clinical decision making, and supporting medical management in oncology. RWD encompasses different data types that are not collected in conventional randomised controlled trials, including but not limited to medical records, hospital data and demographic and social indicators. Researchers can leverage large real-world datasets with machine learning tools to advance the concept of personalised therapies by, for instance, identifying specific treatments' unique biomarkers [20]. Statistical modelling and machine learning techniques recognise and quantify real-world data patterns, contributing to understanding cancerous diseases' basis and trajectories. In this way, real-world evidence provides valuable inputs to inform and improve clinical pathways [21]. To fully optimise the opportunities that RWD present, researchers need to transform machine learning into discovery tools by taking on complex problems such as censored survival analysis [15], learning from pharmacological time-series data [22], inclusively data that are missing at random [23] and data that are not missing at random with informative missingness.

2.1 Causal modelling

RWD are observational, regularly retrospective, and biased by clinical practice, experimentation with patients is limited, and hence, there are multiple challenges in modelling. Often the questions that we want to answer when it comes to RWD are not predictive but causal. Furthermore, if we use the current RWD to predict the most beneficial treatment choice, the best-case scenario could match the current medical practice [24]. To go beyond clinical

practice and recognise the heterogeneity in treatment response, we need to change the question to a causal one. Based on a causal model, we might guide treatment decisions; however, individuals may respond differently to a treatment. Therefore, we need to design how to use observational RWD to answer cause and effect queries.

The causal modelling procedure aims to separate causal patterns from mere associations [24–28]. To develop a causal model, we use a causal structure, representing the directionality of cause and effect in a system. From the generated causal model, we can reason about the impact of interventions or distribution changes by eliciting potentially observed outcomes. Consequently, when assessing the intervention's effect, we assume that the causal model is known and precisely compute the situations that have not occurred, or "counterfactuals outcomes" [29–31]. To be explicit in the difference between causation and association, consider the following simple example adapted from [32] represented in figure 2.1: a population with some individuals treated and some individuals not treated. On the one hand, association compares the observed outcome on the population's treated part against the observed outcome on the population's not-treated part. On the other hand, causation compares the hypothetical outcome of treating the entire population against the hypothetical outcome of not treating the entire population, i.e. the counterfactual outcomes.

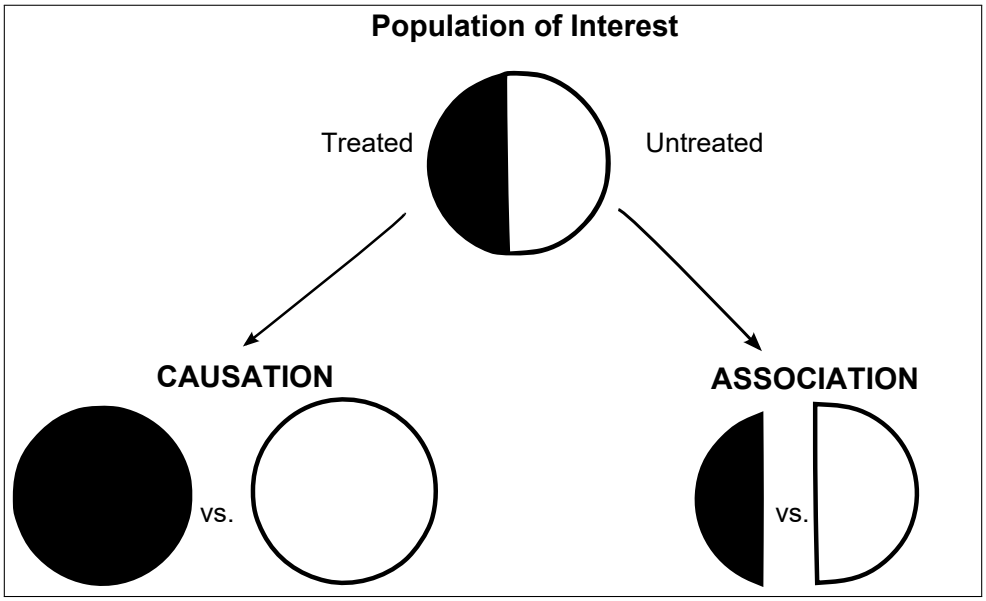


Figure 2.1: Illustration of the difference between association and causation: association compares the outcome observed on the population's treated part versus the outcome observed on the population's not-treated part. Causation compares the hypothetical outcome of treating the entire population versus the hypothetical outcome of not treating the entire population.

To accurately answer these questions, we need to formulate them as struc-

tural causal model (SCM) have equivalent formulations via causal graphs [33–35] and structural equations [36, 37]. Classically, the SCM framework does not consider feedback between variables, which is especially true for cross-sectional studies, and outlines the SCM's causal graph via DAG, first developed by [33] and refined by [38]. [1] DAGs typically have two components: nodes and arrows (directed edges). The nodes represent variables (unobservables), and the arrows causal relationships between them. Analogously, we can derive the SCM's structural equation. In the simplest possible case, an SCM has two random variables X and Y, its DAG takes the form $\mathrm{X} \rightarrow \mathrm{Y}$, and its structural equation is given by:

$$\begin{aligned} X &:= f_X(U_X) \\ Y &:= f_Y(X, U_Y) \end{aligned} \tag{2.1}$$

where X is the cause of Y, and U_X, U_Y are independent and identically distributed (i.i.d) random variables. Generally, Eq. 2.1 is solved by regressing X on Y. Adding another node Z, and two arrows from Z to Y and to X respectively,such that $\mathrm{X} \leftarrow \mathrm{Z} \rightarrow \mathrm{Y}$, gives another simple SCM known as the "fork" diagram [40], its structural equation given by:

$$\begin{aligned} W &:= f_W(U_W) \\ A &:= f_A(U_A, W) \\ Y &:= f_Y(U_Y, W) \end{aligned} \tag{2.2}$$

where W is a common cause of A and Y and generates a spurious correlation. We say that the random variable W *confounds* the impact of A on Y. Once we recognised the confounds in an SCM, we can de-confound each algorithmically using the back-door criterion [41]. To de-confound Eq.2.2 we need to collect data on W to correspondingly adjust, and show that $\mathrm{A} \perp\!\!\!\perp \mathrm{Y}|\mathrm{W}$, read A is independent of Y given W. Consequently, as long as we know the causal graph, we can learn the treatments' causal effect. Unfortunately, the causal graph is not generally known [35]. Moreover, several SCM may give rise to the same probability distribution and the same dataset, an inverse problem discussed in [35, 42]. Research in causal inference is about being transparent about our assumptions, such as drawing them graphically, so peers understand them and minimise the number of assumptions we need.

[1] Feedback may exist in nature, especially when considering the time-domain, e.g. on time series [39], DAG is still a useful construct to outline complex SCM.

Confounding bias

Confounders are variables that affect the treatment decision and also affect the outcome independently of treatment, i.e. not through its impact on treatment. Let us return to the toy SCM from equation 2.2, where there are three sets of random variables: **W**, including all confounding factors and possibly high-dimensional; A indicates a binary treatment; Y, the outcome of interest, e.g. survival time. In this SCM, **W** is a confounder of the impact of A, which takes values in $[a, a']$, on Y. If our interest is in evaluating the impact's strength of treatment A on outcome Y in the presence of confounder **W**, we obtain:

$$\mathbb{E}\left[\mathrm{Y}^{a} - \mathrm{Y}^{a'}\right] = \mathbb{E}_{W}\left[\mathbb{E}\left[Y|A = a, W\right] - \left[Y|A = a', W\right]\right] \tag{2.3}$$

also known as the adjustment formula, or the back-door criterion [38]. The notation Y^{a} indicates the outcome we would observe if we set the treatment to A=a, and the notation $\mathbb{E}$ the expected value and denoted what to expect, on average, over many repetitions.

Using the adjustment formula 2.3, one can identify causal effects such as the average treatment effect (ATE), given that one measures all the confounders **W**. Therefore, to de-confound one needs to collect data on **W**. The causal modelling framework may be complex, including any number of variables to describe more complicated scenarios. Still, only four kinds of confounding arrangements can ever arise in a DAG, see depicted in figure 2.2 adapted from [43].

Fork : The confounder figure 2.2a is a variable that is a common cause of two others [43]. We have seen an example of this junction known as the fork already back in equation 2.2. Our interest is in measuring the impact of A on Y in the presence of confounders **W**. **W** causes a spurious association between A and Y unless adjusted using the adjustment formula 2.3. To evaluate whether there are unmeasured confounding variables, we need to cooperate with clinicians and leading experts to learn what factors affect treatment decisions and measure all confounding forks.

Collider : A collider C, shown in figure 2.2b is a common result of A, and Y. Conditioning on C creates a statistical (spurious) association between A and Y, such that $A \perp\!\!\!\perp \mathrm{Y}$, but $A \not\perp\!\!\!\perp \mathrm{Y}|\mathrm{C}$. A significant inferential effect of the collider is selection bias [44]. Unmeasured variables can also create colliders [40, 43].

Mediator : A mediator M, shown in figure 2.2c mediates the association between A and Y. Conditioning on M removes dependency between A and Y, such that $A \perp\!\!\!\perp \mathrm{Y}|\mathrm{M}$. A significant inferential effect of controlling for M

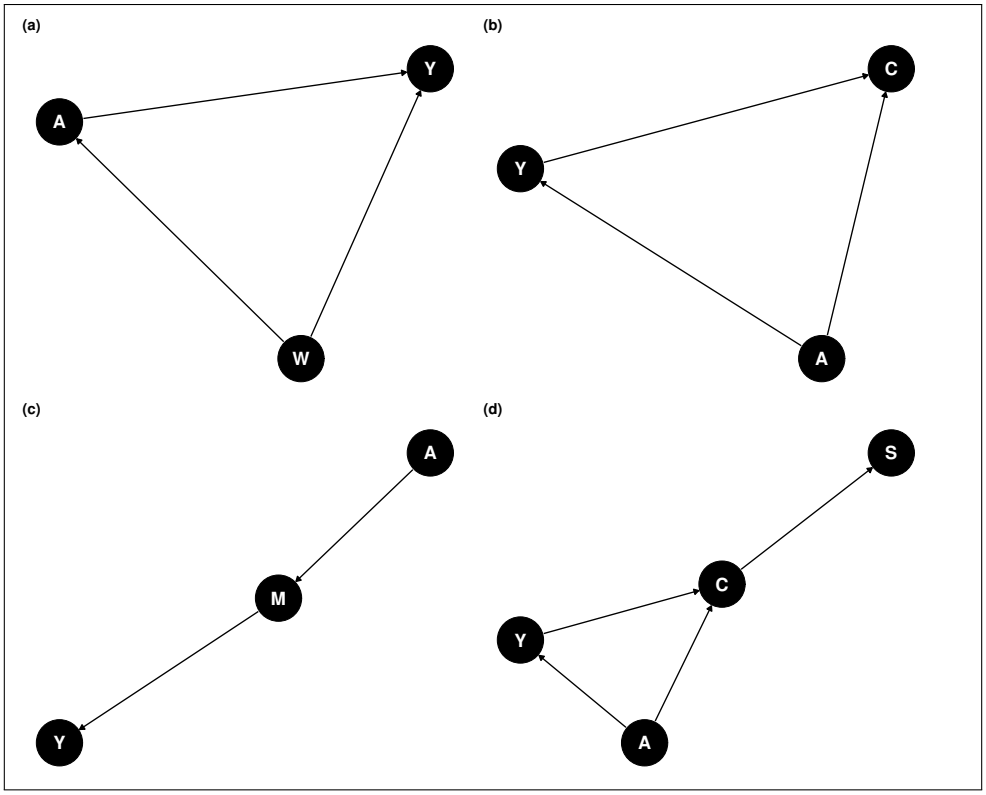


Figure 2.2: Examples of typical SCM's causal graphs: **(a)** represents the structure of confounding where W is a common cause of *A* and Y; **(b)** a collider C is a common result of W, and Y. Conditioning on C creates a statistical (spurious) association between W and Y. A significant effect of the collider is selection bias. **(c)** A mediator M mediates the association between W and Y. A significant inferential effect of controlling for M is post-treatment bias. **(d)** S is a descendant of C, and conditioning on S is like conditioning on C.

is post-treatment bias, controlling for the consequence of treatment [43]. For example, post-treatment bias is an inferential threat in RWD analysis [45] if covariate adjustment is not principled in not including any post-treatment variables or treatment consequence.

Descendant : The descendant S shown in figure 2.2d is a fourth variable resulting from either a confound, mediator or collider. Conditioning on S is like weakly conditioning on its parent variable. Descendants are akin to proxies when information is not available on the confounder [43].

There is a framework that unites all these examples called the back-door criterion [34, 38, 43]. The back-door criterion's general idea is to solve the causal graph given by a DAG and discover the causal impact of some exposure on some outcome. Then we need to shut all back-door paths from that exposure to the outcome. In experimental studies, one shuts back-doors by intervening on variables, e.g. randomisation. In observational studies and RWD, we need some criteria for which variables to include and shut the back-door paths.

2.1.1 Causal inference techniques utilised

Upmost related to our work is recent literature bringing together machine learning techniques and causal inference for modelling treatment interventions in RWD applications [46–48], which include applications of the central causal techniques selected in our work:

1. Average treatment effect: This allows us to define a causal population-level impact. It is helpful to determine the difference in the mean causal effect of treatment over the whole population of interest in RWE studies.

2. Conditional average treatment effect: This allows us to stratify a population into subpopulations by a covariate of interest in RWE studies. We can conduct biomarker-based treatment effect modification or heterogeneity of treatment effects.

3. Outcome modelling allows us to estimate treatment effects from observed data in RWE studies by estimating the model parameters conditional on the outcome.

4. The propensity score allows us to estimate treatment effects in a complementary way to the outcome modelling via inverse probability weighting.

5. Causal bounds and sensitivity analysis estimate the impact of unobserved confounding bias on our treatment effects estimates.

Section 3.2 describes the central assumptions to allow identifiability of causal effects, section 3.3 the method to build statistical estimands with data, and section 3.4 parameter estimation and identification of treatment effects.

2.2 Missing data

One of the biggest challenges when working with RWD is how to handle missing data. Missing data appear as gaps in the dataset that hide meaningful values for analysis. Hence, excluding the underlying value of missing data may completely invalidate the results. There are several other practical implications when missing data are present; for example, it can lower the power and affect the precision of parameter estimates' confidence intervals.

Missing data analytical methods date back to the 1960s and 1970s when [23, 49, 50] introduced the concept. [49] described a systematic approach using a factored likelihood estimation for simple problems with missing data. [23], on the other hand, established the concept of missing at random where the central idea was to model the missingness mechanism. Before the seminal work of [23, 51, 52], the tendency was to handle missing data with simple ad-hoc methods: discarding the data with any missing values, plug-in the mean or last-observation carried forward. However, [40, 53] brought attention to the concerns with off the shelf handling of missing data methods such as squandering information in excluding all cases with any missing values, create confounds, and model misspecification.Missing data imputation for large RWE datasets is a topic of active research [54, 55]. The field focus' is in complex models, such as latent class models for categorical data [56], bagging of regression trees [57], random forest [58] and artificial neural networks [59].

2.2.1 Missingness mechanisms

In missing data, the observed covariates of interest X are incomplete. The missingness mechanism R^X places the missing values in X^{obs} masking the actual value X, i.e. following [54] R^X is a random variable taking values in $[0, 1]$, so that:

$$X_i^{obs} = \begin{cases} X_i & \text{if } R_i^X = 0 \\ \text{n.a.} & \text{if } R_i^X = 1 \end{cases}$$

Rubin [23] defined three possible scenarios for missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). All types of missingness can be classified usefully into this taxonomy, which indicates the most appropriate procedure to make an unconfounded inference [60].

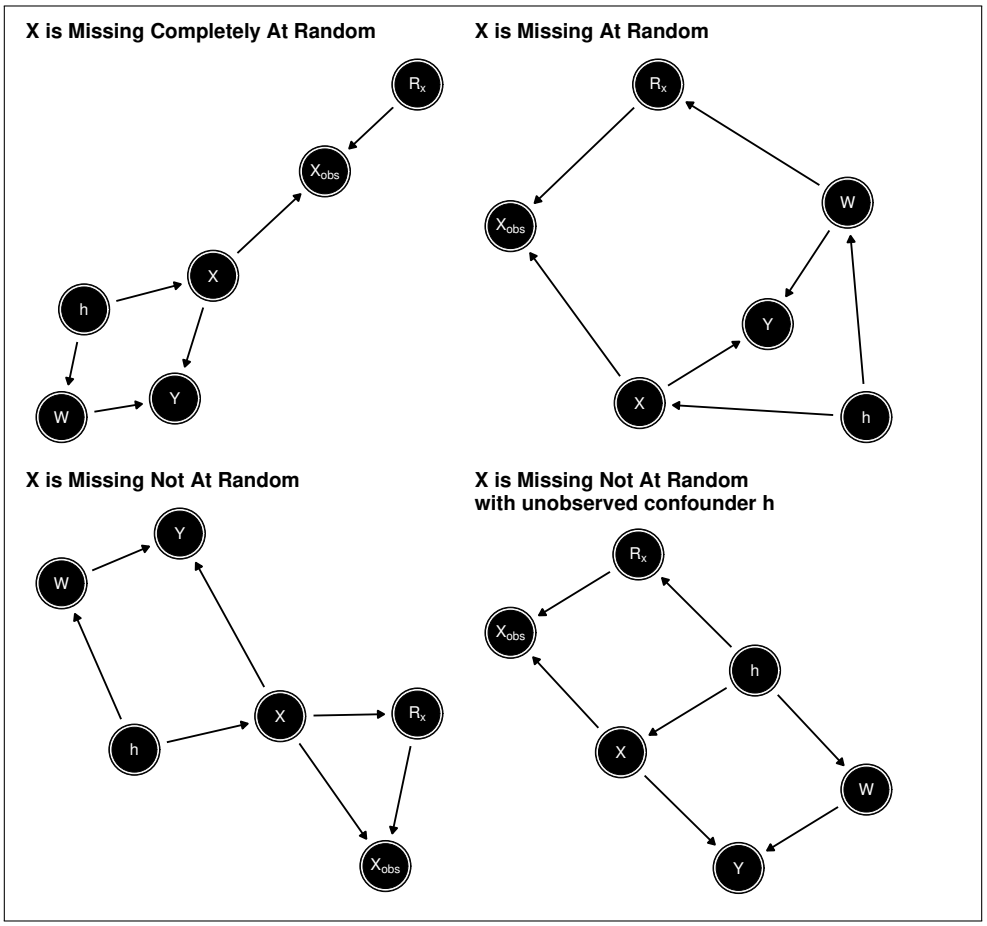


Figure 2.3: Illustration of missingness mechanisms: Missingness Complete at Random (MCAR), Missingness at Random (MAR), and Missingness Not at Random (MNAR).

MCAR

We say that the data are MCAR if the complete dataset X do not influence the mask R^X, i.e. X are independent of R^X. In MCAR, the missingness mechanism R^X is ignorable because no causal back-door paths exist from the observed variable X^{obs} to Y. Therefore, we can condition on X^{obs} for direct effect. Imputation is not mandatory for unbiased estimation, yet it adds precision [60]. In large sample theory, we may test for MCAR from the data by conducting Little's test [61], which compares the change in the empirical mean of measured variables X^{obs} if removing cases with missing values. Since in MCAR, the missingness mechanism R^X is independent of the actual value X, Little's test assumes that R^X does not change the overall distribution of X^{obs}. Summing up, a non-significant result of Little's test suggests that the MCAR assumption is valid. For our running example, see figure 2.3 top left, the SCM of MCAR is given by:

$$\begin{aligned} X^{obs} &:= f_{X^{obs}}\left(R^X, X\right) \\ R^X &:= f_{R^X}\left(U_{R^X}\right) \\ X &:= f_X\left(U_X, h\right) \\ W &:= f_W\left(U_W, h\right) \\ h &:= f_h\left(U_h\right) \\ Y &:= f_Y\left(U_Y, W, X\right) \end{aligned} \tag{2.4}$$

where the missingness mechanism R^X is i.i.d and h is an unobserved confounder that does not impact the missingness mechanism.

MAR

In MAR, other observed variables W are influencing the missingness mechanism. The consequences of MAR are diverse but addressable by controlling for other known variables and imputation [40]. MAR is different to MCAR in terms of its consequences. We need to impute to avoid polluting other variables with the associated missingness pattern [60]. MAR is closely related to the concept of confounding bias, see the section 2.1, where we collect data on W.

Nevertheless, the MAR assumption may not always hold. For example, a hidden variable may confound the missingness mechanism. Such circumstances propmpt the modeller to conduct sensitivity analyses, the topic of section 3.5. For our running example, MAR opens a causal back-door path from X^{obs} to Y that needs conditioning on W to d-separate and impute to de-bias estimates. For our running example, see figure 2.3 top right, the SCM of MAR is given

by:

$$\begin{aligned} X^{obs} &:= f_{X^{obs}}\left(R^X, X\right) \\ R^X &:= f_{R^X}\left(U_{R^X}, W\right) \\ X &:= f_X\left(U_X, h\right) \\ W &:= f_W\left(U_W, h\right) \\ h &:= f_h\left(U_h\right) \\ Y &:= f_Y\left(U_Y, W, X\right) \end{aligned} \tag{2.5}$$

where the missingness mechanism depends on measured variables W. In analysing RWD, it is helpful to assume MAR because it allows handling missing data with general-purpose imputation algorithms, the central topic of chapter 4.

MNAR

In MNAR, unobserved values impact the missingness mechanism. These unobserved values can be missing values of partially observed variables or unobserved hidden variables. The MNAR scenario is a case of confounding bias, where we have not got data available on the confounder. Therefore, if MNAR is present, it means that X^{obs} may induce a systematic bias [60]. Possible resolutions are to model the MNAR mechanism from prior knowledge, collect more data on the missing variable, or collect information on other child variables, such as the number of visits or disease severity, to model MNAR as a MAR scenario. For our running example, we show two MNAR situations where the missingness mechanism depends on X itself, see figure 2.3 bottom left, and another that arises through unobserved variables h, see figure 2.3 bottom right.

2.3 Censored data and time-to-event analysis

After handling missing data in section 2.2, we will discuss modelling treatment interventions on patient outcomes, such as survival time. Given its longitudinal nature, RWD is often used for various types of time-to-event analyses [62]. The time-to-event analysis also referred to as survival analysis, is a modelling approach used for estimating the time to an event of interest in a population-based or a sample from that population. Time-to-event methods are instrumental when some samples are censored. Censoring occurs when the event of interest does not occur in the window of observations. Censoring may, for example, occur if:

1. The database release is before the event happened.

2. The patient is lost to follow-up.

3. The event record is upon consultation on a calendar basis.

Example 1 is administrative right censoring, 2 competing risks, and 3 interval censoring. In the right censoring setting, some individuals' observed times T^* are lower than their actual event time T. For example, figure 2.4.a shows that enrollment to database release is shorter than the time-to-event for some patients. Because from right-censored survival data, we can not estimate the mean time-to-event directly, $\mathbb{E}\left[\hat{T}\right]$, we favour estimands that can accommodate censoring such as survival, risk and hazard. The *survival probability*, or the

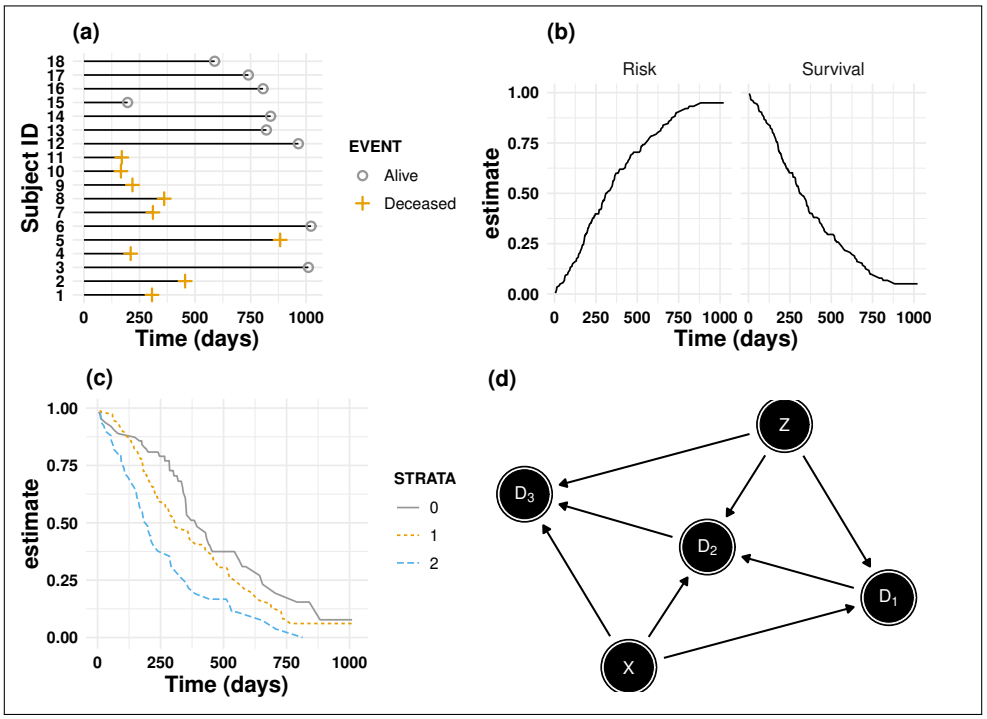


Figure 2.4: Illustration of conventional survival analysis concepts: **(a)** illustrates the patient time for 18 patients in a time-to-event study, alive patients at last-follow-up are right-censored; **(b)** outlines the estimates of risk or cumulative and survival, which are respectively monotonic increasing and monotonic decreasing; **(c)** shows conditional survival curves on a treatment level; **(d)**, a directed acyclic graph (DAG) depicting the selection bias incurred in analysing hazard ratios, here X is a baseline exposure and Z a set of baseline covariates, D_1, D_2, and D_3 are event indicators for the discrete times 1,2, and 3.

survival at time t, is the proportion of patients that have survived beyond time t, and is given by:

$$S(t) = \Pr[T > t] \tag{2.6}$$

A survival curve is a plot of the survivals at each observation time until the last follow-up (database release) over time. Figure 2.4.b outlines such a survival

curve, which starts at $\Pr[T > 0] = 1$ at time 0, to then decreases monotonically, that is a survival curve stays flat or decreases. Similarly, we can define a risk function, or cumulative incidence, see figure 2.4.b. The cumulative incidence describes the proportion of patients that have developed the event before t, and it is defined as one minus the survival probability:

$$1 - \Pr[T > t] = \Pr[T < t] \tag{2.7}$$

Therefore, we say that the cumulative incidence increases monotonically or that it is strictly non-decreasing.

A conventional approach to measure treatment effects is to compare survival under each treatment group. A statistical procedure to compare survival curves is the log-rank test [63], which assesses if the curves differ significantly. However, in RWD with confounds, a comparison of survival curves may not have a causal interpretation. We need to model the survival function $S(t)$ instead. It turns out that modelling the event rate, or hazard, is more straightforward than directly modelling the survival.

The conventional definition of the hazard function $h(t)$ is the event's immediate risk [64]. If we assume that time measurements are continuous, we can use calculus to obtain the equality:

$$h(t) = -\frac{d}{dt}\log(S(t)) \tag{2.8}$$

However, time measurements are always discrete from finite data samples, e.g. days, weeks, months. In this case, it is more intuitive to use the discrete-time hazard [2] $\Pr[T = t | T > t - 1]$, which is the proportion of patients who develop the event among those who have not developed it before t, such that:

$$h(t) = \frac{S(t_{d-1}) - S(t_d)}{S(t_{d-1})} \tag{2.9}$$

where $S(t_d)$ is the survival at the dth measurement. Hence, the hazard may increase or decrease over time because its numerator and its denominator are time-varying, see equation 2.9. Its numerator is the number of events at dth time step. Its denominator is the number of patients who have not developed the event at $d - 1$, also known as the risk set, which changes with time by definition. Hence, time-varying effects on the hazard scale are worth considering even when analysing survival data from cross-sectional studies, i.e. not considering time-varying treatments. For example, [15] considers the interaction of baseline covariates with time-to-follow-up to adjust for the time-varying confounding built-in in hazard models.

[2] The relation between discrete-time hazard and discrete-time survival is similar to the continuous case.

A common approach to comparing treatment effects in RWD across one or more covariates is to model the indiviudal hazard [65]. For example, consider the Cox proportional hazard model [66], given by:

$$h_i(t) = h_0(t) \exp\left[\sum_{p=1}^{P} \beta_p x_{ip}\right] \tag{2.10}$$

where $h_0(t)$ is the baseline hazard, x_{ip} is an array of individual P individual covariates, and β are the p regression coefficients, or the $\log-$*hazard ratios*. The hazard ratio is a popular measure of the association of treatment with patient outcome [67]. The hazard ratios are obtained from $\exp\left[\beta_p\right]$. When comparing treatment groups, x_p is a categorical variable, and its hazard ratio is precisely a ratio of hazards, e.g. the treatment arm's hazard between the control arm's hazard. A hazard ratio lower than one indicates that the treatment is beneficial, such that the treatment group survival is longer on average, while a hazard ratio higher than one indicates that the treatment is harmful, such that the treatment group survival is shorter on average. Regularly hazard ratios are interpreted as a relative risk [68] but are not the same. Relative risk reduction refers to the total amount of events at the end of the study, while the hazard ratio represents an average of the difference in survival across all the follow-up time points.

2.3.1 Parametric hazard modelling

As explained above, survival analysis models the time between an index date and an event of interest. For the baseline hazard $h_0(t)$ that may vary in time, one can evaluate canonical parametric distributions, such as exponential and Weibull, which baseline hazard distributions are given by:

$$\begin{aligned} \text{Exponential} \ : \quad & h_0(t) = 1 \\ \text{Weibull} \ : \quad & h_0(t) = \gamma t^{\gamma - 1} \end{aligned} \tag{2.11}$$

where γ denotes the Weibull shape parameter.

In analysing real-world survival data often is appropriate to make weaker parametric assumptions about the underlying structure of the baseline survival and hazard functions. Indeed, clinicians usually prefer non-parametric Kaplan-Meier survival estimates [69] or semi-parametric Cox models [66] because they make weak assumptions about the baseline hazard. Flexible parametric models for survival analysis became popular with the seminal study of Royston, Parmar et al. [70]. The Royston-Parmar model allows us to model unstructured log-hazard function via baselines-splines (B-splines). The hazard function for the

B-splines model is given by:

$$h_i(T_i) = \exp\left(B(T_i; \boldsymbol{\theta}, \boldsymbol{k_0}) + \eta_i\right) \tag{2.12}$$

where η_i is the linear predictor, and $B(T_i; \boldsymbol{\theta}, \boldsymbol{k_0})$ denotes a B-spline with regression coefficient $\boldsymbol{\theta}$ and knots $\boldsymbol{k_0}$ evaluated at time t.

Time-varying effects hazard modelling

Furthermore, because the hazard is, by definition, time-varying, a hazard ratio is also time-varying. For example, the hazard ratio might in the first months of follow-up be significantly greater than one, indicating that the treatment is harmful, but at the end of the study turn significantly lower than one, implying that treatment is beneficial.

Non-proportional hazard modelling allows the regression coefficients to be time-dependent, i.e. time-varying. As shown in equation 2.8, the hazard function is time-varying. Hence, the hazard ratio is also time-varying, which is remarkably relevant in the presence of proportional hazard violations [71]. Extending the regression hazard model from Equation 2.10, the hazard for the time-varying effects model is given by:

$$h_i(t) = h_0(t) \exp\left[\eta_i(t)\right] = h_0(t)\lambda_i(t) \tag{2.13}$$

where the linear predictor $\eta_i(t)$ and the link function $\lambda_i(t)$ may vary with time. [72] proposed a model for the time-varying effects via B-splines, which allows a smooth estimation of the time-dependent population parameters as depicted in figure 2.5. For RWD analysis, the time-varying nature of hazard ratios is relevant because confounding can occur by conditioning on the risk set at time t, see figure 2.4.d. [71, 73, 74] described such a circumstance in a Women's Health Initiative study of hormone therapy's effect in reducing coronary heart disease in post-menopausal women. They noted that susceptible women were developing coronary heart disease soon after hormone therapy initiation because hormone therapy acted as a selection force of unsusceptible women to coronary heart disease, thereby downwards biasing the conventional analysis's hazard ratio estimates. A series of papers [15, 75] discussed methodological approaches to tackle the apparent paradox of hazard ratio change with time in observational studies.

The counterfactuals approach of [71] emulated a trial using observational data. As [71] had access to the randomised clinical trials data, they could compare with the trial's treatment effect estimates. [76] use time stratified hazard ratios to evaluate the impact on survival of intrinsic breast tumour subtypes. These studies used an arbitrary cut-off time to stratify by the time

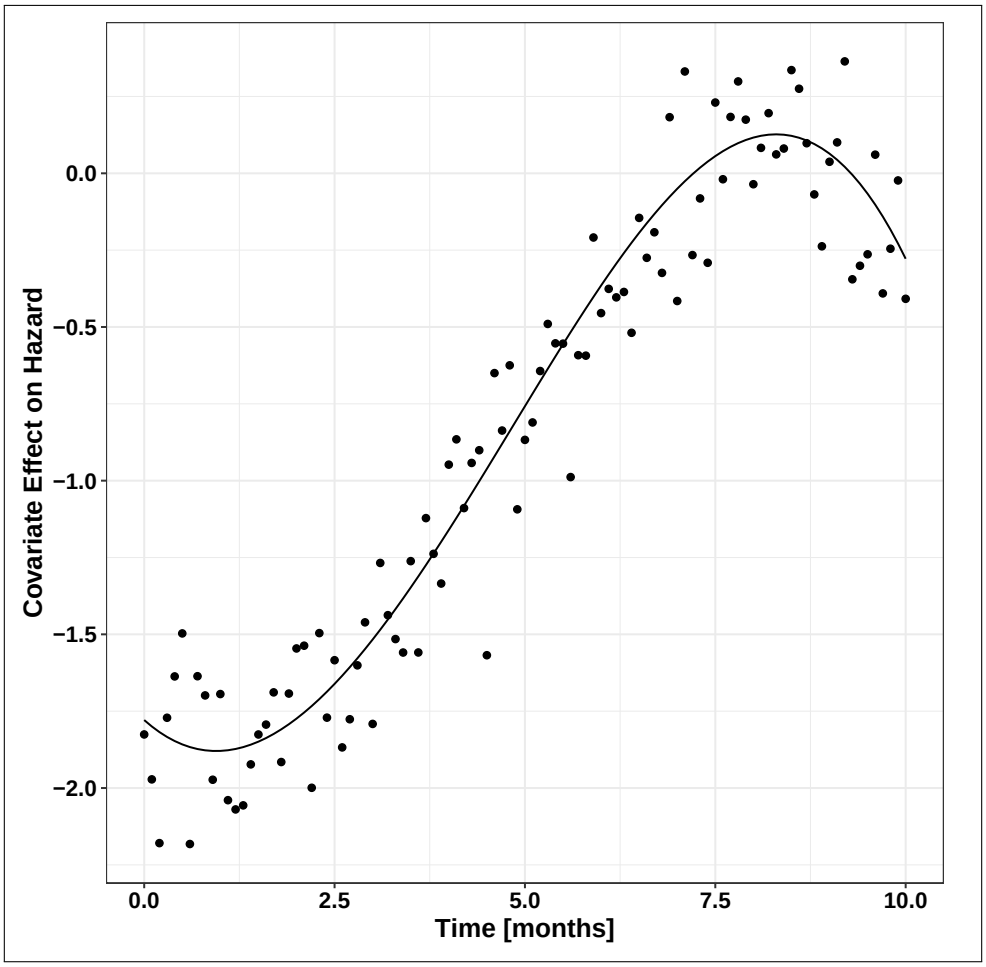


Figure 2.5: Illustration of B-splines for time-varying effects (TVE) hazard modelling using synthetic data.

of follow-up. [15] suggested a more principled approach by modelling time-to-follow up with a flexible parametric function and estimate product terms of baseline exposure and time-to-follow up. A time-stratified approach was suitable in these studies since the baseline exposures were not time-varying. However, in RWD, treatment often is varying. Consequently, it is hard to relate baseline exposure to survival.

2.4 Summary and perspectives

The rapid adoption of real-world clinical data presents unprecedented opportunities for accelerating drug development in oncology by analysing cancer cohorts for biomarker defined populations. However, we identify four methodological challenges in analysing RWE studies:

1. The data is biased by clinical practice. We need a better methodology to estimate unbiased parameter estimates because conventional summary statistics cannot predict treatment effects even in large samples.

2. Missing data is a universal problem in RWE studies. Missing data appear as gaps in the dataset that hide meaningful values for analysis.

3. Given its longitudinal nature, time-to-event is a typical analysis performed in analysing RWD. However, because RWD is observational, considering time-varying treatment effects is essential to avoid selection bias.

4. Analysing real-world clinical data poses the methodological challenge of interpreting time-varying treatment effects. In most real-world analyses, linear models are unreliable because of model misspecification and non-constant treatment effects resulting in biased inference.

To answer these questions, we develop a causal inference framework in chapter 3. Chapter 4 tackles the missing data problem and introduces a new imputation algorithm, MITABNET, conquering state-of-the-art imputation algorithms performance. In addition, it presents a standardised evaluation of new algorithms that allows advancing machine learning research in missing data imputation algorithms. Chapter 5 provides a new approach for efficient estimation of treatment effects in the presence of non-linear interactions with biomarkers using GP survival regression. Finally, chapter 5 presents a Bayesian time-varying effects model, benchmarked to conventional time-fixed and time-varying effects models with several synthetic and real-world examples.

Chapter 3

Causal Models and Statistical Methods

This chapter illustrates how to perform causal inference for RWE studies. As noted previously, causal inference is not a specific modelling tool but rather considering a particular type of question, such as the causal effect of a treatment. Nonetheless, there is a toolkit for causal inference. The chapter introduces a causal inference framework that allows us to discuss average treatment effects, treatment effects modification, and individualised treatment effects. Since these methods' applications are ultimately on real-world studies in clinical research in oncology, various examples illustrate the ideas' implementation. The first section of this chapter covers inferring treatment effects on some outcome in conventional RWD analyses. Central to these studies is the concept of counterfactual outcomes, which allows us to define the causal treatment effect. To estimate from data the counterfactual outcomes, however, we need some assumptions or conditions. The second section of the chapter is devoted to describing the four main causal assumptions: positivity, no interference, consistency and exchangeability. These allow us to define a causal estimand and turn it into a statistical estimand through association. The section will focus on the Neyman – Rubin causal modelling framework [77, 78] since this is the conventional approach within biomedical observational studies for more than 30 years [79]. In the subsequent section, we will discuss estimation by turning the statistical estimand of treatment effects into an estimate using data. The bounds and sensitivity analyses section will focus on constituting a framework in which we can conduct causal inference in situations where we prefer to weaken the causal assumptions.

3.1 A framework for defining treatment effects

Often in analysing real-world data, we are interested in study the *main treatment effect* measured by a specific outcome, e.g. survival time, in a population of interest. Let us consider the toy example depicted in figure 3.1. The graph involves three sets of random variables: $\mathbf{W}$, including all confounding factors and possibly high-dimensional; A indicates the treatment, e.g. binary treatment; Y, the outcome, or primary end-point of interest. Its structural equation is given by:

$$\begin{aligned} W &:= f_W(U_W) \\ A &:= f_A(W, U_A) \\ Y &:= f_Y(W, A, U_Y) \end{aligned} \tag{3.1}$$

The left-hand sides are the variables we might have data for $\mathbf{W}$, A, and Y, and get their value from the functions at the right-hand side, where U_W, U_A, U_Y are unknown i.i.d. random variables.

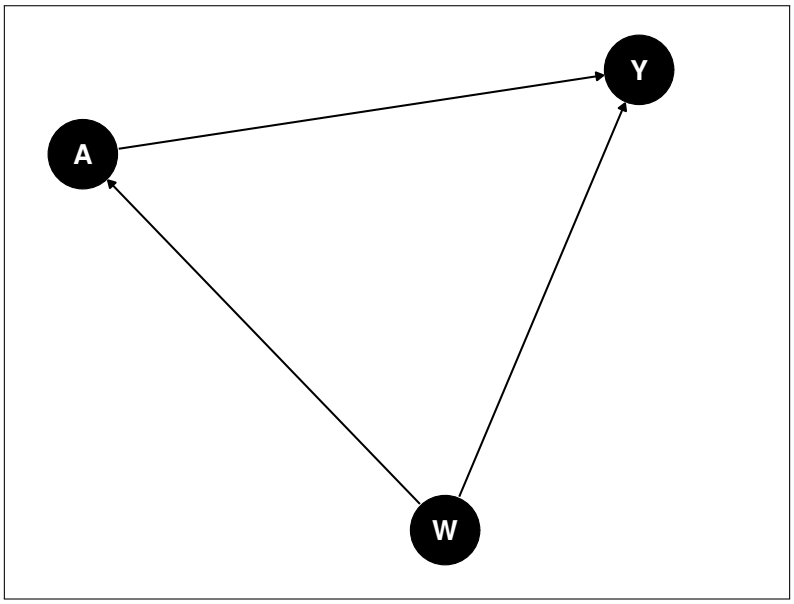


Figure 3.1: Example DAG with three sets of random variables: $\mathbf{W}$, including all confounding factors; A indicates the treatment; Y, the outcome.

We can define a potential outcome as the outcome we would see under each of the treatments. For example, $\mathrm{Y}^0(w_i)$ is the potential outcome under the treatment $A = 0$; $\mathrm{Y}^1(w_i)$ is the potential outcome under the treatment $A = 1$. If $A = 0$ indicates no treatment, one also may refer to $\mathrm{Y}^0(w_i)$ as the control outcome, and $\mathrm{Y}^1(w_i)$ as the treated outcome.

Counterfactual outcomes, or contrary-to-the-fact, are outcomes that would

have been observed had treatment been different. For example, if the ith individual was under treatment $A = 1$, the counterfactual outcome is $\mathrm{Y}^0(w_i)$. Alternatively, if the ith individual was under treatment $A = 0$, the counterfactual outcome is $\mathrm{Y}^1(w_i)$. Counterfactual outcomes are related to potential outcomes. Both terms are used interchangeably in the literature [80, 81]. However, we note that they have slightly different meanings. Counterfactuals outcomes make explicit reference to the fact that the data collection is earlier than the research purpose. Hence, they are more appropriate for retrospective RWD, and we will refer to the counterfactual outcome framework moving forward. We say that treatment had a causal effect on Y if the counterfactual outcome Y^1 differs from Y^0, i.e. there is only a causal effect if $\mathrm{Y}^1 \neq \mathrm{Y}^0$.

3.1.1 Individualised treatment effects

Following [32, 38] we define the causal effect of a treatment on an individual to be the individualised treatment effect (ITE) for individual w_i, given by:

$$\mathrm{ITE}(w_i) = \mathbb{E}_{\mathrm{Y}^1|w_i}\left[\mathrm{Y}^1|w_i\right] - \mathbb{E}_{\mathrm{Y}^0|w_i}\left[\mathrm{Y}^0|w_i\right] \tag{3.2}$$

where w_i denote the individual covariates, and $\mathrm{Y}^1, \mathrm{Y}^0$ are the counterfactual outcome under treatment $A = 1, A = 0$, respectively. For ITE, one computes a difference in expectations of $\mathrm{Y}^1(w_i)$ from $\mathrm{Y}^0(w_i)$. Practically, a cross-over study estimates individualised treatment effects under the assumptions that no confounding variables exist between the periods, and the design is balanced [82]. Nevertheless, in clinical research in oncology, cross-over studies are limited, i.e., we will only see one counterfactual outcome for each individual.

3.1.2 Average treatment effects

Following [32, 38] we can estimate, with certain assumptions, a population level (average) causal effects. The ATE is the difference in the mean causal effect of treatment over the whole population, given by:

$$\begin{aligned}\mathrm{ATE} = & \\ & \mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] = \\ & \mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y}|A = 1, W\right] - \mathbb{E}\left[\mathrm{Y}|A = 0, W\right]\right] = \\ & \mathbb{E}_{w\sim p(w)}\left[\mathrm{ITE}(w)\right]\end{aligned} \tag{3.3}$$

where Y^1 is the average value of Y if everyone is on treatment A=1, and Y^0 is the average value of Y if everyone is on treatment A=0. W is the set of confounders, $p(w)$ is the distribution of confounders, and ITE is the individualised treatment effect.

3.1.3 Conditional average treatment effects

The conditional average treatment effect (CATE) is very similar to ATE, but following [32, 38] we condition on a set of variables of interest X, such that:

$$\begin{aligned} &\text{CATE} = \\ &\qquad \mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0 | X = x\right] = \\ &\qquad \mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y} | A = 1, X = x, W\right] - \mathbb{E}\left[\mathrm{Y} | A = 0, X = x, W\right]\right] \end{aligned} \tag{3.4}$$

where we are conditioning on X to be x. CATE estimation allows us to investigate treatment modification for some set of variables X.

Biomarker-based treatment effect modification is an example of conditional average treatment effects, also known in the literature as the heterogeneity of treatment effects [17]. In biomarker-based treatment effect modification, we define a subpopulation by a covariate of interest X, e.g. a predictive biomarker that provides information on the likelihood that the patient benefits from the treatment. Often in machine-learning literature [81, 83], the X are all the observed covariates $\mathbf{W}$ so that the CATE are the ITE. Yet we will distinguish between CATE for treatment modification and ITE for individualised treatment effects, respectively.

3.1.4 Other measures of treatment effects

Other measures of causal effects on the population we might be interested in include:

1. $\mathbb{E}\left[\frac{\mathrm{Y}^1}{\mathrm{Y}^0}\right]$
2. $\mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0 | A = 1\right]$
3. $\mathbb{E}\left[y'(t)\right]$

where 1, 2, and 3 are the causal relative risk, the treatment effect on the treatead (TET) and the average treatment effect varying with time t. For binary outcomes, instead of a difference in counterfactual outcomes, we take a ratio, i.e. the causal relative risk [1]. Like the CATE, the treatment effect on the treated is the average individualised treatment effects for a sub-population, particularly the treated sub-population. Time-varying effects allow treatment effects to be dynamic and can reveal change over time in treatment effects.

[1] For any nonlinear functional of the causal effect, such as the relative risk or the hazard, we resource to modelling to estimate the individual treatment effects because we never observe both $\mathrm{Y}^1(x_i)$ and $\mathrm{Y}^0(x_i)$.

3.2 Causal inference conditions

To estimate counterfactual outcomes from RWD, we need to make some *assumptions* about the observed data. To estimate unbiased treatment effects, we also need to make some untestable assumptions about the set of measured variables on the population. We will need to accompany our analyses with sensitivity analysis techniques to assess the robustness to violations of our untestable assumptions. What follows is an account of the conditions for causal inference, or causal assumptions, in analysing treatment effects from RWD.

The theory around causal inference conditions that we present builds on *Causal Inference: What If* by Hernan and Robin [32], *Causal Inference for Statistics, Social and Biomedical Science* by Imbens and Rubin [84], and *Introduction to Causal Inference* by Neal [85, 86]. The section on confounds builds on *Statistical Rethinking* by McElreath [40] and *Causality* by Pearl [38]. We refer the reader to those text for further theory around causal inference.

3.2.1 Exchangeability

Exchangeability means that there are no unmeasured confounding factors, i.e. all confounding factors are present in the RWD. Under exchangeability the counterfactual outcomes Y^{a} are conditionally independent of the treatment decision given the observation on individual w_i, such that:

$$\left(\mathrm{Y}^0, \mathrm{Y}^1\right) \perp\!\!\!\perp A|W \tag{3.5}$$

where $\left(\mathrm{Y}^0, \mathrm{Y}^1\right)$ are independent of A given $\mathbf{W}$. Therefore, we have exchangeability conditional on a sufficient adjustment set of covariates $\mathbf{W}$. In analysing observational RWD, the exchangeability condition might be unrealistic, and there may be situations where there is unmeasured confounding. A violation of exchangeability arises if there are hidden variables h that impact the counterfactual outcomes, and critically h also impacts the outcome. In this case, the counterfactual outcomes Y^0 and Y^1 are not conditionally independent of the treatment decision given the observation on individual w_i, formally given by:

$$\left(\mathrm{Y}^0, \mathrm{Y}^1\right) \not\perp\!\!\!\perp A|W \tag{3.6}$$

These hidden confounders may affect treatment decision and patient outcome. A real-world example of hidden confounders are treatment guidelines concerning comorbidities not recorded in our RWD. To evaluate whether the problem setting is correct and exchangeability holds, we as data analyst need to cooperate with clinicians to learn what factors affect treatment decisions and measure any confounds.

In our work on RWE studies for cancerous diseases, exchangeability holds if there is no unobserved confounding, i.e. all covariates are measured. Therefore, it remains an assumption. In the study of missing data, the concept of non-exchangeability motivates the study of missingness mechanisms and the scenario where the MAR assumptions are adequate. For clinical outcome interpretations of survival data, weaker assumptions than exchangeability, such as causal bounds, allow for more credibility in the results of the causal survival analyses. Analytical methods to assess robustness to hidden confounds and exchangeability violations include sensitivity analyses [87], which we explain in section 3.5.

3.2.2 No interference

The No interference assumption means that the counterfactual outcome for the individual i, which could be a function of all the treatments in the population of interest $1, \ldots, n$, is a function of only the treatment t_i, i.e. the treatment assignment on one patient does not affect another patient's counterfactual outcome, such that:

$$Y_i\left(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n\right) = Y_i\left(a_i\right) \tag{3.7}$$

the no interference assumption is equivalent to the no interaction principle [88], where patients do not interfere with each other. No interference forms part of the stable unit treatment value assumption (SUTVA) from [89]. Violations of no interference will not allow us to write counterfactual outcomes for the individual x_i in terms of only that individual's treatment. In medical research, a violation of no interference may happen in vaccine trials for contagious disease, where one individual's vaccination may impact other individuals outcome, e.g. via herd immunity [90]. However, in our analysis of RWD for oncology, we will assume that the no interference assumption holds throughout because it is plausible that one individual's treatment does not impact other individual's outcomes for cancer patients.

3.2.3 Consistency

Consistency means that if treatment $A = a$ implies that the observed outcome is the counterfactual outcome given by Y^a, such that:

$$A = a \Rightarrow Y = \mathrm{Y}^a \tag{3.8}$$

Consistency implies well-defined treatments and outcomes. Although it may seem that consistency is simple to accomplish, that may not necessarily always be the case. For example, one conventional analysis of observational real-world

data is inconsistent. In particular, [71] discuss the analyses that divide each individual's follow-up time into individual time, obtaining exposed person-time or unexposed person-time, comparing incidence rates between them. Such an analysis may be inconsistent because while analysing RWD in terms of unexposed person-time and exposed person-time, we have lost the relation to the causal question that is not in terms of person-time but a population followed-up a pre-determined length of time. Hence, consistency may not hold on RWD cohort survival analyses that do not consider the crucial elements of study design. Another phrase for violating this aspect of the consistency assumption is no multiple versions of treatment, which also forms part of the SUTVA from [89]. The consistency assumption holds in our work on RWE studies by delimiting and defining the study population by treatment, for example, first line, second line, and outcome, for example, survival time or duration of treatment.

3.2.4 Positivity

The positivity assumption means that there is always some stochasticity in the treatment decision, such that:

$$\Pr[A = a|\mathbf{W} = w] > 0 \ \forall \ a, w \tag{3.9}$$

where $\Pr[A = a|\mathbf{W} = w]$ is the probability that treatment A takes value a given covariates $W = w$. Overlap, or common support, is another perspective on the positivity assumption; see figure 3.2, where we define the overlap between the distribution $P(\mathbf{W}|A = 0)$ and $P(\mathbf{W}|A = 1)$ in a binary treatment example. No overlap of these distributions means severe positivity violation. Partial overlap suggests no positivity violation among the covariates where there is overlap. However, among the covariates where there is no overlap, then there is partial positivity violation. Between a confounder's levels, an overlap is also necessary for treatment modification by a variable X on subpopulation analysis. More generally, positivity assumes that the probability of receiving treatment for each individual, known as the propensity score [91] and denoted as $e(A|\mathbf{W})$, is bounded between 0 and 1, such that:

$$\epsilon < e(A|\mathbf{W}) < (1 - \epsilon) \tag{3.10}$$

for some $0 < \epsilon < 1$. Positivity violations occur, for example, when patients only receive treatment $A = 1$ and never receive treatment $A = 0$. Then, we can never determine the counterfactual of what would have happened if patients had received treatment $A = 0$. The positivity assumption is testable from the data. With few categorical variables as confounders, we could summarise

frequencies by strata. However, RWD is finite, and the strata may be high dimensional, requiring resourcing on modelling to extrapolate by smoothing over the strata [92]. In actual RWD obtained from clinical practice, positivity may not hold because there are clinical guidelines. However, positivity may hold if the treatment choice is not apparent, such as in second-line treatments where clinicians might try different medicines [93]. Positivity may even hold with established clinical guidelines if training varies between clinicians from various locations generating stochasticity across clinicians. To test for positivity in our RWE studies, we test for positivity in the distribution of covariates in the study population using the propensity score $e(A|\mathbf{W})$ and frequency tables.

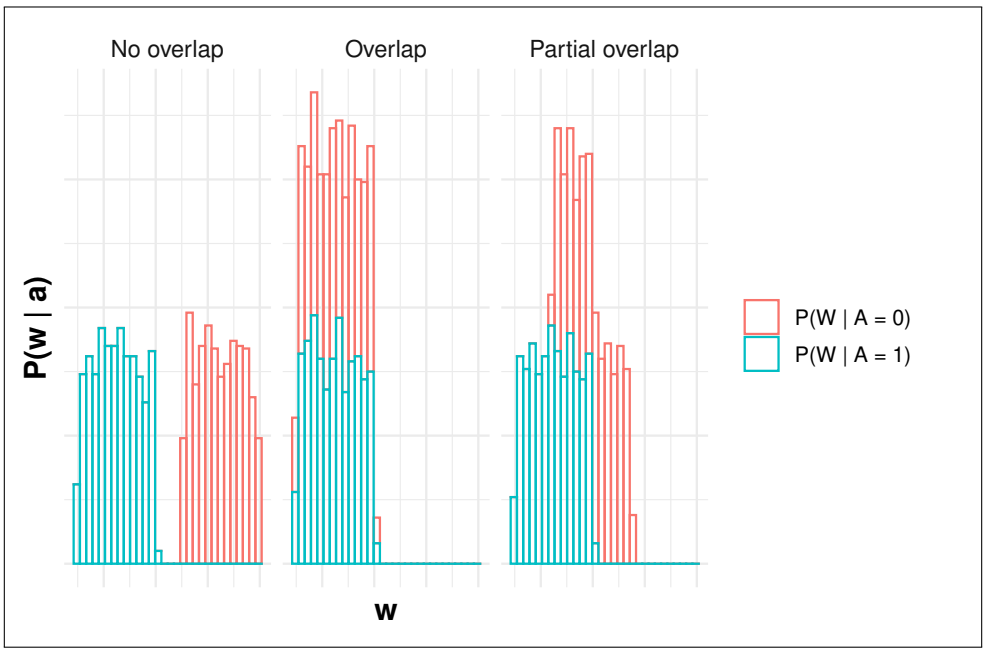


Figure 3.2: Visualisation of another perspective of positivity, overlap, the distribution $P(W|A = a)$ for binary treatment. No overlap means severe positivity violation. Partial overlap suggests no positivity violation on the confounders where there is overlap, but there is no overlap on the confounders, then there is severe positivity violation. Complete overlap suggests no positivity violation.

The exchangeability-positivity trade-off

The exchangeability-positivity trade-off [85, 86, 94] says that the more covariates we condition on, the more likely we have exchangeability. However, the more covariates we condition on, the worst positivity becomes. To see this, we consider the conditional distribution again from figure 3.2 adapted from [94], where we have partial overlap and study its supports when increasing the number of dimensions. Let us assume that each dimension's overlap is constant,

denoted by O, where $0 < O < 1$. Then, the total overlap decreases exponentially with the number of dimensions d by O^d.

3.2.5 Treatment effects revisited

Now that we concluded the four main causal assumptions, we can derive the proof for obtaining an unbiased estimate for the ATE. The relevance of this derivation adapted from [38] is that it nicely ties together the main causal assumptions used in this thesis: exchangeability, positivity and consistency.

Firstly, we use the No Interference assumption to justify that these counterfactual outcomes are a function of each individuals' treatment, and no other's treatment:

$$\text{ATE} = \mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] \tag{3.11}$$

then, we use linearity of expectations to give:

$$\mathbb{E}\left[\mathrm{Y}^1\right] - \mathbb{E}\left[\mathrm{Y}^0\right] \tag{3.12}$$

Furthermore, we use the law of iterated expectations to give:

$$\mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y}^1|W\right] - \mathbb{E}\left[\mathrm{Y}^0|W\right]\right] \tag{3.13}$$

which is necessary to allow the application of conditional exchangeability, now we use conditional exchangeability and positivity, such that:

$$\mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y}^1|A=1,W\right] - \mathbb{E}\left[\mathrm{Y}^0|A=0,W\right]\right] \tag{3.14}$$

Finally, we use consistency:

$$\mathbb{E}_W\left[\mathbb{E}\left[Y|A=1,W\right] - \mathbb{E}\left[Y|A=0,W\right]\right] \tag{3.15}$$

where we make clear that applying consistency allows us to link the counterfactual outcomes to the observed outcomes. Tying all this together we obtain:

$$\begin{aligned}
\text{ATE} =& \mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] = \\
& \mathbb{E}\left[\mathrm{Y}^1\right] - \mathbb{E}\left[\mathrm{Y}^0\right] = \\
& \mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y}^1|W\right] - \mathbb{E}\left[\mathrm{Y}^0|W\right]\right] = \\
& \mathbb{E}_W\left[\mathbb{E}\left[\mathrm{Y}^1|A=1,W\right] - \mathbb{E}\left[\mathrm{Y}^0|A=0,W\right]\right] = \\
& \mathbb{E}_W\left[\mathbb{E}\left[Y|A=1,W\right] - \mathbb{E}\left[Y|A=0,W\right]\right]_\square
\end{aligned} \tag{3.16}$$

3.3 Statistical inference

So far, this chapter has focused on the conceptual, non-statistical aspects of causal inference, see Sec. 3.1. However, from RWD and, indeed, also from randomised trials, we only have access to a random sample from the population of interest. Therefore, we must construct a framework for estimating the treatment's causal effects from observed data. Statistical inference concerns itself with estimating a function (or parameter) of the population, formally denoted as the *estimand* θ by taking a sample from the population and using an estimator, which is a rule that takes data and yields a numerical value for the estimand. This numerical value for a particular sample is the *estimate* $\hat{\theta}$ that is our best guess of the estimand θ for that sample. Perhaps, most importantly, we can use the sample from the population to compute the uncertainty on the estimate $\hat{\theta}$, for example, by calculating the standard error of $\hat{\theta}$ and computing a 95% Wald confidence interval [95] that contains θ in 95% of random samples.

3.3.1 Probability and inference

A fundamental question in statistics is the definition of *probability*. Frequentist defines probability as fundamentally related to the frequencies of repeated events. For Bayesians, probability instead is fundamentally related to the certainty or uncertainty about the conditions for the events. The consequence is that frequentist analyses variations of data and derived quantities in terms of fixed model parameters, while Bayesians analyse variations of beliefs about parameters in terms of fixed observed data. In the present thesis, we often opt for the Bayesian definition of probability, yet we analyse our statistical methods' frequentist properties, too.

The following section will address different aspects of probability theory, primarily Bayes' theorem and its role in scientific discovery, machine learning, and real-world data analysis. Bayes' rule and Bayesian modelling will play a crucial role in the RWE study analysis conducted in Chapter 5. Hence, we introduced here the topic.

The concepts presented in this section builds on *Bayesian data Analysis* by Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin [96], *Bayesian Statistics* by Lambert [97] and *Statistical Rethinking* by McElreath [40]. The information theory subsection has been adapted from *Information Theory and Statistics* from Kullback [98]. We refer the reader to those text for further theory around Bayesian inference.

The general situation where Bayes' theorem is relevant is when we have a hypothesis, observe new evidence, and want to know the probability that the hypothesis holds given the evidence. A geometrical idea underlying Bayes' theorem is that new evidence restricts possibilities' space, see figure 3.3 adapted

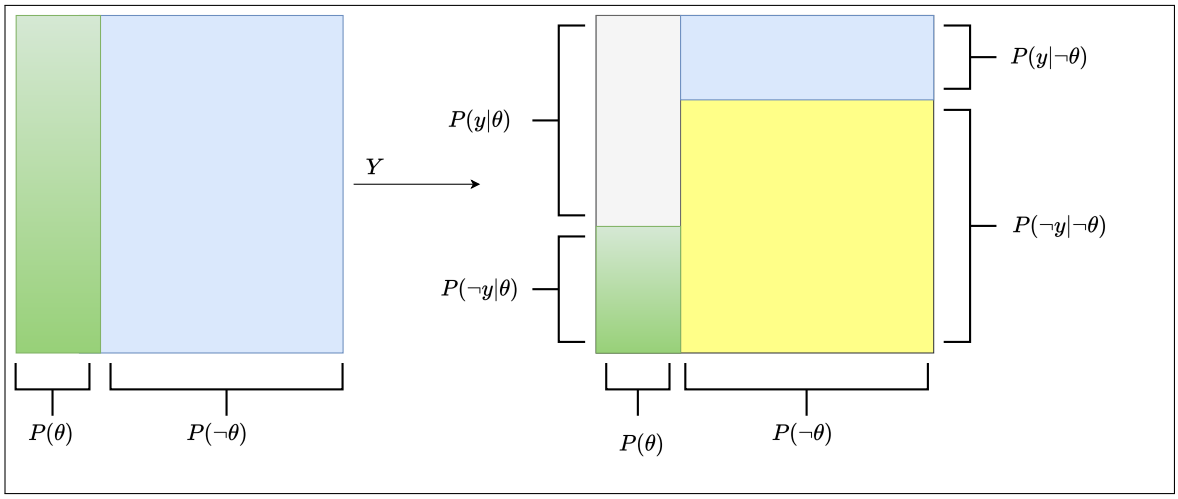


Figure 3.3: Visualisation of Bayes' rule: new evidence Y restricts the space of possibilities, and the ratio we need to consider $p(\theta|y)$.

from [99]. Bayes' rule is the mathematical formula that defines the *posterior density* as the ratio we need to consider to update one's beliefs about a hypothesis θ, given some observed data y, and takes the form:

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)} \tag{3.17}$$

There are three essential quantities in equation 3.17. The quantity $p(\theta)$, denoted as the "prior", is the probability that the hypothesis is true (before any evidence). The quantity $p(y|\theta)$, denoted as the "likelihood" or "data probability", is the probability of seeing the evidence if the hypothesis is true, i.e. the likelihood of the experimental new data given that the hypothesis holds. Similarly, to apply Bayes' rule, we need to consider the denominator $p(y)$, which is the unconditional probability of the evidence given by:

$$p(y) = p(\theta)p(y|\theta) + p(\neg\theta)p(y|\neg\theta) \tag{3.18}$$

Concerning the term $p(y|\neg\theta)$, the probability of the evidence given that the hypothesis does not hold, i.e., how much of the complementary space includes the evidence in figure 3.3, is, in practice, hard to estimate [100]. Moreover, with the data y constant, we can derive an equivalent form of equation 3.17, which yields the unnormalised posterior density given by:

$$p(\theta|y) \propto p(y|\theta)\,p(\theta) \tag{3.19}$$

Because $p(y)$ does not depend on θ, equation 3.19 determines the shape of the posterior density. We regard equation 3.19 as a weighted geometric average [2],

[2]We say weighted geometric average because we are multiplying $p(y|\theta)$ and $p(\theta)$

which is sensitive to small values of either $p(y|\theta)$ or $p(\theta)$.

The posterior density, $p(\theta|y)$, is proportional to the product of the probability of the data and the prior probability. Intuitively, collecting more data will make the posterior probability distribution narrower and resembling the data probability mass, a process known as Bayesian updating, see figure 3.4. We will use Markov chain Monte Carlo (MCMC) methods [101] for approximating the posterior distribution of complex models in practice. The engine we will use is Stan [102], which implements the No-U-Turn-Sampler extension of the Hamiltonian Monte Carlo (HMC) algorithm [103].

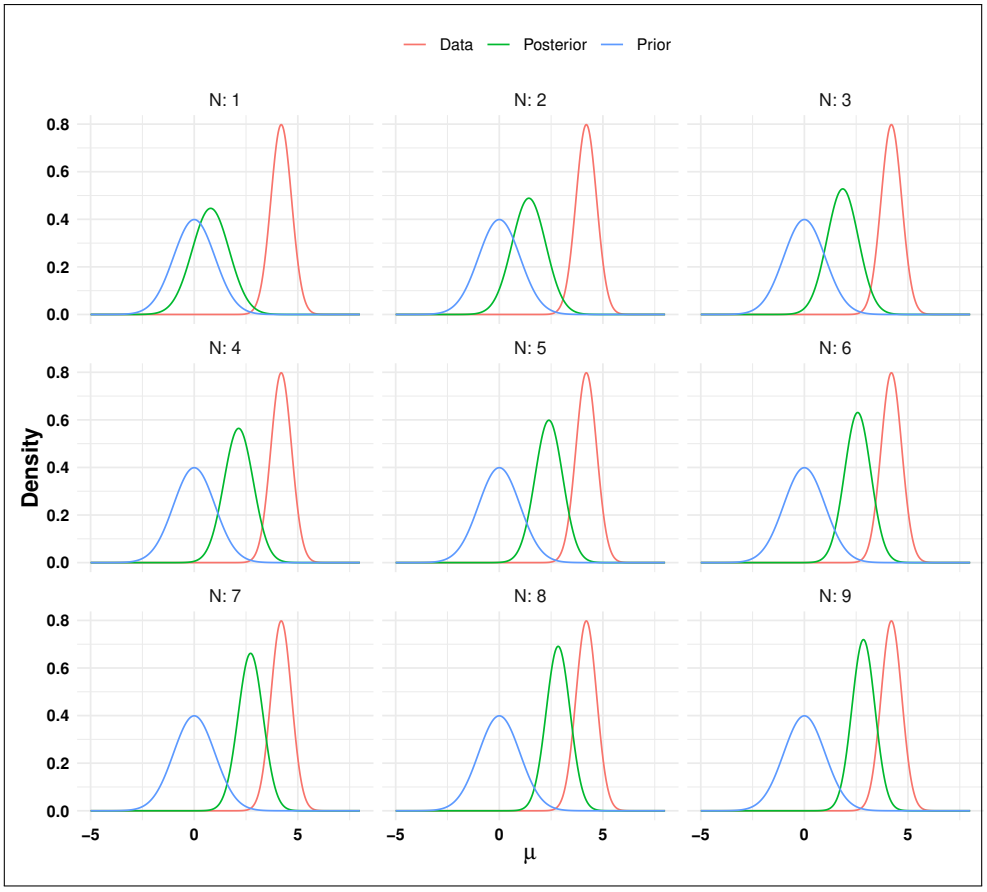


Figure 3.4: Graphical toy illustration of Bayesian updating for the mean of a Normal distribution: Before the first observation, the model has a prior set for μ given by Normal$(0, 1)$. After each data point arrives, the model updates transforming it into a posterior. Observing a sample contains information about the parameter of interest μ.

3.3.2 Entropy and accuracy

The general idea of model evaluation is to assess the model's ability to generalise to out-of-sample by forecasting accurate predictions [104]. To do so, we appeal to information theory, a framework in which we derive a rigorous and principled

method to *uncertainty* in a probability distribution p, and its reduction by learning an outcome. Formally, to measure the uncertainty in a probability distribution, $H(p)$, we can use information entropy [105], given by:

$$H(p) = -\mathbb{E}\left[\log(p_i)\right] = -\sum_{i=1}^{n} p_i \log(p_i) \tag{3.20}$$

where $p_i \log(p_i)$ is the log-probability of an event. Intuitively, entropy is a measure of the information in a distribution. In model evaluation, we score our predictive model, q, on its predictive accuracy. We aim to minimise the difference between the entropy of the model and the data. To do so, we use the Kullbak-Leibler divergence [106], given by:

$$D_{KL}(p, q) = \sum_{i} p_i \left(\log(p_i) - \log(q_i)\right) \tag{3.21}$$

Considering we score q's accuracy using the divergence, or distance in log-probability from p to q, averaging over the events' frequency. Notably, although D_{KL} is a distance, it is not symmetric. Therefore, $D_{KL}(p, q) \neq D_{KL}(q, p)$, see figure 3.5 adapted from [104].

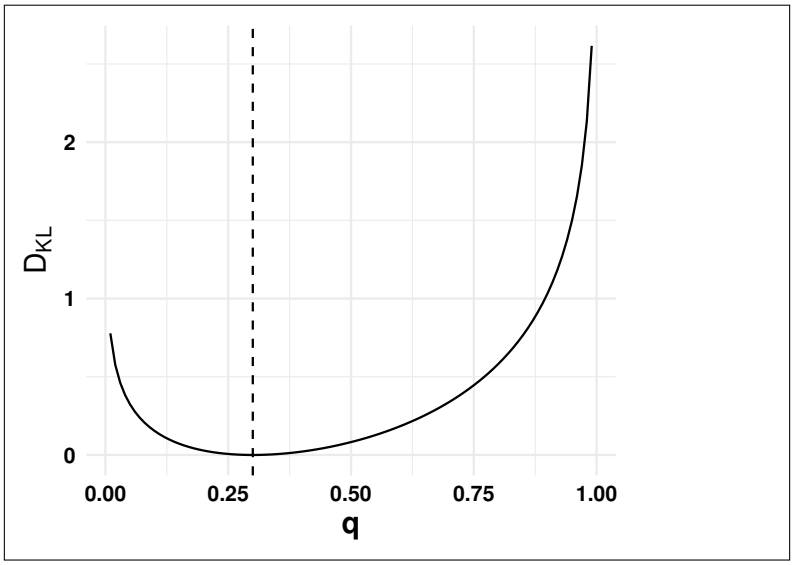


Figure 3.5: Illustration of the asymmetric property of D_{KL}: we calculate D_{KL} using equation 3.21 for a grid of alternative models p. $D_{KL} = 0$, denoted by a dotted line, corresponds to $p = q$.

In practice, we do not know how to compute precisely equation 3.21, because p is unknown. However, we do not need p when comparing models. In this case, we only need the log-score, given by:

$$S(q) = \sum_{i} \log(q_i) \tag{3.22}$$

where $\log(q_i)$ denotes the log-probability for the ith event. The log-score is straightforward to compute as the sum of log probabilities for each observation i. Notably, the R^2 is a special case of log-score [107]. Conventionally, we use the deviance instead of the log-score, which we obtain by scaling the log-score by -2. On the deviance scale, smaller deviance values are better, suggesting less KL divergence.

In the Bayesian approach, instead of a single log-score, we have a distribution of log-scores, or log-pointwise-predictive-density (lppd), given by:

$$\text{lppd}(y, \Theta) = \sum_i \log \frac{1}{S} \sum_s p(y_i|\Theta_i) \tag{3.23}$$

where for each data point i, we are taking the average log-probability of that observation denoted by $p(y_i|\Theta_i)$, conditional on the parameters Θ_i, and average over samples S.

Arguably, we aim to obtain a higher log-score, lppd, and lower deviance in making *out-of-sample* predictions. Ideally, we use a training set, holding out a validation set to evaluate the model prediction. Similarly, we can conduct cross-validation [108]. Moreover, even if we do not have an independent set to evaluate our model, we can use heuristics to predict out-of-sample error by applying information criteria [109].

3.3.3 Deep neural networks

Deep learning is a part of machine learning, a part of the broader field of artificial intelligence that has revolutionised statistical inference. We can define deep learning as a technique to extract patterns from data with deep neural networks (DNN). As the name suggests, the brain arrangement inspired the computational models that are artificial neural networks (ANN). DNN comprise numerous connected neurons, each of which computes a simple operation. We can use DNN to study complex functions directly from data by carefully setting the network's parameters.

The following section will address different aspects of deep learning, describing ANN' architecture and the backpropagation algorithm. The deep learning theory that we present builds on *Deep learning* by Goodfellow, Bengio, and Courville [110], and *Understanding machine learning: From theory to algorithms* by Shai and Shai [111]. We refer the reader to those textbooks for further theory around deep learning. However, because deep learning is a field of current active research, often innovations are in conferences proceeding. We explore more recent innovations on DNN and their practical applications to RWD analysis in Chapter 4 and 5.

Neural Network's architecture

A fundamental idea of deep learning is data representation [110]. In particular, deep learning builds on the idea of repeated composition by taking iterated simple transformations of the data, gradually abstracting meaningful patterns. The insight of deep learning is that we can get from inputs to outputs gradually. We describe DNN' network architecture as a stack of computational units, or neurons, that form a graph. We can describe the depth of architecture via the path's depth from the inputs to the computational graph's outputs. Figure 3.6 shows a simple toy ANN's architecture. The first stack of our ANN, or input layer, is given by the raw input data, the last stack of our ANN is the output layers, all the stacks in between are the hidden layers. The way the networks operates activations in one layer determine the activations on the next layer.

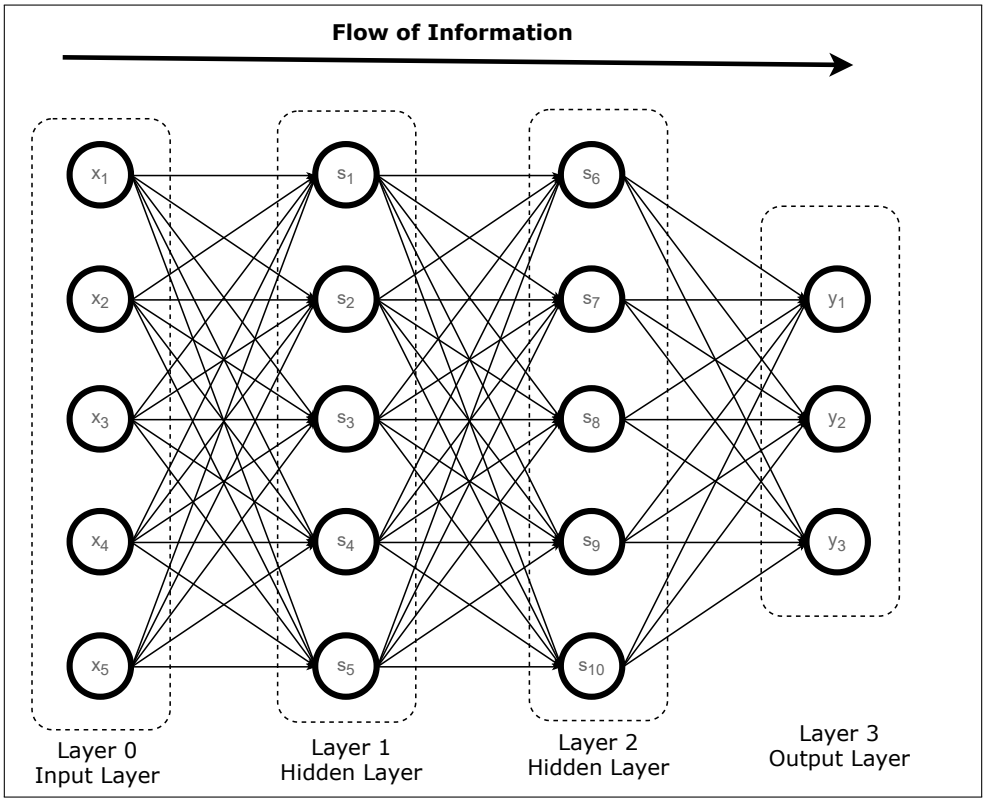


Figure 3.6: An artificial neural network (ANN) with information flowing from left to right. The first layer represents the ANN's inputs, the middle or hidden layers represent neurons or computational units that act on the input, the last layer represent the ANN's output.

Neurons' activation

For an ANN to capture the data patterns, it has to adjust its parameters θ, which conventionally includes the weights $\overrightarrow{w}$ of the connections between

subsequent layers neurons and the biases b that determine each neuron's activity. Each neuron's activation is a composed weighted sum of the inputs, i.e., a function $g(x, \theta)$, its output given by:

$$o = g\left(\overrightarrow{w} \cdot \overrightarrow{x} + b\right) \tag{3.24}$$

A fundamental part of equation 3.24 is the function g, which may involve non-linear transformations such as sigmoid or rectified linear units [112, 113], given by $\text{ReLU}(x) = \max(0, x)$. Therefore, we can see that a neuron is a function that represents its inputs using possibly a non-linear transformation [113]. Every other neuron in the network has its weight and bias associated. The entire network, therefore, is also a complex function with numerous parameters θ.

Cost function

We define the cost function $J(\theta)$ as an average error function. Conventionally, training proceeds by minimising the average *empirical* error function:

$$\begin{aligned} J(\theta) = & \\ & \mathbb{E}_{§, y \sim \hat{p}_{data}} \left[L\left(f\left(x; \theta\right), y\right)\right] = \\ & \frac{1}{N} \sum_{i=1}^{N} L\left(f\left(x^{(i)}; \theta\right), y^{i}\right) \end{aligned} \tag{3.25}$$

where N are the number of training examples. The cost function $J(\theta)$ is a level of complexity on top of the network function, as its inputs are the numerous network's parameters θ, and reduces its output to a scalar depending on the many training data examples [113].

Optimisation

It rapidly becomes infeasible to compute the input that minimises $J(\theta)$ explicitly. Instead, a more flexible tactic is to start at a random initial input and compute the gradient of $\nabla_\theta J(\theta)$, i.e. the direction of steepest increase of $J(\theta)$. Iteratively, minimising $\nabla_\theta J(\theta)$ allows us to approach some *local minimum* of the function [113]. In principle, with numerical methods to approximate $\nabla_\theta J(\theta)$, we can use optimisation algorithms [114] to optimise $J(\theta)$ by updating θ. However, there are many possible valleys, depending on the random input we started, and there is no guarantee that the local minimum is the global minimum, i.e. the smallest possible value of the cost function.

Backpropagation

Backpropagation is the algorithm for computing the gradient $\nabla_\theta J(\theta)$ by determining how a training example adjust the parameters θ not simply in terms of gradient direction but also in terms of the relative proportion that determines the steepest decrease to the cost $J(\theta)$. The notation in defining backpropagation uses indexed to denote the layer. For example, $w^l_{j,k}$ denotes a weight on layer l, from the neuron k to the neuron j. Mathematically, the impact of how sensitive the cost C is to the weights $w^l_{j,k}$ is the derivative of J for $w^l_{j,k}$, which takes the form:

$$\frac{\partial J}{\partial w^l_{j,k}} = a^{l-1}_k \delta^l_j \tag{3.26}$$

where we apply the chain rule. Similarly, the derivative of C for the bias b^l_j takes the form:

$$\frac{\partial J}{\partial b^l_j} = \delta^l_j \tag{3.27}$$

The term δ^l is given by:

$$\delta^l = \left(\left(w^{l+1}\right)^T \delta^{l+1}\right) \cdot \sigma'\left(z^l\right) \tag{3.28}$$

where z^l is the weighted sum given by:

$$z^l = w^l a^{l-1} + b^l \tag{3.29}$$

where b^l_j is the bias of neuron j in layer l and a^{l-1} is the activation of neuron j in layer l given by equation 3.24. We can write more compactly:

$$\delta^L = \nabla_a J \cdot \sigma'\left(z^L\right) \tag{3.30}$$

Since the full cost function, $\frac{\partial C}{\partial \theta^L}$ involves averaging over all the training examples. Its derivatives require averaging equation 3.26. We have then one component of the gradient vector, which comprises the partial derivatives from the cost function for all networks' trainable parameters.

Moreover, the same principle applies to measure the sensitivity to the previous activation a^{L-1} by computing the derivative for the previous layer, i.e., back-propagating. Hence, following [113] we intuitively define backpropagation as propagating backwards the chain rule. Backpropagation is the workhorse of how state-of-the-art DNN learn, such as those implemented with TensorFlow and Torch libraries [13, 115].

3.4 Estimation of treatment effects

Most of the scepticisms around causal inference on observational non-randomised studies [116, 117], argue that:

1. In observational studies, each subject characteristics influence its treatment selection.

2. Baseline characteristics can differ systematically between treatment groups.

The following section will address different aspects of treatments' causal effects estimation. We introduced the classical concepts of outcome regression modelling, inverse probability weighting and doubly robust methods to estimate unbiased treatment effects. The general idea in treatments' causal effects estimation will be to take a causal estimand and turn it into a statistical estimand through association, then turn the statistical estimand into an estimate through estimation.

3.4.1 Outcome modeling

Outcome modelling, also known as covariate adjustment, also related to the concepts of parametric g-formula, standardisation, single learners, or response surface modelling [83], is a very natural way to estimate treatment effects.

Recall the causal diagram in figure 3.1. The toy problem included some treatment A, a set of confounders $\mathbf{W}$, and an outcome of interest Y. Prominently, from RWD, we have confounds $\mathbf{W}$ that impact the treatment assignment mechanism and determine the outcome. Under the consistency assumption, the counterfactual outcome Y^a was the value that Y would take when setting A to A by intervention.

The g-formula approach [118] allows us to estimate the expected value of the counterfactual outcome Y^a given a sufficient adjustment set of covariates $\mathbf{W}$, and takes the form:

$$\mathbb{E}\left[Y^a\right] = \sum_{\mathbf{w}} \mathbb{E}\left[Y|A = a, \mathbf{W} = \mathbf{w}\right] P(\mathbf{W} = \mathbf{w}) \tag{3.31}$$

where on the left-hand side, there is a causal estimand, and on the right-hand side, there is a statistical estimand. Naturally, the g-formula is a standardisation formula allowing us to compute unbiased methods-of-moments for the distribution of the counterfactual outcomes Y^a, where the treated and control populations are different for the set of confounders $\mathbf{W}$. We read the "g" in g-formula meaning generalisation because the g-formula generalises the adjustment formula we defined in 2.3.

Intuitively, outcome modelling is capturing the mechanism that assigns Y by modelling the data generating process $f(A, \tilde{\mathbf{w}})$, and averaging over the covariate $\mathbf{W}$ distribution.

Adopting parametric models indexed by the finite number of parameters θ, following [47], we define the Bayesian g-formula by drawing samples for the counterfactual outcomes after conditioning on the observed data o, such that:

$$p(\tilde{y}_a|o) = \int\int p(\tilde{y}|a, \tilde{\mathbf{w}}, \theta)p(\tilde{\mathbf{w}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|o)d\boldsymbol{\theta}d\tilde{\mathbf{w}} \tag{3.32}$$

We integrate over the observed confounder $\tilde{\mathbf{w}}$ and the uncertainty on the model parameters θ. Therefore, we estimate the counterfactual outcomes Y^a from their posterior predictive draws, $p(\tilde{y}_a|o)$. Finally, we can compute causal estimands, for example, the ATE by comparing the means of $p(\tilde{y}_1|o)$ and $p(\tilde{y}_0|o)$.

Model misspecification

Let us briefly consider the modelling choices of what $f(A, \tilde{\mathbf{w}})$ should be. We consider the following hypothetical linear structural causal model from [119] to be the ground truth of our running example, given by:

$$\begin{aligned} A &:= \alpha_1 W \\ Y &:= \beta_1 W + \gamma A \end{aligned} \tag{3.33}$$

where $\mathbf{W}$ is a sufficient adjustment set, A is a binary treatment, Y is the outcome, α is the regression coefficient of W on A, β the regression coefficient of W on Y, and γ the regression coefficient of A on Y. Importantly, we consider no interactions, i.e. the SCM is entirely linear in $\mathbf{W}$ and A. We know that from equation 3.2 the ITE, for this counterfactual outcome model takes the form:

$$\begin{aligned} \text{ITE}(w_i) = & \\ & \mathbb{E}\left[\text{Y}^1(w_i) - \text{Y}^0(w_i)\right] = \\ & \mathbb{E}\left[(\beta_1 w_i + \gamma) - (\beta_1 w_i)\right] = \\ & \gamma \end{aligned} \tag{3.34}$$

where we plug-in the counterfactual outcomes $\text{Y}^1(w_i)$ and $\text{Y}^0(w_i)$ according to the assumed form in equation 3.33. Similarly, the ATE, see equation 3.3, is given by the average of ITE over all the individuals w_i, such that:

$$\text{ATE} = \mathbb{E}_{p(w)}\left[\text{ITE}(w_i)\right] = \gamma \tag{3.35}$$

We conclude that the ATE value is also equal to the γ term, i.e. the regression coefficient for A *is* the ATE.

A well-known vulnerability of assuming a linear model for treatment effects is that the linear model may be misspecified [119]. For example, let us consider that the ground truth data generating process $f(A, \tilde{\mathbf{w}})$, instead takes the form:

$$\begin{aligned} A &\coloneqq \alpha_1 W + \alpha_2 W^2 \\ Y &\coloneqq \beta_1 W + \beta_2 W^2 + \gamma A \end{aligned} \tag{3.36}$$

where $\alpha_2 W^2$ and $\beta_2 W^2$ are quadratic terms in the regression of covariates on treatment and outcome, i.e. the counterfactual outcomes are still linear on treatment. The target causal estimand is again the ATE, given by:

$$\text{ATE} = \mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] = \gamma \tag{3.37}$$

If following [119] we hypothesised a model from equation 2.2 instead, and we only adjusted for $\mathbf{W}$, given the relation between $\mathbf{W}$, W^2 and A in the ground truth data generating process from equation 3.36, it is straightforward to show that our estimate of the ATE takes the form:

$$\hat{\gamma} = \gamma + \beta_2 \frac{\mathbb{E}[wa]\,\mathbb{E}\left[w^2\right] - \mathbb{E}\left[a^2\right]\mathbb{E}[wa]}{\mathbb{E}[wa]^2 - \mathbb{E}\left[w^2\right]\mathbb{E}\left[a^2\right]} \tag{3.38}$$

where we have a bias term that can be arbitrarily large or small depending on β_2.

In conventional statistical analysis, the regression coefficients play a central role, sometimes interpreted with causal lens [120]. The rationale for the long-standing tradition of paying attention to the regression coefficients is to interpret the prediction problem in terms of the feature of relevance being a treatment, the treatment being linear for the counterfactual outcome, and the treatment's coefficient as having a relationship with the ATE. Moreover, that also explains the rationale behind estimating confidence intervals to measure the impact of sampling variability. In contrast, in data-driven applications with high-dimensional models, it is well-known that there are redundancies between the features that can bias estimates towards zero [121].

Other severe limitations of the linear parametric regression approach are that the treatment causal effect is constant among all individuals, as we showed in equation 3.34, which can be an unnatural assumption in most cases. For an extended critique of regression coefficients' application as having a relationship with the ATE, see, [120] and references within. We conclude that we can get the wrong results if we make the wrong assumptions about the estimator's form. Furthermore, Chapter 5 explore methods that can help in avoiding model misspecification.

3.4.2 Inverse probability weighting

In RWD, the probability of receiving a treatment depends on confounders, variables that affect the outcome. Expert doctors may prescribe the optimal treatment given the conditions of each patient, for example, more potent treatments for sicker patients. As we described above, confounders may potentially bias treatment effect estimates. Under conditional exchangeability and positivity, we can reweight the data using the expectation of treatment assignment to match a randomised experiment better [122]. We already defined the expectation of treatment assignment and denoted it as the propensity score back in equation 3.10.

Propensity score The propensity score is the probability of taking treatment $A = a$, given the sufficient adjustment set of covariates $\mathbf{W}$, and is given by:

$$e(\mathbf{W}) = \Pr(A = a|\mathbf{W}) \tag{3.39}$$

where $e(\mathbf{W})$ describes the propensity for taking treatment, i.e. how likely the patient i is to take treatment given $\mathbf{w}_i$. The propensity score theorem [91] asserts that under conditional exchangeability and positivity, given $e(\mathbf{W})$ implies unconfounded treatment effects, mathematically is written as:

$$Y(A) \perp\!\!\!\perp A|\mathbf{W} \Rightarrow Y(A) \perp\!\!\!\perp A|e(\mathbf{W}) \tag{3.40}$$

figure 3.7 adapted from [123] shows a visual DAG proof of the propensity score theorem.

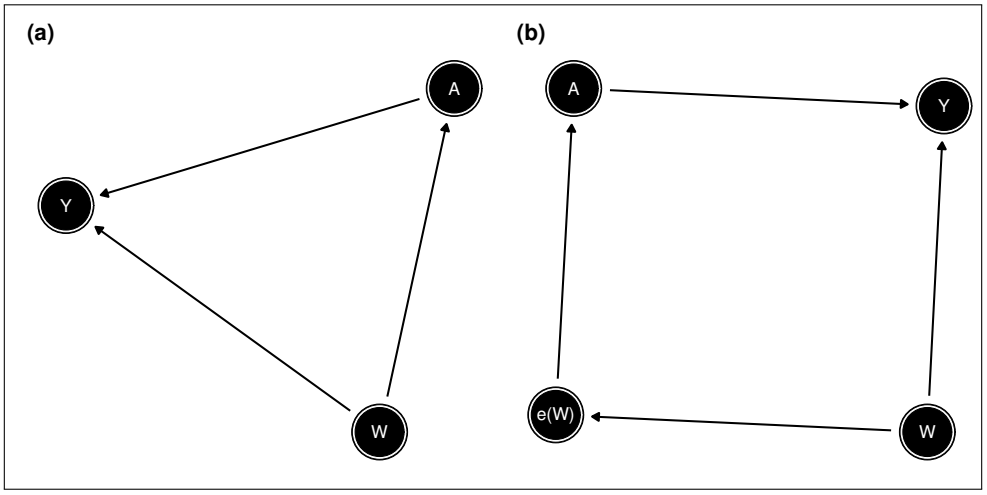


Figure 3.7: DAG proof of the propensity score theorem. (a) $\mathbf{W}$ is a sufficient adjustment set of covariates. (b) Conditioning on the propensity score $e(\mathbf{W})$ block the back-door path $A \leftarrow \mathbf{W} \rightarrow \mathrm{Y}$.

The inverse probability weighting (IPW) general idea is reweighing the data to match a randomised experiment better [123, 124]. To do so, IPW weights down individuals that were very likely to receive the treatment and weights up individuals that were very unlikely to receive the treatment. To do so, we reweight using the reciprocal of the propensity score $e(\mathbf{W})$. Notably, after reweighting, treatment A does not depend on $\mathbf{W}$ anymore. Hence, by multiplying by the propensity score's reciprocal, we have removed $\mathbf{W}$'s impact on A. The IPW estimand takes the form:

$$\mathbb{E}\left[Y^a\right] = \mathbb{E}\left[\frac{I(A=a)Y}{P(a|W)}\right] \tag{3.41}$$

where $I(A = a)$ denotes an indicator that treatment A takes value a, and $P(a|W)$ denotes the unknown propensity score. Under exchangeability, the IPW estimand is equivalent to the g-formula estimand from equation 3.31. However, the statistical estimand suggests a different estimation method. Applying IPW to the binary treatment example, a frequentist ATE estimator takes the form:

$$\widehat{ATE}_{IPW} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i A_i}{\hat{e}\left(w_i\right)} - \frac{y_i\left(1-A_i\right)}{1-\hat{e}\left(w_i\right)}\right] \tag{3.42}$$

Even if $\mathbf{W}$ is high-dimensional, the propensity score $e(\mathbf{W})$ is only one dimensional, i.e. a scalar. However, we do not have access to the $e(\mathbf{W})$ directly, which means it needs a model. Hence, the high-dimensionality and model misspecification obstacles shift from modelling the outcome Y to modelling $e(\mathbf{W})$. However, we can model for $e(\mathbf{W})$ without predefined assumptions on its parametric form, for example, a neural network.

3.5 Bounds and sensitivity analyses

In the previous section, we discussed estimation of treatment effect assuming exchangeability, i.e., we observed all the confounders. We have noted previously in introducing the concept of exchangeability, see Section 3.2, that in RWD, there may exist hidden variables h. Although h are confounders by affecting the potential treatment outcomes and critically impacting the treatment decision, we have not measured them in our RWD see figure 3.8. In this case, the adjustment formula presumes that we should also adjust for h, such that:

$$\begin{aligned}\mathbb{E}\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] = \\ \mathbb{E}_{\mathbf{W}}\left[\mathbb{E}\left[Y|A=1,\mathbf{W},h\right] - \left[Y|A=0,\mathbf{W},h\right]\right] \approx \\ \mathbb{E}_{\mathbf{W}}\left[\mathbb{E}\left[Y|A=1,\mathbf{W}\right] - \left[Y|A=0,\mathbf{W}\right]\right]\end{aligned} \tag{3.43}$$

However, we can not adjust for h because the RWD did not collect it. Therefore, we would be limited to adjust for $\mathbf{W}$ and speculate that by doing so, we are roughly approximating the treatment's effect.

This section evaluates how conclusions about treatment's effect might have changed in the presence of hidden confounding factors, h. We will first tackle the problem with minimal assumptions and the no-assumptions bound approach, allowing us to find an interval that contains the treatment's effect. In chapter 5, we will demonstrate how we can tighten this interval by appending the optimal treatment selection assumption. The second part of the section is on sensitivity analysis, wherein we cast the treatment's effect estimate as a function of the robustness to hypothetical h and its sensitivity to variations of the input.

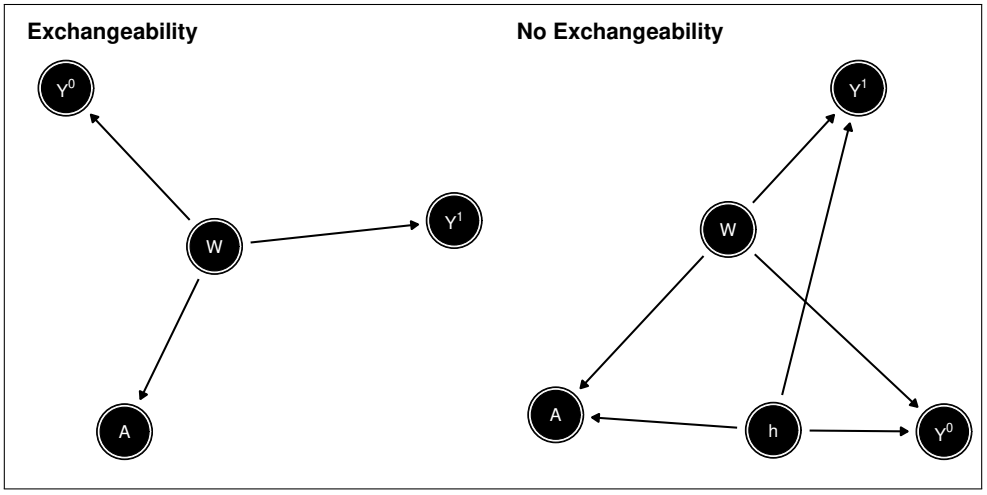


Figure 3.8: Left: exchangeability, Y^0 and Y^1 are conditionally independent of A given $\mathbf{W}$. Right: no exchangeability, Y^0 and Y^1 are not conditionally independent of A given $\mathbf{W}$ because the confound h opens a back-door path from A to Y^0 and Y^1.

3.5.1 Bounds

The work on bounds is motivated mainly by the idea that the assumption of no unobserved confounding is unrealistic. The theory around causal bounds that we present builds on the work of Manski [125, 126]. Neal [85, 86, 127] presents a straightforward adaptation of Manski's work on causal bounds. Chapter 5 develops the theory of causal bounds for survival analyses in RWD. Here, we review the background necessary for developing survival causal bounds.

A helpful concept defined in [125] is the law of decreasing credibility, which states: *'the credibility of inference decreases with the strength of the assumptions maintained'*. Although exchangeability is very useful because it allows us to estimate the treatment's effect by using the adjustment formula

from equation 3.16, it is a relatively extreme assumption. Moreover, ignoring random variability, we consider estimating the average treatment's effect as a point. However, we could make weaker assumptions than exchangeability, trading precision for credibility in our inference by bounding the treatment's effects.

We can regard bounding the treatment's effect as a trade-off between precision and credibility [86]. If we are willing to use strong assumptions such as exchangeability, we can estimate the treatment's effect very precisely, as a point, but because we used such strong assumptions, our conclusions are not that credible. In contrast, we can get more credible conclusions by making weaker assumptions. However, in exchange, we have to trade-off precision by identifying an interval rather than a point.

A trivial bound on treatment's effect is straightforward to obtain if counterfactual outcomes are naturally bounded [127]. For example, if our potential outcomes Y^0 and Y^1 are survival probabilities, we know that the bounds are between 0 and 1. Then we know that:

$$-1 \leq Y_i^1 - Y_i^0 \leq 1 \tag{3.44}$$

We consider that we take the maximum value of the outcome, which is 1, and the minimum value of the outcome, which is 0. Therefore, we know that the treatment's effect bounds are:

$$-1 \leq \mathbb{E}\left[Y_i^1 - Y_i^0\right] \leq 1 \tag{3.45}$$

From equation 3.45, following [127] we can obtain the trivial bound with an interval of length 2. We can generally use β denoting *imphimum* or lower bound as the minimal counterfactual outcome and α denoting *supremum* or upper bound as the maximal counterfactual outcome, such that:

$$\alpha - \beta \leq \mathbb{E}\left[Y_i^1 - Y_i^0\right] \leq \beta - \alpha \tag{3.46}$$

where we can obtain a trivial bound of length $2\,(\alpha - \beta)$.

In order to derive more interesting bounds, we need to use the observational-counterfactual decomposition [85, 86]. We will define it in terms of the average treatment's effect. First, we will use linearity of expectation to get:

$$\mathbb{E}\left[Y(1) - Y(0)\right] = \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \tag{3.47}$$

Then, we separate each counterfactual outcome into two components each,

conditioning on treatment and marginalising it out:

$$\begin{aligned} &P(T=1)\mathbb{E}\left[Y^1|T=1\right] + P(T=0)\mathbb{E}\left[Y^1|T=0\right] \\ -&P(T=1)\mathbb{E}\left[Y^0|T=1\right] - P(T=0)\mathbb{E}\left[Y^0|T=0\right] \end{aligned} \tag{3.48}$$

Then, using consistency, we can obtain:

$$\begin{aligned} &P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\mathbb{E}\left[Y^1|T=0\right] \\ -&P(T=1)\mathbb{E}\left[Y^0|T=1\right] - P(T=0)\mathbb{E}\left[Y|T=0\right] \end{aligned} \tag{3.49}$$

Tying all together, we obtain the observational-counterfactual decomposition [127], with observational terms and counterfactual terms, which takes the form:

$$\begin{aligned} &\mathbb{E}\left[Y(1)-Y(0)\right] = \\ &\qquad \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] = \\ &\qquad P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\mathbb{E}\left[Y^1|T=0\right] \\ &\qquad - P(T=1)\mathbb{E}\left[Y^0|T=1\right] - P(T=0)\mathbb{E}\left[Y|T=0\right] \end{aligned} \tag{3.50}$$

No assumptions bound

The no-assumptions bound only assumes that the counterfactual outcomes are bounded [127]. In the no-assumptions bound, we apply the observational-counterfactual decomposition from equation 3.50. We obtain bounds for the decomposition's counterfactual parts because we can not compute them from the observational distribution. We plug in the observational quantities to the bounds because we can compute them from the observational distribution. For the upper-bound, the counterfactuals' positive quantity takes the maximum outcome possible, α, and the counterfactuals' negative quantity takes the minimum outcome possible, β. Then, the upper-bound takes the form:

$$\begin{aligned} &\mathbb{E}\left[Y(1)-Y(0)\right] \leq \\ &\qquad P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\alpha - \\ &\qquad P(T=1)\beta - P(T=0)\mathbb{E}\left[Y|T=0\right] \end{aligned} \tag{3.51}$$

The reverse reasoning works for the lower-bound. The counterfactuals' positive quantity takes the minimum outcome possible β and the counterfactuals' negative quantity takes the maximum outcome possible α. Then, the lower-

bound takes the form:

$$\begin{aligned}\mathbb{E}\left[Y(1)-Y(0)\right] \geq \\ P(T=1)\mathbb{E}\left[Y|T=1\right]+P(T=0)\beta- \\ P(T=1)\alpha-P(T=0)\mathbb{E}\left[Y|T=0\right]\end{aligned} \tag{3.52}$$

It is straightforward to calculate the length of the interval of the no assumptions bound by subtracting the lower bound from the upper bound, giving:

$$P(T=0)\alpha+P(T=1)\alpha-P(T=1)\beta-P(T=0)\beta=\alpha-\beta \tag{3.53}$$

The no assumptions interval length, without making any additional assumptions, is half of the trivial bound interval length from equation 3.46. In our running example, where the potential outcomes are survival probabilities naturally bounded between 0 and 1. The no assumptions interval length is 1, which is half of the naive bound. However, the no assumptions bound is still unsatisfactory because it will always contain zero, i.e. it does not allow us to estimate the sign of the treatment's effect. Below, we will introduce another assumption that will allow us to construct bounds that can precisely estimate the sign of the treatment's effect.

Optimal treatment selection bounds

A pertinent consideration in analysing real-world data is that the data's originators are expert doctors prescribing treatment. In this setting, patients' treatment selection is optimal, a fact we can use to improve our bounds framework to precisely estimate the sign of the treatment effect. For example, in analysing RWD for Oncology, we must consider that the doctor selects the best treatment for each patient. Recently, [128] described a related scenario where the convalescent plasma's impact of coronavirus(COVID-19) patients statistical significance in an observational RWD study disappeared in a randomised clinical trial (RCT).

The optimal treatment selection (OTS) assumption [86, 125–127] says that we have expert doctors prescribing the treatment, and the doctor *always* prescribes the best available option for each patient. In this scenario, the doctor gives the best treatment for each individual. Therefore, this "Perfect Doctor" acts as a confounder of the treatment effects.

Formally, the OTS assumption says that if an individual is in a given treatment group $A_i = a$, her expected potential outcome under this treatment is best or equal to her expected potential outcome in another treatment group $A_i = \neg a$. Furthermore, if an individual is not in a given treatment group $A_i = \neg a$, her expected potential outcome in the treatment group $A_i = a$ would

be lower. Mathematically:

$$A_i = a \Rightarrow Y_i^a \geq Y_i^{\neg a},\ A_i = \neg a \Rightarrow Y_i^{\neg a} \geq Y_i^a \tag{3.54}$$

Therefore, the general OTS assumption implies two inequalities: the direct OTS assumption and the contrapositive OTS assumption.

For simplicity, let us consider the conventional treatment-control setting $a \in [0, 1]$. Then, the first inequality implied by the OTS assumption is that the expected potential outcome under treatment in the control group is worst or equal to the expected potential outcome under no treatment in the control group, which by consistency is the observed outcome in the control group, such that:

$$\mathbb{E}\left[Y^1|A = 0\right] \leq \mathbb{E}\left[Y^0|A = 0\right] = \mathbb{E}\left[Y|A = 0\right] \tag{3.55}$$

The second inequality implied by the contrapositive of the OTS assumption $A_i = 1 \Leftarrow Y_i^0 \leq Y_i^1$, is that the expected potential outcome under treatment in the control group is less or equal to the expected potential outcome under treatment in the treatment group, which by consistency is the observed outcome in the treatment group, such that:

$$\mathbb{E}\left[Y^1|A = 0\right] \leq \mathbb{E}\left[Y^1|A = 1\right] = \mathbb{E}\left[Y|A = 1\right] \tag{3.56}$$

We can use equation 3.55 to derive an upper bound with the OTS assumption applying the observational-counterfactual decomposition from equation 3.50, such that:

$$\begin{aligned}
&\mathbb{E}\left[Y(1) - Y(0)\right] = \\
&\qquad P(T = 1)\mathbb{E}\left[Y|T = 1\right] + P(T = 0)\mathbb{E}\left[Y^1|T = 0\right] \\
&\qquad - P(T = 1)\mathbb{E}\left[Y^0|T = 1\right] - P(T = 0)\mathbb{E}\left[Y|T = 0\right] \leq \\
&\qquad P(T = 1)\mathbb{E}\left[Y|T = 1\right] + P(T = 0)\mathbb{E}\left[Y|T = 0\right] \\
&\qquad - P(T = 1)\beta - P(T = 0)\mathbb{E}\left[Y|T = 0\right] = \\
&\qquad P(T = 1)\mathbb{E}\left[Y|T = 1\right] - P(T = 1)\beta
\end{aligned} \tag{3.57}$$

Similarly, we can get a lower bound. To do so, we flip the inequality from equation 3.54, such that:

$$\mathbb{E}\left[Y^1|A = 0\right] \leq \mathbb{E}\left[Y^0|A = 0\right] = \mathbb{E}\left[Y|A = 0\right] \tag{3.58}$$

We then plug in equation 3.58 to obtain:

$$\begin{aligned}\mathbb{E}\left[Y(1)-Y(0)\right] = \\ & P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\mathbb{E}\left[Y^1|T=0\right] \\ & - P(T=1)\mathbb{E}\left[Y^0|T=1\right] - P(T=0)\mathbb{E}\left[Y|T=0\right] \geq \\ & P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\beta \\ & - P(T=1)\mathbb{E}\left[Y|T=1\right] - P(T=0)\mathbb{E}\left[Y|T=0\right] = \\ & P(T=0)\beta + P(T=0)\mathbb{E}\left[Y|T=0\right]\end{aligned} \tag{3.59}$$

Tying all together, we obtain the complete bound:

$$\begin{aligned}\mathbb{E}\left[Y^1 - Y^0\right] &\leq P(T=1)\mathbb{E}\left[Y|T=1\right] - P(T=1)\beta \\ \mathbb{E}\left[Y^1 - Y^0\right] &\geq P(T=0)\beta + P(T=0)\mathbb{E}\left[Y|T=0\right]\end{aligned} \tag{3.60}$$

Likewise, we can use equation 3.56 to derive another set of upper and lower bound with the contrapositive OTS assumption:

$$\begin{aligned}\mathbb{E}\left[Y^1 - Y^0\right] &\leq \mathbb{E}\left[Y|T=1\right] - P(T=1)\beta - P(T=0)\mathbb{E}\left[Y|T=0\right] \\ \mathbb{E}\left[Y^1 - Y^0\right] &\geq P(T=1)\mathbb{E}\left[Y|T=1\right] + P(T=0)\beta - \mathbb{E}\left[Y|T=0\right]\end{aligned} \tag{3.61}$$

Because both intervals follow from the OTS assumption, we can choose from which set of intervals take the lower and upper bound, thereby obtaining a more precise interval that potentially determines the treatment's effect sign.

3.5.2 Sensitivity analysis

In conducting sensitivity analysis, we aim to ascertain if our inference is robust to the decisions made in obtaining it. Such decisions comprise data inputs and assumptions. Hence, we distinguish two sensitivity analyses approaches: sensitivity to hidden unobserved confounders and variations of the inputs.

Sensitivity analysis of omitted variable bias

Sensitivity analysis of omitted variable bias [84] constitutes a modelling framework that quantifies the bias for gradual levels of exchangeability violations. As mentioned above, in estimating treatments effects, it is helpful to assume exchangeability because it allows us to obtain unbiased estimates. Sensitivity analysis of omitted variable bias questions how far from exchangeability we have to move to see the results we saw in the data. In a sensitivity analysis of omitted variable bias, we conventionally allow for exchangeability conditional on the hidden covariate h, such that:

$$\left(\mathrm{Y}^0, \mathrm{Y}^1\right) \perp\!\!\!\perp A|W, h \tag{3.62}$$

However, as we have noted before, we can not estimate the impact of h because we have not collected data on them. Instead, we conventionally assume that h is a confounder and compute the treatment effects for different coefficients of h. Beyond linear settings, presented sensitivity analysis of omitted for logistic regression on the propensity score [87]. More recently, [129] advanced a more flexible approach that does not presume a simple parametric form for the propensity score, allow h to be multivariate and not necessarily binary, only assuming a parametric form for the outcome model.

Sensitivity analysis to variations of the input

Sensitivity analysis to variations of the input is a modelling and simulation framework that quantifies the robustness of our inference to changes in the input data. A local sensitivity analysis (LSA) quantifies the impact of variations of the input in models that use ordinary differential equations [130]. For a model $f(A, \tilde{\mathbf{w}}, \theta)$, and an estimated treatment effect Y^a, we analyse:

$$\left.\frac{\partial f(A, \tilde{\mathbf{w}}, \theta)}{\partial Y^a}\right|_{Y^a} \tag{3.63}$$

or

$$\left.\frac{\partial f(A, \tilde{\mathbf{w}}, \theta)}{\partial \tilde{\mathbf{w}}_i}\right|_{Y^a} \tag{3.64}$$

for one parameter θ_i, or data input x_i at-a-time.

3.6 Summary

This chapter illustrated the tools of causal modelling that we will use in analysing real-world data in the following applications chapters. We start by defining treatment and outcome of interest with consistency. To do so, we need to define inclusion and exclusion criteria to collect data on the defined populations that meet those, ensuring that treatment is well-defined. To ensure exchangeability and thereby allow the estimation of unbiased treatment effects, we need to cooperate with clinicians and use their domain expert knowledge to collect all potential confounders. Testing for positivity ensures that the counterfactuals are practically identifiable from data. In real-world applications, usually, we will need to resource for modelling to smooth over the strata of almost continuous categories and high-dimensional space of the confounders. This chapter showed diverse methods for statistical modelling, including:

1. Frequentist methods to estimate parameter's mode and compute Wald confidence intervals.

2. Bayesian methods to estimate the posterior distribution, prior regularisation, and hierarchical models.

3. Deep learning methods to approximate non-linear associations of inputs and outputs.

Here, we demonstrated how to convert the causal estimand into a statistical estimand either by outcome modelling or IPW. Dependent on the complexity of the actual data generating process, model misspecification is a challenge for estimating the counterfactual outcomes. For our inference to be reliable and have external validity, we might need to weaken our exchangeability assumption. We presented two different methods that allow for weaker assumptions and studied assumptions' impact on our inference, bounds and sensitivity analysis, respectively. We can use the assumptions that the treatment selection is optimal to tighten our bounds and estimate the unbiased sign of the treatment effect. We can apply sensitivity analysis to study the impact of parameter variations using modelling and simulation of the generative data process. Although we present the different methods in sequential order, we recognise that the modelling and analysing workflow may have loops. We may often iterate on the different steps, and, for example, assuring positivity may require expanding the population to ensure that parameters are practically identifiable from data.

Chapter 4

Multiple Imputations for Missing Values in Machine Learning - a comparative study

Missing data is a universal problem in analysing RWE datasets. In RWE datasets, there is a need for understanding which features best correlate with clinical outcomes. In this context, several biomarkers status may appear as gaps in the dataset that hide meaningful values for analysis. Using the Flatiron NSCLC dataset, including more than 35,000 subjects, we compare the imputation performance of six such methods: predictive mean matching, expectation maximisation, factorial analysis, random forest, generative adversarial networks and multivariate imputations with tabular networks. We also conduct extensive synthetic data experiments with structural causal models. Statistical learning from incomplete datasets should consider several imputation algorithms, the impact of missing data, and the distribution shift induced by the imputation algorithm. For our synthetic data experiments, tabular networks had the best overall performance. Methods using tabular networks can become part of the data integration techniques for data cleaning in RWE studies.

This chapter focuses on developing a new multivariate imputation algorithm that performs multiple imputations and supports mixed data types. We define the causal mechanism of missing data and explain the rationale for considering imputation algorithms in real-world data. Following best practices in developing a new algorithm for imputation [131], we aim to find an accurate imputation algorithm that provides unbiased parameter estimates and covers the uncertainty on the parameter estimates determined from sampling and missing data variance.

We implement the tabular networks (TABNET) approach [19] within the multivariate imputations by emplying a chained equations framework [132]. Hence, we named our new imputation algorithm MITABNET. We conduct a comparative study of MITABNET with several state-of-the-art imputation algorithms. We show improvements in imputation accuracy and parameter estimation in censored survival analysis. Besides, by using MITABNET, we obtain the benefit of performing interpretable imputations, allowing us to investigate which values had more bearing for the imputation process. To conduct the head-to-head comparison of MITABNET and state-of-the-art imputation algorithms, we conduct several simulation studies under different missing data mechanisms and real-world scenarios.

4.1 Background

Missing data are a universal problem in structured tabular datasets arising from RWE datasets. As Little put it, the best resolution for handling missing data is not to have missing data [133]. However, analysing RWE datasets poses the challenge of handling missing values. Indeed, missing data are found not only in observational RWE datasets but also in controlled clinical trials [134]. There are many practical implications when missing data are present; for example, it can lower the power and affect the precision of parameter estimates' confidence intervals, leading to biased estimates. In RWE datasets, there is a need for understanding which biomarkers best correlate with clinical outcomes. A substantial difficulty in this context is that several biomarkers status may appear as gaps in the dataset that hide meaningful values for analysis. Hence, excluding the underlying value of missing data may completely invalidate the results.

4.1.1 Multiple imputations

A conventional ad-hoc method to handle missing data is the complete case analysis, to delete any rows or columns with missing variables. The problem with complete case analysis is that it squanders information reducing the sample size considerably. Imputation algorithms are general strategies that replace missing values with plausible values. Nevertheless, replacing missing values with static values cannot be correct in general. After all, imputed values are estimated, not observed. Therefore, it is often more appropriate to apply a random variable approach to represent missing values. In his seminal paper, Rubin [135] proposed multiple imputations (MI) for survey non-responders to tackle the uncertainty in the missing values that single imputation cannot represent with a point estimate. The general idea of MI is to generate multiple

complete datasets, analyse each dataset separately, and summarise the results, see. figure 4.1 and [136]. Imputation algorithms that perform MI must replace missing values with samples of the missing values' joint probability density function. Therefore, the MI approach embraces the uncertainty in the missing values that single imputation with a point estimate cannot represent. Early studies proposing imputation algorithms for MI often apply conventional statistical methods like expectation maximisation [137]. More sophisticated methods, adapting ideas from MCMC [57], dimension reduction [58], ensemble learning [58] and DNN [54], have been proposed. However, there is a lack of literature on validation, and systematic comparison of imputation methods [138]. Furthermore, the importance of considering missingness patterns and the data distribution when comparing methods has received little attention [139]. Neglecting to do this may lead to biased results concerning the relative performance of imputation methods [40].

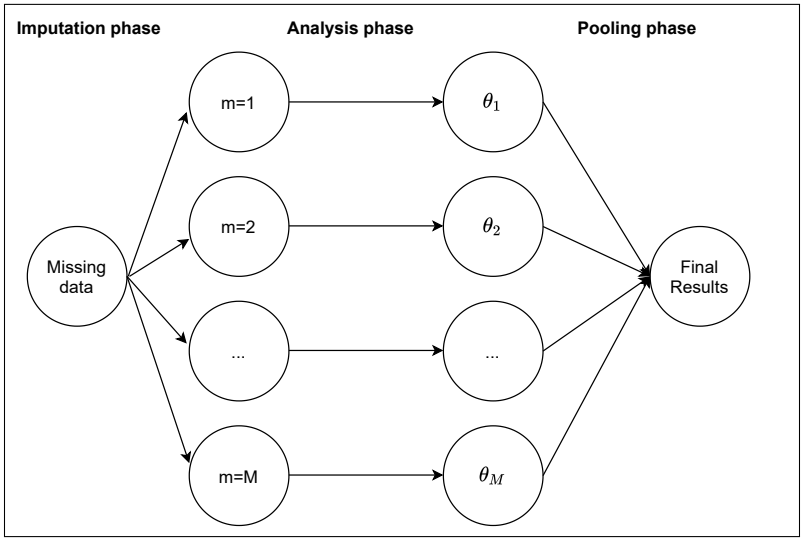


Figure 4.1: Illustration of multiple imputations: in the imputation phase, we generate multiple (M) complete datasets using a stochastic imputation algorithm. Each complete dataset is analysed separately. Therefore, we obtain M θ parameters of interest. Finally, we combine the results to obtain the parameter estimates and the uncertainty on the missing data.

4.1.2 Related work

Different approaches to drawing multiple imputations exist in the literature. Van Buuren et al [140] suggest a Gibbs sampler for the multivariate imputations by chained equations (MICE) algorithm to draw from an approximate posterior distribution after evaluating the "complete data" replacing missing values for placeholders. MICE is an iterative method to use regression strategies such

Algorithm	Imputation	Data types	Reference
MICE	multiple	mixed	[132, 140]
RF	single/multiple	mixed	[58, 138]
PCA	single/multiple	mixed	[56, 141]
EM	multiple	continuous/mixed	[137]
GAIN	single/multiple	continuous	[54]

Table 4.1: Summary of state-of-the-art imputation methods: MICE, multiple imputations by chained equations; RF, random forest; PCA, prinicpa component analysis; EM Expectation Maximisation; GAIN generator-adversarial imputation networks.

as generalised linear model (GLM), predictive mean matching for continuous variables, logistic regression for binary variables, and polytomous logistic regression for categorical variables.

Stekhoven et al [58] suggested multiple imputation random forest (MIRF), a random forest algorithm for missing data imputation that can perform multiple imputations by running the algorithm with different random seeds. In contrast, [138] used the Gibbs sampler from [140] to perform MI with random forest expanding MICE to use non-parametric regression.

Honaker et al [137] built on the expectation-maximisation (EM) approach to impute missing values and performed multiple imputations using a bootstrap-based design. [141] proposed multiple imputation principal component analysis (MIPCA) methods that exploit the global similarity between individuals and the correlation between variables to impute missing datasets. The MIPCA algorithm is flexible enough to perform multiple imputations via a non-parametric bootstrap.

Yoon et al [54] adopted the generative adversarial networks (GAN) framework from [142] to develop a generative adversarial imputation networks (GAIN), which is a non-stochastic neural network constituted by a generator and a discriminator network trained in a zero-sum game. We summarise published algorithms that perform multiple imputations in Table 4.1.

Tabnet

DNN traditionally work with continuous numerical data. For example, for image classification, DNN uses the intensity of each pixel, basically a continuous measurement, a number between 0 and 255. In contrast, tree-based methods such as random forest [58] can handle mixed data types, including discrete categorical and continuous numerical data. The TABNET is a front-end deep learning architecture that can handle mixed data types. At the time of writing, TABNET is a developing algorithm published in the preprint paper attentive interpretable tabular learning [19].

The general idea of TABNET is that performance and explainability are both valuable in neural networks. The main concepts of TABNET comprise attention transformer blocks, which perform instance-wise feature selection that allows having a built-in explainability, feature transformer block, which is a multilayer perceptron with gated linear unit activation sharing layers across different levels, and sequential steps, which mimics ensembling and increases model capacity.

4.1.3 Contributions

This chapter makes several contributions that we summarise as follows:

1. We review the imputation techniques mentioned above, explain their differences in detail, and give recommendations depending upon the missing data problem at hand.

2. We develop a new type of imputation algorithm, which we call MITABNET, expanding the Gibbs sampler from [140] to perform MI with TABNET [19].

3. We modify the architecture of TABNET [19] to perform multiple imputations allowing us to replace missing values with samples of the missing values' joint probability density function estimated with MITABNET.

4. We propose a systematic approach for comparison of imputation methods on RWE datasets.

5. We apply this approach to compare the model performance of seven imputation methods. The six methods are EM, predictive mean matching (PMM) with MICE, bootstrap-based MIPCA, MIRF, GAIN and a method that uses MICE with MITABNET.

6. We conduct a comparative study of the state-of-the-art imputation algorithms in simulations and RWE data benchmarks with clinical oncology applications.

7. We provide Python and R implementations of MITABNET and all the discussed methods.

Table 4.2: Biomarker status and characteristics of the population cohort of NSCLC patients followed-up for the present study.

Characteristic	**N**	**N = 35,012**
ALK	28,583	971 (3.4%)
Unknown		6,429
EGFR	31,124	5,196 (17%)
Unknown		3,888
KRAS	17,025	4,778 (28%)
Unknown		17,987
BRAF	16,426	847 (5.2%)
Unknown		18,586
PD-L1	17,353	6,052 (35%)
Unknown		17,659
Time at risk	35,012	288 (112, 648)
Deceased	35,012	23,773 (68%)

4.2 Methods

4.2.1 RWE dataset analysed

The RWE data source used in the present chapter was the NSCLC Flatiron database[143], a dataset of de-identified patient-level electronic medical records in the United States spanning 280 community practices seven sizeable academic research institutions. In RWE datasets, clinical interest is often on biomarkers that help identify sub-populations that most benefit the targeted treatments. For NSCLC, clinical practice guidelines recommendations include testing the genomic biomarkers EGFR, ALK, KRAS, BRAF, and immunotherapy marker PD-L1. The Flatiron cohort analysed consists of patients who received a diagnosis of advanced NSCLC. The inclusion criteria are patients aged ≥ 18, pathological confirmation of NSCLC obtained from tumour cytology or biopsy, documented diagnosis of unresectable stage III-IV NSCLC, and at least one biomarker status of EGFR, ALK, KRAS, BRAF, or PD-L1. The dataset analysed includes 35,012 individuals, see Table 4.2.

This chapter analyses the impact of biomarker status ALK, BRAF, EGFR, KRAS, and PD-L1 on real-world survival analysis. Even though one can use unknown status as a predictive marker in a multivariate survival model, we argue that such an analysis would not be helpful for clinicians seeking to understand the impact of biomarker status on clinical outcomes. Figure 4.2 displays the combinations of missingness for the biomarkers EGFR, ALK, KRAS, BRAF, PD-L1 in the RWE dataset.

The survival analysis had the following relevant parameters: index date and the end date. We define the index date as the start date of treatment anchoring the survival analysis. We define the end date as the death date for patients for whom this is known or the last confirmed activity for patients for

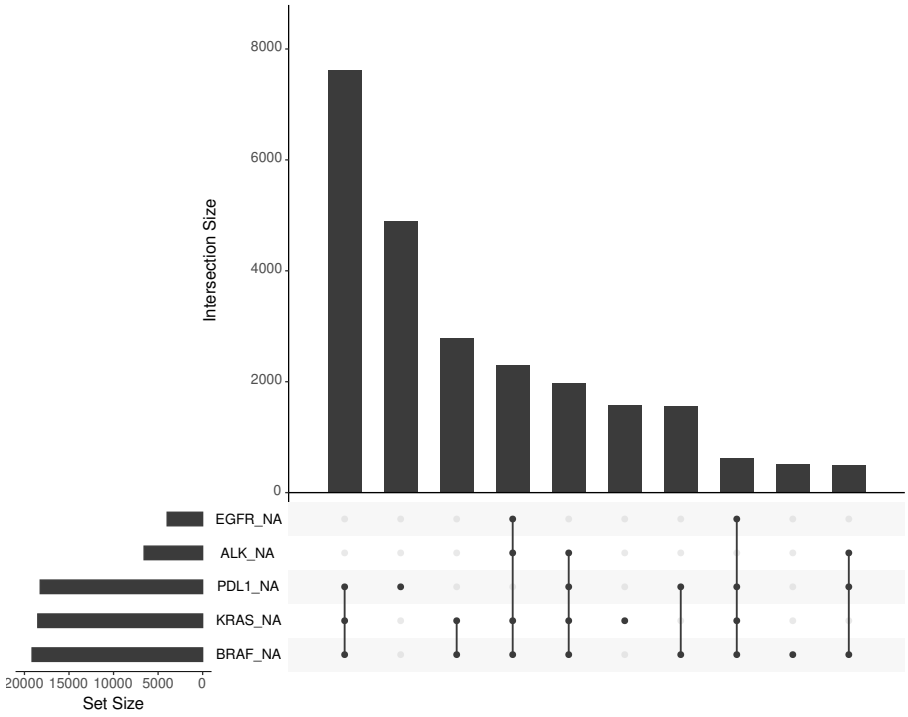


Figure 4.2: EGFR, ALK, KRAS, BRAF, PD-L1 status missingness and its combinations of missingness using the UpSet visualisation method.

whom it is not known. The time at risk is the difference between the end date and the index date.

4.2.2 Benchmark methods for multiple imputations

Different approaches to drawing multiple imputations exist in the literature. The following is a brief description of the battery of algorithms compared in the present chapter.

Expectation-maximisation Honaker et al. [137] built on the EM approach to impute missing values and performed multiple imputations using a bootstrap-based design. The EM algorithm iteratively starts with an expectation step that calculates the likelihood function given by the expected complete data conditional on current parameter estimates. The expectation step is hence, a form of imputation. Then, the maximisation step chooses the model parameters by optimising the likelihood function. For a detailed explanation of the EM imputation algorithm, we refer to [137]. We used the implementation of EM in the R package *Amelia* [137].

Predictive mean matching Gerko et al. [144] suggest a Gibbs sampler for the MICE algorithm to draw from an approximate posterior distribution after evaluating the *complete data* replacing missing values for placeholders.

MICE is an iterative method to use regression strategies such as PMM for any variable. For a detailed description of the MICE -PMM algorithm, we refer to [144]. We use the stable R release package *mice* for PMM [57].

Random forest Stekhoven et al. [58] suggested a random forest algorithm for missing data imputation that can perform MIRF by running the algorithm with different random seeds. The missing random forest algorithm issues predictions for missing values by weighing many relatively uncorrelated trees. [138] implements the random forest algorithm iteratively by fitting the observed values and updating the missing values until meeting a model performance stopping criterion. We use the implementation of missing random forest in the R package *missRanger* [145].

Principal component analysis Josse et al [141] proposed MIPCA methods that exploit the global similarity between individuals and the correlation between variables to impute missing datasets. The MIPCA algorithm starts calculating the MIPCA components and then projects each principal component using the MIPCA prediction. Iterative MIPCA repeats these steps until convergence. The MIPCA algorithm is flexible enough to perform multiple imputations via a non-parametric bootstrap. We estimate the number of components for MIPCA using cross-validation. For a detailed explanation of the regularised iterative MIPCA algorithm, we refer to [141]. We use the regularised iterative MIPCA in the R package *missMDA* [141].

Generative adversarial imputation networks Yoon et al. [54] adopted the generative adversarial framework from [142] to develop a GAIN, which is a non-stochastic neural network constituted by a generator and a discriminator network trained in a zero-sum game. The discriminator attempts to distinguish the imputed values from the actual ones, which predicts the mask. The generator, on the other hand, attempts to deceive the discriminator. The generator inputs are the mask and the original data with missing values that are substituted by noise, e.g., from a Normal random variable. We use the Python implementation of GAIN [54].

4.2.3 Multiple imputations with tabnet

This section describes our approach for generating samples from the missing values' joint distribution with MITABNET. We highlight the two separate components of the algorithm, estimate a specified conditional model with TABNET using dropout, and update one step in the MICE algorithm.

Tabnet with dropout

Tabnet takes the raw input features and applies dropout and batch normalisation. Then, there are consecutive steps that are identical. Each step starts with a feature transformer block, followed by an attentive transformer block that creates the mask, its output is passed to another feature transformer block, which creates both predictions and the input for the next step's attentive transformer block. The predictions are the sum of the step's outputs, passed to a final fully connected layer to resolve any regression or classification problems. Moreover, we can use each step's mask to provide information about the feature's attributes.

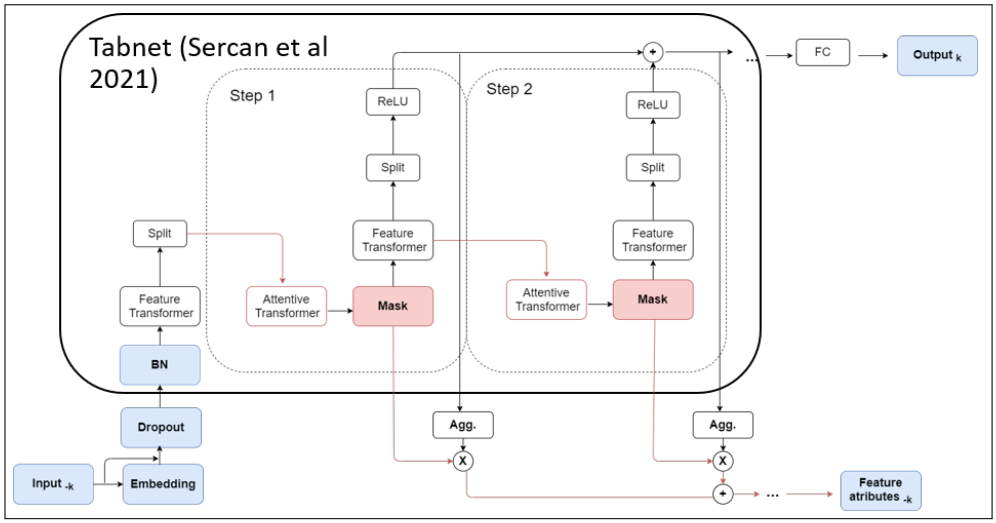


Figure 4.3: Illustrates TABNET's global architecture with the dropout layer. We adopt the TABNET architecture and add dropout to the inputs allowing us to perform multiple imputations.

The main steps and layers in TABNET are as follows:

Embeddings We use an approach similar to that in [146], we apply entity embeddings in TABNET to represent categorical variables. The general idea of embedding is to retain meaning by transposing large vectors to lower-dimensional space. Hence, the categories group the outcome.

Dropout With dropout, we randomly switch-off inputs in each example used to train MITABNET. The implementation is straightforward: it is a regular TABNET with an additional dropout layer applied to the input. The hyper-parameter that controls the rate of dropout is a tuning parameter denoted α.

Batch normalisation Applying batch normalisation normalises the output from the inputs layer per batch, hence the name batch normalisation. The batch normalisation layer has two additional sets of trainable parameters, the mean and the standard deviation.

Feature transformer The feature transformer block comprises four consecutive gated linear units (GLU) blocks. Each GLU block is a fully connected layer, followed by a batch normalisation layer and a GLU activation, given by:

$$GLU(x) = \sigma(x) \odot x \tag{4.1}$$

In addition, there are two shared and two independent blocks. Hence the model shares the first two GLU blocks across decision steps. Moreover, there are skip connections at every GLU block, allowing the training of deeper models.

Attentive transformer The attention transformer block comprises a fully connected layer, batch normalisation layer, and a prior scale. At the initial state, the prior setting is $P_0 = 1$ for all features, which is similar to placing a non-informative prior on the feature attention process. For the following steps, the prior is given by:

$$P_i = \prod_{j=1}^{p} (\gamma - M_j) \tag{4.2}$$

γ is a tuning hyper-parameter. Setting γ close to one allows the model to select different features for each step while setting γ to a higher number will make the model use the same features at all steps. The output of the prior scales is the input to the sparsemax activation function, which projects the probability distribution onto probability simplex such that:

$$p(x) = \underset{p \in \Delta^{K-1}}{\operatorname{argmin}} \parallel p - x \parallel_2^2 \tag{4.3}$$

In brief, the sparse max outputs probabilities that sum to one and induces regularisation.

The cost function of TABNET is given by:

$$J(\theta) = \begin{cases} (x' - x)^2, \text{if } x \text{ is continuous} \\ -\sum_k^K x_k \log\left(x'_k\right), \text{if } x \text{ is categorical} \end{cases}$$

We define the cost function as the squared error for continuous variables and cross-entropy for categorical variables. The optimisation algorithm that we use is the Adams optimiser [12], an extension to stochastic gradient descent explained in section 3.3.3.

Predictions are the output of a fully connected layer which inputs are the sum of the outputs across the steps. Predictions with the X_{-j} inputs for the missing values allows us to perform a single imputation for the features X_j, which for classification are a sample from the outputs of a softmax activation function.

MITABNET

Developing the imputation model was done sequentially, starting with random draws from the observed data X_j^{obs}. The MITABNET algorithm, see the pseudo-code in algorithm 1, is repeated for each $m = 1, \ldots, M$ imputed datasets as follows:

Step I Initialise the complete dataset with random samples from the observed dataset X^{obs}.

Step II For each variable, split the missing and observed datasets, split the observations into train and validation sets, we recommend 80:20, use TABNET to learn the distribution $P(X_j|X_{-j}, R, \theta)$.

Step III Use the trained TABNET to predict the missing values in X_j.

Step I to III are repeated a prespecified number of iterations for each K feature with missing values.

Algorithm 1: Pseudo-code of MITABNET

```
Result: M complete datasets
for m = 1, ..., M do
    initialisation;
    for j = 1, ..., p do
        Random Draw from X_j^obs to fill in initial imputations X_j^0.
    end
    for j = 1, ..., p do
        Define X_{-j}^0 as the currently complete dataset;
        switch X_j do
            case Continuous do
                X_j = TabnetRegressor (X_{-j}^0);
            end
            case Categorical do
                X_j = TabnetClassifier (X_{-j}^0);
            end
        end
    end
end
```

4.2.4 Strategy for comparing methods

A head-to-head comparison of imputation algorithms involves the ability of the algorithms to recover the actual value from an "amputated" dataset. By amputation, we refer to the concept developed in [139], where a simulation algorithm generates the missingness mechanism to obtain datasets that have missing values following a specific pattern. Besides, for inference, a reliable multiple imputations algorithm needs to preserve parameter estimates' moments having low bias, high coverage, and distributional characteristics for the multiply imputed datasets.

Synthetic data generation

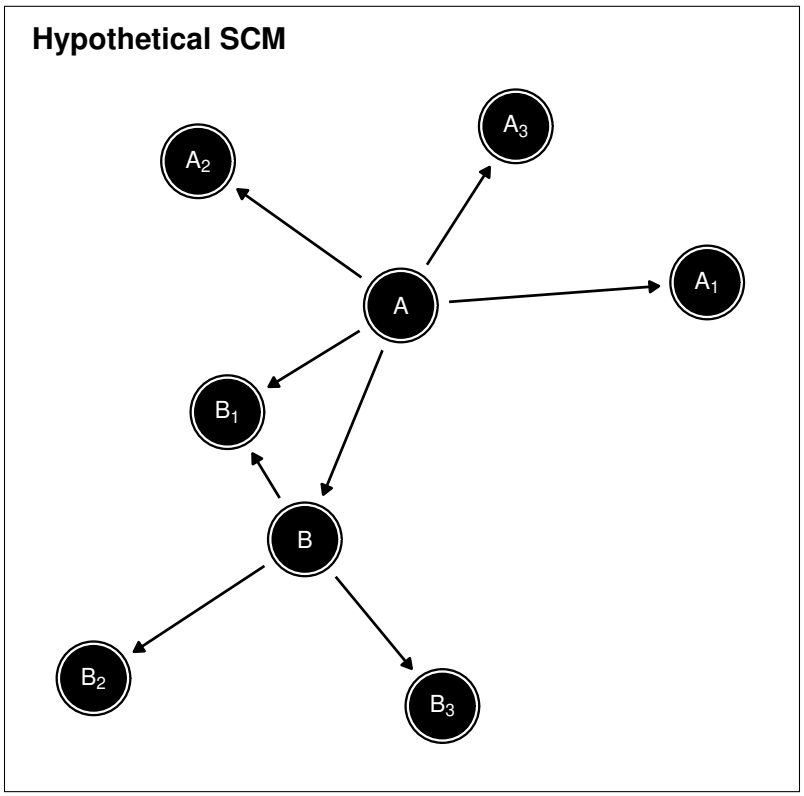


Figure 4.4: Hypothetical DAG depicts the running example structural causal model (SCM).

The following running example is a standard SCM with real-world applications in gene networks [147] used here for demonstration, see Figure 4.4. Let A and B be latent variables, i.e. not observed. A causes B. A also causes three manifest variables A_1, A_2, A_3 and partially causes B_1. B causes three manifest variables B_1, B_2, B_3. The variables $A_1, A_2, A_3, B_1, B_2, B_3$ are manifest, i.e. observable. We generate sample datasets with:

$$
\begin{aligned}
B &\coloneqq f_B(A) \\
A_1 &\coloneqq f_{A_1}\left(A, U_{A_1}\right) \\
A_2 &\coloneqq f_{A_2}\left(A, U_{A_2}\right) \\
A_3 &\coloneqq f_{A_3}\left(A, U_{A_3}\right) \\
B_1 &\coloneqq f_{B_1}\left(A, B, U_{B_1}\right) \\
B_2 &\coloneqq f_{B_2}\left(B, U_{B_2}\right) \\
B_3 &\coloneqq f_{B_3}\left(B, U_{B_3}\right)
\end{aligned}
\tag{4.4}
$$

Notably, $A_1, A_2, A_3, B_1, B_2, B_3$ have random noise attached. From this model, we adjust the linear coefficients to obtain two different datasets with different levels of correlation between the manifest variables, such that:

1. Dataset type I has a high correlation $\rho \approx 0.8$.
2. Dataset type II has a low correlation $\rho \approx 0.2$.

To understand the impact of sample size on the imputation accuracy of the compared algorithms, we generate various datasets, progressively decreasing the sample size with the geometric sequence: 10000, 5000, 2500, 1250, 625. We generate 200 datasets of each type and proceed to "amputate" values. We amputate using MAR, the most commonly accepted missingness mechanism in analysis RWD [148]. Reproducible code is included in appendix A.1.

For MAR, we use a multivariate missingness simulation method based on a multivariate amputation algorithm [139]. Multivariate amputation's general idea is to define the probability that the nth individual's ith variable is missing conditional on the nth individual's other variables' missingness or observed value such that:

$$
P_i^m = \frac{p^m(i) \cdot N \cdot \exp\left(-\sum_{j \neq i} w_j m_j(n) x_j(n)\right)}{\sum_{l=1}^{N} \exp\left(-\sum_{j \neq i} w_j m_j(l) x_j(l)\right)} \tag{4.5}
$$

$p^m(i)$ corresponds to the proportion of missingness, and the w_j are pre-specified weights. The linear combination of weights and variable's value or missingness gives a weighted sum score determining each variable's missingness probability.

$$
wss_i = \sum_{j \neq i} w_j m_j(n) x_j(n) \tag{4.6}
$$

where wss_i denoted the weighted sum score for the ith variable. We introduce MAR by setting the probability of missingness according to a standard right-

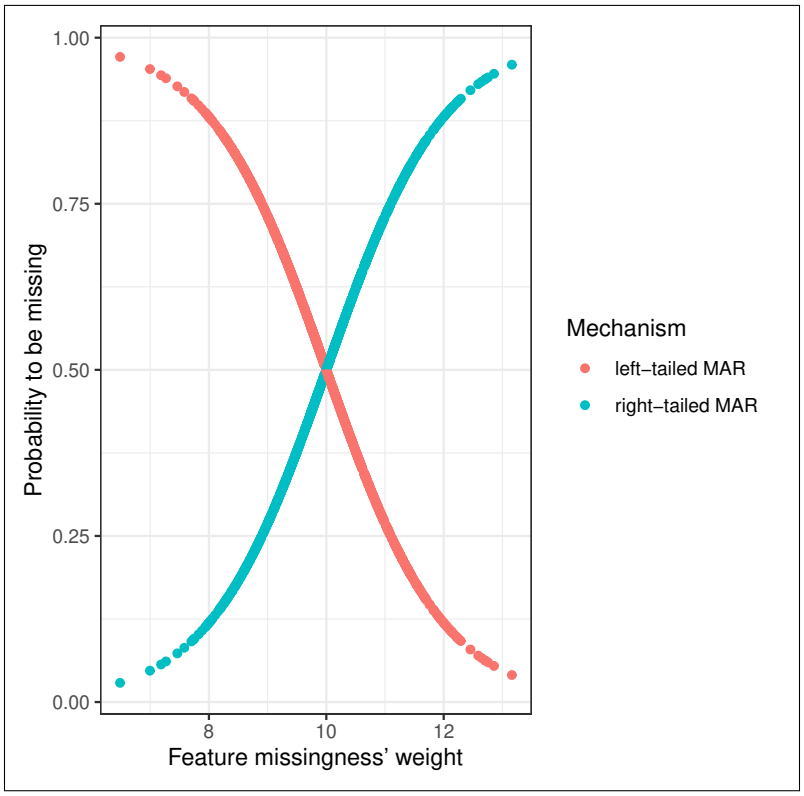


Figure 4.5: Right and left-logistic functions to generate datasets with MAR mechanism.

tailed logistic function, i.e. the likelihood of a given variable missing is positively correlated with the weighted sum wss_i, see Figure 4.5.

Single imputation accuracy

To evaluate the imputation accuracy for single imputations, we report the root mean squared error (RMSE), given by:

$$\sqrt{\sum_{i=1}^{N} \frac{\left(x_i' - x_i\right)^2}{N}} \tag{4.7}$$

where x_i' is the actual amputated value for the ith individual's x variable, and x_i is the imputed value.

Generative survival model

In the context of survival modelling, the survival function $S(t)$ is the probability of surviving to time t and is given by equation 2.6. Our approach to model survival datasets is to define the time-to-event process to be a system of two states (initial and end states), where the survival function (S) is the transition from the initial state to the end state, then given by:

$$\frac{d\mathrm{S}}{dt} = -\mathrm{S} \cdot h(t, \theta, x_i) \qquad \mathrm{S}(0) = 1 \tag{4.8}$$

h denotes the hazard function, and x_i is a vector of individual covariates.

The survival time results from an inverse transform sampling (ITS) process, which is an efficient method to generate independent samples from a given probability density [97]. Equation 4.8 gives the survival function, which is the complementary cumulative distribution function. Hence, the cumulative distribution function is given by:

$$\mathrm{CDF}(T) = 1 - S(T) \tag{4.9}$$

Generally, the CDF for a given time point t is equal to:

$$\mathrm{CDF}(t) = \int_0^t p(u)du \tag{4.10}$$

where we integrate with respect to t and u is a dummy variable. For time $t = 0$ the CDF is 0, and the $S(T)$ is 1. As time approaches infinite the CDF approaches 1, and the $S(t)$ approaches 0. Hence, both the CDF(t) and the $S(T)$ are given in the interval $(0, 1)$. Therefore, given the survival function from

Algorithm 2: Pseudo-code of survival generative model

Result: N individuals, event time T, and censoring indicator d, where E is the administrative time.

for $n = 1, \ldots, N$ **do**
  Generate $u \sim \mathrm{Uniform}(0, 1)$;
  $T = S_T^{-1}(u)$;
  if $T > E$ **then**
    $d \leftarrow False$;
    $T \leftarrow E$;
  end
  else
    $d \leftarrow True$;
  end
end

Equation 4.8 we can generate independent samples with algorithm 2, such that:

1. Generate a random number u from the standard uniform distribution (in the interval $(0, 1)$).

2. Set the event time to T, $S_T(u)$.

3. If the event time occurs during follow-up, set as an observation, otherwise introduce a censored record.

To perform ordinary differential equation (ODE) simulations of survival models, we developed *simtte* [149] an open-source R package that allows for flexible, hierarchical and bespoke survival models, see appendix A.2. For proof of the ITS algorithm, see [96]. Our new R package *simtte* [149] allows specifying $S(t)$ flexibly with ODE and use ITS to sample from the solution of the $S(t)$.

4.2.5 Analysis of interest

The analysis of interest is survival analysis. We used Effron's likelihood [150] to handle tied death times as implemented in the **rms** R package [151]. To be consistent with all imputation methods, we use the same multivariate Cox proportional hazards model [66], given by equation 2.10, where the prognostic index (μ) is given by:

$$\mu_i = \beta_{\text{ALK}} \cdot \text{ALK}_i + \beta_{\text{BRAF}} \cdot \text{BRAF}_i + \beta_{\text{EGFR}} \cdot \text{EGFR}_i + \beta_{\text{KRAS}} \cdot \text{KRAS}_i + \beta_{\text{PD-L1}} \cdot \text{PD-L1}_i \tag{4.11}$$

where β_P are the log-hazard ratios of each P biomarker ALK, BRAF, EGFR, KRAS and PD-L1. The pairs plot matrix in figure 4.6 shows the kernel density estimates of μ in the Flatiron NSCLC data for each imputation method with pairwise scatter plots calculated on the off-diagonal. The pairs plot shows that the different imputation methods' μ are positively correlated but not collinear, indicating that there may be practical differences in the post-imputation prediction performance.

To combine inference in the frequentist Cox proportional hazard model, equation 4.11 above, we calculate a parameter estimate for each imputed dataset (β_m) and use Rubin's rule [152] to average over the estimates. The formula for the point estimate of each parameter estimate ($\bar{\beta}$) is an average over the point estimates of each imputed dataset, such that:

$$\bar{\beta} = \frac{1}{M} \left(\sum_{m=1}^{M} \beta_m \right) \tag{4.12}$$

Furthermore, multiple imputations allow us to compute the total variance, which is helpful for constructing confidence intervals for the parameter estimates. To do so, we must consider the sampling variability within each complete dataset, denoted as σ_w, and related to the conventional estimation of variance, which is given by:

$$\sigma_w = \frac{1}{M} \left(\sum_{m=1}^{M} \sigma_m \right) \tag{4.13}$$

Besides, we must consider the variability between complete datasets, denoted

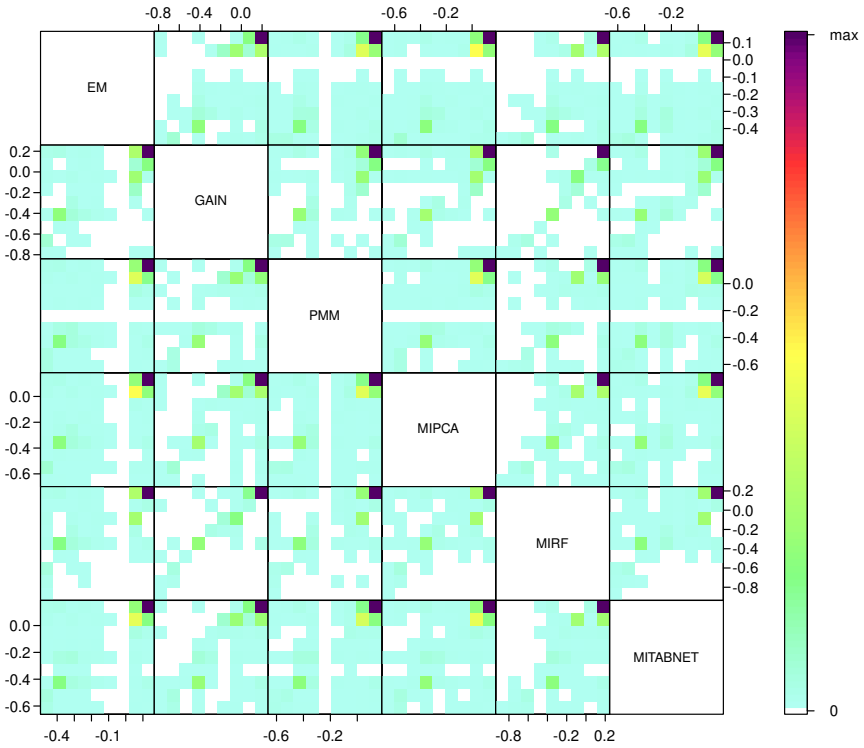


Figure 4.6: Scatter plot matrix of prognostic indexes (μ) and pairwise comparisons using the six different imputation methods on the off-diagonal for the first random imputation sample of the Flatiron NSCLC analytical cohort.

as σ_b, regarded as being the result of missing values, such that:

$$\sigma_b = \sqrt{\frac{\sum_{m=1}^{M} \left(\beta_m - \bar{\beta}\right)^2}{M-1}} \tag{4.14}$$

Finally, to compute the total variance of the parameter estimates, we apply the formula adapted from [153] by combining σ_w and σ_b, such that:

$$\sigma = \sigma_w + \left(1 + \frac{1}{M}\right)\sigma_b \tag{4.15}$$

Using the adjusted formula from [152] to calculate the degrees of freedom [152], it is straightforward to compute a confidence interval for $\bar{\beta}$ with α value, given by:

$$\bar{\beta} \pm t_{\mathrm{df}, \frac{1-\alpha}{2}} \sqrt{\sigma} \tag{4.16}$$

where $\bar{\beta}$ is the pooled estimate obtained from equation 4.12, σ is the total variance obtained from equation 4.15, df is the degrees of freedom and t is the t-statistic.

Imputation models

Imputation models included the five biomarkers status EGFR, ALK, KRAS, BRAF, and PD-L1. Additionally, we use the target survival as a predictor variable as suggested before [58] to improve the data efficiency of the imputation model. The marginal Nelson-Aalen cumulative hazard [154] estimate (H) was perfectly correlated with time at risk (T); see figure 4.7. Hence, we used H to improve the imputation model. The correlation among the biomarkers ranges from 0 to 0.2. To be consistent with all imputation methods, we used the same multivariate Cox proportional hazards model in all cases; see equation 4.11.

Imputations performance

Accuracy use is a convenient yardstick for benchmarking imputation algorithms [54]. However, for inference, our interest is in the distributional characteristics of the multiply imputed datasets, such as preserving parameter estimates' moments having low bias, high coverage of confidence intervals, and the robustness to missingness mechanisms such as MAR. Let a survival dataset given by Y the outcome space comprised by the survival time and the censoring indicator, and X_p the feature space with $1, \ldots, p$ features.

Once we compute $\bar{\beta}$ and σ for each parameter estimate, we evaluate the imputation performance of each imputation algorithm, see table 4.1, by evaluating the following heuristics:

1. **Percentage Bias** : An optimal imputation algorithm should be unbiased.

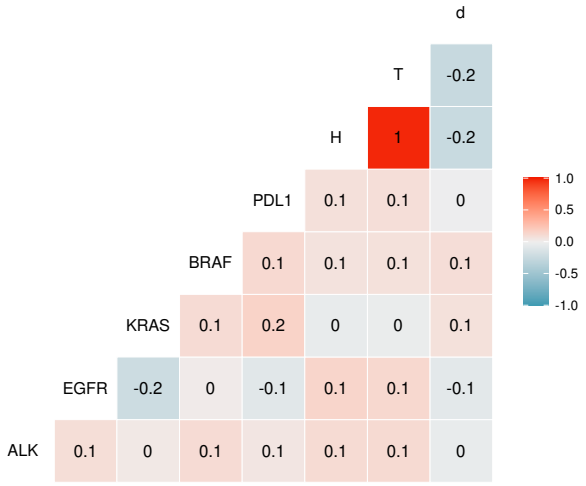


Figure 4.7: Pearson correlations between the biomarker status ALK, BRAF, EGFR, KRAS, PD-L1, cumulative death hazard H, survival time T, and death status in the Flatiron NSCLC dataset.

We compute the parameter estimate $\hat{\beta}$ from the complete dataset before running the amputation algorithm, see section 4.2.4. For each parameter, the percentage bias is given by:

$$100 \cdot \left| \frac{\sum_{p=1}^{P} \left(\bar{\beta}_p - \hat{\beta}_p \right)}{\sum_{p=1}^{P} \left(\hat{\beta}_p \right)} \right| \tag{4.17}$$

where $\hat{\beta}$ is the parameter estimate before amputating the dataset with missing values, and $\bar{\beta}$ is the point estimate obtained with equation 4.12. Therefore, the best imputation algorithm will have a lower percentage bias, with a value of 0 being a perfect imputation model.

2. **Width of Confidence Intervals** : With σ computed with equation 4.15 we compute 95% confidence interval with equation 4.16. A narrow confidence interval that covers the parameter of interest $\hat{\beta}$ are prefered. However, smaller confidence intervals that cover the parameter interest $\hat{\beta}$ indicate sharper inference.

3. **Coverage** : The coverage is the probability that the low (θ_{low}) and upper (θ_{upp}) bounds of the confidence interval include the actual parameter estimate θ. With a 95% confidence interval, we compute if it includes the original parameter of interest $\hat{\beta}$. We repeat this computation 200 times and compute the proportion of times that $\hat{\beta}$ is inside the 95% confidence

interval, which is given by:

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(\widehat{\theta_{low,i}} \leq \theta \leq \widehat{\theta_{upp,i}}\right) \tag{4.18}$$

where I denotes the indicator function, $\widehat{\theta_{low,i}}$ denotes the lower bound of the confidence interval for the ith sample, and $\widehat{\theta_{upp,i}}$ denotes the upper bound of the confidence interval for the ith sample The coverage values range between the interval (0, 1) or in percentage, 0% to 100% — 0% indicating no coverage of the actual parameter and 100% showing a perfect coverage.

4. **Convergence** : For MITABNET we need to monitor convergence. To do so, we use the trace-plot and the $\hat{R}$ metric [155]. In general, the $\hat{R}$ metric evaluates how well Markov chains mixed, within and between chains. Therefore, we suggest it is a helpful statistic for imputation algorithms that rely on the Gibbs sampler, such as MITABNET.

Generation of sample datasets for multiple comparisons

We will not know the resulting criteria for the comparison study based on the data distribution from only the available complete data because of the severity of the missing values in RWE datasets. Instead, one can impute the original data with each of the imputation algorithms (home imputation), which will generate a *concept drift* that we define as the shift in the data distribution induced by the imputation method. Then, one can sample amputated datasets with a MAR mechanism similar to the original data by using the original mask matrix given by the indicators of the cells that were missing in the original data, see figure 4.8. We suggest using a Bernouilli random variable since is the maximum entropy distribution for binary events. We set the probability of missingness p to 0.25 for the initially observed cells and $1 - p = 0.75$ for the initially missing cells. The value of $p = 0.25$ strikes a balance in bootstrapping the missingness pattern from the original dataset while conducting a fair head-to-head comparison among the imputation algorithms (visiting imputation). Note that setting p to 0.5 would yield an MCAR mechanism of missingness. The method has been used for benchmarking imputation algorithms [58].

To illustrate the dependence on the concept drift, we evaluate the imputation performance using three imputation methods on amputated datasets samples of the Flatiron NSCLC cancer biomarker data. Table 4.3 shows the percentage bias results for two such samples. Let us look at the first re-imputed sample. The percentage bias varies depending on the imputation algorithm

Algorithm 3: Algorithm for evaluating the performance of imputation methods in real-world datasets.

```
for each imputation method do
    initial single imputation of X^obs;
end
for each imputation method do
    for each of S random amputations do
        Draw an amputated datasets.
    end
end
for each imputation method as Home do
    for each of S random amputations do
        for each imputation method as Visiting do
            Draw multiple imputations.
            Calculate the single and mutiple criteria on the masked
             values.
        end
    end
end
Compare the different imputation methods in terms of the average over
 the multiply imputed datasets and the spread of the criteria values.
```

Table 4.3: Values of the percentage bias for three imputation methods using two imputed bootstraps from the NSCLC Flatiron dataset.

Method	Sample 1			Sample 2		
	EM	**PMM**	**MITABNET**	**EM**	**PMM**	**MITABNET**
EM	34.7	**18.6**	54	27	**19.2**	46.8
GAIN	53.6	61.6	**8.5**	57.4	69	**4.9**
PMM	90.5	**20.3**	99	79.3	**7.8**	75.8
MIPCA	64.9	**8.5**	66.1	61.1	**14**	83
MIRF	117.3	**24.4**	119.2	96.1	**15.3**	133
MITABNET	**6**	24.5	13.4	**6.9**	28.4	**2.2**

used to obtain the complete dataset. Moreover, EM obtains the lowest percentage bias for the dataset imputed originally with MITABNET, which contrasts with the results from the amputated sample 2, where MITABNET obtained the lowest bias for the dataset imputed originally with MITABNET. The inconsistent results from samples 1 and 2 illustrate that one needs several samples to evaluate the performance of the imputation methods. Another drawback of investigating only one sample is that we do not entirely take advantage of all the available information on the data. Therefore, we sample S amputated datasets and calculate the criteria for each sample, yielding S different values for each of the criteria. We then compute the average of the multiply imputed datasets and their total variance. We sample S independent datasets correlated to the NSCLC Flatiron biomarker data distribution. We

will use S = 200, a number seeking to support uncertainty due to few samples and long computational time.

4.3 Results

4.3.1 Synthetic data experiments

Using equations 4.4 we experiment with increasing sample size: 625, 1250, 2500, 5000 and 10000. In the high correlation setting ($\rho \approx 0.8$), every method performs better than in the low correlation setting ($\rho \approx 0.2$). The trend implies that MITABNET and GAIN consistently outperform each benchmark across sample size and correlation settings. Figure 4.9 shows the RMSE for each imputation algorithms. Figure 4.10 shows that the MICE algorithm converges both in the mean of the parameter estimates $\hat{\theta}$ and the variance σ. In the table 4.4 we report the value $\hat{R}$ that is ≈ 1 confirming that the Markov chains mixed well. We now compare the same benchmarks considering the bias in the

Table 4.4: Convergence of MITABNET in synthetic dataset, $\hat{R}$ statistic

Feature	$\bar{\theta}\hat{R}$	$\sigma\hat{R}$
A1	0.991	1.003
A2	1.006	0.992
A3	1.016	1
B1	1.01	1.03
B2	0.997	1.004
B3	1.02	0.995

estimate of the log-hazard ratio for the synthetic data. The log-hazard ratio $\bar{\beta}$ after imputation should be as near to the original $\hat{\beta}$ as possible. Table 4.5 shows the results of percentage bias for a sample size of 10,000.

Table 4.5: Percentage bias in high and low correlation scenario for a sample size of 10,000.

	Percentage Bias	
model	$\rho = 0.8$	$\rho = 0.2$
EM	11.5 ± 0.6	74.9 ± 1.5
GAIN	10.5 ± 1.3	41.5 ± 1.7
PMM	10.1 ± 1.2	74.8 ± 1.7
MIPCA	15.7 ± 1.3	95.2 ± 1.7
MIRF	10.8 ± 0.7	74.3 ± 1.7
MITABNET	2.6 ± 1.4	36.2 ± 1.4

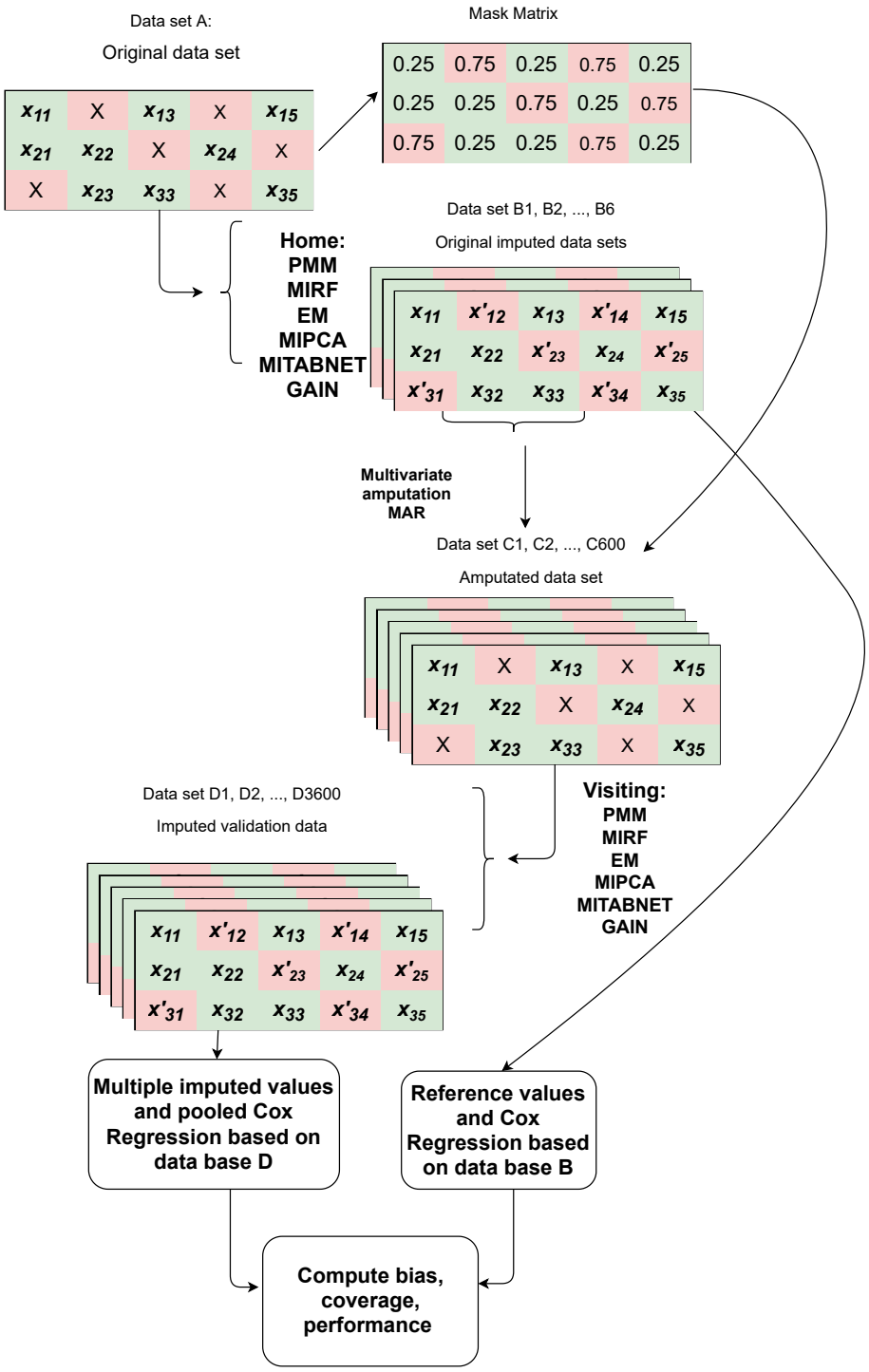


Figure 4.8: Generation of datasets with artificial missingness from a population of patients with NSCLC in the Flatiron database. datasets B1, B2, ..., B6 are imputed datasets with the imputation algorithms (PMM, MIRF, EM, MIPCA, MITABNET, GAIN) serving as host to the comparison or imputing at home. datasets C1, C2, ..., C600 are amputated dataset 100 for each B dataset. datasets D1, D2, ..., D3600 are imputed datasets 6 for each C dataset, the imputation algorithms are visiting.

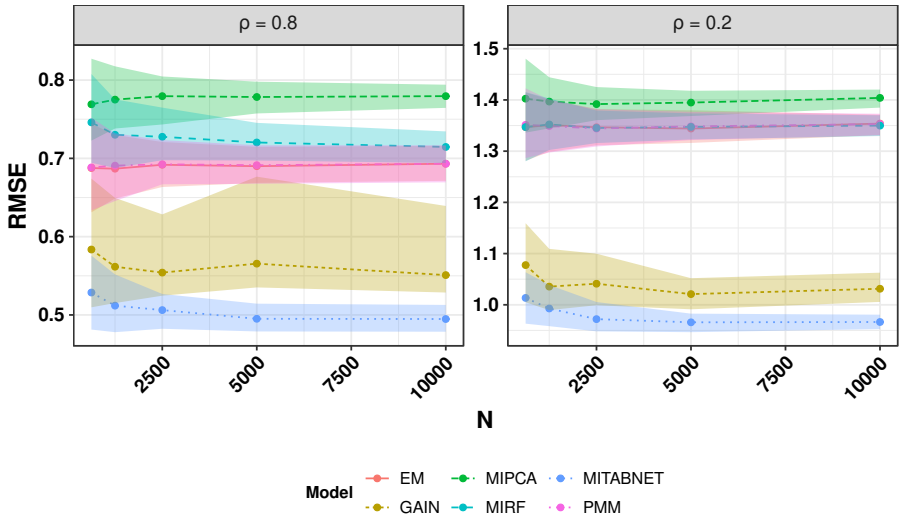


Figure 4.9: Head-to-head comparison of imputation algorithms in post-imputation accuracy in high and low correlation scenario and increasing sample sizes.

4.3.2 Real-world data experiments

MCAR test

We test the common MCAR assumption from the observed data in the following experiment. To do so, we implement Little's test of MCAR [61], and include one variable at a time. We interpret as evidence that the MCAR assumption does not hold a statistically significant result ($p < 1e-5$). The test results depicted in figure 4.11 suggests that MCAR does not hold for the RWE Flatiron NSCLC cohort analysed ($p < 1e-5$).

MAR imputation performance

In the next experiment, we assume that MAR hols, as explained in Section 3.2 the MAR assumption is untestable from the data and relies on the assumption of exchangeability of the missingness mechanism given all the observed variables.

Evaluating the imputation performance for MAR in RWE datasets is difficult since ground truth parameters are often unknown. We, therefore, cannot use MRSE to evaluate the performance of the imputation algorithms on the RWE dataset. In our real-world data experiment, we instead focus on comparing bias and coverage of parameter estimates and the impact of missingness for each imputation algorithm, see section 4.2.5. We also analyse the interval width. For methods that rely on the Gibb sampler, such as MITABNET, the convergence of the methods was visually checked and evaluated, see figure 4.12. In this experiment, we first evaluate the percentage bias of using each imputation algorithm to impute 200 Flatiron NSCLC datasets. We first

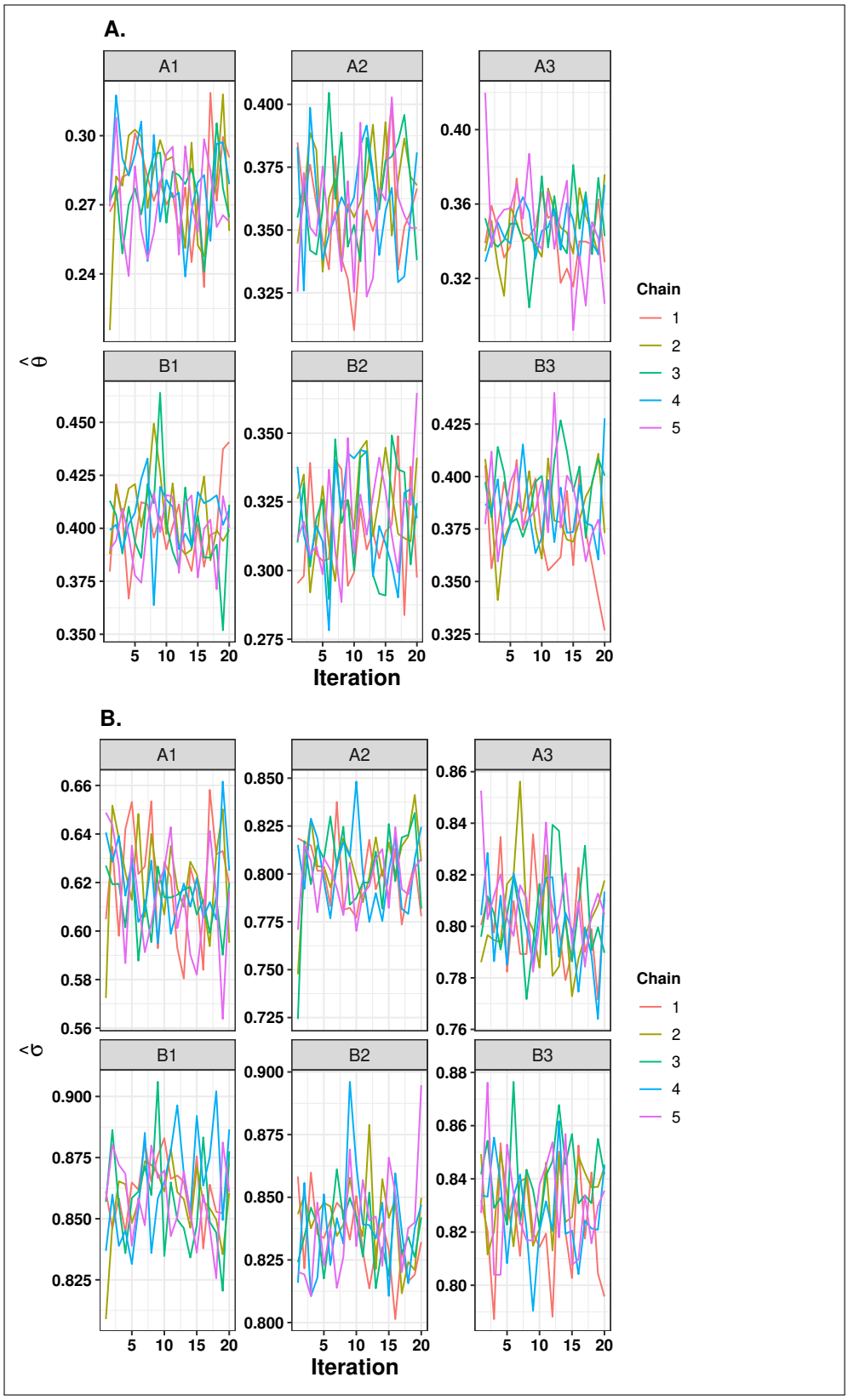


Figure 4.10: Convergence of MITABNET in synthetic dataset. Up: Trace plot of average parameter estimate $\bar{\theta}$. Down: Trace plot of variance in parameter estimate σ.

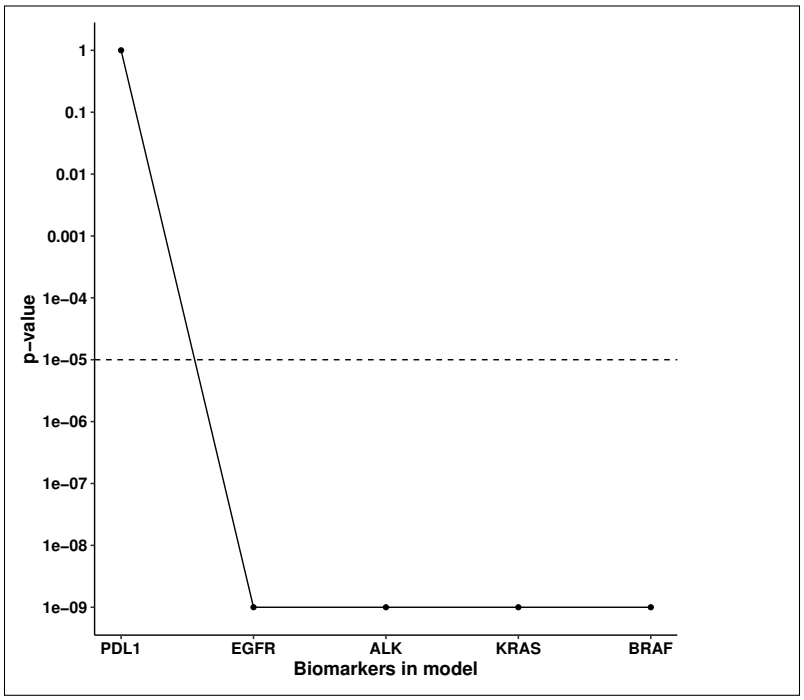


Figure 4.11: The figure depicts Little's test of MCAR results on Flatiron NSCLC RWD. Values less than 1e-5 are statistically significant; the smaller value represented is 1e-9. The test is significant with $p < 1e-5$. Therefore, we can reject that MCAR holds.

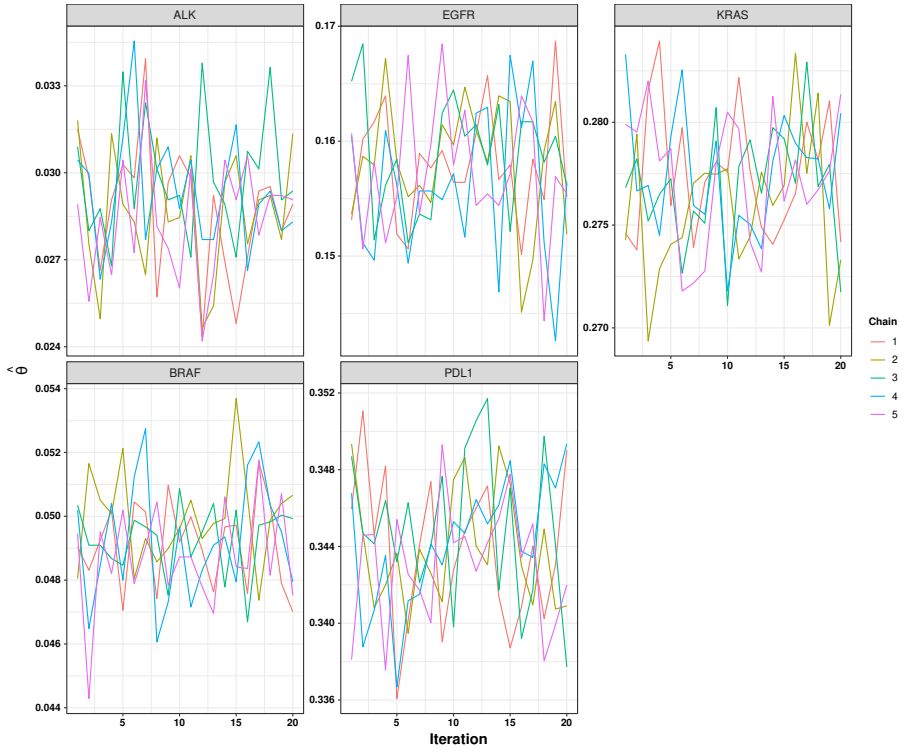


Figure 4.12: Convergence of MITABNET in RWE Flatiron dataset. Trace plot of average parameter estimate $\bar{\theta}$ for each of the biomarker status.

Table 4.6: Percentage bias for each imputation algorithm in the Flatiron NSCLC dataset.

Home	EM	GAIN	MIPCA	MIRF	MITABNET	PMM
EM	33.9 ± 5.8	121.5 ± 111	**13.5 ± 3.6**	28.6 ± 13.6	47.6 ± 4	16.7 ± 3.4
GAIN	**53.1 ± 4.1**	81.9 ± 36	69.3 ± 2.4	66.6 ± 3.4	95.8 ± 90.6	66 ± 5
MIPCA	62.8 ± 7.2	123.3 ± 116.6	26.8 ± 8.1	**8.4 ± 7.8**	71.1 ± 16.2	9.2 ± 4.6
MIRF	106.8 ± 7.1	100.8 ± 87.8	55.2 ± 5.5	**9.1 ± 3.4**	130 ± 15.5	25.8 ± 5.6
MITABNET	**3.8 ± 3.3**	80 ± 74.1	4.3 ± 3.7	57 ± 3.9	9.8 ± 5.7	27.5 ± 4.9
PMM	85.9 ± 8.2	52.5 ± 43.5	47.1 ± 11.7	11 ± 9.6	102.3 ± 19.5	**10.8 ± 7**

perform multiple imputations from the original Flatiron NSCLC dataset for each method, obtaining 5 datasets, generate 200 (40×5) amputated datasets using algorithm 3, in section 4.2.5.

Table 4.6 shows the percentage bias as given by equation 4.17. Each method outperforms row-wise if it has the lowest percentage bias in each imputed dataset. Theoretically, the lowest value is zero, which indicates a perfect model. As explained in section 4.2.5 the imputation performance and the percentage bias depend on the missingness and the concept drift. For instance, the on-diagonal elements of table 4.6 indicate the impact of missingness for each imputation algorithm because it is estimating the values imputed with the same algorithm. The off-diagonal elements indicate the algorithm's additional difficulty in learning the concept drift generated by evaluating the values imputed with a different algorithm. The model is re-imputing a dataset that was imputed originally with another algorithm. The percentage bias does not indicate a superior method, with MIRF (9.1 ± 3.4) and PMM (10.8 ± 7) performing best in on-diagonal elements. MITABNET has a low percentage bias (9.8 ± 5.7), but the best row-wise is EM (3.8 ± 3.3).

Table 4.7 shows the coverage of the 95% confidence interval constructed using Rubin's rules, as given by equation 4.18. The extreme coverage values are 0 and 1, 0 indicating no coverage of the actual parameter and 1 indicating a perfect coverage of 100%. Similarly to the percentage bias experiment, the on-diagonal elements of table 4.7 indicate the impact of missingness on coverage and the off-diagonal the impact of concept drift. The best methods considering the on-diagonal elements of the coverage table 4.7 are PMM (0.9 ± 0.02), MITABNET (0.88 ± 0.02) and MIRF (0.68 ± 0.03).

As we can see in table 4.8, the interval width for each imputation algorithm depends more on the concept drift than on the missingness. In general, an algorithm is best if it has a smaller confidence interval width with higher coverage. Therefore, table 4.8 interpretations need to consider together table 4.7, which showed coverage results. Finally, figure 4.13 depicts the ground truth parameter estimates beta, regarding the different concept drifts and the parameter

Table 4.7: Coverage for each imputation algorithm in the Flatiron NSCLC dataset.

home/visit	EM	GAIN	MIPCA	MIRF	MITABNET	PMM
EM	0.5 ± 0.04	0.38 ± 0.03	**0.72 ± 0.03**	0.56 ± 0.04	0.52 ± 0.04	0.68 ± 0.03
GAIN	0.18 ± 0.03	0.22 ± 0.03	0.32 ± 0.03	0.34 ± 0.03	0.14 ± 0.02	**0.5 ± 0.04**
MIPCA	0.36 ± 0.03	0.4 ± 0.03	0.5 ± 0.04	0.9 ± 0.02	0.4 ± 0.03	**0.94 ± 0.02**
MIRF	0.32 ± 0.03	0.2 ± 0.03	0.4 ± 0.03	**0.68 ± 0.03**	0.22 ± 0.03	0.62 ± 0.03
MITABNET	0.68 ± 0.03	0.12 ± 0.02	0.62 ± 0.03	0.32 ± 0.03	**0.88 ± 0.02**	0.7 ± 0.03
PMM	0.22 ± 0.03	0.22 ± 0.03	0.1 ± 0.02	0.32 ± 0.03	0.24 ± 0.03	**0.9 ± 0.02**

Table 4.8: Interval width for each imputation algorithm in the Flatiron NSCLC dataset.

Home	EM	GAIN	MIPCA	MIRF	MITABNET	PMM
EM	0.09	0.11	0.11	0.13	0.10	0.19
GAIN	0.11	0.11	0.12	0.12	0.13	0.16
MIPCA	0.10	0.12	0.12	0.14	0.11	0.30
MIRF	0.12	0.12	0.14	0.15	0.12	0.20
MITABNET	0.08	0.08	0.09	0.09	0.10	0.16
PMM	0.11	0.12	0.12	0.15	0.12	0.28

estimates considering each algorithm.

4.4 Discussion

Several methods have recently been proposed to perform multiple imputations with missing data for RWE observational datasets [54, 138]. To our knowledge, few studies have systematically compared the statistical properties of the various methods, considering the impact of missing data and the concept drift. To help potential research on RWE datasets choose an imputation method, we have studied six methods that perform multiple imputations.

All methods draw multiple imputations using different but comparable methods. PMM and MITABNET use a Gibbs sampler approach; MIRF uses different random seeds to initialise a random forest; EM and MIPCA use bootstrap-based approach; GAIN uses generative adversarial networks. A DNN powers both GAIN and MITABNET. Hence, multiple imputations may be drawn by applying dropout layers [156] at training and predicting imputations time. To our knowledge, we are the first to investigate the usefulness of TABNET as an algorithm for multiple imputations (MITABNET) and systematically compared it with state-of-the-art methods. MITABNET can become part of the pre-processing step of covariates for RWE dataset analysis, combining the interpretation of multiply imputed datasets for more robust inference. Methods using MITABNET are promising for complex datasets with indirect associations among variables as depicted in Figure 4.4.

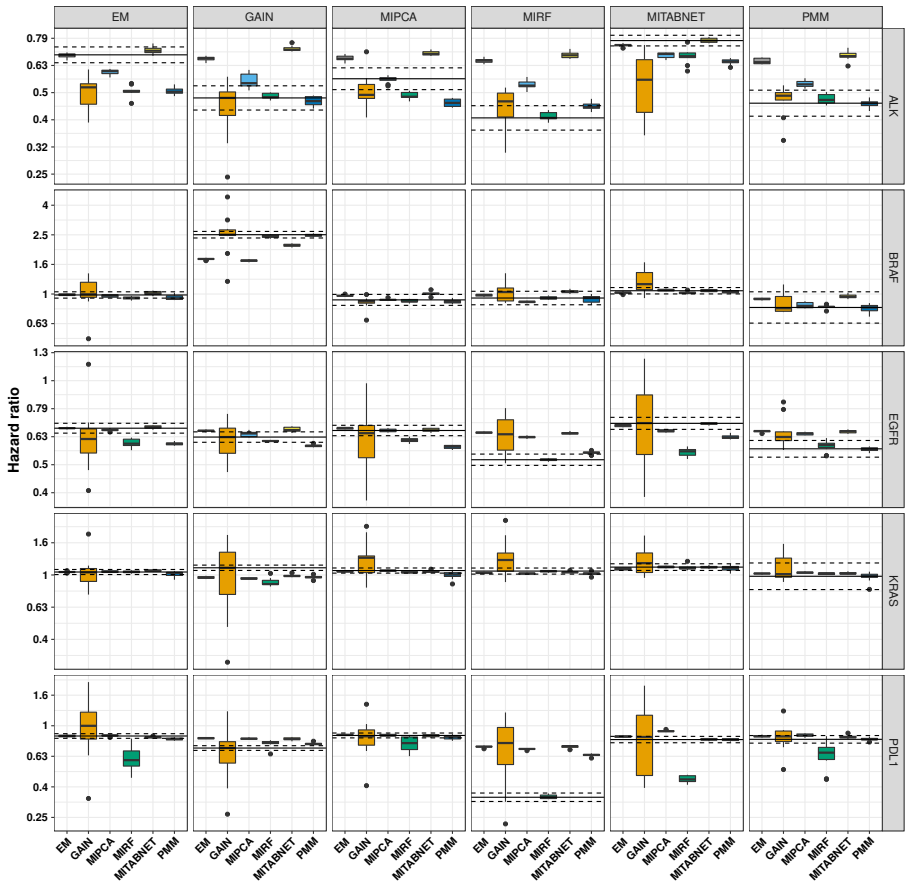


Figure 4.13: Estimates of hazard ratios for ALK, BRAF, EGFR, KRAS, and PD-L1. The solid line indicated the hazard ratio estimated for the imputation algorithm, dashed line 95% confidence interval. Boxplot shows the point estimates of over 100 amputated and re-imputed samples for each imputation algorithm.

In this chapter, we have focused on finding the best imputation method for realistically complex analysis. Our synthetic data experiment used a structural causal model to sample multivariate datasets with different levels of correlation among the observed features. As seen in Figure 4.9 and Table 4.5, all methods perform best when the correlation among variables is high. While MITABNET performed better than other methods for synthetic datasets, the evaluation on RWD Table 4.6 and Table 4.7 did not indicate a superior method by the criteria defined in Section 4.2.5. We suspect that the discrepancy is because of the low correlation seen in the RWD analysed, as depicted in Figure 4.7. Further work may explore how the imputation algorithms perform when adding more correlated features to the imputation model, including demographic variables that might share information with biomarkers such as smoking history, history of malignancies or histology.

Our results agree with previous research showing that the best-case setting for applying off-the-shelve imputation algorithms is the MAR mechanism with a high correlation between variables. The synthetic data experiment found that MITABNET and GAIN outperformed every other algorithm in high and low correlation settings, using accuracy and percentage bias. However, analysing the RWE NSCLC Flatiron dataset did not find conclusive results of the best method considering missingness impact and concept drift. Only three methods, MIRF, MITABNET and PMM, achieved low percentage bias for the scenario where the concept drift was in favour of them, also showing low percentage bias for the impact of missing data $< 20\%$. As seen in Table 4.6, PMM achieved consistently acceptable coverage $> 50\%$, only outperformed under the concept drift of MIRF and MITABNET. On the other hand, PMM also had the most extensive confidence intervals across all imputation algorithms.

We analysed the bias and coverage of parameter estimates after imputing with several imputation algorithms, extending the approach for a standardised evaluation of imputation algorithms from [54, 138], which concluded that MIRF or GAIN result in more accurate imputation and sharper inference than other imputation algorithms. Our synthetic data results indicate that MITABNET may outperform randomised decision trees and GAIN for low and high correlated datasets with the structural causal model used previously by [147], being less biased and hence, prefered for sharper inference.

Limitations

Although we performed the present comparative study with realistically complex analyses and real-world data, it has limitations. The most critical limitation is that our results are dataset-dependent. A limitation of our imputation algorithm is that to avoid an excessive computational burden, we only performed

five multiple imputations for each method in each bootstrap sample, leading to potentially noisy between-imputation variability. For realistic analysis, [153] recommended estimating the number of imputations necessary to produce efficient estimates by conducting a relative efficient analysis of the fraction of missing information [52]. Nevertheless, the default choice for the most popular multiple imputations packages is five [57], and although we evaluated the convergence of the algorithms, it is possible that analysing RWE datasets need more imputations to produce efficient estimates.

Finally, our study focused on MAR missingness patterns. However, an MNAR missing data pattern may be unknown in practice, and results should be generalised with caution. Alternatives to pre-canned algorithms, such as full information maximum likelihood [157], and full Bayesian imputation [40], where the missing values' model assumptions are explicit in the model formulation, may be more appropriate for MNAR settings. However, full information maximum likelihood and fully Bayesian approaches require extra engineering steps to include the missing variables in the model and are out of the scope of this analysis. Algorithms for multiple imputations such as MITABNET work well for MAR and remain the standard approach for handling missing data with imputation algorithms [57, 158].

4.5 Conclusions

The multiple imputations approach is the standard approach for handling missing data in RWE datasets, and we have shown a new method to compare algorithms that perform multiple imputations. MITABNET is a promising algorithm to draw multiple imputations for complex datasets. Conventional methods such as multiple imputations with PMM still perform well for RWE datasets examples, such as the Flatiron NSCLC dataset.

Chapter 5

Bayesian Survival Analysis - a real-world case study

This chapter introduces the field of causal inference for analysing RWE datasets with a time-to-event outcome. We aim to develop a causal design that allows for statistical procedures to elucidate causal questions of interest in RWE datasets, such as understanding how biomarkers modify a treatment effect and the impact of unobserved confounders. To do so, we implement Bayesian computational methods to conduct parameter estimation with finite samples in right-censored survival data. We show improvement over state-of-the-art methods for heterogeneous treatment effects modelling and non-proportional hazards using synthetic data and real-world examples. We apply our new approach to an RWE dataset cohort of advanced advNSCLC patients treated with double-platinum chemotherapy or immunotherapy.

5.1 Introduction

The chapter starts by reviewing the role of immuno-oncology and the RWE dataset analysed. Then, we develop the concepts of average treatment effects and heterogeneous treatment effects using Bayesian survival analysis. Finally, we develop a technique for head-to-head comparison with a utility function based on personalised treatment effects and analyse the sensitivity of our results to unobserved confounding, i.e. violations of exchangeability, defining the OTS bounds for survival outcomes. The following section is a brief clinically-oriented introduction to the field of immuno-oncology (IO).

5.1.1 Clinical background in immuno-oncology

IO focuses on developing therapies that promote an immune response against cancer by counteracting the tumour's mechanism to evade the immune system. A wide variety of mechanisms of action have been investigated [159] with varying

degrees of success. A very intuitive model for understanding IO is the cancer immunity cycle initially described by [160], see figure 5.1. In a very simplified manner, T-cells are the main components of the adaptive immune system that can recognise cancer cell formation and unleash a cytotoxic response that can lead to cancer cell death. Cancer cells undergo spontaneous cell death in the tumour microenvironment, releasing tumour antigens trafficked by dendritic cells, antigen presentation cells (APC) to the lymph nodes. In the lymph node, dendritic cells present the antigen to T cells leading to priming and activation. The T cells travel back to the tumour microenvironment via blood vessels, and in the tumour microenvironment, the activated T cells infiltrate into tumours, recognise cancer cells and unleash a cytotoxic response killing cancer cells. T cells' recognition of cancer cells needs two activation signals, a T Cell receptor and a co-stimulation signal. In the early 1990s [161], researchers recognised that cancer cells are not soliciting immune response because they could not provide co-stimulation. Cancer cells can evade recognition by the immune system, which allows them to survive by expressing PD-L1, which binds to programmed death co-receptor 1 (PD-1) and halts anti-tumour responses. Certain crucial classes within the IO space have resulted in approved medicines, including immune checkpoint inhibitors (ICI) [162], such as PD-1 inhibitors (e. g. pembrolizumab, nivolumab) and PD-L1 inhibitors (e. g. durvalumab). There are several agents available for the management of advNSCLC ; the recommendations include monotherapy and combination therapy. The monotherapy may be chemotherapy or an ICI, while the combination may be a doublet of chemotherapy, such as carboplatin and pemetrexed, or a triplet including ICI [163]. Table 5.1 summarises the results of the pivotal clinical trials that demonstrated the efficacy of ICI in advNSCLC patients prolonging overall survival.

PD-L1 is a biomarker of interest for clinical research in oncology because its expression may be associated with different ICI treatment outcomes in advNSCLC. Clinicians may consider factors including staining intensity and per cent staining, defined as the per cent of stained tumour cells, when assessing PD-L1 status. The approaches to testing, reporting and interpreting PD-L1 results have evolved with the science around PD-L1 and everyday clinical practice [164]. Initial recommendations set a $> 50\%$ staining threshold to determine PD-L1 "positivity". However, data is emerging suggesting that advNSCLC patients with lower per cent staining may respond to ICI [165].

5.1.2 RWE study design

The data source used was the Flatiron database [143], which is a dataset of de-identified patient-level electronic medical records in the United States spanning

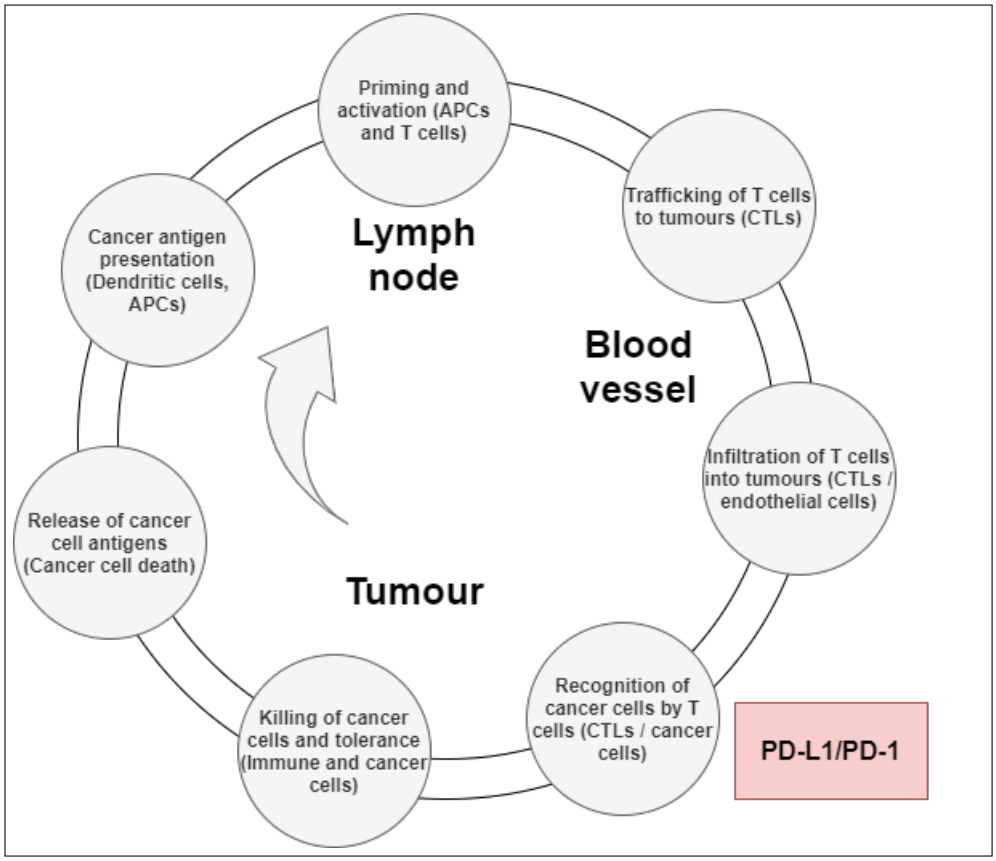


Figure 5.1: The cancer immunity cycle: T-cells are the main components of the adaptive immune system that can recognise cancer cell formation and unleash a cytotoxic response that can lead to cancer cell death.

Table 5.1: Summary of pivotal clinical trials in immune checkpoint inhibitors (ICI) and chemotherapeutics in advanced NSCLC patients on progression-free survival (PFS) and overall survival (OS).

Study	Agent	nivolumab (N)	PFS (months)	OS (months)	Reference
KEYNOTE-24	Pembrolizumab	154	10. 3	21	[166]
KEYNOTE-24	Chemotherapy	151	6	12	[166]
KEYNOTE-189	Pembrolizumab	637	7. 1	20	[167]
KEYNOTE-42	Chemotherapy	637	6. 4	12. 2	[167]
CheckMate-227	Nivolumab	396	5. 6	15. 7	[168]
CheckMate-227	Chemotherapy	396	4. 2	14. 9	[168]
PACIFIC	Durvalumab	476	17. 2	> 24	[169]

280 community practices and seven sizeable academic research institutions.

The inclusion criteria are patients aged ≥ 18, pathological confirmation of NSCLC obtained from tumour cytology or biopsy, documented diagnosis of unresectable stage III-IV advNSCLC. The exclusion criteria are patients participating in a clinical trial for stage III-IV advNSCLC and patients that did not receive any treatment for stage III-IV unresectable advNSCLC.

The primary research questions are the real-world overall survival (rwOS) and the real-world time to treatment discontinuation (rwTTD). The relevant parameters for the time-to-event analysis are the index date and the end date. We define the index date as the start date of treatment anchoring the survival analysis. We define the end date as the death date for rwOS, treatment discontinuation date for rwTTD for patients for whom this is known or last confirmed activity for patients for whom it is unknown. The term *line of therapy* refers to the first eligible drug administration plus other administered drugs within a defined time frame. Following Flatiron Health rules for line of therapy definition [170], we set the time frame to be the first 28 days of starting a line. Figure 5.2 depicts the definition of line of treatment, rwTTD and rwOS. Overall survival that relates to rwOS has been described as a gold

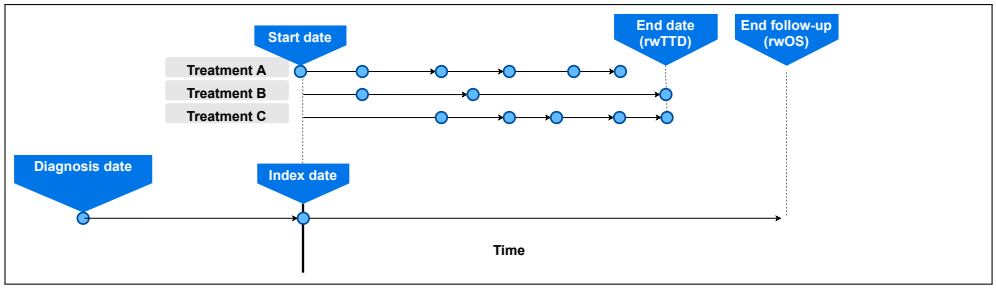


Figure 5.2: Illustration of line of therapy, real-world overall survival analysis (rwOS) and the real-world time-to-treatment discontinuation analysis (rwTTD).

standard primary endpoint to evaluate the efficacy of most drugs, biologics, interventions, or procedures in oncology clinical trials [62]. As an endpoint, it is clearly defined, recorded based on objective assessment, and is a clinically meaningful measure that provides confirmatory evidence that a given treatment extends the life of a patient [62]. However, some studies may have insufficient data to generate sufficiently robust estimates of rwOS (e. g. , when rwOS time is long and thus the RWE studies require extended follow-up). In these cases, rwTTD has proven to be a practical surrogate endpoint for regulatory approval and provides direct clinical benefit evidence. As such, rwTTD is a common endpoint for utilising RWD.

5.1.3 Average treatment effects in survival analysis

As explained above in section 2.3, the survival analysis objective is to analyse treatment's impact on a possibly censored event time of interest, such as rwOS and rwTTD. Its analytical goal is to contrast the distribution of the survival time between treatment groups, i.e. analysing whether the survival times distributions are stochastically longer or shorter between treatment groups.

A traditional approach to survival analysis is modelling the survival function $S(t)$ using the hazard function $h(t)$, which is the limit of the conditional probability that the event of interest will occur and is given by:

$$h(t) = \lim_{dt \to 0} \frac{P\left(t \leq T < t + dt\right)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} \tag{5.1}$$

where $f(t)$ is the event time density function, assuming a continuous density function. Note that the hazard function is, therefore, mathematically an infinite-dimensional parameter space. However, when comparing the survival distribution between treatments, or hazard functions, ideally, we summarise the treatment effect with a single or a limited number of parameters.

The proportional hazard approach [66] summarises the treatment effects using the hazard ratio. The proportional hazard model is given by:

$$h_1(t) = h_0(t) \exp(\beta) \tag{5.2}$$

where $\exp(\beta)$ is the hazard ratio. A hazard ratio < 1 suggests that the treatment stochastically increases survival probability by decreasing the hazard; hence, it is beneficial in preventing the event of interest. Conversely, a hazard ratio ≈ 1 suggests that both treatments impact the hazard similarly, i.e. treatment effects are equivalent. In his seminal paper, Cox [66] noted that the proportional hazard model is an unbiased estimator of hazard ratios under the assumption of proportional hazards. However, non-proportional hazards are common in RWD analysis. A typical violation of the proportional hazards assumptions occurs when two survival curves cross, implying that the corresponding survival functions cross. If the proportional hazard assumption holds, the survival curves proportionally diverge from each other over time, and the average hazard ratio remains constant. However, if the survival curves cross or the separation of survival curves is not constant, the hazard ratio estimate is not constant across time. Figure 5.3 depicts the contraposition of a proportional hazard (PH) and non proportional hazard (NPH) scenario. It is apparent that when the hazard ratio is not constant, the survival curves cross, suggesting NPH.

Additionally, the use of hazard ratios alone may have other disadvantages. The lack of a reference hazard function limits the practical interpretation of an estimated hazard ratio. A hazard ratio may not align well with the visual

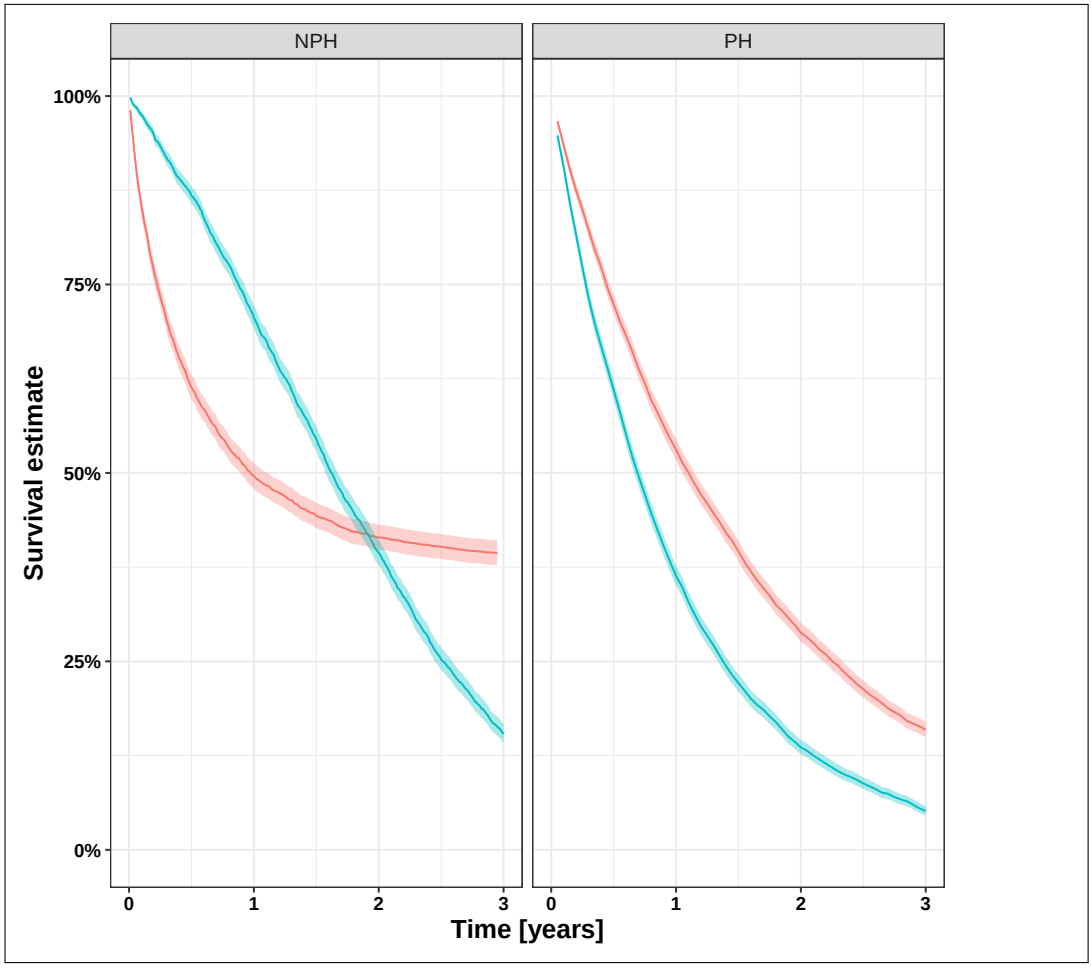


Figure 5.3: Illustration of proportional hazard (PH), where the survival curves separation is constant, and non-proportional hazard (NPH), where the survival curves separation is not constant, and cross.

assessment of survival curves. Moreover, proportional hazard assumption violations make the hazard ratios dependent on the censoring distribution [75], making them more challenging to interpret. Alternatives to the hazard ratio

Table 5.2: Difference between hazard ratios (HR) and restricted mean survival time (RMST(τ)).

	HR	RMST(τ)
Without reference level	Uninterpretable	Interpretable
Censoring distribution	Dependent	Independent
Actual survival time	Independent	Dependent
Study duration	Dependent (implicit)	Dependent (explicit)

approach in an uncensored time-to-event analysis include the mean and median survival time. However, in the presence of censoring, mean and median survival times are not immediately accessible. Instead, following [171] we may define the restricted mean survival time (RMST) as the expectation of the truncated survival time by a timepoint τ, which is given by:

$$\text{RMST}(\tau) = \mathbb{E}\left[\min\left(T, \tau\right)\right] \tag{5.3}$$

Given time-to-event data, one can obtain an estimate $\widehat{\text{RMST}}(\tau)$, which is given by:

$$\widehat{\text{RMST}}(\tau) = \int_0^\tau \hat{S}(t)dt \tag{5.4}$$

The interpretation of the RMST is more straightforward than that of hazard ratios because mathematically it can be shown that it is the average survival time during the interval of time $[0, \tau]$. Besides, a visual assessment of the $\widehat{\text{RMST}}(\tau)$ is also available via the area under the survival curve, see figure 5.4. Moreover, we argue that there are several additional advantages in using RMST over hazard ratios, which we summarise in Table 5.2. RMST has an interpretable reference level, while hazard ratios have not. Hazard ratios do not depend on the actual survival time but only on the ranking of event times. On the other hand, RMST depends on the actual survival times. Both RMST and hazard ratios depend on the study duration, which is *explicit* for RMST, but it is *often overlooked* in reporting hazard ratios.

Counterfactual outcomes for survival analysis

In Section 3.1 we introduced counterfactual outcomes. The following is a running example that demonstrates a counterfactual outcomes framework to compute ATE using RMST. For convenience, we will use the notation Y instead of RMST, however, note that as explained above, we measure the outcome with RMST. Let us consider two possible treatments ($A = 0, 1$). Given a sample

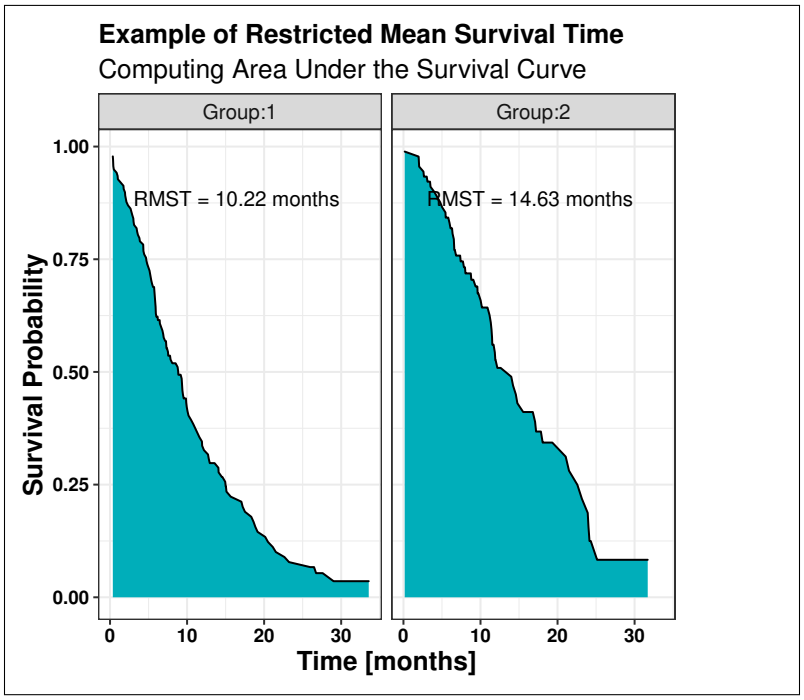


Figure 5.4: Illustration of survival functions for two groups: the area under the survival function is the $\widehat{\text{RMST}}(\tau)$ for each group.

of subjects and a treatment, we have two counterfactual outcomes:

$$Y^0, Y^1 \tag{5.5}$$

representing the RMST outcomes under each treatment. However, in practice taking both measurements is impractical because we only observe one outcome:

$$\mathbb{I}(A = 0)\, Y^0 + \mathbb{I}(A = 1)\, Y^1 \tag{5.6}$$

where $\mathbb{I}$ indicates if the observations belong to the treatment $A = 0$ group or the treatment $A = 1$ group. Therefore, there exists a factual observation and a counterfactual one that we can tackle as a missing value problem. Opportunely, we can adapt the techniques introduced in Chapter 4 for missing value problems. Let us consider a population of four patients and the two treatment options ($A = 0, A = 1$). For clarity, we may disregard sampling variability in this example by considering that each sample comes from a very large hypothetical super-population and represents a large number of identical individuals. Subjects 1 and 2 had treatment 0, and subjects 3 and 4 had treatment 1. The outcome Y is given by:

$$Y := f_Y(W, A, U_Y) \tag{5.7}$$

Table 5.3: Illustration of hypothetical results from a survival study showing counterfactuals.

Subject	Y^1	Y^0	$Y^1 - Y^0$	W
1	2	8∗	-6	1
2	4	10∗	-6	1
3	11∗	7	4	0
4	14∗	5	9	0
Population mean	7. 75	7. 5	1	
Observed mean	3	6		

where U_Y is an i. i. d random variable. We have not measured the counterfactuals, i.e. the denoted $*$. Let us consider that doctors give the best treatment for each individual given unobserved covariates W. In this case, if we only analyse the observations, we would come to the opposite conclusion about the best treatment because of the confounding variable W, i.e. the doctor's action. Therefore, in observational studies, we need counterfactuals to analyse treatment effects. We adopted the perfect doctor example above from [79] to illustrate the need for the counterfactual outcomes framework introduced by Rubin and Neyman [77, 78], which is very helpful for analysing causal questions from RWD. We discuss the problem of confounding by perfect doctors in section 5.2.2 where we compute the OTS bounds defined in section 3.5.1.

To complete our running example for ATE, we need to compute counterfactuals in censoring for survival analysis. To do so, we need a model to compute the expected RMST given the treatment A and the confounders W for each individual. Substituting in table 5.3 the missing counterfactual outcomes, denoted by $*$ with results for the estimate of the $(\widehat{\text{RMST}}(\tau))$ allows us to compute counterfactuals for survival censored outcomes. Interestingly, the raw data do not show in our estimator of treatment effects; such an approach would deliver a biased estimator. Because of censoring, we can not use the observed event times in our dataset. Moreover, to obtain an estimator that is robust to proportional hazard violations, we need to fully model non-proportional hazards, which we define in section 5.2.

5.1.4 From ATE to treatment effects heterogeneity

The term heterogeneity of treatment effects refers to how treatment effects vary with observable characteristics of individuals [172]. Estimating heterogeneous treatment effects implies that there is not a unique ATE but a treatment effect conditional on a set of covariates X, i.e. a CATE. The covariates of interest X might impact the outcome via interactions with treatment. Mathematically, heterogeneous treatment effects are the conditional expectation of the difference

in counterfactual outcomes given that a variable of interest X takes a pre-specified value. Heterogeneous treatment effects are denoted as CATE and are given by:

$$\text{CATE} = \mathbb{E}\left[\text{Y}^0 - \text{Y}^1 | X = x\right] \tag{5.8}$$

Let us consider a hypothetical population stratifying by a variable X into subpopulations with counterfactual outcomes that are different, adapted from [32] and depicted in figure 5.5. The outcome Y is given by: The outcome Y is given by:

$$Y \coloneqq f_Y(W, X, A, U_Y) \tag{5.9}$$

where U_Y is an i. i. d random variable, and the variable X is a variable of interest because it modifies the treatment effect, for example, a biomarker for treatment personalisation.

Further, let us consider a population of eight patients and two treatment options ($A = 0, A = 1$). X denotes the treatment modifier measured for each individual. For clarity, we ignore sampling variability in this example and assume that each sample comes from a very large hypothetical super-population and represents many identical individuals. Table 5.4 shows the results of this thought experiment. In particular, the sign of the CATE is potentially opposite to the observed mean between strata of X. This exercise echoes the analysis of ATE conducted above. To complete our running example for CATE, we need to compute conditional counterfactuals in censoring for survival analysis. To do so, we need a model to compute the ($\widehat{\text{RMST}}(\tau)$) given the treatment A, the biomarker X and the confounders W for each individual. Substituting in Table 5.3 the missing conditional counterfactual outcomes, denoted by $*$ with results for the estimate of the ($\widehat{\text{RMST}}(\tau)$) allows us to compute the CATE, given by equation 5.8.

5.1.5 Related work

There is a vast, fast-growing literature on treatment effects heterogeneity. The study of treatment effect heterogeneity covers several topics, including but not limited to biomarker discovery [17], subpopulation stratification [173], multiple hypothesis testing [174], identifying the highest individualised treatment effects [175], estimating OTS [176], and causal model discovery [37]. Each question has its own set of methods. One crucial question is that of stratification. Early literature on stratification focused on non-parametric analysis methods [32]. Rothman et al. [177] reviewed the most common methods for stratification. More recently, Athey et al. [173] explored low-dimensional parameter estim-

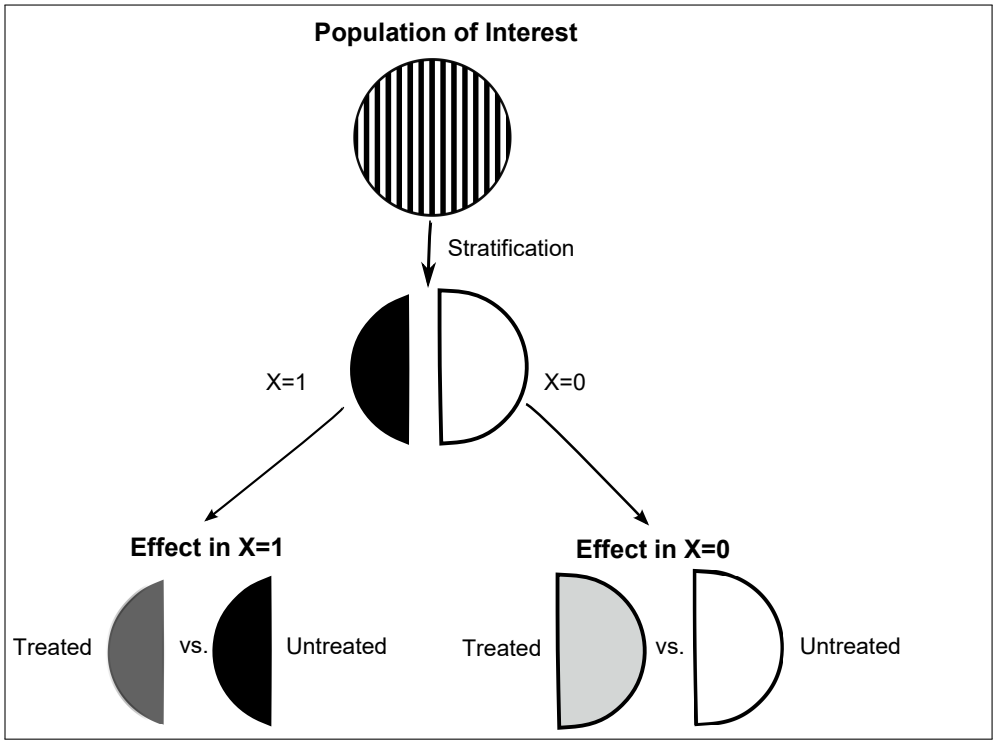


Figure 5.5: Illustration of treatment modifiers by stratifying a hypothetical population into subgroups based on a covariate of interest X the treatment effect is the difference between the outcomes of treated and untreated for each sub-group.

Table 5.4: Illustration of hypothetical results from an observational study showing counterfactuals with stratification by variable X.

Subject	Y^1	Y^0	$Y^1 - Y^0$	A	X
1	2	8∗	-6	1	1
2	10	10∗	0	1	0
3	6	4∗	2	1	1
4	10	6∗	4	1	0
5	2∗	3	-1	2	1
6	8∗	6	2	2	0
7	4∗	5	-1	2	1
8	2∗	10	-8	2	0
Population mean	5.5	6. 5	-1		
Observed mean	7	6	1		
Population mean $\|X = 1$	3. 5	5	-1. 5		
Population mean $\|X = 0$	7. 5	8	-0. 5		
Observed mean $\|X = 1$	4	4	0		
Observed mean $\|X = 0$	10	8	2		

ates via random forest for heterogeneous treatment effects estimation and stratification.

More recent literature extends these methods to ITE. Kunzel et al. [121] suggested a metalearner architectured dubbed X-learner that combines two models, one to predict the CATE, and another to impute ITE by predicting the counterfactual outcomes from all the individual measured covariates, assuming exchangeability, see section 3.2. [121] suggested using the propensity score to weight the two estimators in the X-Learner. Yoon et al. [178] suggested an GAN architecture simulate counterfactual outcomes. Hill et al. [179, 180] showed applications of GP to estimate time-varying effects and spatial correlated effects. However, there is a lack of literature on applying GP regression for estimation of heterogeneous treatment effects in survival analysis. Our research uses Bayesian survival regression for non-proportional hazards and GP regression for heterogenous treatment effects in survival analysis.

5.1.6 Contributions

This chapter makes several contributions that we summarise as follows:

1. Introduces a counterfactual approach to estimate treatment effects in the presence of non-proportional hazards under the Bayesian survival outcome modelling framework.
2. Develops an OTS Bound for non-proportional hazards using RMST.
3. Conducts simulation studies to evaluate the Bayesian survival outcome modelling approach in the non-proportional hazard setting.
4. Applies the Bayesian survival outcome modelling approach and the OTS bounds in several real-world examples.
5. Introduces Weibull GP hazard regression for heterogeneous treatment effects analysis advancing the concept of treatment personalisation.
6. Discusses methods for comparing causal models related to personalisation and treatment effects heterogeneity.
7. Lastly, we provide Stan implementations of our Weibull GP hazard regression model.

5.2 Methods

In analysing RWE datasets, we need to distinguish between the data we observe and the data we would like to have i.e. the factual and counterfactual outcomes.

The following section introduces the dataset analysed and the research questions for our RWE study. Next, we explain the modelling techniques and the sensitivity analysis method.

5.2.1 Dataset Analysed

Our RWE study is a retrospective cohort study assessing the rwOS during or immediately after starting chemotherapy or immune therapy. A secondary objective was to assess the rwTTD as the start of a subsequent line of therapy or death and the correspondence between rwTTD and rwOS. The data source used was the Flatiron database [143], which is a dataset of de-identified patient-level electronic medical records in the United States spanning 280 community practices and seven sizeable academic research institutions.

In addition to the inclusion-exclusion criteria described in Section 5.1.2, the study excludes therapy lines where maintenance therapy is clinically relevant, e. g. given for an indication approved as maintenance therapy. The analysis includes patients that are in the first line of therapy for the following immunotherapy, chemotherapy and combination treatments: carboplatin, pembrolizumab, pemetrexed (CPP); carboplatin, pemetrexed (CP); durvalumab (D); N; and pembrolizumab (P). The follow-up for the advNSCLC cohort is four years.

Table 5.5 summarise the time to overall survival, treatment discontinuation and PD-L1 per cent staining for the first-line. Note the unbalanced treatment groups in the first line cohort: CPP (N = 3,198), CP (N = 857), D (N = 108), N (458), and P (N = 2,945). Class imbalance is typical for RWE datasets obtained from daily clinical practice and hints using the techniques explained below to estimate treatment effects. The advNSCLC cohort also includes baseline and longitudinal individually measured covariates, see table 5.6. We study which covariates may be confounders of treatment effects and those that may be mediators of treatment effects (post-treatment bias). We classify the following covariates as likely confounders and attempt to adjust for them using the below-mentioned modelling techniques. Let us denote the potential confounders by W, which comprise:

1. Histology: squamous cell carcinoma (SCC), non squamous cell carcinoma (NSCC), and not otherwise specified (NOS).

2. Smoking: a binary variable indicating if the individual ever smoked.

3. Gender: male or female.

4. Ethnicity: asian, hispanic or latino, black or african american, white or caucasian, and other race.

5. Body weight: standardised.

6. Biomarker status of ALK, EGFR, KRAS, and BRAF.

7. Age at diagnosis.

8. Patient performance status as measured by the Easter Cooperative Oncology Group (ECOG) performance status.

Numbers at risk

Table 5.7 summarises the individuals at risk by year of follow-up for the rwTTD and rwOS cohorts. We can see that patients drop out earlier in the rwTTD cohort and that the maximum follow-up time is near four years for both cohorts.

Handling of missing data

Back in section 4.2, we comprehensively explained and investigated imputation methods for handling missing data. We discussed the three possible mechanisms of missing data: MCAR, MAR, and MNAR. For our first analysis, we assume that a MAR mechanism is in play, which means that missing PD-L1 status values are not caused by the actual PD-L1 per cent staining values, but other measured variables can explain them. Our results in section 4.3 suggested that PD-L1 status is low correlated ($\rho \leq 0.2$) with other variables in the advNSCLC Flatiron dataset.

All the rest of the measured individual variables also have varying levels of missingness. Hence, to not squander information, we use multiple imputations to handle missing variables in the confounders. To analyse multiple imputations in Bayesian modelling, we obtain Bayesian estimates, i.e. the posterior distribution over parameters, for each imputed dataset (θ_m) and combine the posterior distributions to average over the estimates. Combining results after multiple imputations is helpful because it visualises uncertainty in imputations via credible intervals and the combined posterior predictive distribution.

5.2.2 Bayesian hazard regression modelling

Following [72], we suggest a Bayesian outcome regression model using a parametric model that estimates the hazard, i.e. the immediate risk of death, rather than the survival function directly. The proportional hazard regression model is given by:

$$h_i(t) = h_0(t) \exp[\eta_i] = h_0(t)\lambda_i \tag{5.10}$$

Table 5.5: Patient outcome, PD-L1 per cent staining, EGFR, KRAS, ROS1, BRAF status for the followed-up cohorts: Carboplatin, Pembrolizumab, Pemetrexed (CPP); Carboplatin, Pemetrexed (CP); Durvalumab (D), Nivolumab (N), and Pembrolizumab (P).

AdvNSCLC	CPP, N = 3,198[1]	CP, N = 857[1]	D, N = 108[1]	N, N = 458[1]	P, N = 2,945[1]
OS time	266 (120, 539)	280 (130, 551)	350 (181, 545)	233 (110, 567)	280 (101, 607)
OS status	1,881 (59%)	566 (66%)	49 (45%)	326 (71%)	1,749 (59%)
TTD time	63 (62, 84)	63 (42, 84)	99 (56, 226)	112 (56, 266)	154 (63, 357)
Unknown	288	123	19	66	386
TTD status	2,623 (90%)	694 (95%)	73 (82%)	332 (85%)	1,915 (75%)
Unknown	288	123	19	66	386
PD-L1 staining					
0%	827 (26%)	272 (32%)	23 (21%)	163 (36%)	79 (2. 7%)
≤ 1%	373 (12%)	129 (15%)	9 (8. 3%)	71 (16%)	37 (1. 3%)
1%	208 (6. 5%)	61 (7. 1%)	8 (7. 4%)	28 (6. 1%)	67 (2. 3%)
2%-4%	134 (4. 2%)	46 (5. 4%)	8 (7. 4%)	11 (2. 4%)	39 (1. 3%)
5%-9%	209 (6. 5%)	49 (5. 7%)	9 (8. 3%)	37 (8. 1%)	86 (2. 9%)
10%-19%	244 (7. 6%)	62 (7. 2%)	1 (0. 9%)	35 (7. 6%)	80 (2. 7%)
20%-29%	205 (6. 4%)	36 (4. 2%)	9 (8. 3%)	23 (5. 0%)	66 (2. 2%)
30%-39%	117 (3. 7%)	30 (3. 5%)	2 (1. 9%)	19 (4. 1%)	77 (2. 6%)
40%-49%	87 (2. 7%)	17 (2. 0%)	4 (3. 7%)	7 (1. 5%)	50 (1. 7%)
50%-59%	101 (3. 2%)	26 (3. 0%)	5 (4. 6%)	13 (2. 8%)	341 (12%)
60%-69%	106 (3. 3%)	18 (2. 1%)	4 (3. 7%)	6 (1. 3%)	271 (9. 2%)
70%-79%	102 (3. 2%)	25 (2. 9%)	9 (8. 3%)	13 (2. 8%)	323 (11%)
80%-89%	137 (4. 3%)	16 (1. 9%)	4 (3. 7%)	8 (1. 7%)	364 (12%)
90%-99%	248 (7. 8%)	44 (5. 1%)	5 (4. 6%)	16 (3. 5%)	721 (24%)
100%	100 (3. 1%)	26 (3. 0%)	8 (7. 4%)	8 (1. 7%)	344 (12%)
EGFR					
−	2,732 (95%)	712 (93%)	77 (93%)	330 (96%)	2,323 (98%)
+	135 (4. 7%)	55 (7. 2%)	6 (7. 2%)	12 (3. 5%)	58 (2. 4%)
Unknown	331	90	25	116	564
KRAS					
−	1,277 (61%)	306 (63%)	42 (76%)	164 (74%)	892 (60%)
+	807 (39%)	176 (37%)	13 (24%)	59 (26%)	607 (40%)
Unknown	1,114	375	53	235	1,446
ROS1					
−	2,665 (99%)	669 (100%)	70 (98. 6%)	302 (99. 7%)	2,122 (99. 8%)
+	6 (0. 2%)	2 (0. 3%)	1 (1. 4%)	1 (0. 3%)	4 (0. 2%)
Unknown	527	186	37	155	819
BRAF					
−	2,243 (94. 4%)	505 (94. 6%)	69 (94. 5%)	245 (97. 2%)	1,632 (94. 3%)
+	132 (5. 6%)	29 (5. 4%)	4 (5.5%)	7 (2. 8%)	99 (5. 7%)
Unknown	823	323	35	206	1,214

1 Statistics presented: Median (IQR); n (%)

Table 5.6: Patient histology (SCC: squamous cell carcinoma, NSCC : non-squamous cell carcinoma, NOS: NSCLC not otherwise specified), smoking history, gender, race, body weight and age for the followed-up cohorts: Carboplatin, Pembrolizumab, Pemetrexed (CPP); Carboplatin, Pemetrexed (CP); Durvalumab (D), Nivolumab (N), and Pembrolizumab (P).

AdvNSCLC	CPP, N = 3,198[1]	CP, N = 857[1]	D, N = 108[1]	N, N = 458[1]	P, N = 2,945[1]
Histology					
NSCC	4,023 (95%)	1,321 (96%)	800 (50%)	596 (55%)	3,272 (66%)
NOS	147 (3. 5%)	44 (3. 2%)	67 (4. 2%)	48 (4. 4%)	210 (4. 2%)
SCC	46 (1. 1%)	17 (1. 2%)	728 (46%)	445 (41%)	1,500 (30%)
Smoking history					
Yes	3,742 (89%)	1,226 (89%)	1,513 (95%)	1,009 (93%)	4,617 (93%)
Gender					
Female	1,953 (46%)	682 (49%)	713 (45%)	486 (45%)	2,390 (48%)
Race					
White	2,809 (76%)	927 (75%)	1,118 (78%)	749 (76%)	3,529 (79%)
Asian	63 (1. 7%)	27 (2. 2%)	22 (1. 5%)	8 (0. 8%)	80 (1. 8%)
Black	410 (11%)	120 (9. 7%)	150 (10%)	109 (11%)	400 (9. 0%)
Other	433 (12%)	157 (13%)	148 (10%)	118 (12%)	446 (10%)
Unknown	499	149	156	105	524
Body weight (Kg)	73 (62, 86)	74 (63, 87)	74 (62, 86)	73 (60, 87)	72 (61, 85)
Unknown	1,118	410	355	389	1,329
Age	68 (61, 75)	69 (62, 76)	69 (61, 75)	72 (64, 78)	72 (64, 79)
1 Statistics presented: Median (IQR); n (%)					

Table 5.7: Numbers at risk for the rwTTD and rwOS cohort.

Cohort	0 year	1 year	2 year	3 year	4 year
rwTTD	3331	656	149	25	1
rwOS	3410	1660	595	157	5

where η_i is the linear predictor and λ_i is the link function for individual ith. For the baseline hazard $h_0(t)$ that may vary in time, we evaluate canonical parametric distributions, such as exponential and Weibull, which baseline hazard distribution are given by:

$$\begin{aligned} \text{Exponential} \ : \quad & h_0(t) = 1 \\ \text{Weibull} \ : \quad & h_0(t) = \gamma t^{\gamma-1} \end{aligned} \tag{5.11}$$

where γ denotes the Weibull shape parameter.

Likelihood

Let T_i be a random variable indicating the observed time and d_i be a censoring indicator, denoting observed events by $d_i = 1$ and right-censoring by $d_i = 0$. Allowing for right censoring, the data probability $p(\mathcal{D})$ is the likelihood of this survival model, given by:

$$p(\mathcal{D}_i|\boldsymbol{\theta}) = [h_i(T_i|\boldsymbol{\theta})]^{I(d_i=1)} \times [S_i(T_i|\boldsymbol{\theta})]^{I(d_i \in \{0,1\})} \tag{5.12}$$

where θ are the model parameters. For example, for the Weibull hazard model the likelihood takes the form:

$$p(\mathcal{D}_i|\boldsymbol{\lambda_i}, \boldsymbol{\gamma}) = \gamma T^{\gamma-1} {\lambda_i}^{I(d_i=1)} \times \exp\left(-T_i^{\gamma}\lambda_i\right)^{I(d_i \in \{0,1\})} \tag{5.13}$$

Regression Coefficients Priors

We denote the prognostic index for the effects of the time-constant covariates by:

$$\eta_i(t) = \beta_0 + \sum_{p=1}^{P} \beta_p w_i \tag{5.14}$$

where β_0 is the intercept parameter, and β_p are the regression coefficient. We evaluate the following prior distributions for the regression coefficients: vague prior, flat prior distribution; weakly informative prior distribution (WIP) with Normal and Student's t distributions; specific informative prior distribution

(SIP). Mathematically, the prior distributions for β_0 and β_p are given by:

$$
\begin{aligned}
&\text{Vague}: \\
&\beta_0 \sim \text{Normal}(0, 20);\ \beta_p \sim \text{Normal}(0, 2.5) \\
&\text{Normal WIP}: \\
&\beta_0 \sim \text{Normal}(0, 1);\ \beta_p \sim \text{Normal}(0, 0.5) \\
&\text{Student WIP}: \\
&\beta_0 \sim \text{Normal}(0, 1);\ \beta_p \sim \text{Student's t}(3, 0, 1) \\
&\text{Normal SIP}: \\
&\beta_0 \sim \text{Normal}(0, 1);\ \beta_p \sim \text{Normal}(0.5, 0.2)
\end{aligned}
\tag{5.15}
$$

Weibull shape parameter prior For the Weibull hazard model, one also needs to specify a prior for the shape parameter. The shape parameter allows the Weibull model to change the rate and fit closer to real-world measurements. A value near 1 implies an exponential distribution. Therefore, instead of using a flat prior on the linear scale of shape, which would not be a flat prior on the probability space, we choose a scaled prior strictly positive, such as the half-Normal, the half-student and the exponential distribution, which are given by:

$$
\begin{aligned}
&\text{Half-Normal}: \\
&\gamma \sim \text{Half-Normal}(1, 1) \\
&\text{Half-Student}: \\
&\gamma \sim \text{Half-Student}(3, 1, 1) \\
&\text{Exponential}: \\
&\gamma \sim \text{Exponential}(1)
\end{aligned}
\tag{5.16}
$$

Note that we make the Half-Normal and the Half-student to peak at one by design. Figure 5.6 show that the probability density for these priors.

5.2.3 Gaussian process Weibull hazard regression

PD-L1 per cent staining is a proxy for the PD-L1 expression in the tumour cells that patients with similar tumours have. However, PD-L1 expression is a continuous variable. Note that not any patient has the same PD-L1 expression. Researchers have begun studying the clinical relevance of PD-L1 per cent staining instead of simply using a binary positive or negative interpretation

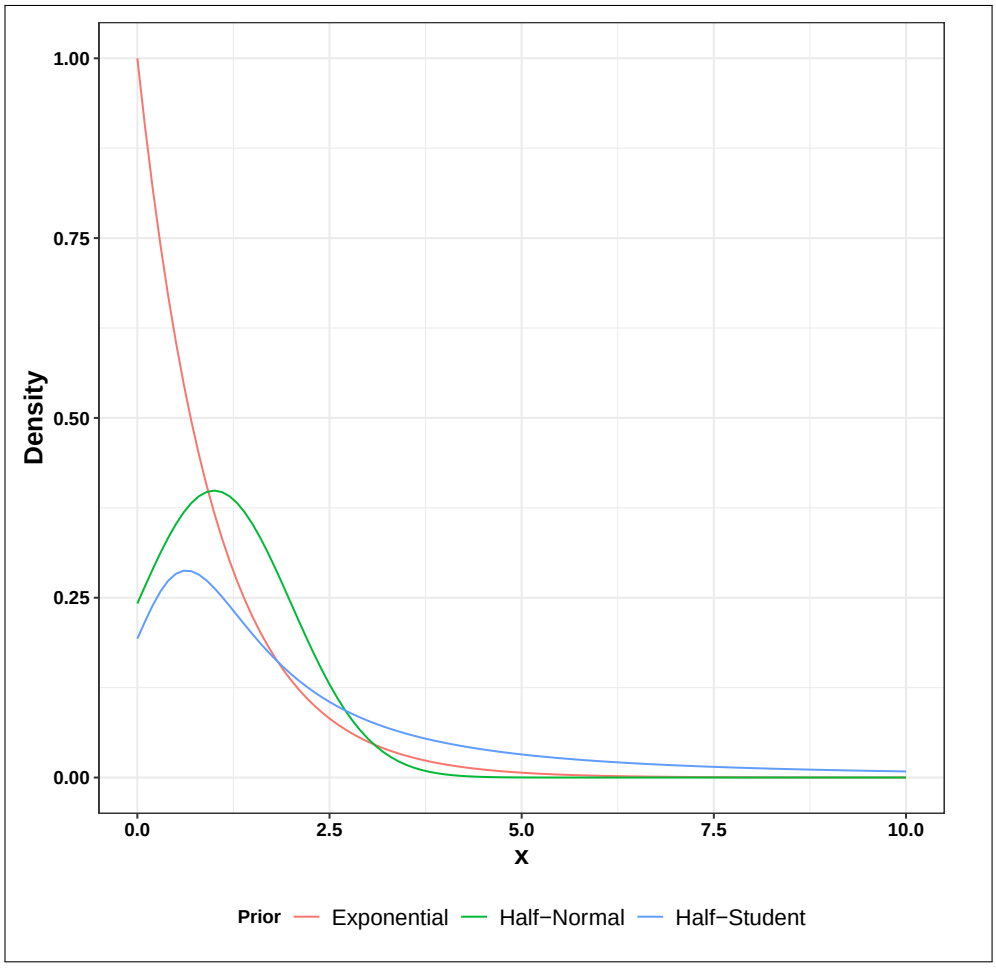


Figure 5.6: Probability distributions for chosen priors for Weibull's γ parameter.

of PD-L1 status [164]. Part of it is related to margin of error and difficulties in defining a threshold for predicting treatment response. Moreover, new treatment guidelines have since shifted the staining threshold values associated with PD-L1 biomarker positive status to a $\geq 1\%$ framework for many use cases [163].

Although there is no obvious cutpoint in continuous variables such as PD-L1 expression, PD-L1 per cent staining close values may potentially share interactions with ICI treatment. For example, before the data arrives, we know that PD-L1 50% and PD-L1 60% are more similar PD-L1 expression levels than PD-L1 1%. We want to exploit pooling between proximate PD-L1 expression levels than between distant ones. It is possible to discretise PD-L1 expression and build interaction with treatment, such as everybody with a PD-L1 50% staining and treatment CP has the same intercept. However, nothing in the interaction model informs that PD-L1 50% is more similar to PD-L1 60% than PD-L1 1%, because all the discretised variables in varying effects models are unordered. Mathematically, we can define the expression levels to be continuous categories, and a standard approach for continuous categories is GP regression [40].

In regression, we are interested in estimating the association between variables. Equation 5.10 shows that one can use a linear function to fit the data to a hazard model using a log transformation. However, with higher-order polynomials, one can obtain a closer fit to the data points. As the data gets complex, one may need a higher-order polynomial to obtain a satisfactory fit. Although higher-order polynomials can fit the data, they do not generalise new covariate measurements, a classical problem in machine learning known as overfitting the original data, see section 3.3. The difficulty is in choosing the functional form to use for regression.

In practice, an infinite number of functions can provide a good fit for a given set of data. GP assign each of the different functions a probability, and the mean over the probability distribution provides the most credible fit to the data. Hence, a GP is a probabilistic method that tackles the uncertainty for the predicted prognostic index, given by a prior over function $P(f)$ used for Bayesian regression. Similar to a multivariate Gaussian distribution, parameterised by a mean vector μ and covariance matrix Σ, a GP is parameterised by:

$$\mathrm{P}(f) = \mathrm{GP}\left(\mu(x), K(x|\theta)\right) \tag{5.17}$$

where x is the input, in our use case, the measured PD-L1 per cent staining. μ is the mean function, i.e. the contribution of the GP to the prognostic index. Sampled functions from the GP will recover μ, μ is defined for $\in \mathbb{R}$. K is the covariance function or kernel, and θ are the parameters specific to the kernel.

K is a covariance kernel applied to all pairwise datapoints, which determines the variation in the functions of the GP. K must produce a positive definite matrix for the input x.

The time-to-event observations may be modelled with a suitable survival model. We suggest a Weibull likelihood, see Equation 5.13, since the Weibull distribution assumptions are uncomplicated and have proven to do well in practice [67]. The observations of a GP Weibull hazard model are given by:

$$y \sim \text{Weibull}\left(\text{P}(f), \gamma\right) \tag{5.18}$$

Since we are not using a Gaussian likelihood, we use MCMC to fit the model. We use the advanced No-U-Turn-Sampler for HMC as implemented in Stan Math [181], see Appendix A.5.

Gaussian process prior

The GP prior is a multivariate Gaussian prior. The Gaussian distribution is appealing because one can model the covariance between observations instead of the mean and have exact predictions [40]. We set μ to zero, implying that $\text{P}(f)$ are offsets on the hazard model. We expect that individuals with similar PD-L1 expression will have similar outcome. We do not have a priori any reason to believe that the impact of PD-L1 expression on the hazard will be more complex, such as cyclical. Therefore to model K, we can use the conventional exponentiated quadratic kernel, also known as radial basis function (RBF) kernel [11]. The resulting covariance matrix is given by:

$$K\left(x|\alpha,\rho\right)_{i,j} = \alpha^2 \exp\left(-\frac{1}{2\rho^2}\sum_{d=1}^{D}\left(x_{i,d} - x_{j,d}\right)^2\right) \tag{5.19}$$

where $x_{i,d} - x_{j,d}$ is the difference in PD-L1 per cent staining between individuals i and j with treatment d. For clarity, we denote the difference in PD-L1 per cent staining by Δx_{PD-L1}. α^2 is the maximum covariance, or amplitude, for two observations with the same PD-L1 expression. ρ is the length scale, or volatility, and denotes the rate of decline in correlation with Δx_{PD-L1}, such as for small ρ the covariation between individuals with different PD-L1 expression may be higher. Since decay is non-linear, the squared distance Δx^2_{PD-L1} encapsulates the prior information that for individuals with similar PD-L1 expression, the correlation is higher and declines rapidly with Δx_{PD-L1}, see Figure 5.7. Figure 5.8 depicts the analytical steps that we perform for the survival analyses with GP Weibull hazard models. We model the heterogeneous treatment effect with a GP prior from a survival dataset stratified by PD-L1 and treatment. We adjust the outcome model for individual characteristics such as additional

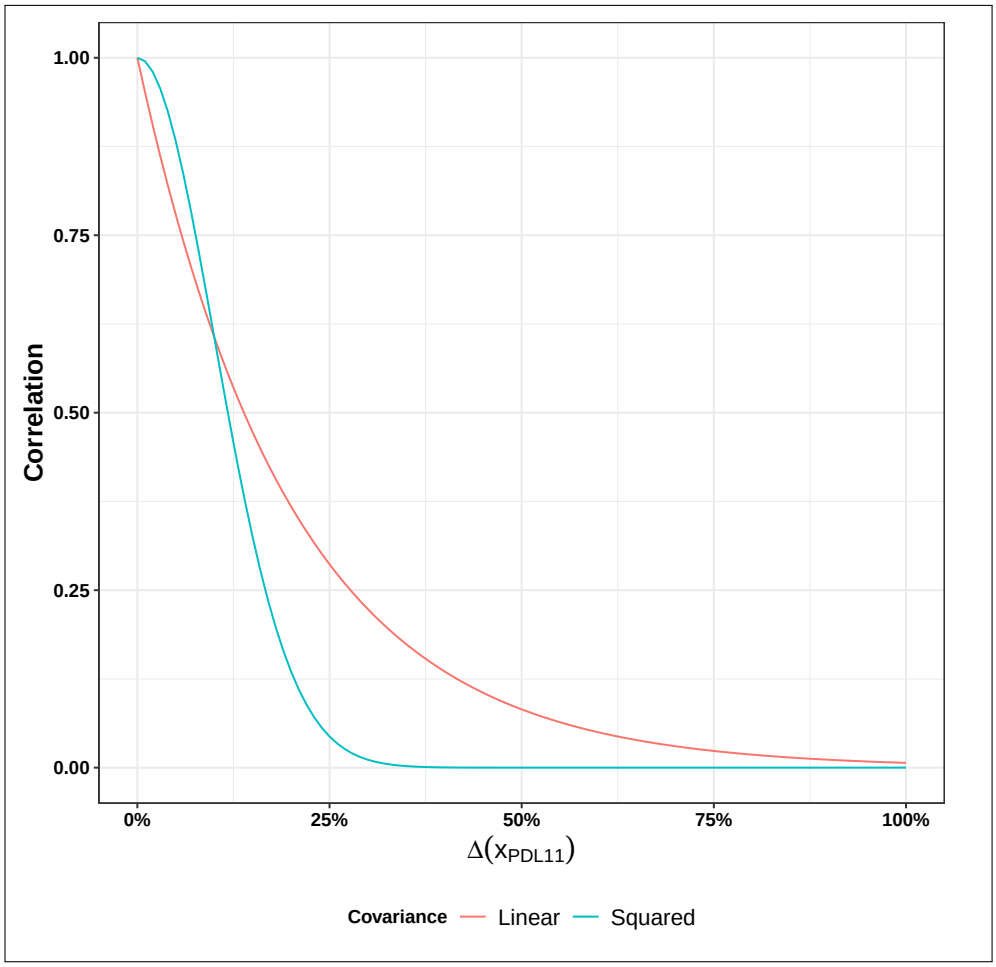


Figure 5.7: Correlation between individuals with different PD-L1 expression using linear exponentiated and squared exponentiated kernels.

biomarkers, age and race, see table 5.6. We update GP prior distribution using the observed data to obtain a posterior prediction of the rwTTD and rwOS that we summarise with RMST.

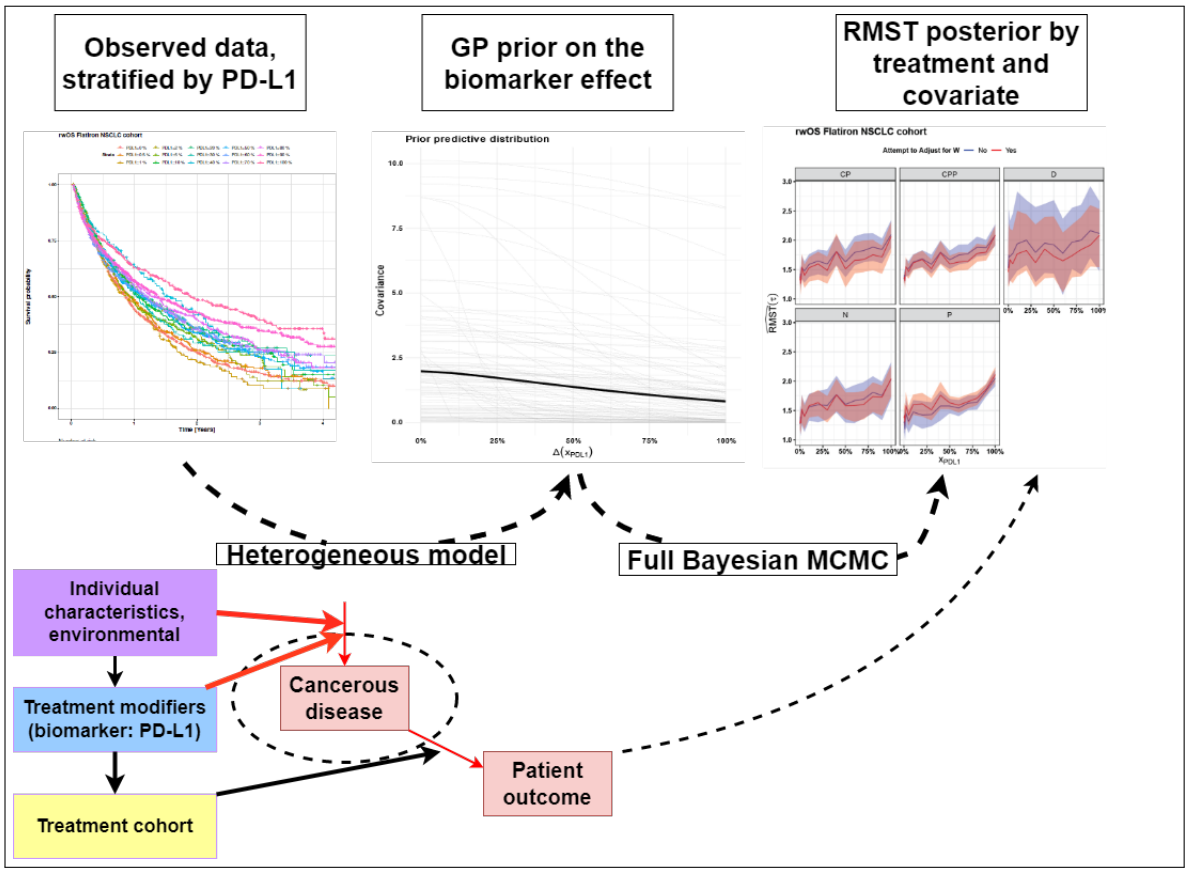


Figure 5.8: It depicts a graphical summary of methods for computing Weibull GP, which assumes a heterogeneous treatment effect on survival time from the observed survival data stratified by PD-L1 and treatment.

Methods to benchmark the GP Weibull

Cox penalised splines non-linear model The Cox penalised splines model is a regression hazard model that estimates the impact of a categorical, continuous covariate non-linearly [182]. Let us consider PD-L1 per cent staining to be categorical, continuous. We can estimate the parameters of the smoothing splines by the maximum likelihood approach choosing the degrees of freedom that improve the fit to the dataset as evaluated by the Akaike information criterion (AIC) [98]. We fit the Cox hazard model maximising the penalised partial likelihood as implemented in the R *survival* package [183], which is our benchmark for the baseline Weibull GP model that only regresses on PD-L1 per cent staining.

Bayesian Weibull interaction effects The Weibull GP builds up an extended model that measures the influence of treatment conditional on PD-L1 per cent staining. The interaction effects model is a conventional approach to measuring the influence of a covariate conditional on another covariate [40]. Therefore, to benchmark the Weibull GP model, let us consider a Weibull baseline hazard model where the effect of treatment depends upon PD-L1 per cent staining, its link function given by:

$$\lambda_i = \exp\left[\alpha_{x_{PD-L1}[i]} + \beta_{x_{PD-L1}[i]} A_i\right] \tag{5.20}$$

where $x_{PD-L1}[i]$ indicates the PD-L1 per cent staining for individual i.

5.2.4 Bayesian flexible parametric hazard models

Recall that the hazard function for the flexible parametric B-spline model is given by equation 2.12. Hence, the survival function for the B-splines model is given by:

$$S_i(T_i) = \exp\left(-\int_0^{T_i} h_i(u)du\right) \tag{5.21}$$

A significant computational burden of the B-splines model is that the survival function has no tractable analytical solution. Hence, it requires a costly numerical integration for each iteration in the HMC algorithm. Brilleman et al. [72] studied this problem and proposed a more convenient form of the Royston-Parmar model where instead of B-splines, one can use non-negative splines (M-splines) to calculate the survival function in closed form. From a computational perspective, using the M-spline formulation is advantageous because one can pre-compute the integral of the M-splines (I-splines), which is handy for integrating the hazard to obtain the survival function. The M-splines hazard model is given by:

$$\text{M-splines} \;:\; h_0(t) = \sum_{l=1}^{L} \theta_l M_l\left(t; \mathbf{k}; \delta\right) \tag{5.22}$$

where $\mathbf{k}$ denotes the knots, $\boldsymbol{\theta}$ the coefficients, and δ the degree of the M-splines. Since the M-splines coefficients $\boldsymbol{\theta}$ must sum to one, we use a *simplex* vector of values that must sum to one and give $\boldsymbol{\theta}$ a Dirichlet uniform prior, given by:

$$\boldsymbol{\theta} \sim \text{Dirichlet}\,(1) \tag{5.23}$$

We evaluate cubic M-splines with 5 and 10 degrees of freedom, respectively, as a compromise between underfitting and overfitting errors [40].

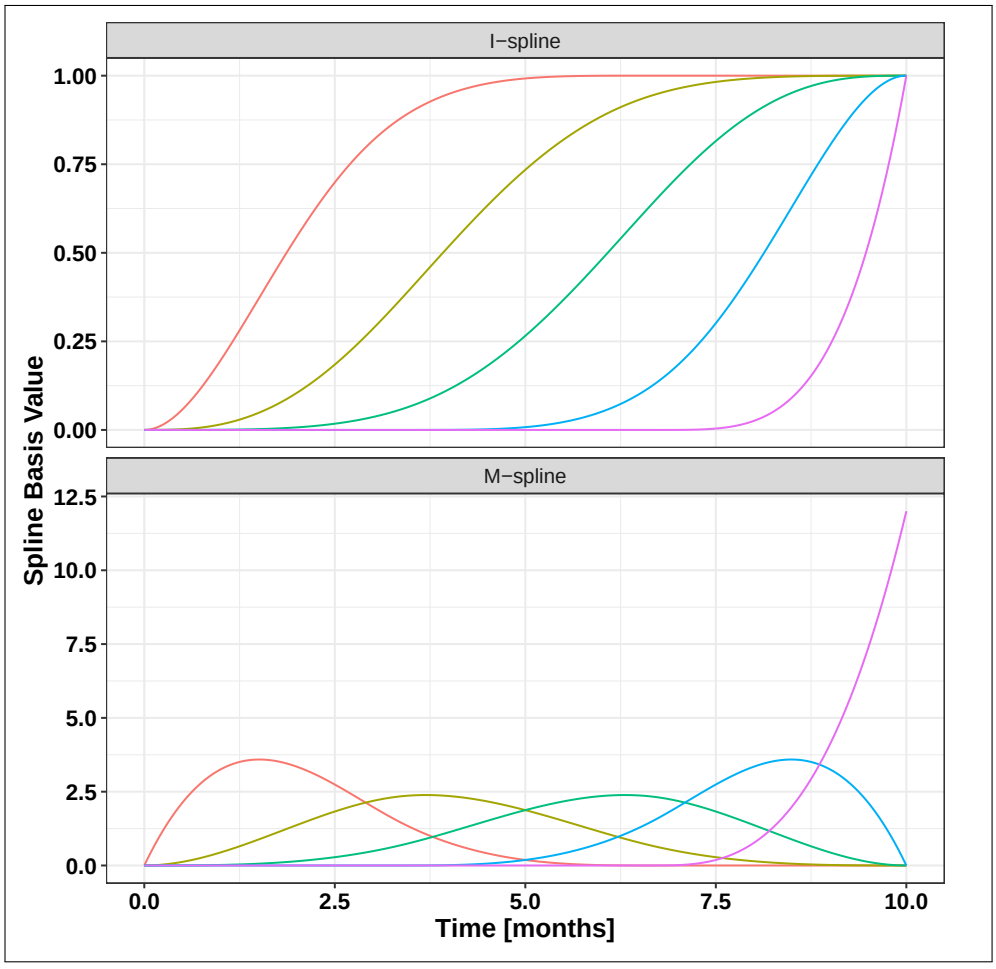


Figure 5.9: Illustration of M-splines and I-splines basis.

5.2.5 Bayesian non proportional hazard models

Building on the above idea of having a flexible parametric hazard model, we model treatment effects varying with time. The time-dependent regression coefficients are given by:

$$\beta_p(t) = \theta_{p0} + \sum_{m=1}^{M} \theta_{pm} \mathrm{B}\left(t; k; \delta\right) \tag{5.24}$$

where θ_{p0} is the covariate effect at time zero, $\mathrm{B}\left(t; k; \delta\right)$ denotes a B-spline with regression coefficients $\theta_{p,m}$, k knots and degree δ. Because our RWE study design's exclusion criteria exclude patients with resectable advNSCLC, see Section 5.2.1, we do not have reasons to believe that the time-varying effects vary rapidly, such a surgical procedure that would cause an accelerated change in the hazard. Therefore, we assume a smooth change of the hazard at the population level.

To encapsulate our assumption, we set the parameters of the B-splines to evolve smoothly by using a random walk prior distribution of the form:

$$\begin{aligned} \theta_{p,1} &\sim \mathrm{Normal}(0, 1) \\ \theta_{p,m} &\sim \mathrm{Normal}(\theta_{p,m-1}, \tau_p) \\ \tau_p &\sim \mathrm{Exponential}\left(1\right) \end{aligned} \tag{5.25}$$

where τ is a hyper-parameter for the standard deviation in the adaptive prior for the B-splines regression coefficients $\theta_{p,m}$. For regularisation, we assume a weakly exponential informative prior [96] for τ.

5.2.6 Standardised survival curves

Back in section 3.1 we introduced the fundamental problem of causal inference that counterfactuals are never observable outcomes and discussed several workarounds using causal assumptions, such as excahgeability and positivity. We showed that using consistency, the counterfactual outcomes are the expected outcome values under treatments $A = a, a'$, such that:

$$\mathbb{E}\left[Y^a\right], \mathbb{E}\left[Y^{a'}\right] \tag{5.26}$$

We introduced the parametric g-formula [118], which under the exchangeability assumption gives an unbiased estimate of the treatment effect. Here, we extend the parametric g-formula to survival analysis to estimate the ATE with observational RWE survival data and the CATE for heterogenous treatment effects.

Recall section 3.4, where adopting parametric models indexed by the finite

number of parameters θ, following [47], we defined the Bayesian g-formula by drawing samples for the counterfactual outcomes after conditioning on the observed data o, such that:

$$p(\tilde{y}_a|o) = \int \int p(\tilde{y}|a, \tilde{\mathbf{w}}, \theta)p(\tilde{\mathbf{w}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|o)d\boldsymbol{\theta}d\tilde{\mathbf{w}} \tag{5.27}$$

where we integrate over the observed confounder $\tilde{\mathbf{w}}$ and the uncertainty on the model parameters θ. Hence, for a finite sample N we estimate the standardised survival curve such that:

$$p(\widehat{S}_a(t)\,|\tilde{\mathbf{w}}, \theta) = \frac{1}{N}\sum_{i=1}^{N} p(\widehat{S}(t)\,|a, \tilde{w}_i, \theta) \tag{5.28}$$

where we estimate the counterfactual survival curves $S_a, S_{a'}$ by drawing samples from the posterior distribution evaluated for each i individual in the population of interest. We can visualise the full posterior counterfactual distribution or summarise it conveniently using point estimates such as the mean, median and quantiles. Finally, for summarising treatment effects, we can integrate the posterior survival function up to a pre-specified time τ to compute the counterfactual $\text{RMST}(\tau)_{a,a'}$, as described in equation 5.4.

For CATE estimation, we can condition on a variable of interest X when computing standardised survival curves. A population estimate of the standardised survival curve conditional on X is given by:

$$p(\widehat{S}_a(t)\,|\tilde{\mathbf{w}}, \theta, X = x) = \frac{1}{N_{X=x}}\sum_{i \in X=x} p(\widehat{S}(t)\,|a, \tilde{w}_i, \theta, X = x) \tag{5.29}$$

where we condition on the variable X to take the value x. For example, X can be the PD-L1 biomarker taking the 50% staining value. Similarly, we can compute the posterior counterfactual distribution for the RMST up to a pre-specified time τ given covariate $X = x$, denoted by $\text{RMST}(\tau)_{a,a'|X=x}$.

IPW non-parametric estimates

As explained in section 3.4, IPW is an alternative method to estimate treatment effects. IPW estimation of survival curves is a popular method to estimate treatment effects in survival analysis [124] because it allows for direct non-parametric modelling of the baseline survival function. In conventional IPW estimation, one first estimates the propensity scores and uses them to compute a weighted Kaplan and Meier maximum likelihood (KM) method.

Although standardisation and IPW estimation may yield by definition similar results [32], we take a pragmatic view and use conventional IPW to check deviations of the model predicted estimates from the non-parametric IPW

estimate. However, one must consider that IPW does not include time-varying covariate effects in the computation of the propensity score; hence, deviations can be misleading. Still, we consider it a helpful starting point for model checking.

We use the conventional multinomial log-linear regression model with a vanilla neural network [184], which is the maximum entropy model for multi-label classification to estimate the propensity scores. For obtaining IPW estimates of survival, we compute adjusted survival curves by adapting the KM method [185] weighting the individual contributions by the inverse of the propensity score. A disadvantage of using IPW survival is that estimation of confidence intervals for multiply imputed datasets are unavailable, and the authors are not aware of any appropriate method to compute the between imputations variance from survival curves. In contrast, Bayesian outcome regression modelling supports combining posterior distributions to visualise within and between imputation uncertainty. For IPW, one only has immediately available the average counterfactual of the survival estimate $\bar{S}(t)^a$, and the average $\text{R}\bar{\text{M}}\text{ST}(\tau)$, and bootstraped 95% confidence interval.

5.2.7 Model comparison: accuracy and utility

This chapter aims to estimate treatment effects, ATE and CATE. To do so, we focus on treatment and covariate effect modelling, adjustment by confounders, circumventing several types of potential bias, and assessing unobserved confounding bias. However, for the modelling part, we need to avoid the underfitting and overfitting pitfalls, see section 3.3. Hence, we want our models to perform well on out-of-sample data, not simply reproducing the training data but learning valuable patterns for predicting new unseen observations, such as counterfactuals. Using regularising priors, such as the weakly informative priors defined in equation 5.15 is a helpful Bayesian technique to shield against overfitting. However, we can improve on it by conducting a Bayesian model comparison [109], as explained below.

Prediction model performance comparison

A simple definition of accuracy for model-based prediction of survival data would be the difference between actual and predicted time-to-event. However, as discussed above, events of interest, such as death or disease progression, are sometimes not observed during the study, i.e. , censored. Hence, the above definition of accuracy would throw away costly data on subjects who have survived up to the follow-up. Instead, we use a more precise definition of accuracy based on the event's probability and survival up to the last observed time, including censored observations. That is the probability of the data for

deriving the likelihood. Let T be a random variable designating the event time, and d a censoring indicator. A censored individual ($d = 0$) will contribute to the likelihood via the survival probability (S) up to the censored time; otherwise an event ($d = 1$) will contribute via the hazard function (h) evaluated at the time of the event. Computing the accuracy amounts to multiplying all of the individual likelihoods to get the joint probability, which is given by:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} [h_i(T_i|\boldsymbol{\theta})]^{I(d_i=1)} \times [S_i(T_i)|\boldsymbol{\theta}]^{I(d_i \in \{0,1\})} \tag{5.30}$$

The joint likelihood allows us to benchmark competing models up to a proportional constant, although there are some caveats because we aim to evaluate the out-of-sample accuracy. For example, the widely applicable information criterion (WAIC) introduces a penalty term for selection bias. The WAIC is given by:

$$-\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{S}\sum_{j=1}^{S} p(\mathcal{D}_i|\boldsymbol{\theta}_j) + \frac{1}{N}\sum_{i=1}^{N} V_j^S \log p\left(\mathcal{D}_i|\boldsymbol{\theta}_j\right) \tag{5.31}$$

The first term is a measure of fit over all the posterior samples. The second term is the penalty term measures the degree of uncertainty in parameter values. If it is vast, it indicates that the model is too complex, suggesting overfitting.

Consider again the data $\mathcal{D}_1, \ldots, \mathcal{D}_n$, where the fundamental unit of observation is a subject. We collapse the data probability within each subject $\mathcal{D}_1, \ldots, \mathcal{D}_N$ and decompose the likelihood into a product of subject-wise likelihoods, given by:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathcal{D}_i|\boldsymbol{\theta}) \tag{5.32}$$

After the subject-data $\mathcal{D}$ have been observed, we can predict a new subject given a prior distribution $p(\theta)$ and a posterior predictive distribution, given by:

$$p(\tilde{\mathcal{D}}|\mathcal{D}) = \int p(\tilde{\mathcal{D}}_i|\theta)p(\theta|\mathcal{D})\, d\theta. \tag{5.33}$$

Therefore, the expected log-predictive density (elpd) for a new subject is given by:

$$\text{elpd} = \sum_{i=1}^{N} \int p_t(\tilde{\mathcal{D}}_i) \log p(\tilde{\mathcal{D}}_i|\mathcal{D})\, d\tilde{\mathcal{D}}_i, \tag{5.34}$$

where $p_t(\tilde{\mathcal{D}}_i)$ is the likelihood of the left-out-subject $\tilde{\mathcal{D}}_i$. Estimating $p_t(\tilde{\mathcal{D}}_i)$ via cross-validation approximated by Pareto-smoothed importance sampling [186] allows for fast and optimal targeted predictions. We use the survival elpd

leave-one-subject-out cross-validation ($\text{elpd}_{\text{loso}}$), given by:

$$\text{elpd}_{\text{loso}} = \sum_{i=1}^{n} \log p(\mathcal{D}_i|\mathcal{D}_{-i}), \tag{5.35}$$

where $\mathcal{D}_i$ denotes the group of observations of subject i. It is an extension of the method proposed by [187], where similar to [109], the data division is subject-wise.

Causal model performance comparison

Similarly to evaluating model-based predictions, comparing models for causal inference needs a yardstick to assess model performance. Conventionally, to compare and evaluate causal models, researchers use synthetic data. By using a generative model, one can simulate survival outcomes using the technique explained in section 3.2. The helpfulness of synthetic data relies on the fact that the technique allows generating data on the factual and counterfactual outcomes and, therefore, allows for comparing how accurate are the models in estimating the treatment effect and the effect modifiers, respectively. See section 5.3.1 for such a synthetic data comparison for proportional and non-proportional hazard models.

However, for the causal inference task, the definition of accuracy in real-world data is truly unobservable because, as explained above, the counterfactual outcomes are never available. Therefore, the likelihood of the causal survival model which is given by:

$$p(\mathcal{D}|\boldsymbol{\theta}, A) = \prod_{i=1}^{N} \left[h_i(T_i^{A=a}|\boldsymbol{\theta})\right]^{I(d_i=1)} \times \left[S_i(T_i^{A=a}|\boldsymbol{\theta})\right]^{I(d_i \in \{0,1\})} \tag{5.36}$$

where $A = a$ is a treatment indicator, is inaccessible for modelling and conventional train-test splitting criterion or cross-validation.

Instead, we need to define a different utility function that is computable. Such a heuristic should be determinable for all survival models and incorporate model usefulness in predicting treatment interventions. We propose a head-to-head model comparison based on the utility of the model for OTS. We assume exchangeability and positivity and hence based the heuristic on observable characteristics of individuals. The key is that the model has to estimate how the treatment effects vary to make accurate predictions of the outcome and be helpful for subpopulation stratification in treatment individualisation. Mathematically, the utility function is given by:

$$\mathbf{U(x)} = \sum_{i=1}^{N} \mathbb{E}_{\mathrm{Y}^1|x_i}\left[\mathrm{Y}^1|x_i\right] - \mathbb{E}_{\mathrm{Y}^0|x_i}\left[\mathrm{Y}^0|x_i\right] \tag{5.37}$$

where $\mathbf{U(x)}$ is the utility of separating the individualised factual and counterfactual outcomes. That is conditional on a set of covariates of interest X, such as the biomarker PD-L1 expression, equation 5.37 maximises the accuracy of predicitng the CATE, such that:

$$arg \max_{A \in \{a,a'\}} \left\{ \mathbb{E}\left(Y^a | X - Y^{a'} | X \right) \right\} \tag{5.38}$$

The heuristic $\mathbf{U(x)}$ is only helpful for comparing models that handle interactions with treatment because models that fail in estimating how treatment varies with covariates will give a value of zero, by definition. Therefore, we only apply $\mathbf{U(x)}$ for comparing causal models for heterogeneous treatment effects. To evaluate equation 5.37 we use a stratified cross-validation seeking to ensure that each fold is representative of the target defined in equation 5.38.

5.2.8 Optimal treatment selection bounds for survival end-points

As explained above, the ATE is the expected difference between the counterfactual outcome under treatment $A = a$ and treatment $A = a'$, such that:

$$\mathbb{E}\left[Y^a - Y^{a'} \right] \tag{5.39}$$

Hence, applying equation 5.28, we define the ATE for survival outcomes as the expected difference in the counterfactual $\text{RMST}(\tau)^{A=a,a'}$ between treatments $(A = a, a')$. We estimate the survival ATE by drawing samples from the posterior distribution for the difference in counterfactual RMST for a population with measured confounders $\mathbf{w}$, after conditioning on the observed data, such that:

$$\widehat{\text{ATE}} = \int_0^\tau \hat{S}^{A=a}(t|\tilde{\mathbf{w}}, \boldsymbol{\theta}) - \hat{S}^{A=a'}(t|\tilde{\mathbf{w}}, \boldsymbol{\theta}) dt \tag{5.40}$$

where $\hat{S}^{A=a}(t|w, \boldsymbol{\theta})$ is a parametric estimate of $S^{A=a}(t|w, \boldsymbol{\theta})$.

Nevertheless, as noted in section 3.5 the exchangeability assumption is strong. Hence, the credibility of the results may be weak. To assess the impact of exchangeability violations on one's results, one can consider OTS bounds on the ATE. The OTS assumption says that a "perfect doctor" *always* prescribes the best treatment available for each patient. We argue that for RWD, the OTS assumption may hold, and hence, it is a valuable technique for estimating causal bounds on the ATE with weaker assumptions than exchangeability [86, 125, 126]. For an in-depth explanation of the OTS assumption, we refer to section 3.5, below we define OTS bounds for survival end-points.

Let us consider an OTS bound for the comparison of two treatments. Let $\text{RMST}(\tau)^{A \in a,a'}$ be the RMST under each treatment. According to the OTS

assumptions, the following inequalities hold:

$$\begin{aligned} A_i = a \Rightarrow \text{RMST}(\tau)_i^a \geq \text{RMST}(\tau)_i^a, \\ A_i = a' \Rightarrow \text{RMST}(\tau)_i^{a'} \geq \text{RMST}(\tau)_i^{a'} \end{aligned} \tag{5.41}$$

The first inequality is known as the direct OTS inequality. The second inequality is the contrapositive OTS inequality. Let us consider that 0 days in time is the worst possible RMST outcome, and the maximum time of follow-up τ is the best possible RMST outcome. From equation 5.41, we can show that the upper and lower direct OTS bounds are given by:

$$\begin{aligned} \mathbb{E}\left[\text{RMST}(\tau)^a - \text{RMST}(\tau)^{a'}\right] \leq P(A = a)\mathbb{E}\left[\text{RMST}(\tau)|A = a\right] \\ \mathbb{E}\left[\text{RMST}(\tau)^a - \text{RMST}(\tau)^{a'}\right] \geq P(A = a')\mathbb{E}\left[\text{RMST}(\tau)|A = a'\right] \end{aligned} \tag{5.42}$$

Similarly, from equation 5.41, we can show that using the contrapositive OTS we can derive a second set of upper and lower OTS bounds, which are given by:

$$\begin{aligned} \mathbb{E}\left[\text{RMST}(\tau)^a - \text{RMST}(\tau)^{a'}\right] \leq \mathbb{E}\left[\text{RMST}(\tau)|A = a\right] - P(A = a')\mathbb{E}\left[\text{RMST}(\tau)|A = a'\right] \\ \mathbb{E}\left[\text{RMST}(\tau)^a - \text{RMST}(\tau)^{a'}\right] \geq P(A = a)\mathbb{E}\left[\text{RMST}(\tau)|A = a\right] - \mathbb{E}\left[\text{RMST}(\tau)|A = a'\right] \end{aligned} \tag{5.43}$$

Because both sets of OTS bounds derive from the same OTS assumption, we can derive a narrow interval for the OTS bounds. Figure 5.10 depicts the analytical steps that we perform for computing ATE and causal OTS bounds on the ATE. We develop a Bayesian non-proportional hazard model from the survival data, which we use to compute the ATE via the Bayesian g-formula. Finally, we compute the OTS assumption's causal bounds to estimate the robustness of the ATE to unobserved confounding.

5.3 Results

5.3.1 Synthetic data experiments

The following sub-section presents a synthetic data experiment to compare the proportional hazard and the non-proportional hazard models in the presence of time-varying effects and observed confounders.

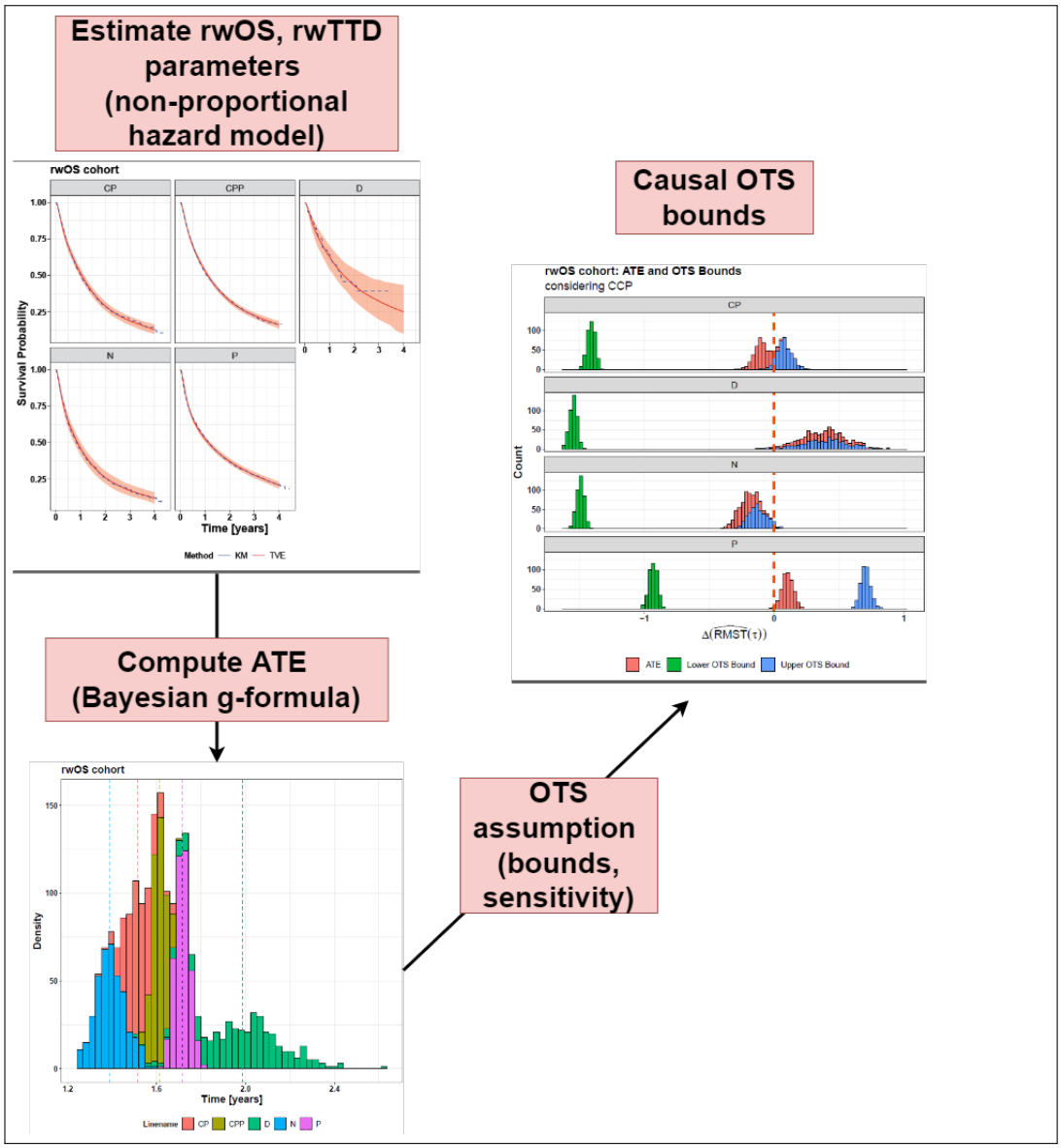


Figure 5.10: It depicts a graphical summary of the methods used for computing ATE and causal OTS bounds for rwTTD and rwOS.

Experimental settings

For our first set of experiments, we use the synthetic data generation model to validate the performance of the Bayesian non-proportional hazard model under randomised and observational data scenarios. Let us assume the SCM depicted in figure 5.11. A denotes a treatment variable impacting a right-censored survival outcome Y, and two sets of confounders W_1, W_2 act as forks affecting both treatment A and outcome Y. The structural equation of this hypothetical scenario is given by:

$$\begin{aligned} W_1 &:= f_{W_1}\left(U_{W_1}\right) \\ W_2 &:= f_{W_2}\left(U_{W_2}\right) \\ A &:= f_A\left(U_A, W_1, W_2\right) \\ Y &:= f_Y\left(U_Y, A, W_1, W_2\right) \end{aligned} \tag{5.44}$$

We draw the two sets of confounds W_1 and W_2 adapting the data generating process described in section 4.2.4 to allow for relatively low correlation $\rho \approx 0.2$ between and within the observable variables of the sets W_1 and W_2.

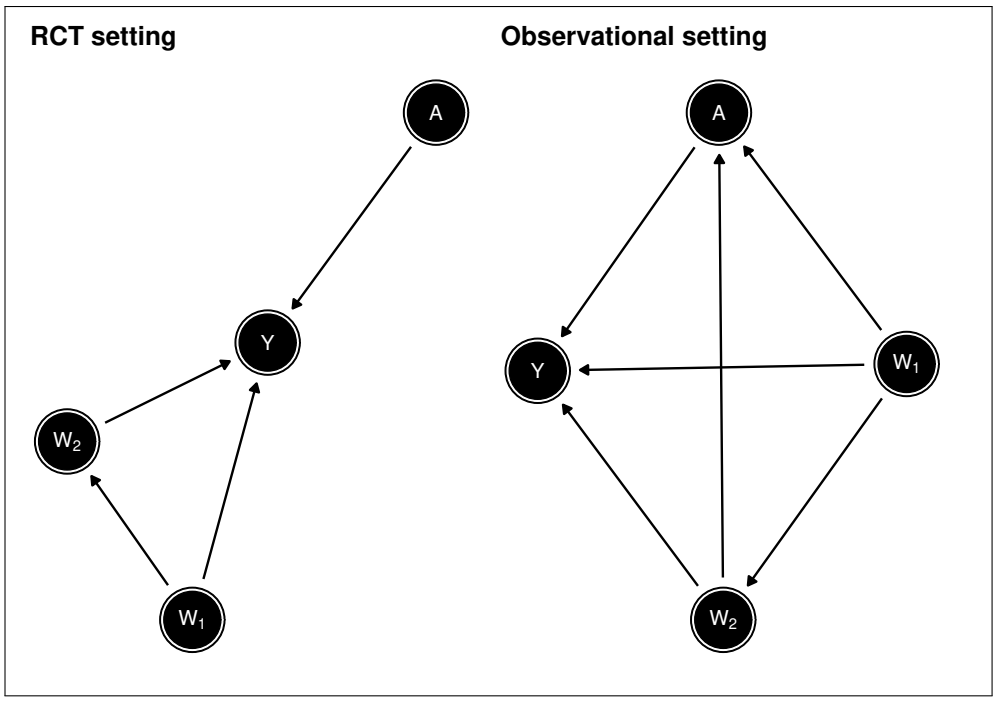


Figure 5.11: Hypothetical SCM's DAG for randomised clinical trial (RCT) and Observational non-randomised simulation scenarios.

Treatment assignment Let the treatment variable A take on three different values for the synthetic data experiment: $A = 1, 2, 3$. We set our non-randomised

(Observational) simulation experiments such as the set of confounders W_1 and W_2, denoted by W, to impact the treatment assignment variable A. To do so, we define the following multinomial logistic transformation given by:

$$\begin{aligned}
\Pr(A = 1) &= \frac{\exp(\beta_1 \mathrm{W})}{1 + \sum_{k=1}^{K-1} \exp(\beta_k \mathrm{W})} \\
\Pr(A = 2) &= \frac{\exp(\beta_2 \mathrm{W})}{1 + \sum_{k=1}^{K-1} \exp(\beta_k \mathrm{W})} \\
\Pr(A = 3) &= \frac{\exp(\beta_3 \mathrm{W})}{1 + \sum_{k=1}^{K-1} \exp(\beta_k \mathrm{W})}
\end{aligned} \tag{5.45}$$

We simulate each treatment assignment using a multinomial random variable with size one, given by $\text{Multinomial}(1, Pr(A = a))$ since it is the maximum entropy distribution for multiclass variables. For the randomised clinical trial scenario, we sample treatment assignment independently from $W1, W2$, using a multinomial random variable with size one, given by $\text{Multinomial}\left(1, \frac{1}{3}\right)$, assuring class balance and exchangeability.

Generation of event times Let us assume that the event times are Weibull distributed, $y \sim Weibull(\lambda(t), \gamma)$. We set the shape parameter of the Weibull distribution to $\gamma = 1.1$, generating a baseline hazard that increases with time. The constant-time log-link function is given by:

$$\log(\lambda_i) = \eta_i = \beta_0 + \sum_{p=1}^{P} \beta_p w_i \tag{5.46}$$

where the regression coefficients β are set to appropiate values. We define the non-proportional hazard via a time-varying treatment effect weighted by the set of parameters ϕ. To simulate event times we use the following non-linear hazard model:

$$\begin{aligned}
\frac{dS}{dt} &= - \cdot \lambda \cdot \gamma \cdot t^{\gamma - 1} \cdot \exp(\phi_1 \cdot A_2 \cdot t) \cdot \exp(\phi_2 \cdot A_3 \cdot t) \\
S(0) &= 1
\end{aligned} \tag{5.47}$$

The model is non-linear on the impact of treatment A; hence, the hazard function between treatment arms is non-proportional. We integrate equation 5.47 using an LSODA ODE solver [188] and use ITS to simulate event times, see algorithm 2, for a synthetic population of $N = 10,000$ and administrative censoring set to five months.

Outcome regression models for the synthetic dataset We present a head-to-head comparison of three parametric survival models for simulated

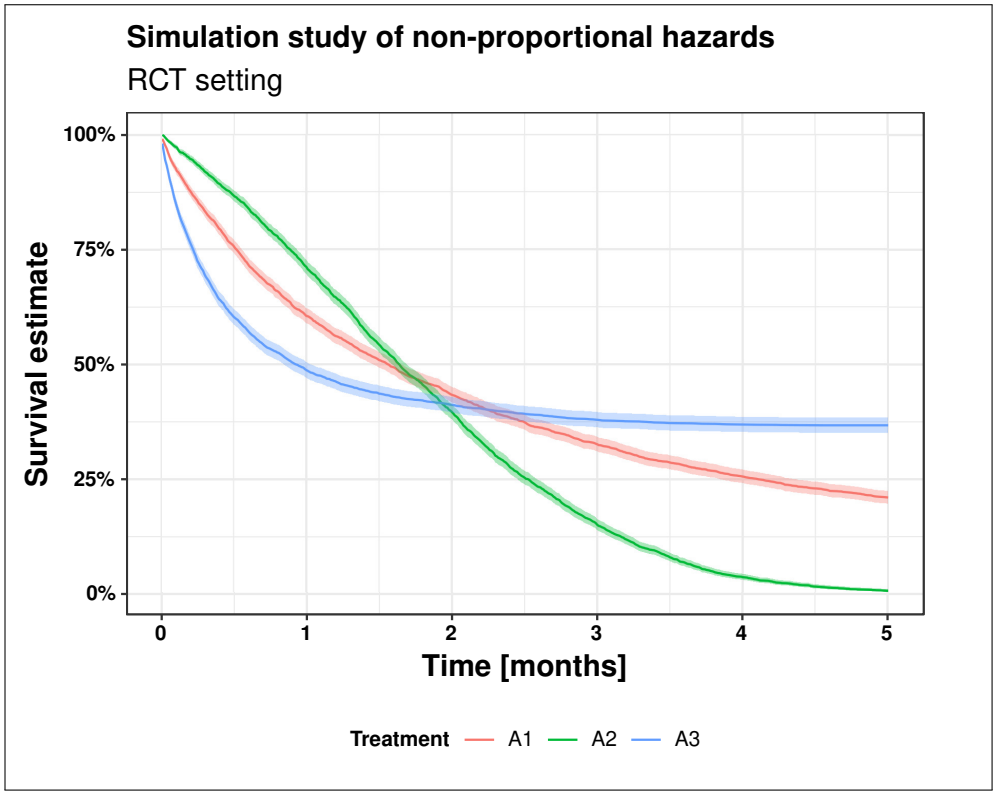


Figure 5.12: The figure shows the non-parametric survival estimate and 95% confidence interval for the scenario of non-proportional hazards impacting survival in a simulated randomised clinical trial (RCT), where exchangeability holds.

observational data, where the model of the baseline hazard (h_0) is a Weibull baseline model. The unadjusted model (Unadjusted) is a proportional hazard model regressed on treatment variables, whose hazard is given by:

$$h(t) = h_0(t) \exp(\beta_2 A_2 + \beta_3 A_3) \tag{5.48}$$

The PH adjusts for the set of confounders $W1$ and $W2$ by regressing the hazard on treatment, $W1$ and $W2$, whose hazard function is given by:

$$h(t) = h_0(t) \exp(\beta_{A2} A2 + \beta_{A3} A3 + \beta_{W1} W1 + \beta_{W2} W2) \tag{5.49}$$

The NPH includes a time-varying effect component for the treatment variables with natural cubic B-splines that model the impact of A on the hazard as a smooth function of time, as explained in the section 5.2.

Model comparison on synthetic data

Figure 5.12 illustrates the survival curve plot for the simulated randomised

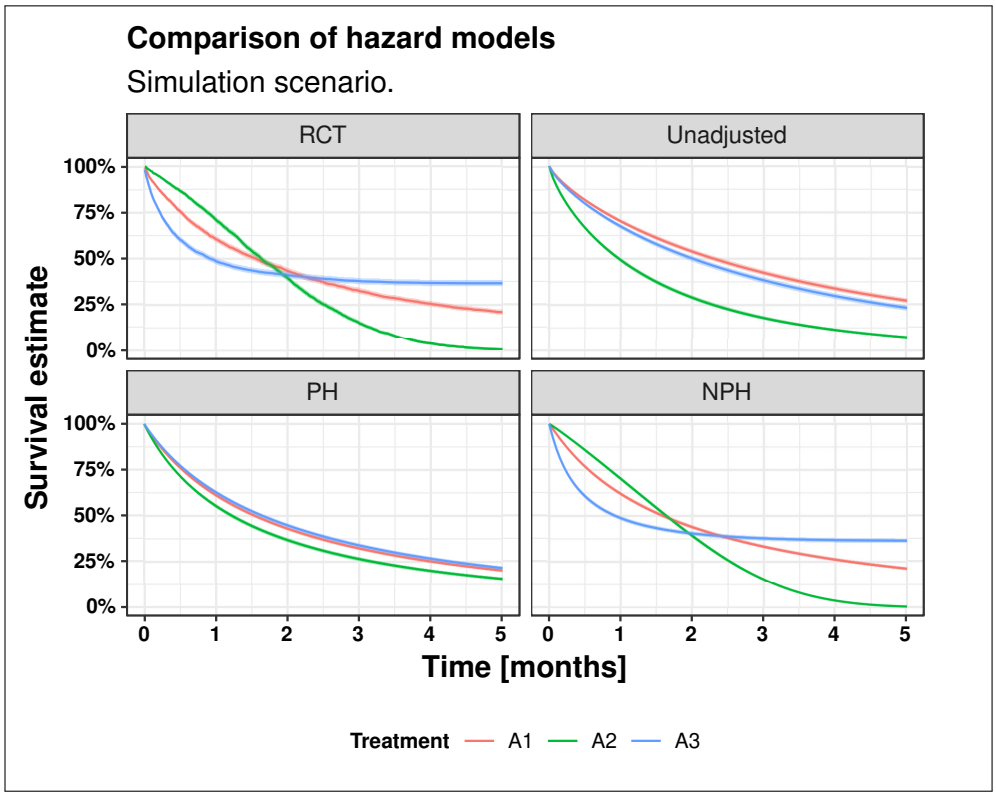


Figure 5.13: Comparison of survival estimates for the NPH, PH and Unadjusted models. Only NPH can recover the non-linearity observed in the non-parametric estimate from the simulated randomised clinical trial (RCT). NPH: Non-proportional hazards. PH: Proportional hazards. Unadjusted: Unadjusted proportional hazards.

Table 5.8: Comparison of of the probability mass included within the simulated RCT confidence intervals for the RMST(τ) counterfactual outcomes. NPH: Non-proportional hazards. PH: Proportional hazards. Unadjusted: Unadjusted proportional hazards.

Model	A1	A2	A3
Unadjusted	0	0	0
NPH	0. 8625	1	0. 9925
PH	1	0	0. 9625

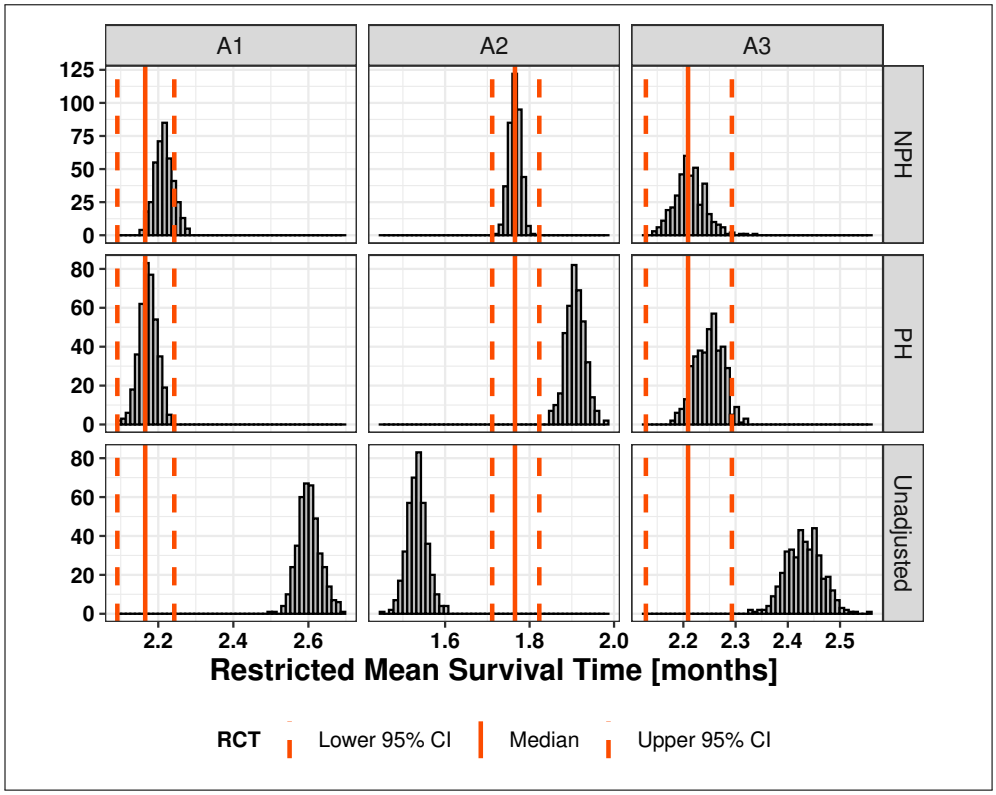


Figure 5.14: Comparison of the restricted mean survial time estimates (RMST(τ)) for NPH, PH and Unadjusted models. Vertical lines show RMST(τ) for the RCT. NPH: Non-proportional hazards. PH: Proportional hazards. Unadjusted: Unadjusted proportional hazards. RCT: randomised clinical trial.

clinical trial. We obtained estimates for the survival curve using the non-parametric KM method [69] and the asymptotic 95% confidence interval.

Figure 5.13 illustrates the survival curve plots for the simulated follow-up time, constituting the survival estimates, the model predicted mean, and 95% credible intervals [40] comparing the randomised clinical trial setting (RCT) and the observational setting. Neither the PH nor the Unadjusted model can capture the non-linearity of the survival curves and non-proportionality of the survival ratios seen in the simulated RCT data. We see visually that the NPH model can recover the actual survival curves seen in the RCT data.

Figure 5.14 illustrates the counterfactual RMST for all models. For the RCT setting, vertical lines show the mean RMST (solid line) and the 95% confidence intervals (dashed lines) for each treatment arm. For the causal models, we show 400 samples from the posterior distribution after determining convergence and unimodality. Table 5.8 compares the models by evaluating the proportion of the posterior distribution included within the RCT 95% CI. In summary, the Unadjusted model does poorly in recovering the actual counterfactual outcomes. The NPH outperforms the PH model for non-proportional survival ratios.

5.3.2 RWD experiments

The following sub-section presents our real-world experiments:

1. We compare the predictive performance of the canonical parametric baseline hazards exponential and Weibull.
2. We report the results for the CATE on PD-L1 expression modelled with the GP are shown compared to current state-of-the-art interaction models.
3. The predictive performance of the flexible parametric Royston-Parman survival models is compared to canonical parametric survival models.
4. We experiment with post-treatment bias.
5. We report the results for the ATE of the PH and NPH models on the Flatiron RWE dataset.
6. We compute the causal OTS bounds for the ATE.

Baseline hazard model

Let us examine the results for the predictive performance of the canonical

Table 5.9: Comparison of fit to data in the rwTTD cohort for canonical baseline hazard models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
Weibull	0. 0	0. 0	-1799. 0	94. 7
Exponential	-71. 9	12. 7	-1870. 9	103. 2

Table 5.10: Comparison of fit to data in the rwOS cohort for canonical baseline hazard models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
Weibull	0. 0	0. 0	-7113. 0	45. 9
Exponential	-49. 7	8. 4	-7162. 7	44. 9

baseline models exponential and Weibull. Tables 5.9 and 5.10 compare the baseline survival model for rwTTD and rwOS, respectively. The first column of the comparative tables is the difference in expected log-predictive density (Δ ELPD) from the best model. The second column is the standard error for the contrast Δ ELPD. The third column is the value of the ELPD leave-one-subject-out cross-validation (ELPD LOSO), whose standard error showing in the fourth column. The results suggest that for both rwTTD and rwOS, the Weibull baseline hazard model obtains a better fit to the dataset ($\Delta\text{ELPD} = -49.7 \pm 8.4$).

Figure 5.15 depicts the KM and the Weibull baseline survival functions comparing the RMST(4 years) for the rwTTD and rwOS cohorts. We can see that the Weibull model is in good agreement with the non-parametric approach ($\Delta\text{RMST(4 years)} < 0.1$ years).

Regression prior predictive checks

Let us review the results for the prior predictive checks (PriorPC), a Bayesian workflow technique to assess the suitability of the prior distribution. We consider the PriorPC for the Weibull parametric survival model with time-constant covariates comprising the confounders listed in section 5.2. In short, we sample from the prior distribution to obtains draws for the survival function and compute the RMST. Let us set τ to 4 years, approximately the maximum survival time of the dataset analysed. Figure 5.16 shows the results as 400 samples for the RMST(4 years) density. The results suggest bumps on the edges for the RMST(4 years) for the vague prior. The specific informative prior may not be adequate for our use case. The Student-t WIP and Normal WIP produce similar RMST(4 years), and for simplicity, we will use the Normal

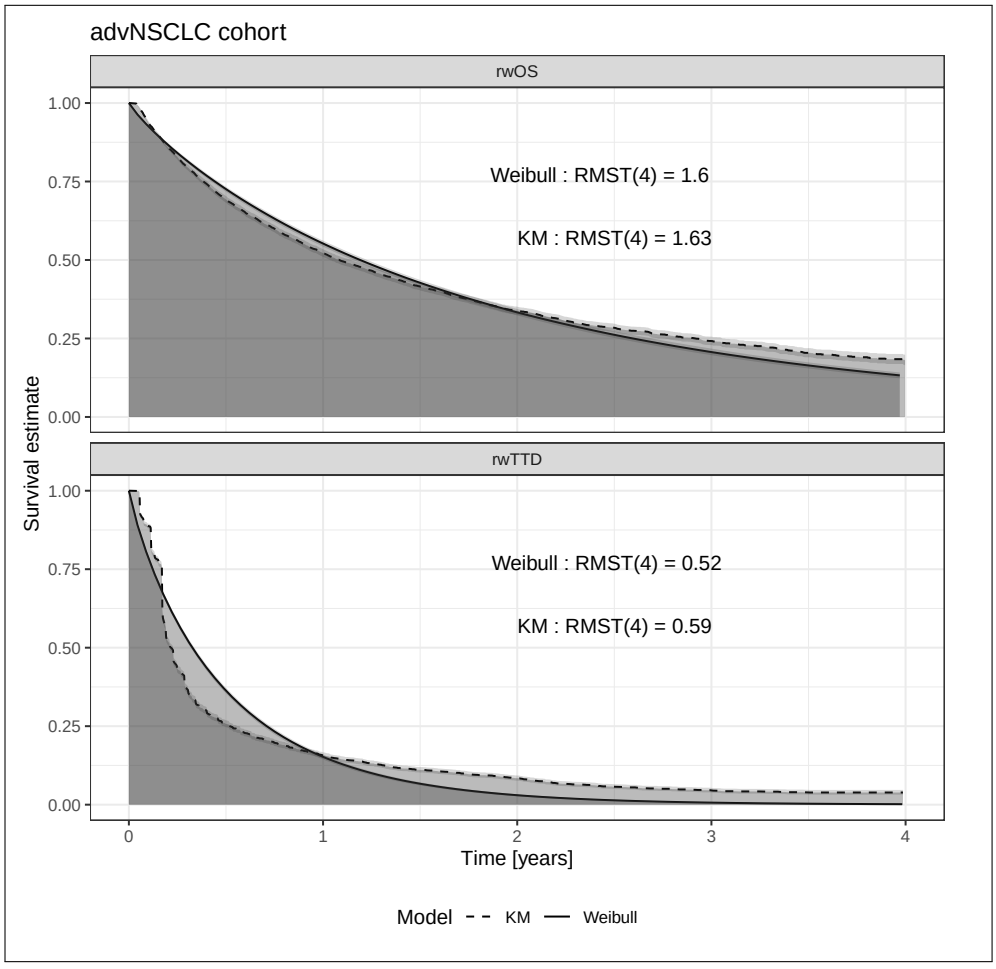


Figure 5.15: Kaplan and Meier non-parametric and Weibull parametric estimates of the survival function for the rwTTD cohort (bottom) and rwOS cohort (top).

WIP for the time-constant covariates in the rest of the analysis.

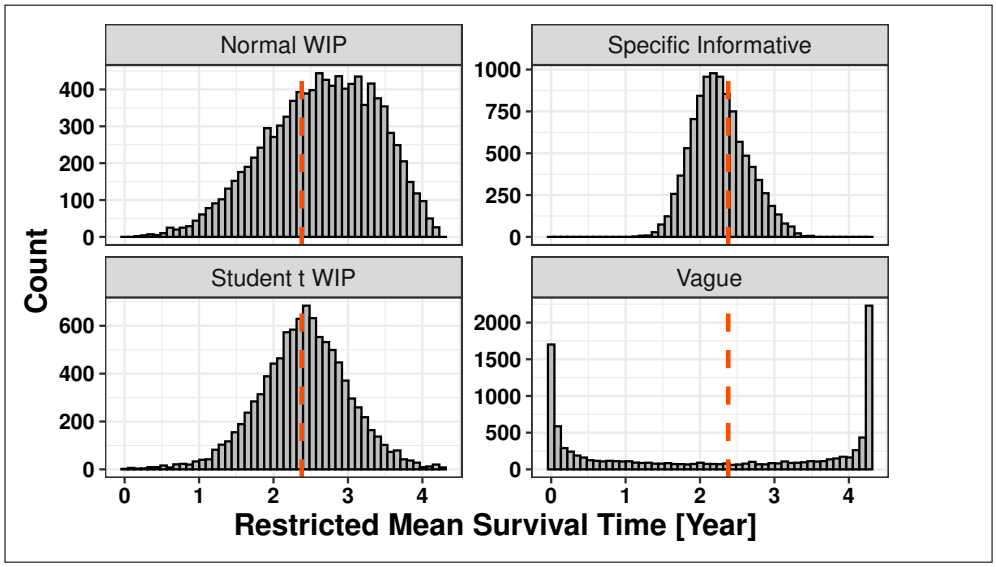


Figure 5.16: Illustration of prior predictive checks (PriorPC) for vague prior, weakly informative prior (WIP), Normal WIP and Student's t WIP, and specific informative prior (SIP) on the hazard scale.

Comparison of non-linear effects models for PD-L1 per cent staining impact on survival

Continuing to analyse the survival dataset marginally on PD-L1 expression, we compare the inference of the GP Weibull model and a Cox penalised splines model. Figure 5.17 and figure 5.18 visualises the term contribution of PD-L1 per cent staining in the prognostic index for the Weibull GP model and the Cox proportional hazard model in the rwTTD and rwOS, respectively. We can see that both models are in good agreement in the mean predictions.

In addition, the Weibull GP model produces a 95 % credible interval that we can use to quantify the uncertainty about the model predictions. The R^2 correlation between the $\hat{f}(x_{PDL1})$ from the Cox proportional hazard model and the Weibull GP model is 0. 973 for the rwTTD cohort and 0. 903 for the rwOS cohort. We see that the non-linear effect of PD-L1 is decreasing, implying that higher PD-L1 per cent staining is associated with longer time-to-event in the rwOS and rwTTD cohorts, which is in good agreement with the non-parametric KM analysis above but is harder to interpret.

Covariance between PD-L1 per cent staining impact on survival

The parameters of the GP are not very easy to interpret. In particular, α^2 and ρ define a covariance on the log scale and only have meaning in combination. To understand the covariance implied by the model with individuals with

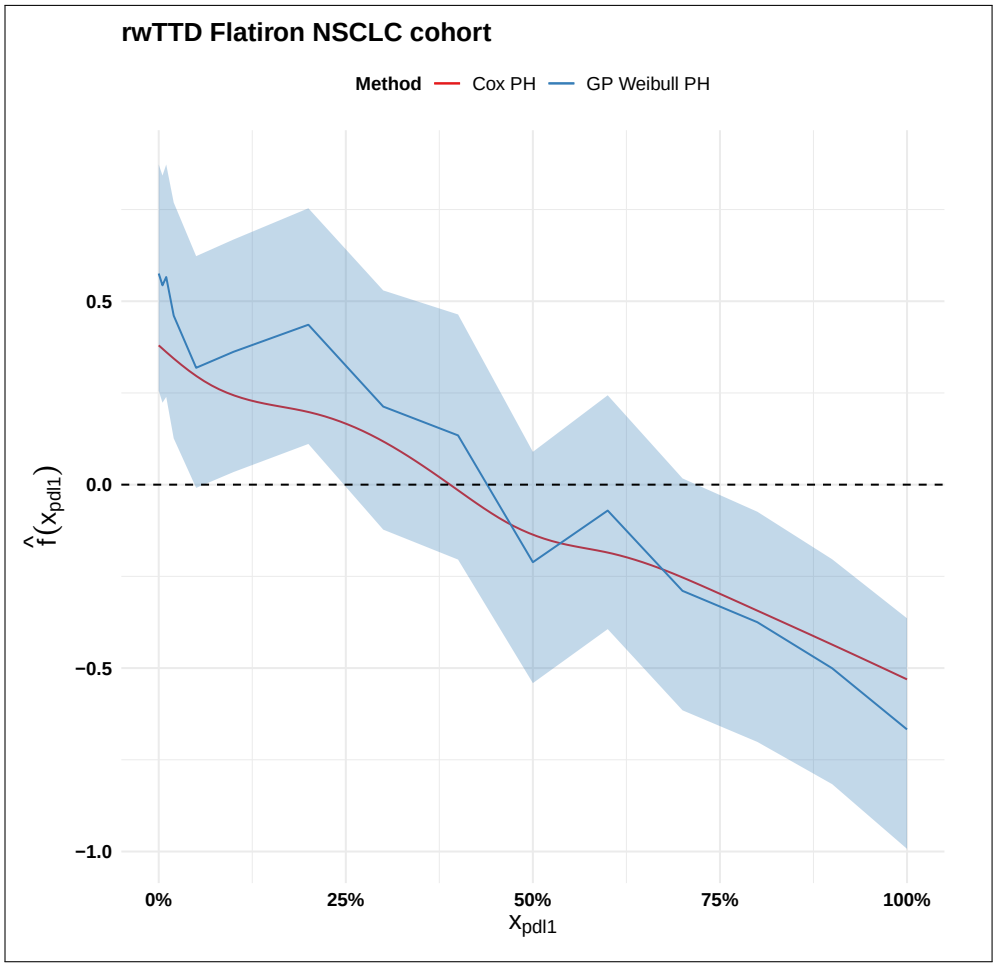


Figure 5.17: Cox penalised splines and GP Weibull log-linear link function, rwTTD.

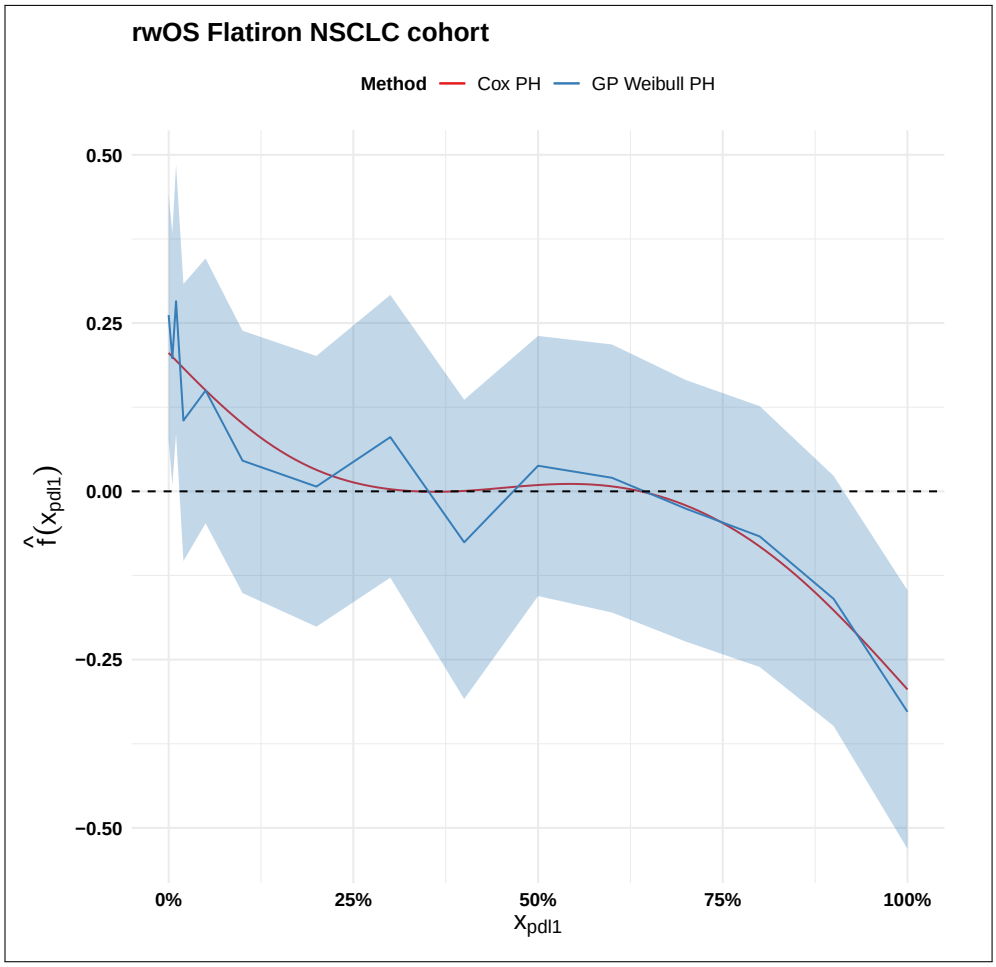


Figure 5.18: Cox penalised splines and GP Weibull log-linear link function, rwOS.

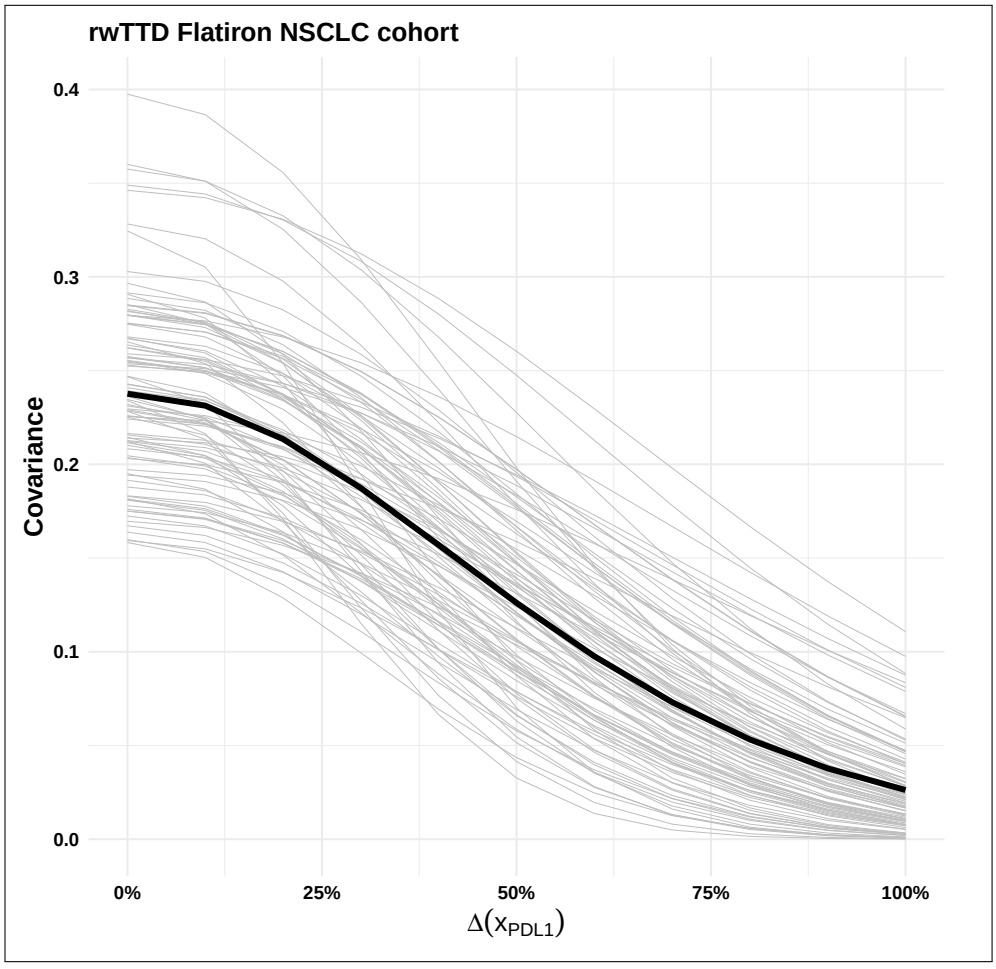


Figure 5.19: Draws from the posterior distribution of covariance functions, rwTTD. Patients with similar PD-L1 expression have similar rwTTD time but not identical.

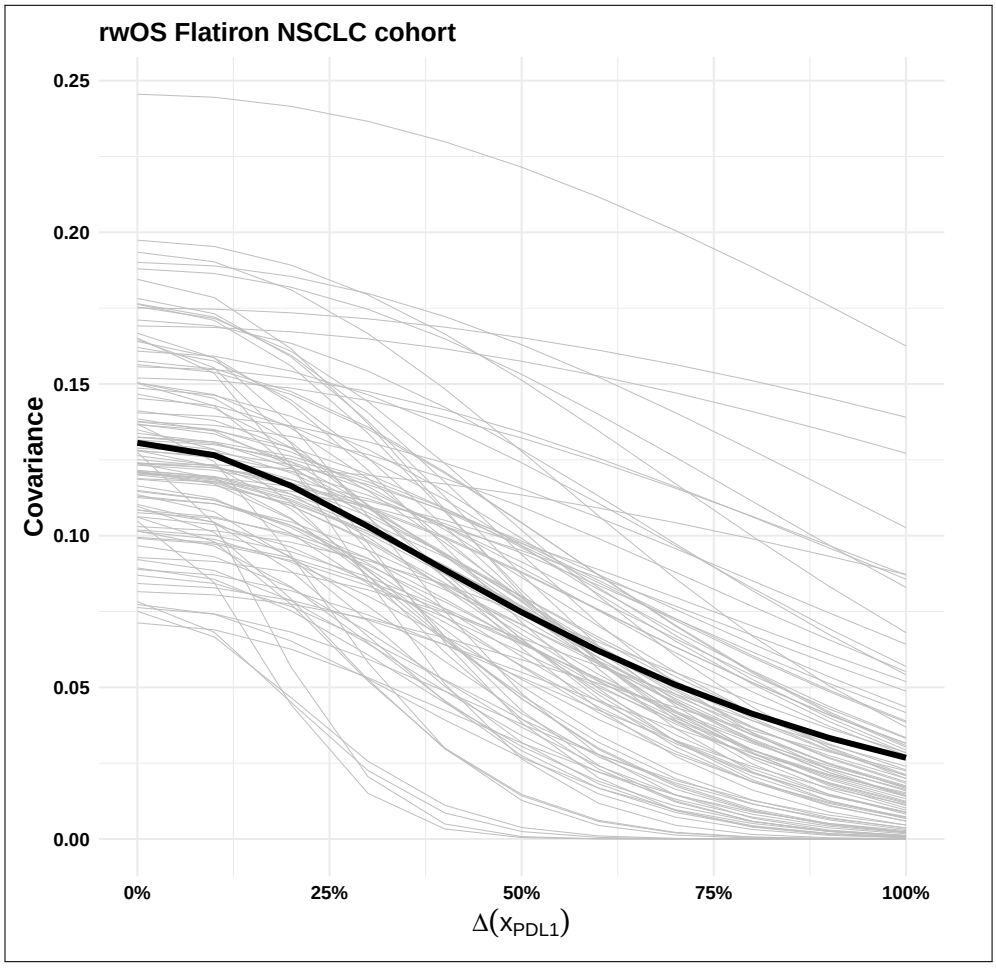


Figure 5.20: Draws from the posterior distribution of covariance functions, rwOS. Patients with similar PD-L1 expression have less similar rwOS than rwTTD from figure 5.19.

different PD-L1 expressions, we sample from the posterior distribution of α^2 and ρ and compute the covariance with Δx_{PD-L1}, see section 5.2.3 for the covariance function definition.

Figure 5.19 for rwTTD and figure 5.20 for rwOS depict 100 draws from the posterior distribution of covariance functions, where the dark line is the mean covariance over 2000 draws. On the x-axis, we plot Δx_{PD-L1}, the difference between PD-L1 per cent staining, and on the y-axis, the covariance. We can see that the covariance decreases with Δx_{PD-L1} more rapidly for rwTTD than for rwOS. Patients with similar PD-L1 expression have similar outcomes but not identical. Patients with similar PD-L1 expression have less similar rwOS than rwTTD. In general, there is little residual covariance between patients with differences of 50% PD-L1 expression or greater. For rwTTD, the covariance is 0.25 for two individuals with the same PD-L1 expression ($\Delta x_{PD-L1} = 0$), and decreases to almost 0 for two individuals with Δx_{PD-L1} of 100 %, and at $\Delta x_{PD-L1} = 50\%$ is approximately 0.1. For rwOS, the maximum covariance is approximately 0.125 for individuals with the same PD-L1 expression and decreases to approximately 0. 025 for individuals with "positive" PD-L1 (100% PD-L1 staining) considering individuals with "negative" PD-L1 (0% PD-L1 staining).

CATE and confronting confounding

So far, we have discussed a model that regresses only on PD-L1 per cent staining. However, the goal of Weibull GP is to regress PD-L1 per cent staining by treatment group. In the next experiment, we fit the Weibull GP varying treatment effects model and analyse the impact of the potential confounders W, listed in the section 5.2. Previously, we discussed that the coefficient $\hat{f}(x_{PDL1})$ are a bit opaque. For regressing the Weibull GP on treatment groups, the coefficients of the GP are less interpretable, as we absorb the intercept on the first treatment group for identifiability. Therefore, for interpretations, we focus on the predictive posterior distribution for the RMST. We set τ to four years because it is approximately the maximum follow-up time for the dataset analysed.

We compute the Bayesian parametric g-formula for the CATE stratified by PD-L1 per cent staining and plot the median and 95% credible intervals. In addition, we fit a second model where we additionally regress on the confounders defined in the section 5.2, denoted by W. Figure 5.21 for rwTTD and Figure 5.22 for rwOS depict the expectations of RMST(4 years) for each treatment group in blue, the unadjusted model, and in red, the model that attempts to adjust for W.

For rwTTD, figure 5.21 does not reveal significant changes in the inference for the CATE after adjusting for W. For rwOS, Figure 5.22 suggests that W is

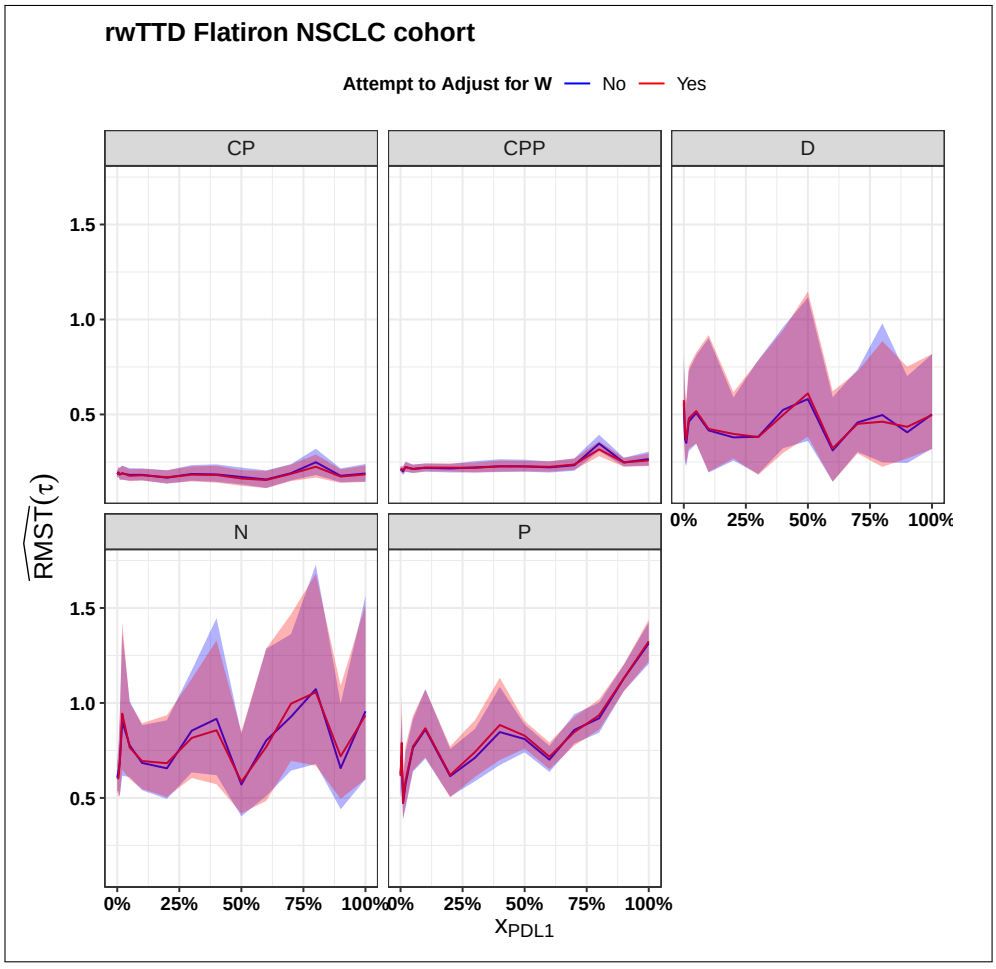


Figure 5.21: RMST(4 years) for unadjusted and adjusted model, rwTTD.

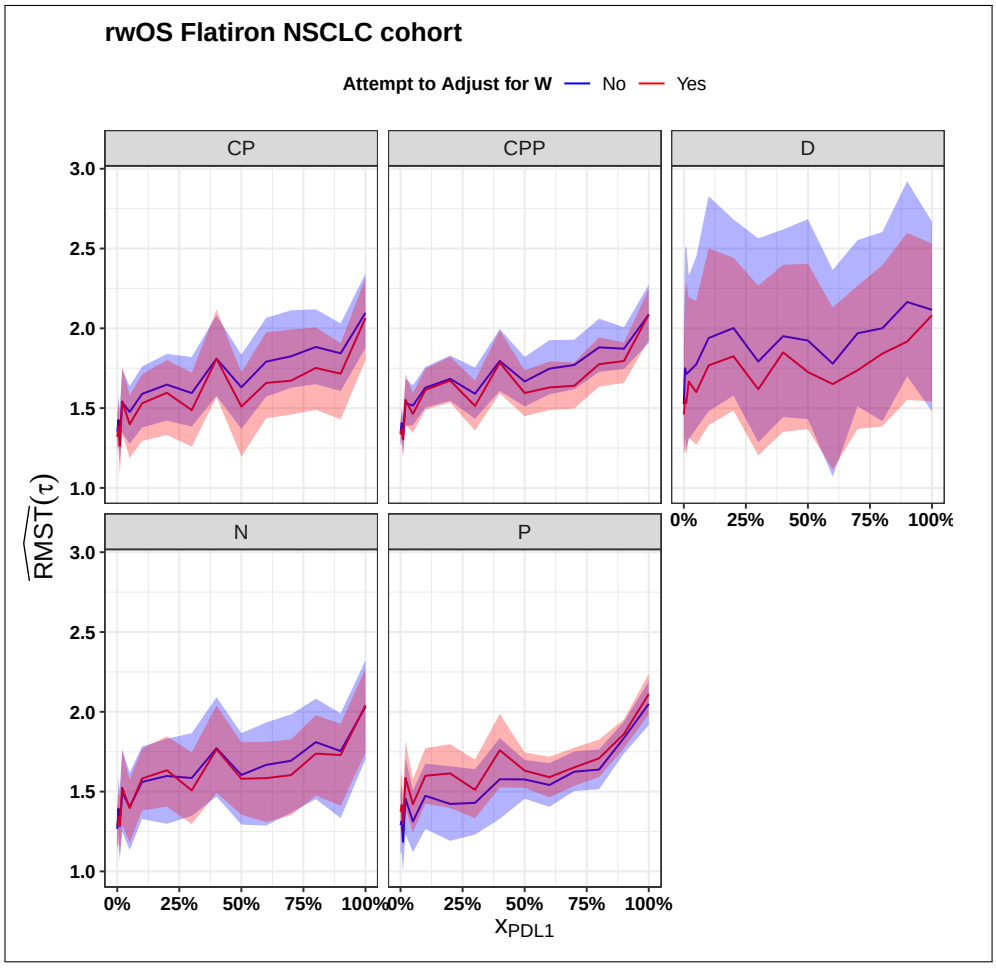


Figure 5.22: RMST(4 years) for unadjusted and adjusted model, rwOS.

Table 5.11: Comparison of fit to data in the rwTTD cohort for Weibull GP model and interaction model. ELPD: expected log-predictive distribution. LOSO: leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
GP	0. 0	0. 0	-371. 3	98. 5
Interaction	-7. 6	2. 9	-378. 8	99. 1

Table 5.12: Comparison of fit to data in the rwOS cohort for Weibull GP model and interaction model. ELPD: expected log-predictive distribution. LOSO: leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
GP	0. 0	0. 0	-7042. 9	46. 8
Interaction	-16. 6	6. 6	-7059. 5	47. 9

somewhat confounding the effect of P, adjusting for W effectively decreases the RMST(4 years) for each treatment except for P. The effect of P is increased by 0. 1 years in the range of 10% to 50% PD-L1 expression. The 95% credible interval is the broadest for D treatment because the sample size is smaller.

Model comparison Let us now review the predictive performance of the GP Weibull model and how it compares with the benchmark interaction model to estimate heterogeneous treatment effects. Table 5.11 for rwTTD and table 5.12 for rwOS suggest that the GP Weibull model has better predictive performance than the interaction model. A higher ELPD LOSO cross-validation value implies that the GP Weibull model has better out-of-sample performance and is potentially more helpful for issuing predictions. Figures 5.23 for rwTTD and figure 5.24 depict the CATE given the PD-L1 per cent staining. We can see that the GP Weibull model predicts smoother counterfactual outcomes.

The GP outperformed the interaction model (rwTTD, $\Delta\text{ELPD}_{CV} = -7.6 \pm 2.9$; rwOS, $\Delta\text{ELPD}_{CV} = -16.6 \pm 6.6$). Using the Weibull model, some estimates of the survival function under-predicted the survival estimate using the KM methods, see figure 5.29. Likely, the low number of individuals at risk at the end of the follow-up contributes less to the likelihood of the Weibull model, see equation 5.13.

Flexible parametric hazard model

Let us review the results for the flexible parametric modelling. We conduct a head-to-head comparison of predictive performance for the M-splines model and the Weibull model. Table 5.13 and Table 5.14 suggest that for both rwTTD and rwOS, a more flexible baseline hazard model, such as the M-splines, obtains

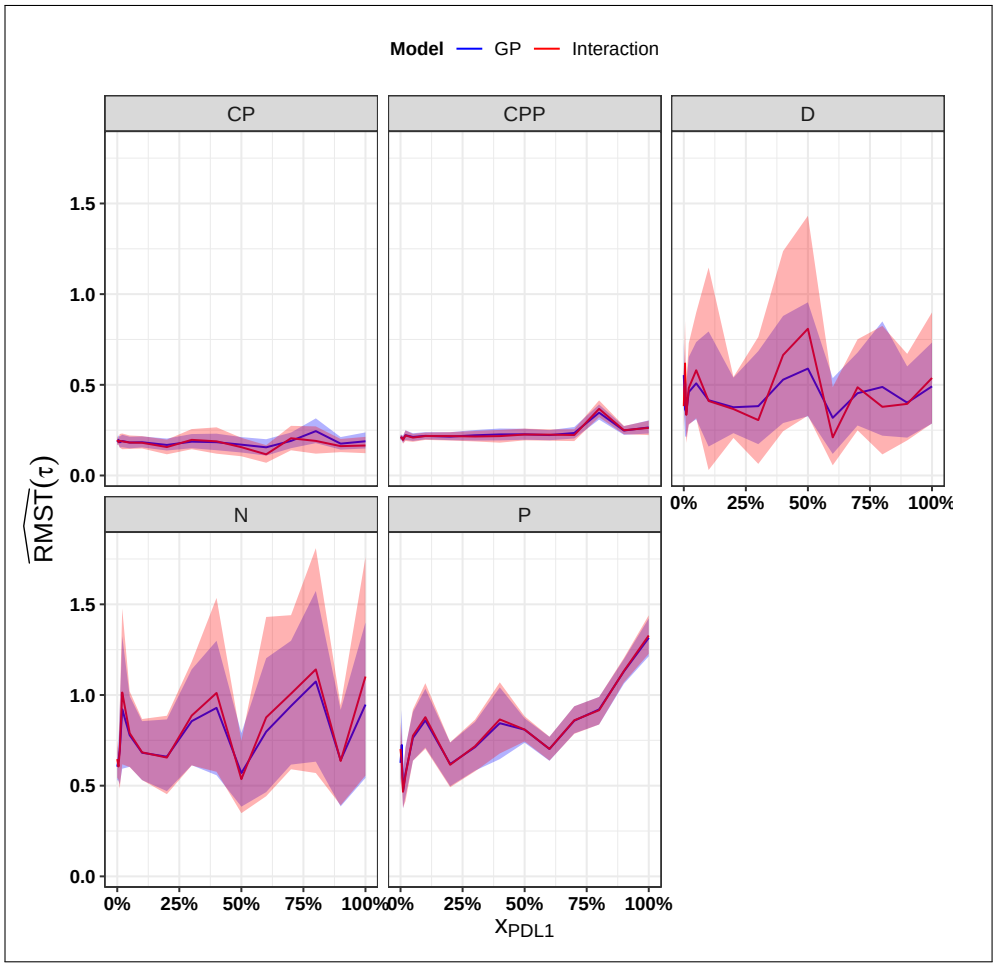


Figure 5.23: RMST(4 years) for Weibull GP and Interaction model, rwTTD.

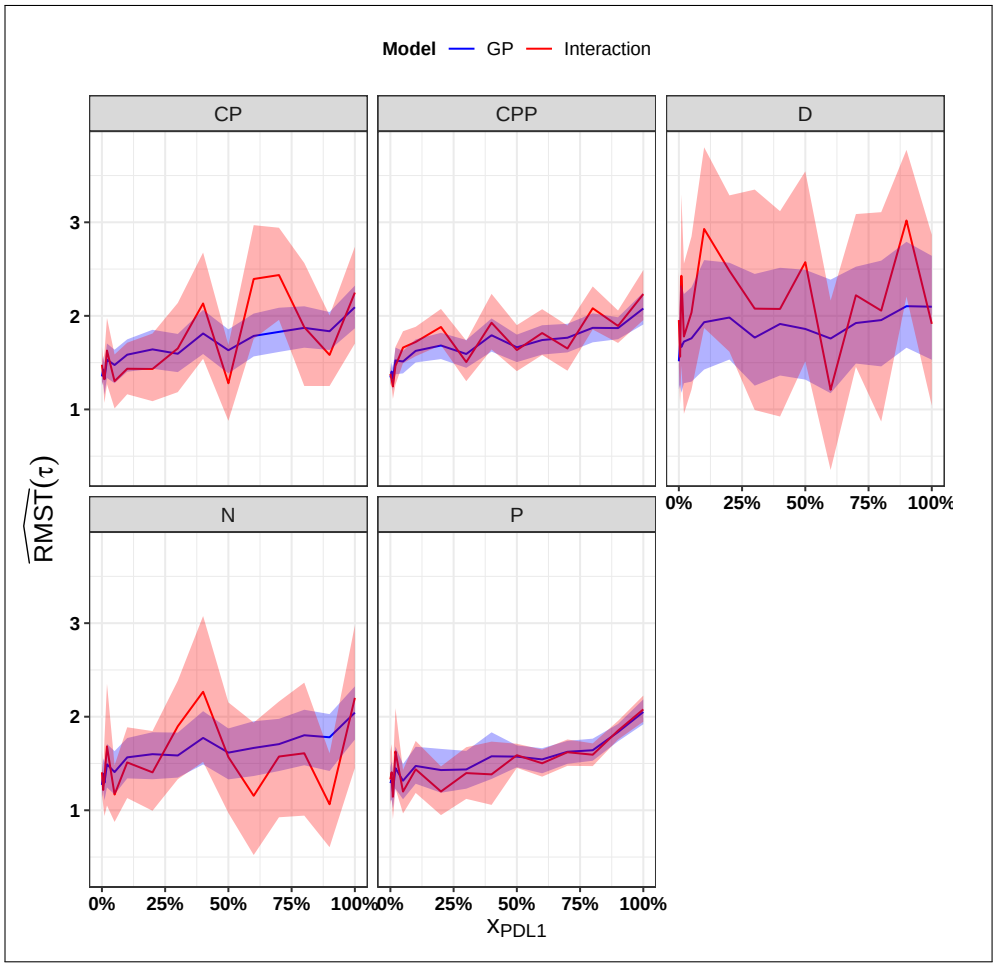


Figure 5.24: RMST(4 years) for Weibull GP and Interaction model, rwOS.

Table 5.13: Comparison of fit to data in the rwTTD cohort for canonical baseline hazard models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
MS	0. 0	0. 0	-687. 4	103. 4
Weibull	-1111. 5	30. 9	-1799. 0	94. 7

Table 5.14: Comparison of fit to data in the rwOS cohort for canonical baseline hazard models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
MS	0	0	-6887. 55	52. 30
Weibull	-225. 48	16. 4	-7113. 03	45. 85

a better fit to the dataset.

Using ELPD, we have shown that the M-splines model obtains a better fit ($\Delta\text{ELPD}_{CV} = -225.5 \pm 16$). Likely, the M-splines is more flexible and can learn more complex survival distributions. Besides, in agreement with previous research [72] the visual predictive checks (VPC) in figure 5.29 suggest, more accurate survival functions are obtained using the M-splines approach.

Post-treatment bias

As mentioned in section 2.1 in observational studies such as RWE datasets, one must be cautious about what variables are included in the survival regression model. Post-treatment bias results from conditioning on variables that are a consequence of the treatment, and are on the path to the outcome of interest. For example, cancer therapy may impact neutrophils count. At the same time, neutrophils count may be a mediator of patient outcome, see figure 5.25. It has been argued that an initial impact on neutrophils may be positively associated with treatment response [189]. Therefore, adjusting for neutrophils count measured later to the start of treatment may alter the treatment effect estimate, $\widehat{\text{RMST}}(\tau)$. We present the results for two flexible M-splines survival models in figure 5.26: the first model regresses treatment on time-to-event, the second includes the first reading of neutrophils count after treatment. We see that all the estimates of $\widehat{\text{RMST}}$(4 years) have higher uncertainty after conditioning on neutrophils post-treatment, which is likely because of the correlation between treatment and neutrophils count. In addition, we see that the double-platinum (CP) $\widehat{\text{RMST}}$(4 years) is the one that changes the most, potentially because CP impacts neutrophil count more extremely.

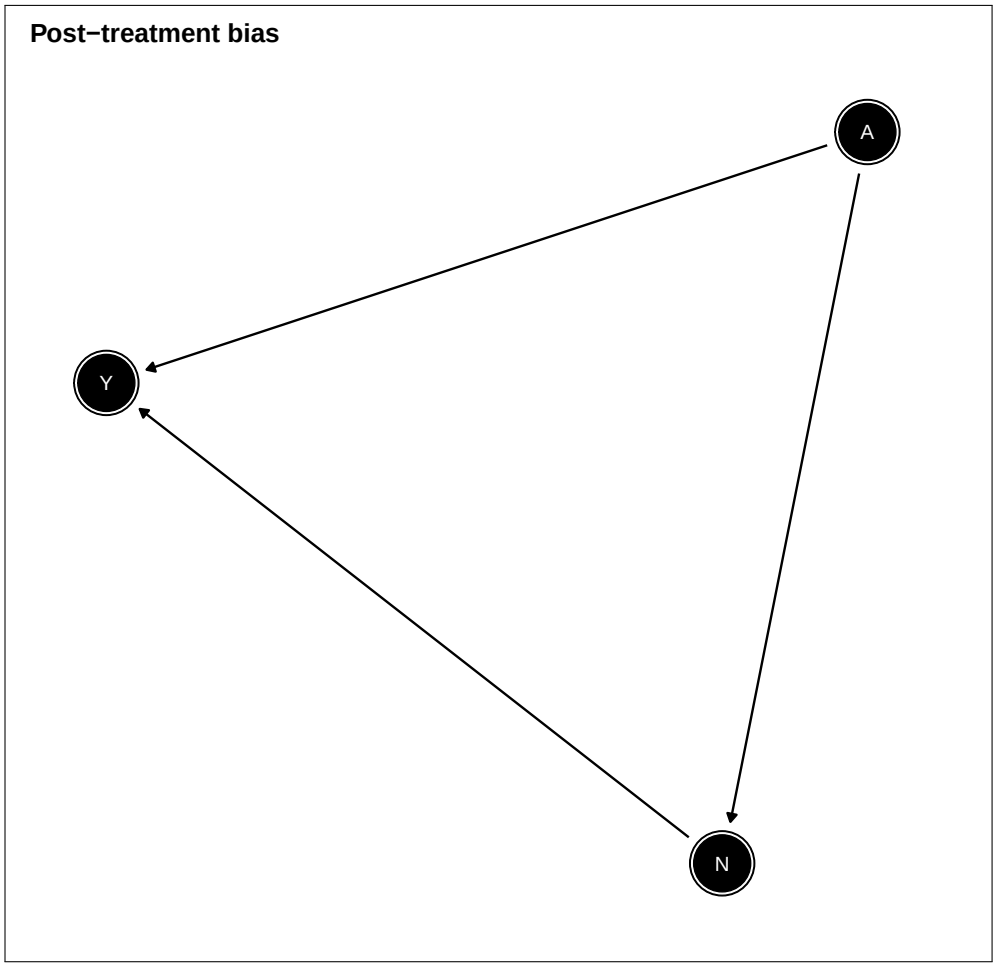


Figure 5.25: This DAG illustrates the post-treatment bias mechanism. Y, outcome; N, post-treatment Neutrophil absolute count; A, treatment variable.

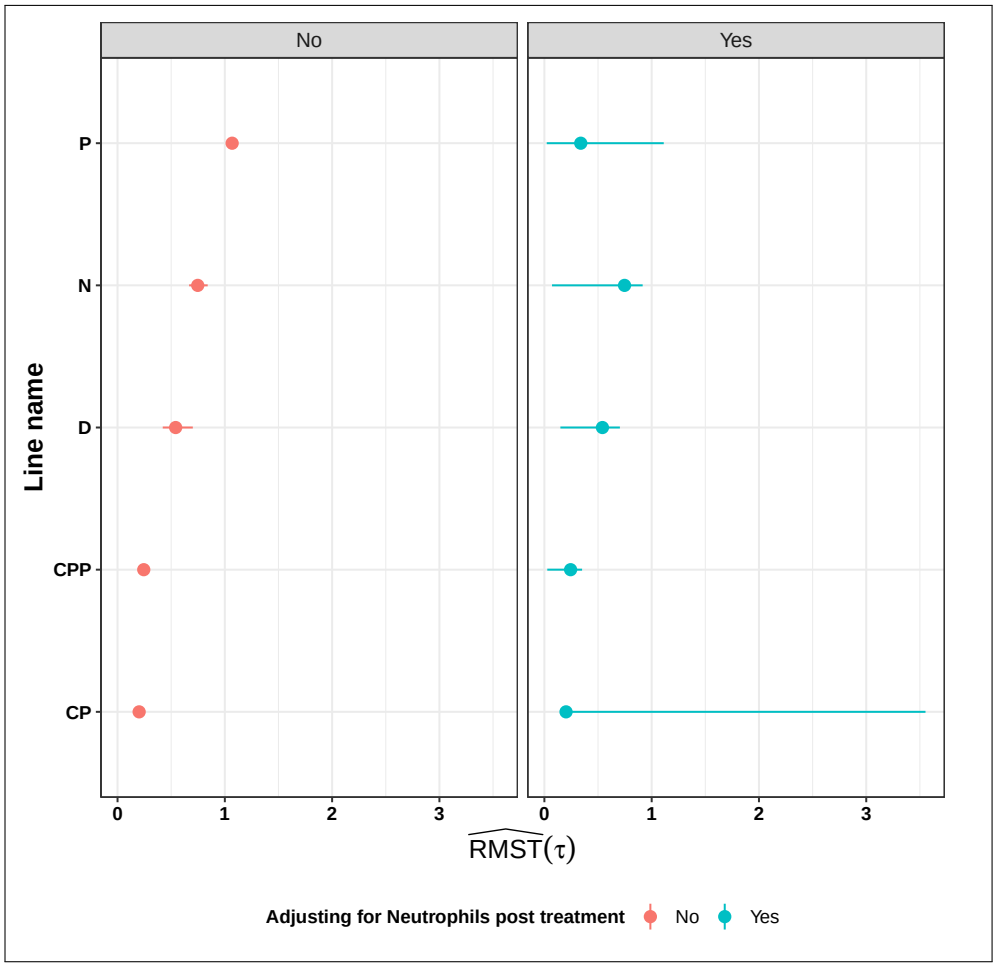


Figure 5.26: Post-treatment bias, : RMST(4 years), rwOS, for adjusting for Neutrophils count post-treatment.

Time varying effects model

Let us review now the results for the time varying effect (TVE) model. We implemented the model described in the section 5.2, using B-splines to estimate the treatment effects varying with time. Above, we have seen that TVE is superior to PH modelling in synthetic data. Let us consider now the RWE dataset and the fit to the RWE dataset. We use a flexible parametric M-splines model of the baseline hazard function and compare the PH and the TVE approach. Figure 5.27 and Figure 5.28 shows the estimate for the hazard

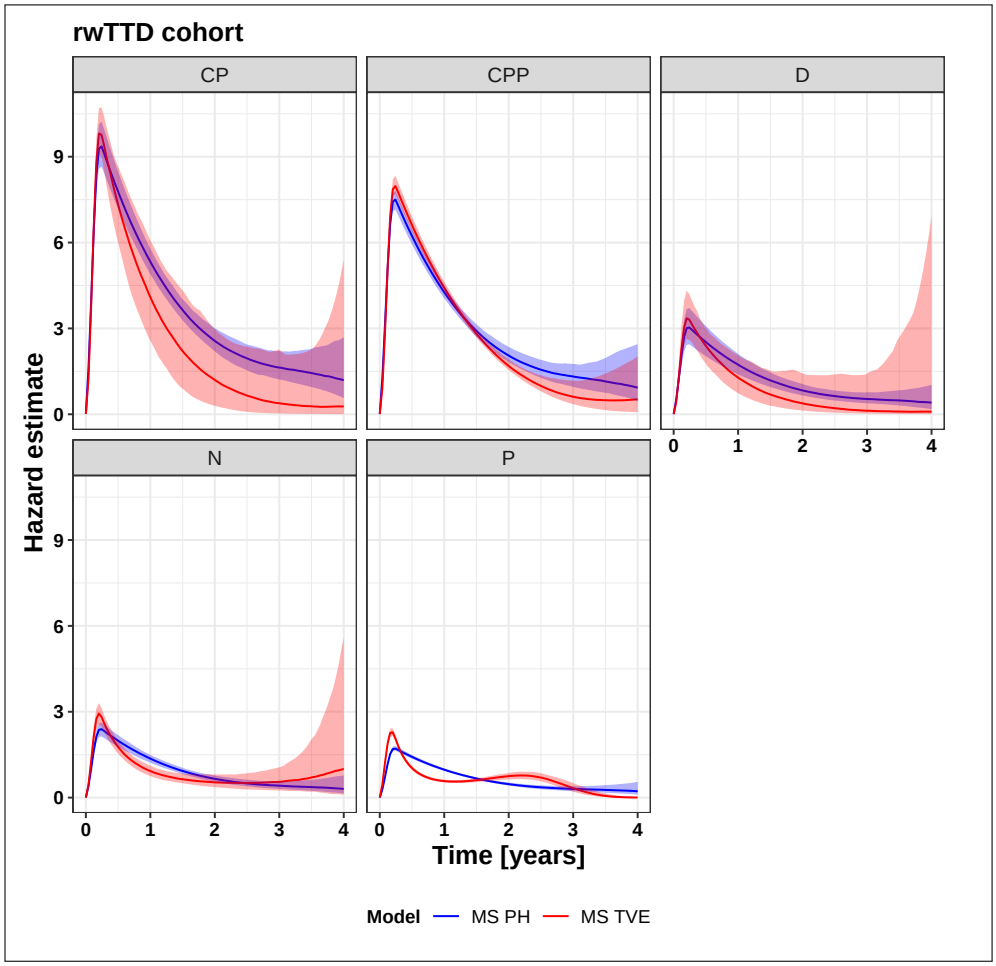


Figure 5.27: Hazard estimate, rwTTD, comparison of M-splines proportional hazard (MS PH) and M-splines time-varying effects (MS TVE).

function for the rwOS and rwTTD cohort, respectively. The results suggest that the TVE may provide a better fit, especially for P, where the TVE estimate crosses the PH estimate of the hazard.

Comparison of standardised survival and IPW estimate In the next experiment, we compare the Bayesian parametric g-formula and the non-parametric frequentist IPW estimate. The yardstick is the prediction of the

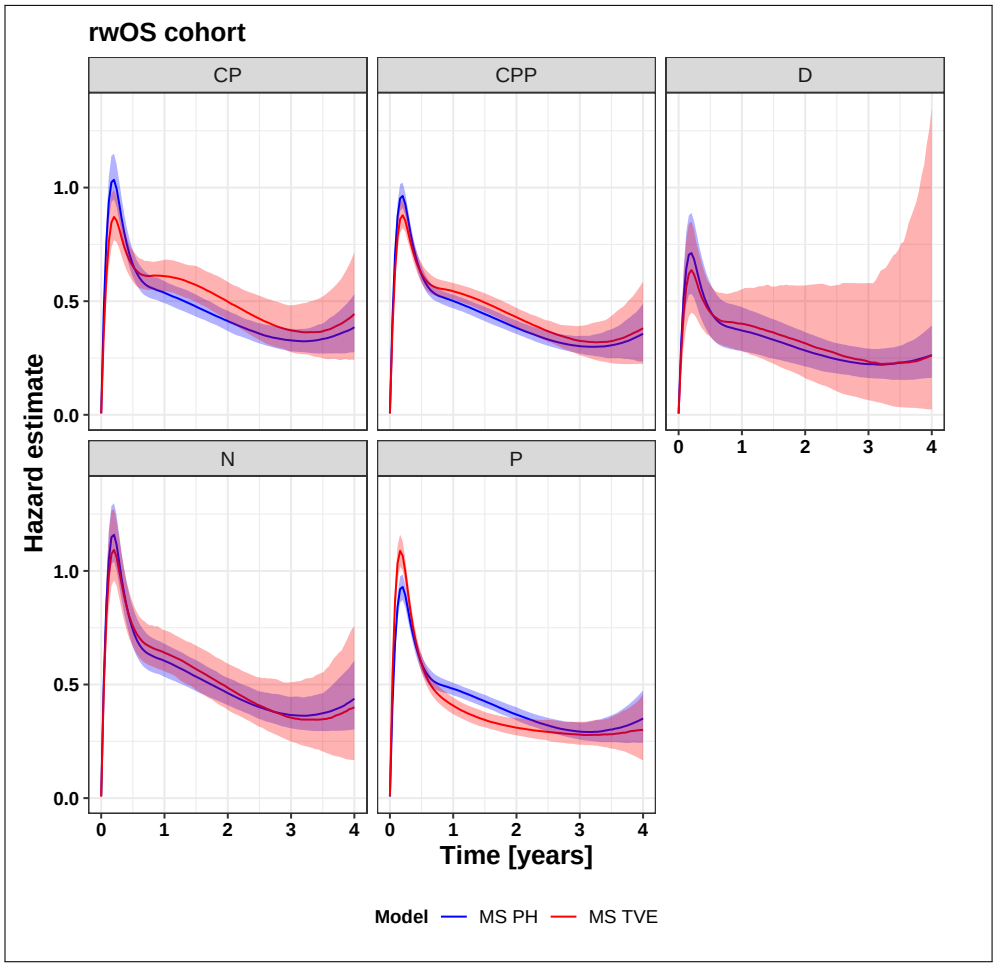


Figure 5.28: Hazard estimate, rwOS, comparison of M-splines proportional hazard (MS PH) and M-splines time-varying effects (MS TVE).

survival curve for each method. For the Bayesian parametric g-formula, we sample draws from the posterior distribution of the TVE model, integrate the survival function up to τ and summarise the results. Figure 5.29 and Figure 5.30 depicts the survival estimates for the rwTTD and rwOS cohorts, respectively. We can see that both methods are in good agreement. Tables

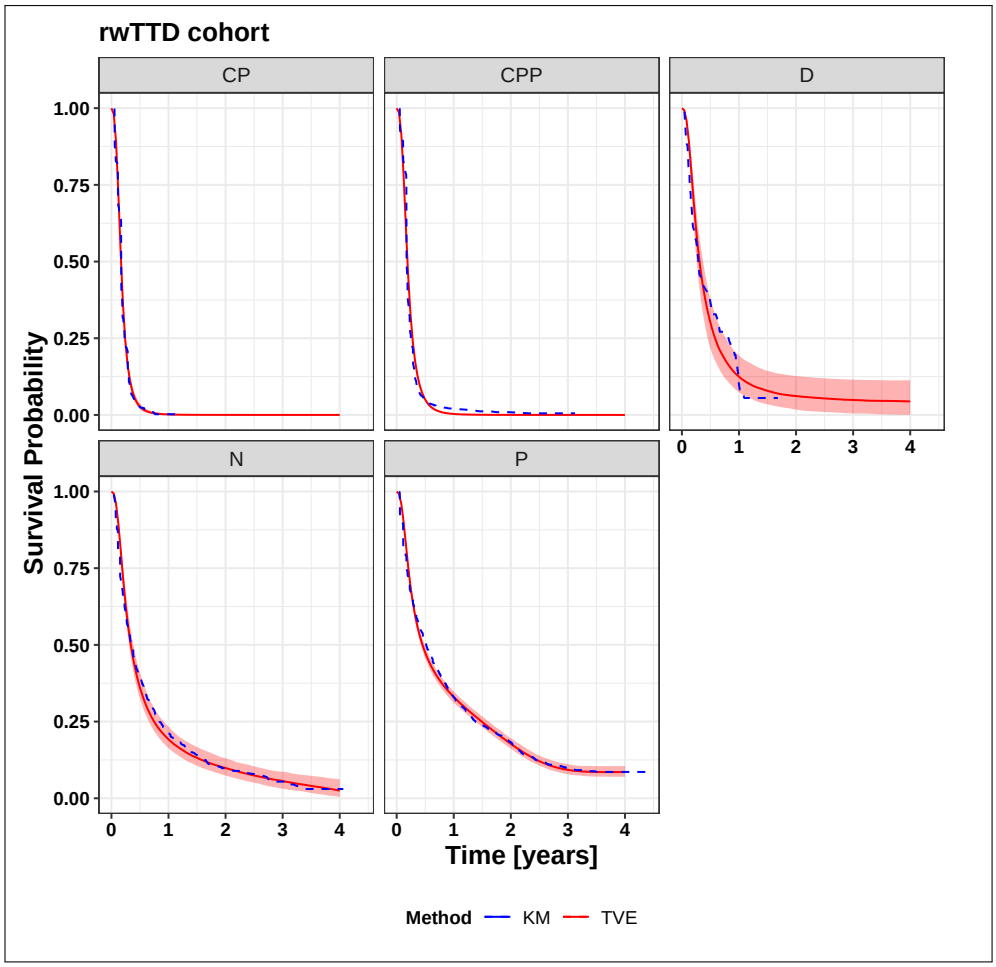


Figure 5.29: Survival estimate, rwTTD. KM : Kaplan-Meier with inverse probability weighting, TVE: Bayesian Time-varying Effect model.

5.15 and 5.16 compare the TVE model for rwTTD and rwOS, respectively. As explained in section 5.2.5, we compare a TVE approach with quadratic (degree = 2) and cubic B-splines (degree = 3). For rwTTD, the TVE approach with quadratic B-splines obtains better out-of-sample predictions. For rwOS, the time fixed effect (TFE) approach gives good predictions, suggesting that the TVE approach might overfit. Therefore, we choose the best model for rwTTD to be a TVE model with quadratic B-splines, and for rwOS, a TFE model with flexible baseline hazard via M-splines with 10 degrees of freedom.

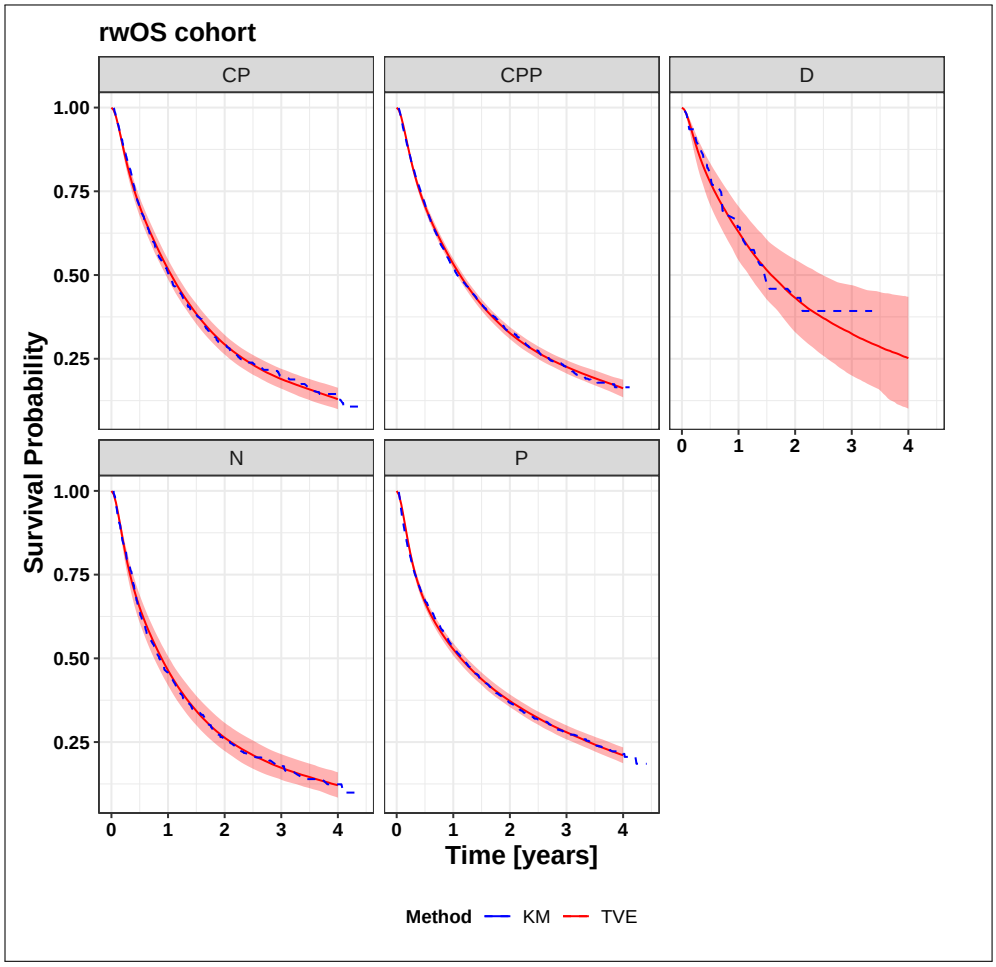


Figure 5.30: Survival estimate, rwOS. KM : Kaplan-Meier with inverse probability weighting, TVE: Bayesian Time-varying Effect model.

Table 5.15: Comparison of fit to data in the rwTTD cohort for tested time-varying effects models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
TFE	0. 0	0. 0	-279990. 4	524. 8
B-splines (degree = 2)	-656. 4	104. 5	-280646. 8	514. 5
B-splines (degree = 3)	-777. 3	98. 7	-280767. 6	516. 4

Table 5.16: Comparison of fit to data in the rwOS cohort for tested time-varying effects models. ELPD: expected log-predictive distribution. LOSO: Leave-one-subject-out cross-validation. SE: Standard error. TVE: Time-varying effect. TFE: Time fixed-effect.

Model	Δ ELPD	SE Δ	ELPD LOSO	SE ELPD LOSO
B-splines (degree = 2)	0. 0	0. 0	-269851. 7	817. 9
TFE	-64. 0	18. 8	-269915. 6	818. 0
B-splines (degree = 3)	-71. 9	11. 0	-269923. 6	818. 1

Counterfactual survival outcomes

In the next experiment, we contrast the counterfactual RMST(4 years) estimate via Bayesian parametric g-formula with the TVE model. Figure 5.31 for rwOS and figure 5.32 for rwTTD depict samples from the posterior predictive distribution standardised by treatment line. Our experiment with synthetic

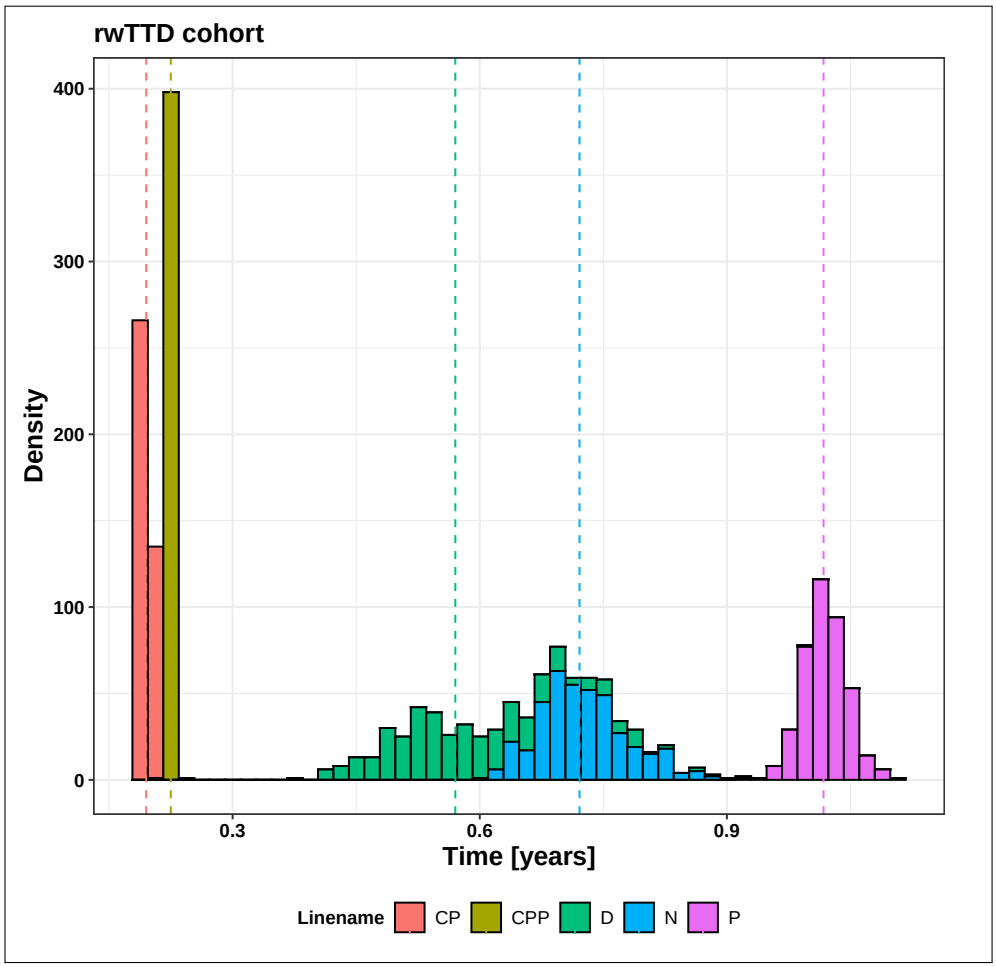


Figure 5.31: Real-world RMST(4 years) in the rwTTD cohort by treatment line.

data showed that Bayesian non-proportional hazard outperforms conventional proportional hazard models (ΔBias > 0.1). The observed reduction in bias makes the Bayesian g-formula attractive to the analysis of observational data from both Bayesian and frequentist perspectives. For the RWE dataset, the truth is not available; hence, one must use heuristics to justify the modelling choices that suggest using the non-proportional hazard approach with B-splines.

OTS Bounds on the ATE

In the next experiment, we compute the Bayesian ATE using the TVE model. We define the ATE as the expected difference in RMST(4 years),

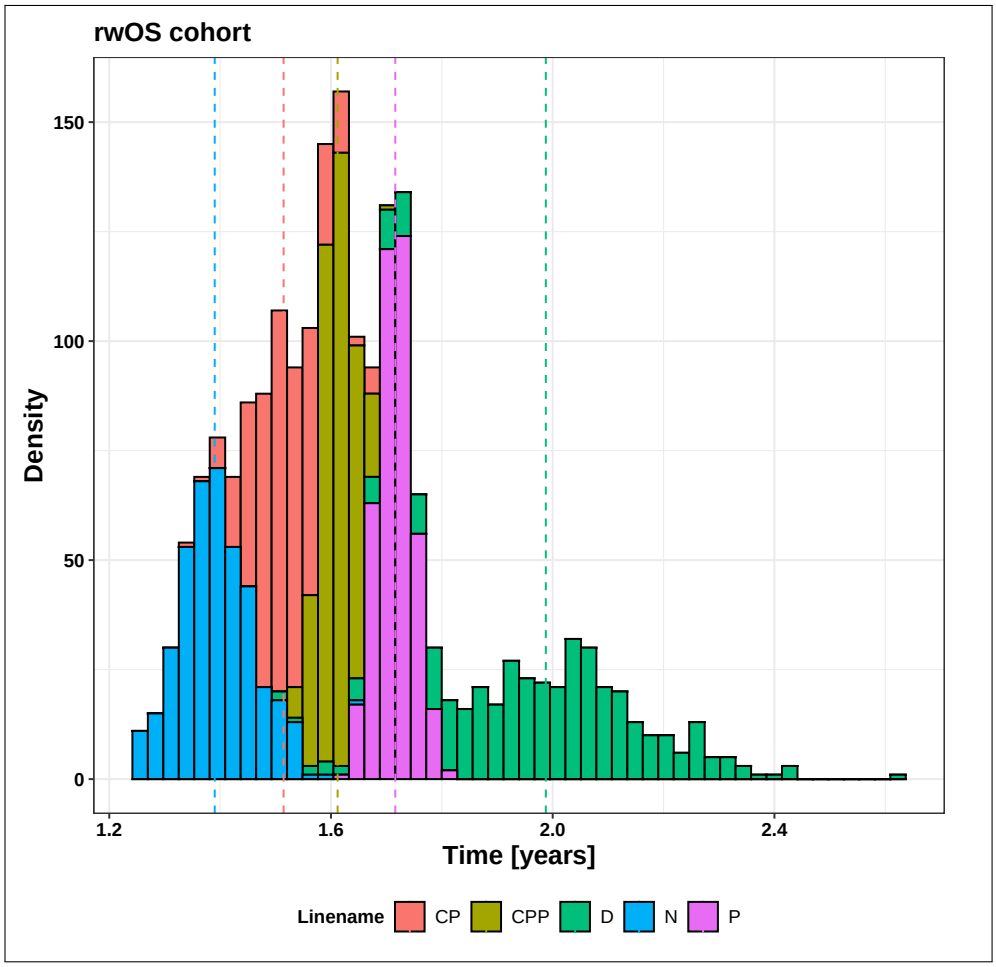


Figure 5.32: Real-world RMST(4 years) in the rwOS cohort by treatment line.

ΔRMST(4 years) from the combination CPP treatment line. Note that any posterior density for ATE that is greater than zero suggests that the treatment positively impacts the patient outcome because it increases the RMST(4 years) considering CPP. On the other hand, if the ΔRMST(4 years) posterior distribution overlays the zero dashed lines, we can not identify the ATE sign from the RWE dataset, and it implies that the RMST(4 years) is similar to the CPP.

Finally, we weaken the assumptions of our modelling by evaluating the OTS bounds. Doubling down on our Bayesian approach, we can sample draws from the ATE posterior and each OTS bounds, obtaining a credible interval for the ATE and the OTS bounds. Figure 5.33 for rwTTD and figure 5.34 for rwOS visualises the results and shows that under exchageability violations and OTS assumption, it is unlikely we could identify the sign of the effect for any of the treatments, except the positive P monotherapy impact on rwTTD and perhaps the negative N monotherapy impact on rwOS.

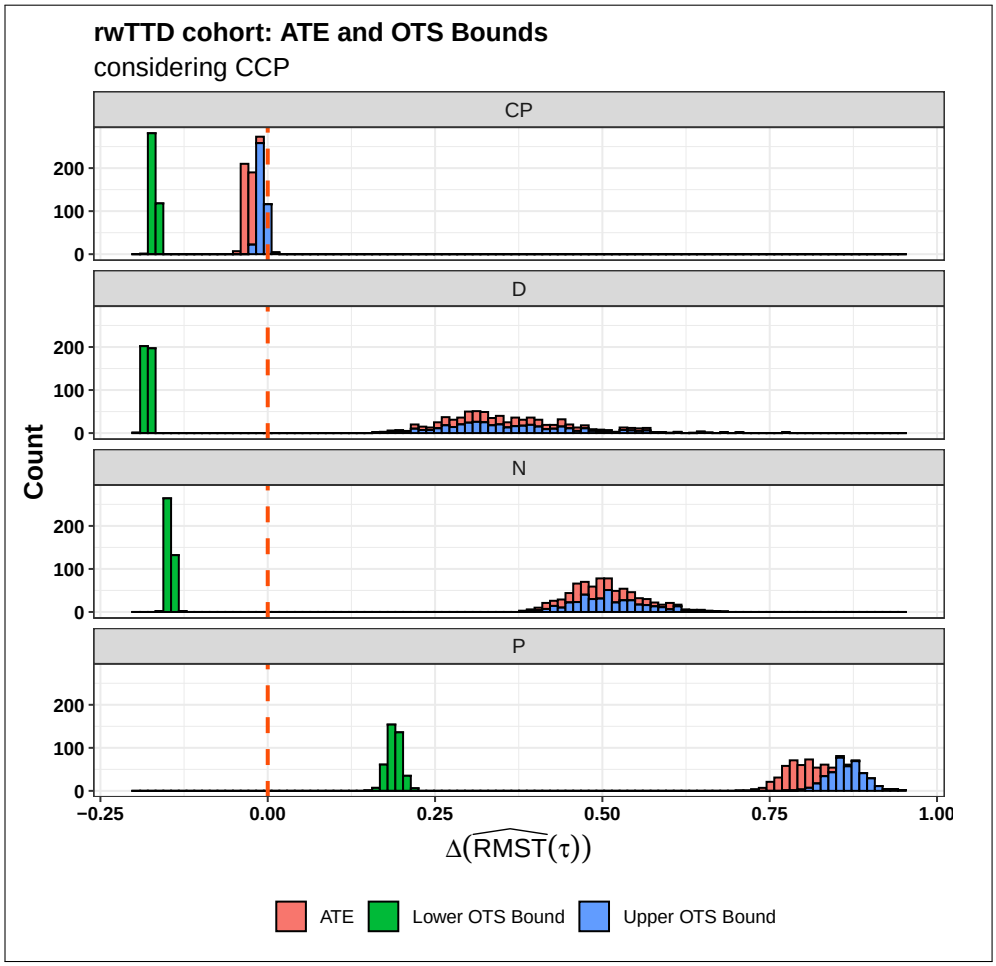


Figure 5.33: Real-world ATE with considering CPP treatment line and OTS Bounds in the rwTTD cohort. Dashed line highlighting the null effect.

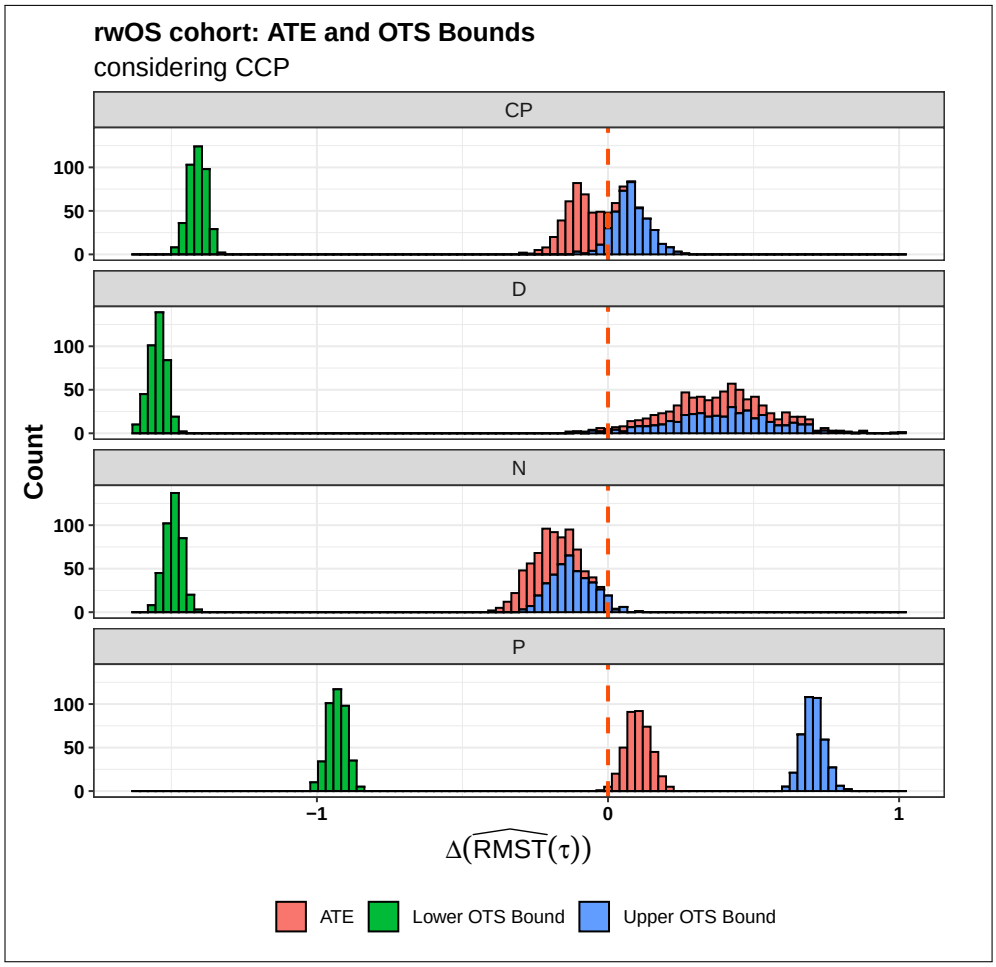


Figure 5.34: Real-world ATE with considering CPP treatment line and OTS Bounds in the rwOS cohort. Dashed line highlighting the null effect.

5.4 Discussion

In this chapter, we proposed a Bayesian framework for causal survival analysis, and applied the Bayesian survival modelling approach to a unique RWE dataset comprising a large retrospective cohort of more than 10,000 US Flatiron advNSCLC patients in ICI or chemotherapy treatment. Intending to estimate treatment effects heterogeneity, we model the varying treatment effect of immunotherapy with PD-L1 expression using a GP. All information regarding the covariance in PD-L1 expression was included to examine the helpfulness of the survival GP approach. The aim was to develop a general model for future analyses of RWD to predict treatment effects in distributed biomarker-defined populations.

GP regression has much application by modelling the covariance in a high-dimension distribution. Interactions can be helpful to understand treatment modifications by binary biomarkers. However, for biomarkers that are measured continuously as PD-L1, we demonstrated that GP is accurate and may be helpful to advance the concept of personalised therapies. For example, research by [190] suggests that for tumour mutational burden (TMB), the average percentage of mutations in a sample of cancerous cells may predict treatment response in ICI therapy. The clinical advNSCLC dataset analysed did not include TMB. However, new RWE datasets named clinical-genomic datasets comprised several genomic markers, including TMB. TMB, as defined above, is generated as a percentage or a distributed variable, hence, applying the survival GP approach presented in this work could be helpful for treatment effects interpretations.

Several RCT have demonstrated the efficacy and safety of ICI and immuno-chemotherapy. Table 5.1 summarised the recent pivotal trials for IO in advN-SCLC. Starting with the initial approval of P in 2016 based on the KEYNOTE-24 trial with a median progression-free survival (PFS) of 10.3 months for P and six months for chemotherapy [166]. Similarly, KEYNOTE-189 demonstrated the efficacy of triplet immuno-chemotherapy CPP with overall survival of 20 months for the intent to treat population [191]. Similar to our RWE study, [191] found that ICI increases the survival across all the expression levels of PD-L1, [166] demonstrating the utility of ICI in untreated locally advanced or metastatic advNSCLC patients, including SCC histology, see table 5.6.

The results from the figures 5.29 and 5.30 combined with the biomarker status summary in the table 5.5 suggest that ICI therapy is effective when regarding as confounders the advNSCLC biomarkers: ALK, EGFR, KRAS, and BRAF. Using the section 3.2 positivity condition, the remark about considering biomarker status a confounder is valid. However, from a clinical point of view, with targeted therapies for ALK, EGFR, and BRAF, it is usual to start with

targeted therapies [93]. In our RWE study, the most balanced biomarker was the KRAS status, which at the time of writing has no approved targeted therapy, as seen in table 5.5.

RCT datasets [168] have previously suggested a non-proportional hazard effect for ICI, i.e. the survival curves cross at around six months. Our results in Table 5.16 support a non-proportional hazards model for the rwOS cohort in the present study, $\Delta(ELPD) = -64.0 \pm 18.8$. Moreover, our work highlights the benefit of using the Bayesian non-proportional approach for modelling the non-proportional time-varying effect flexibly with B-splines, supported by the results in table 5.16. In addition, the present study demonstrated the use of RMST to summarise the ATE in comparing treatment effects. For example, the figure 5.31 shows the ATE as the difference in RMST between each treatment and the combination of CPP, suggesting that there are significant differences if the exchangeability condition of the section 3.2 holds between treatment groups. Note that any distribution for ATE greater than 0 indicates that the treatment positively impacts the patient outcome because it increases the RMST.

Finally, the present study is completed by evaluating the OTS bounds and weakening the exchangeability assumption. We highlight the benefits of using Bayesian survival analysis to estimate the posterior distribution for the causal OTS bounds. We, therefore, can sample draws from the ATE posterior and each of the OTS bounds, obtaining a credible interval for the ATE and the OTS bounds. Figure 5.33 visualises the results and shows that under exchageability violations and OTS assumption, it is unlikely we could identify the sign of the effect for any of the treatments in the rwOS cohort, and only the positive P monotherapy impact on rwTTD. Figure 5.35 explores a possible explanatory reason for the difference observed between rwTTD and rwOS cohorts. In short, a proportion of patients that discontinue a treatment will start the second line of therapy with IO, which may increase their overall survival.

Most research on survival analysis such as the development of RMST [192] has been carried out in the context of randomised clinical trials. However, we have argued the benefits of using RMST in RWD, such as interpretability and benchmarking. In particular, RMST is a helpful measure of treatment effects for assessing the assumptions of exchangeability. OTS bounds with RMST have analytical appeal because it allows for more credibility and interpretability in the results. To this end, we developed an OTS bound analysis for the ATE in Bayesian survival analysis using RMST. We demonstrated that Bayesian survival modelling makes causal bounds uncertainty estimation possible, suggesting multiple future directions in analysing RWD advancing the concept of personalised medicines towards uncertainty quantification and policy optimisation.

A limitation of the present study is that it does not model the subsequent lines of treatment in the rwOS cohort. Future research may investigate the application of the GP survival approach, Bayesian g-formula and OTS bounds considering the switch on treatment lines through the patient follow-up. Figure 5.35 explores the scenario where the patient switch from the first line to the second line. A new time-dependent model may summarise that scenario by helping in informing a mechanistic model of the impact of PD-L1 expression on rwOS.

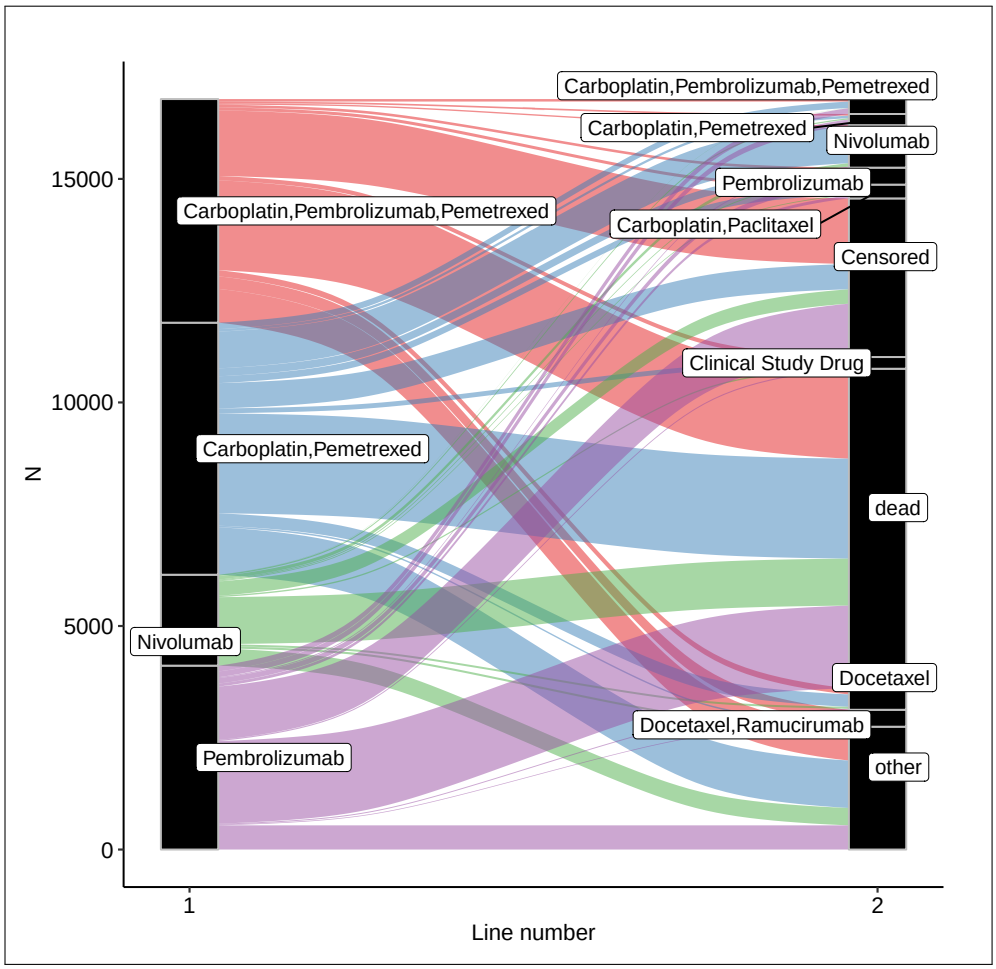


Figure 5.35: The alluvial plot depicts the dynamic treatment strategies from the first line to the second line in the present RWE study. Censored: remains in the first line. Dead: dead event occurred.

5.5 Conclusions

This chapter argued that causal models of survival analysis are necessary to interpret RWE studies. Besides, it demonstrated Bayesian techniques for survival analysis that are bespoke for analysing RWD where the frequentist framework of null hypothesis testing is not standard. It introduced a counterfactual

approach to estimate treatment effects in the presence of non-proportional hazards under the Bayesian survival outcome modelling framework, developed an OTS Bound for non-proportional hazards using RMST.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

This thesis has demonstrated new methods for RWE generation and RWD interpretations. We have defined the causal framework for drawing inferences about treatment effects from observational, non-randomised clinical RWD by using characterisation by example to introduce the application of RWE in clinical oncology. Our primary aims stated in the section 1.2 were to address characteristics of RWD analysis, including:

1. The systematic treatment of missing data.
2. The prediction of treatment effects for biomarker-defined populations.
3. The explanation of treatment discontinuation in the presence of unrecognised factors.

These research aims have been achieved by conducting a comprehensive review of the basic principles of causal inference and RWE study design, providing specific examples of RWD analyses with missing data and applying GP regression methods for estimating heterogeneous treatment effects. Figure 6.1 depicts a summary of our overall aims, results and achievements.

The use of large datasets from RWE studies offers an opportunity for oncology researchers. However, we have identified challenges in analysing clinical data from RWE studies: missing data, heterogeneous treatment effects and unobserved confounding. We proposed:

1. MITABNET, a new machine-learning algorithm to perform MI, and a new method to compare imputation algorithms in RWE studies.
2. Weibull GP for modelling heterogeneous treatment effects in survival data.
3. OTS bounds to assess unobserved confounding.

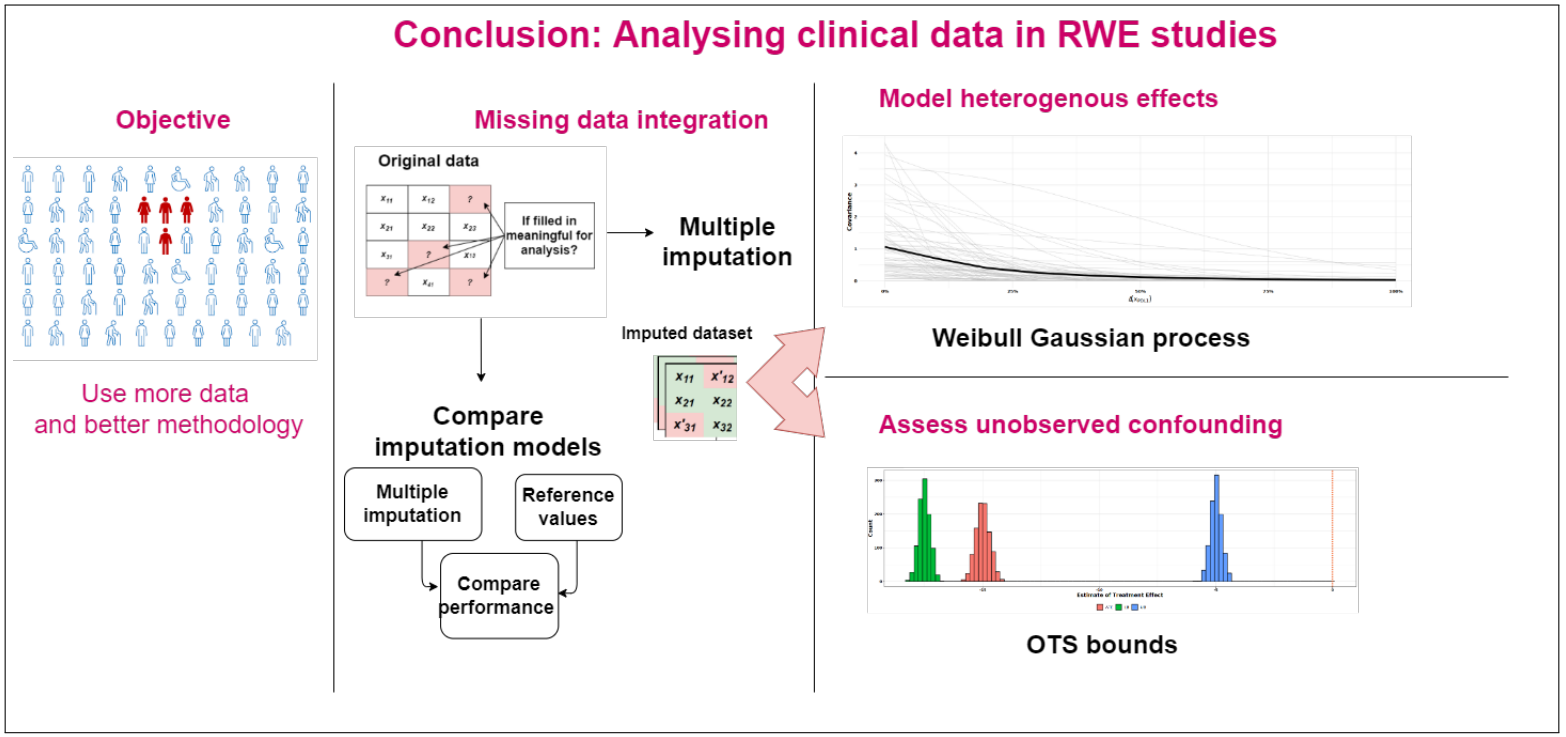


Figure 6.1: Summary of the overall results and conclusions: Only a small percentage of cancer patients are enrolled in clinical trials, we proposed using clinical RWE studies to use larger datasets. However, we need accurate methods for handling missing data, such as multiple imputations, and evaluate the imputations methods. Besides, the GP survival model allows us to interpret the heterogeneous treatment effects advancing the concept of individualised treatment effects. Finally, we must consider unobserved confounding, i.e. no exchangeability, quantifiable via OTS bounds.

The rest of this chapter summarises the core conclusions that can be drawn from the work presented in this thesis and provides concluding remarks on the modelling/data analysis experiments conducted. The central section discusses the applications to RWD analysis examining the methods and results used. Finally, we give suggestions for future research.

RWE studies can be critical in drug development with rich insights provided from a more inclusive patient population. RWE presents an opportunity to bridge the gap between research and clinical practice, specifically, by sourcing new biomarkers (for example, clinical-genomics RWE datasets), stratifying early patient populations (for example, biomarker-based treatment modification), and identifying unmet medical needs (for example, rare cancerous diseases). To reason about these clinical questions, we need to be able to construct unobserved counterfactual outcomes. For example:

- What would have been the treatment response for a sub-population with a different biomarker expression?
- What if the population had been treated with a combination of targeted therapy and chemotherapy?

To tackle these questions, we have aligned this thesis with the causal inference framework, which proposes that one can simulate the counterfactual outcomes to predict the impact of treatment interventions or explain the reason for past treatment failure. In chapter 3, we defined a causal estimand, i.e the treatment effect (section 3.1), specifying our causal assumptions via a causal model (section 3.2), then turn it into a statistical estimand which we can use to estimate the parameter of interest with data (section 3.3 and 3.4). Consequently, we can simulate the counterfactual outcomes using the generative model, thereby predicting the impact of treatment on the population of interest.

Section 2.2 showed that one could use causal lenses to understand the *missing data* problem in RWD applications and decide the best procedure in order to make an unbiased inference. Moreover, under the exchangeability (non-unobserved confounding) assumption, we can use off-the-shelf machine learning algorithms to perform multiple imputations, the central topic of chapter 4. For censored *survival data*, chapter 5 expanded the use of the RMST in RWD, which is more suitable and interpretable than the conventional hazard ratio, see section 5.1.3. Furthermore, one can use RMST effectively to communicate the results of multi-level models for heterogeneous treatment effects, such as survival GP models, see section 5.2.3. We have demonstrated the applicability of our new methods with synthetic data experiments and several real-world experiments, see section 5.3.

The comparative study of missing data imputations in chapter 4 starts with the assumption of MAR, i.e., no unobserved variable impacts the missingness

mechanisms. Using the RWE Flatiron NSCLC dataset, including more than 35,000 subjects, the imputation performance of six state-of-the-art algorithms to perform multiple imputations was compared, see section 4.3. The battery of algorithms compared is not exhaustive. Instead, the focus is on the methodology to benchmark imputation algorithms in RWE datasets with missing data. To our knowledge, this work is the first to study the concept of the data distribution shift induced by the imputation algorithm and the impact of missing data on the performance of the imputation algorithms, see algorithm 3.

Additionally, we have expanded the TABNET approach [19] for developing MITABNET that automatically handles non-linearities and feature selection in the imputation model, see section 4.2.3. Assuming MAR, however, has limitations. There is a connection here with the final chapter: the credibility of an inference decreases with the strength of the assumptions maintained, see section 5.2.8. MAR, though weaker than MCAR, is a strong assumption. Critically, for handling missing values in RWD, one may need to consider unmeasured variables. In this case, one may use past medical history to make informed assumptions about the possible values and conduct an appropriate form of sensitivity analysis.

Moreover, one can use full maximum likelihood or Bayesian imputation to set the probability statement about the missingness mechanism. However, for handling the missingness of the binary biomarker status of ALK, EGFR, KRAS, BRAF and PD-L1, the combinatorial nature of the setting explodes to $2^5 = 32$ possible imputation models. The appeal of machine learning techniques for multiple imputations is to handle the missing data problem as a pre-processing step. In this context MITABNET appears to outperform state-of-the-art methods in complex synthetic datasets (percentage bias = $2.6 \pm 1.4\%$), see table 4.5.

However, for the Flatiron NSCLC dataset, using our new benchmark method for imputations in RWD, MITABNET was met by state-of-the-art imputation algorithms such as PMM, EM and MIRF. The results suggest that for imputing binary biomarkers in the Flatiron NSCLC dataset, one can use PMM with limited loss in imputation accuracy (percentage bias = 10.8%, coverage = 90%), see table 4.6 and table 4.7. On the other hand, for more complex datasets with indirect associations among the covariates, a more elaborate imputation algorithm such as MITABNET may perform more accurate imputations, and hence would be preferred for sharper inference.

Moving on to the real-world case study of ICI in a cohort of 5,000 NSCLC patients treated with ICI, see section 5.1.2 for the study design. The study aimed to understand the impact of PD-L1 expression on treatment response. PD-L1 per cent staining is a proxy for the PD-L1 expression in the tumour cells that patients with similar tumours have. Moreover, although there is

no obvious cut-point in continuous variables such as PD-L1 expression, close PD-L1 per cent staining values may potentially share interactions with ICI treatment because they promote an immune response against cancer [161]. A survival GP to model the covariance in varying treatment effects with PD-L1 was presented, see section 5.2.3. Essentially the survival GP model predicted all the interactions between treatment and PD-L1 per cent staining, setting a prior joint distribution for all of the model parameters. For the survival function, we assumed a Weibull model since the Weibull distribution assumptions are uncomplicated and have proven to perform well in practice [67]. It was shown that the Weibull GP outperformed the conventional survival interaction model in identifying the interaction effects between PD-L1 and ICI treatment with better prediction of treatment response in cross-validation stratified by PD-L1 per cent staining ($\Delta\text{ELPD}_{CV} = -7.6 \pm 2.9$), see table 5.11.

It would be straightforward to develop a survival GP model with another baseline hazard distribution, including flexible parametric distribution with M-splines. Modifying the link function modelling with the GP would involve a similar approach with the new baseline hazard model. On the other hand, modelling the time-varying effects of the survival GP model would be difficult, and it is likely that there is not enough information to learn more elaborate distributions than the treatment varying effects model. Subsequently, the real-world experiments considered time-varying effects on the interaction of treatment and biomarker status. These did not outperform or significantly change the predictions of treatment interventions ($\Delta\text{ELPD} = -1.0 \pm 2.5$).

Our real-world experiment with the Flatiron NSCLC ICI dataset also considered several definitions of accuracy and utility for predictive and causal survival analysis. Our results agree with previous research [40] that the probability of the survival data best describe the definition of accuracy in survival predictive modelling. On the other hand, we have shown that this definition of accuracy does not satisfy the causal survival analysis approach because the counterfactual outcomes are not immediately available from the data, see section 5.2.7. The comparison of causal models, in general, is a topic of active research and does not yet have a definite established answer. From proposals of weighting a cross-validation scheme with the propensity score [81] to using measures of cumulative elasticity in the literature [193] aspects of this debate are still open. For heterogeneous treatment effects models, we have proposed a stratified cross-validation scheme where one stratifies by treatment and biomarker stratum while performing splits of the data. The ingenuity of the approach is that under the no unmeasured confounding assumption and overlap (positivity), the utility function evaluates how well the model predicts treatment interventions, stratifying by a variable of interest, such as a biomarker. Therefore, this approach can benchmark causal models in their

utility for personalised treatment decisions and biomarker discovery. We have demonstrated such an approach in comparing the survival GP model with the survival interaction model in stratifying the Flatiron NSCLC ICI dataset by PD-L1 expression.

Following this, we compared the proportional hazard ratio and the RMST to summarise treatment effects in the Flatiron NSCLC ICI survival dataset. Our results show that proportional hazard ratios' limitations are most acute when the proportional hazard assumption is not sound. For instance, if treatment acts as a selection force for non-susceptible patients, the time-varying nature of the hazard function is relevant in modelling the treatment effect. With synthetic data experiments, we have shown that the non-proportional hazard approach is less biased and preferred for sharper inference (bias < 0.01), see figure 5.14. To model non-proportional hazards, we adopted a B-splines modelling of the time-varying effects [72].

Moreover, we have further shown that an ODE system can describe the non-linear hazard model. The B-splines approach fits the data well for the analyses presented here. However, the ODE approach may be helpful for more complex survival datasets with various end-points forming a Markov process potentially providing greater mechanistic insight [6].

Next, our experiments have examined the differences between standardisation and IPW in predicting the survival function for several ICI treatments. We have seen that under exchangeability, the two estimators yield similar inferences ($RMST(\tau) = 0.9$), see figure 5.29. However, with standardisation and the application of the Bayesian g-formula in survival analysis, our results have shown that one can quantify the uncertainty in the survival model predictions. Moreover, the posterior distribution can be updated when more data arrive in the RWE database. Finally, but no less importantly, we have shown how to apply the OTS bounds in survival analysis using the RMST and computing the posterior predictive distribution, see figure 5.33 and figure 5.34. The results suggest that class unbalance tends to move the treatment effect estimate near the upper OTS bound, which is likely because treatments considered optimal by prescribing doctors are most frequent in the RWE dataset. These causal bounds are a particular case of sensitivity analysis to unobserved confounders. We suggest cooperating with clinicians to assess all confounders in the RWE dataset and conducting sensitivity analysis based on their recommendations.

In conclusion, RWD complements RCT by bridging the research and clinical practice gap with large datasets measuring many biomarkers. However, to go beyond current clinical practice and recognise the heterogeneity in treatment response, RWD analyses need not focus on predictive questions but causal ones. Since missing data are typical in RWE datasets, we need to consider model-based or algorithms to impute missing values. We have studied six different

methods that perform multiple imputations to help potential research on RWE datasets in choosing an imputation method. Given its longitudinal nature, RWD are often used for various types of time-to-event analyses. We have applied a causal framework to survival analysis and shown the limitations of inferring treatment effects from observational RWD. Since RWD is not controlled in contrast to RCT, the model's bias in treatment effect estimates is expected to increase. When comparing the results of the synthetic RCT and RWD experiments, it was shown that adjusting for confounders and time-varying effects was sufficient in order to obtain unbiased estimates. However, given these limitations, RWE studies predictions need to be carefully interpreted, possibly, using small sample RCT datasets when available or assessing the models performance in predicting treatment interventions in biomarker-defined populations.

6.2 Future Work

The adoption of RWE has grown dramatically in the last decade in particular for the evaluation of drug safety and effectiveness [143], thereby modernising the approach to clinical research in oncology. However, more studies are needed to understand how can we better leverage RWE to support drug development. There are several open directions in analysing RWD for clinical research, the methods we described in the present thesis, and extending these to new settings. The following section encompasses several extensions to this work. First, covering limitations and alternative approaches to the RWE studies conducted in this thesis. Second, the development of a causal inference methodology. Third, exploration of different areas of application for RWE in clinical research in oncology.

6.2.1 Extensions of the RWE studies presented

Expanding the RWD study of missing data: In chapter 4 we performed a comparative study of missing data, where we attempted to perform multiple imputations using the survival outcome and the binary status ("positive" or "negative") of five commonly tested biomarkers in NSCLC: ALK, EGFR, KRAS, BRAF and PD-L1. However, it would be interesting to see how the imputation algorithms perform when adding more features to the imputation model, including demographic variables that might share information with biomarkers such as smoking history, history of malignancies or histology.

Analysing the impact of longitudinal variables on imputation algorithms: Another possible extension of the algorithms for imputation of

missing data is to consider variables measured longitudinally, such as neutrophils, lymphocytes and albumin, which also have significant missing values and are sparse in RWD. Moreover, longitudinal observations, such as haematology data, that vary over time may share information with the treatment interventions that occur across time. In particular, past observations may impact current treatment interventions, which may impact future observations. To this end, the model could, in principle, impute missing values while being used to predict dynamic treatment interventions.

Tackling the MNAR scenario: We demonstrated the application of tests of MCAR applicable to any clinical dataset, which suggested that the MCAR assumption is untenable in our examples of RWE studies in section 4.3.2. The MAR framework is, therefore, more appropriate for the RWE dataset analysed under the assumption of exchangeability of the missingness mechanism, as explained in section 2.2.1. However, as studied in section 3.5 the exchangeability assumption may be relatively extreme, and an MNAR missing data pattern may be unknown in practice, and results should be generalised with caution. The algorithms presented for multiple imputations, including MITABNET, need further work to be helpful in an MNAR framework. One key concept in MNAR is the need for a model that helps explain the unobserved confounding. Suppose there is a small sample of cases representing the MNAR mechanism. In that case, a potential solution is to use up-sampling [194] in the MI algorithm 1 to adapt the machine learning algorithms to the MNAR framework.

Adding new genomic biomarkers to model the varying effect of ICI: In the study of survival data in chapter 5, we attempted to model the varying effect of ICI by PD-L1 expression. However, research by [190] suggests that for TMB, the average percentage of mutations in a sample of cancerous cells may predict treatment response in ICI therapy. The clinical NSCLC dataset analysed did not include TMB. However, new RWE datasets named clinical-genomic datasets comprised several genomic markers, including TMB. TMB, as defined above, is generated as a percentage or a distributed variable, hence, applying the survival GP approach presented in this work could be helpful for treatment effects interpretations.

Modelling the hazard function with ODE and GP: On the development side, time-varying effects modelling is an important field of research that bridges the gap between the modelling and statistical analysis of survival data. In particular, it resolves the problem of too simple proportional hazard models that may induce bias in RWE studies. In this work, we expanded on the B-splines model for non-proportional hazards. However, as mentioned

above, non-linear ODE models and GP models of the time-varying hazard function may potentially generalise better than the B-splines approach and are fascinating areas of applied research.

Adding time-varying covariates: Longitudinal time-varying covariates are of increasing mechanistic importance and may inform the clinical development of new investigational drugs. In RWE studies, the treatment strategy may change depending on evolving characteristics. For example, treatment discontinuation may happen because a contraindication occurs, where the patient may stop treatment and decide with the doctor whether to switch to an alternate treatment. These sustained treatment strategies are clinically relevant in screening test interpretations, and treatment interventions [195]. Application of the parametric g-formula, see equation 3.32 in dynamic treatment strategies can help in the task of sequential decision making. To extend the RWE study to include time-varying covariates, one must specify the RWE study design (consistency), measure covariates adequately (exchangeability), and apply an appropriate method under the exchangeability assumption.

Construct OTS bounds for CATE from censored survival data: Finally, in this work, we used the OTS bound to assess the impact of non-observed confounding for the ATE. However, one could generalise the OTS bound to determine the effect of non-observed confounding on the CATE. Similarly, a time-varying causal bound may be essential for interpreting dynamic predictions of treatment interventions.

6.2.2 Development of causal inference

Comparison of causal models: The field of research related to causal inference has had a significant upswing in the last decade, with particular interest in wrapping the recent advances in machine learning within a causal framework. For example, adversarial attacks showed that neural networks might be unreliable in the real world in terms of recognising objects by tricking the state-of-the-art Inception V3 Google AI algorithm [196] into predicting that a cat was an avocado, a baseball was a coffee machine [35, 197] etc. Although complex predictive models identify patterns in the data, they may not be suitable for recognising "good" patterns for RWD analysis. Hence, an area of active research is the comparison of causal models [81, 193]. In particular, further development of this technology with applications to survival analysis and clinical data will help the applicability of RWE studies.

Causal graph discovery: Most of the methods described in this thesis focused on the simple setting where the causal model was assumed to be known.

For example, we had the set of covariates W, a treatment A and the patient outcome Y. We recognised that the treatment effects affected the patient outcome, and the question was its strength. Moreover, the development of causal inference includes causal discovery, for example, learning the causal graph from data in genetic networks that may enable $N = 1$ studies, which help in drug development for rare cancerous diseases.

6.2.3 Examining RWE in new areas

Over the last decade, a variety of studies have studied the problem of applying machine learning techniques to RWE studies for advancing medicine [143]. We see that RWE studies can support drug development, informing trial design, and biomarker sourcing. As mentioned above, an exciting area of research is the modelling of gene networks to individualised treatment and the development of therapies for rare cancerous diseases. Such gene networks may model the genomic and transcriptomic processes that underlie therapeutics at the molecular level. With the rapid advancements in DNA sequencing technology and the wealth of anonymised RWD available for basic and applied research, it may be possible to develop more personalised treatments in the next decade.

Bibliography

[1] Carlos Traynor, Tarjinder Sahota, Helen Tomkinson, Ignacio Gonzalez-Garcia, Neil Evans, and Michael Chappell. Imputing biomarker status from rwe datasetsa comparative study. *Journal of Personalized Medicine*, 11(12):1356, 2021.

[2] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. DNN multiple imputations: Improved imputation accuracy in RWE datasets. *Proceedings of Pharmacokinetics UK conference, Online*, 10-12 November 2021.

[3] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian time-varying effect models- a rwe study in oncology. *To submit to Pharmacometrics and Systems Pharamcology*, 2022.

[4] C. Traynor, T. Sahota, I. Gonzalez-Garcia, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian model comparison of survival models. *To submit to Proceedings of Pharmacokinetics UK conference, Canterbury*, 2-4 November 2022.

[5] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Multi-state model of disease progression and overall survival in nsclc. *Proceedings of the Population Approach Group in Europe, Stockholm*, 11-14 June 2019.

[6] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Bayesian multistate models in pharmacometrics. *Proceedings of Pharmacokinetics UK conference, Stratford upon Avon*, 6-8 November, 2019.

[7] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. Elastic net applications to target sourcing in metastatic breast cancer patients. *Proceedings of Pharmacokinetics UK conference, Manchester*, 21-23 November 2018.

[8] C. Traynor, T. Sahota, H. Tomkinson, N. Evans, and M.J. Chappell. A model-based prediction of survival outcome for breast cancer patients

stratified integrative genomics. *Proceedings of the WIN Oncology symposium, Paris*, 25-26 June 2018.

[9] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35:1–9, 2016.

[10] Charles C Margossian. A review of automatic differentiation and its efficient implementation. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1305, 2019.

[11] Tadej Ciglarič, Rok Češnovar, and Erik Štrumbelj. Automated opencl gpu kernel fusion for stan math. In *Proceedings of the International Workshop on OpenCL*, pages 1–6, 2020.

[12] Umberto Michelucci. Tensorflow: Advanced topics. In *Advanced Applied Deep Learning*, pages 27–77. Springer, 2019.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[14] Fei Dong, Humayun Irshad, Eun-Yeong Oh, Melinda F Lerwill, Elena F Brachtel, Nicholas C Jones, Nicholas W Knoblauch, Laleh Montaser-Kouhsari, Nicole B Johnson, Luigi KF Rao, et al. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PloS one*, 9(12):e114885, 2014.

[15] Miguel A Hernan. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

[16] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

[17] Peter L Bonate and Danny R Howard. *Pharmacokinetics in Drug Development: Problems and Challenges in Oncology, Volume 4*, volume 4. Springer, 2016.

[18] Biwei Huang, Kun Zhang, and Bernhard Schölkopf. Identification of time-dependent causal model: A gaussian process treatment. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[19] Sercan. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, May 2021.

[20] Elodie Baumfeld Andre, Robert Reynolds, Patrick Caubel, Laurent Azoulay, and Nancy A Dreyer. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiology and drug safety*, 29(10):1201–1212, 2020.

[21] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela van der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

[22] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

[23] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[24] David Sontag. Lecture 14: Causal inference, part 1. In *Machine Learning for Healthcare—MIT Course No. 6.871/HST.956*. Cambridge MA, 2021. MIT OpenCourseWare.

[25] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[26] Samantha Kleinberg. *Why: A guide to finding and using causes*. " O'Reilly Media, Inc.", 2015.

[27] Robert J Rovetto and Riichiro Mizoguchi. Causality and the ontology of disease. *Applied Ontology*, 10(2):79–105, 2015.

[28] Roger Kerry, Thor E Eriksen, Svein A Noer Lie, Stephen Mumford, and Rani L Anjum. Causation in evidence-based medicine: in reply to s trand and p arkkinen. *Journal of evaluation in clinical practice*, 20(6):985–987, 2014.

[29] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.

[30] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994.

[31] Barbra A Dickerman and Miguel A Hernán. Counterfactual prediction is not only for causal inference. *European Journal of Epidemiology*, 35(7):615–617, 2020.

[32] Miguel A Hernán and James M Robins. Causal inference, 2010.

[33] Sewall Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):320, 1920.

[34] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[35] Ludvig Hult. What is causal inference, and why should data scientists know? In *PyCon Sweden*. 2019.

[36] Richard Netemeyer, Peter Bentler, Richard Bagozzi, Robert Cudeck, Joseph Cote, Donald Lehmann, Roderick McDonald, Timothy Heath, Julie Irwin, and Tim Ambler. Structural equations modeling. 2001.

[37] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[38] Judea Pearl. *Causality*. Cambridge university press, 2009.

[39] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using structural equation models. *arXiv preprint arXiv:1207.5136*, 2012.

[40] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[41] Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

[42] Ludvig Hult and Dave Zachariah. Inference of causal effects when adjustment sets are unknown. *arXiv preprint arXiv:2012.08154*, 2020.

[43] Richard McElreath. Lecture 6: The haunted dag and the causal terror. In *Statistical Rethinking Winter 2019*. 2019.

[44] Peng Ding and Luke W Miratrix. To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57, 2015.

[45] Vanessa Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime data analysis*, 25(4):593–610, 2019.

[46] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

[47] Alexander P Keil, Eric J Daza, Stephanie M Engel, Jessie P Buckley, and Jessie K Edwards. A bayesian approach to the g-formula. *Statistical methods in medical research*, 27(10):3183–3204, 2018.

[48] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

[49] Theodore W Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the american Statistical Association*, 52(278):200–203, 1957.

[50] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[51] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[52] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[53] Andrew Gelman. Missing-data imputation. *Data analysis using regression and multilevel/hierarchical models*, pages 529–543, 2006.

[54] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

[55] Rungang Han, Rebecca Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020.

[56] Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[57] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[58] Daniel J Stekhoven and Peter Bühlmann. Missforestnon-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[59] Ramiro D Camino, Christian A Hammerschmidt, and Radu State. Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666*, 2019.

[60] Richard McElreath. Lecture 18: Missing data. In *Statistical Rethinking Winter 2019*. 2019.

[61] Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988.

[62] James J Driscoll and Oliver Rixe. Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials. *The Cancer Journal*, 15(5):401–405, 2009.

[63] David G Kleinbaum and Mitchel Klein. Kaplan-meier survival curves and the log-rank test. In *Survival analysis*, pages 55–96. Springer, 2012.

[64] Frank E Harrell. Regression modeling strategies. *BIOS*, 330(2018):14, 2017.

[65] Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

[66] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[67] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[68] Rinku Sutradhar and Peter C Austin. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of epidemiology*, 28(1):54–57, 2018.

[69] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[70] Ana Quartilho, Daniel M Gore, Catey Bunce, and Stephen J Tuft. Royston- parmar flexible parametric survival model to predict the probability of keratoconus progression to corneal transplantation. *Eye*, 34(4):657–662, 2020.

[71] Miguel A Hernan, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Meir J Stampfer, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6):766, 2008.

[72] Samuel L Brilleman, Eren M Elci, Jacqueline Buros Novik, and Rory Wolfe. Bayesian survival analysis using the rstanarm r package. *arXiv preprint arXiv:2002.09633*, 2020.

[73] Karin B Michels and JoAnn E Manson. Postmenopausal hormone therapy: a reversal of fortune, 2003.

[74] Francine Grodstein, Joann E Manson, and Meir J Stampfer. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *Journal of Women's Health*, 15(1):35–44, 2006.

[75] Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of clinical Oncology*, 32(22):2380, 2014.

[76] Katie M O'Brien, Stephen R Cole, Chiu-Kit Tse, Charles M Perou, Lisa A Carey, William D Foulkes, Lynn G Dressler, Joseph Geradts, and Robert C Millikan. Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study. *Clinical Cancer Research*, 16(24):6100–6110, 2010.

[77] Guido W Imbens and Donald B Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

[78] Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

[79] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[80] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

[81] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[82] Peter Armitage and Michael Hills. The two-period crossover trial. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 31(2):119–131, 1982.

[83] Ahmed M Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

[84] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[85] Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

[86] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[87] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[88] David Roxbee Cox. Planning of experiments. 1958.

[89] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

[90] Emilia Gvozdenovic, Lucio Malvisi, Elisa Cinconze, Stijn Vansteelandt, Phoebe Nakanwagi, Emmanuel Aris, and Dominique Rosillon. Causal inference concepts applied to three observational studies in the context of vaccine development: from theory to practice. *BMC medical research methodology*, 21(1):1–10, 2021.

[91] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[92] Andrew Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

[93] Nasser Hanna, David Johnson, Sarah Temin, Sherman Baker Jr, Julie Brahmer, Peter M Ellis, Giuseppe Giaccone, Paul J Hesketh, Ishmael Jaiyesimi, Natasha B Leighl, et al. Systemic therapy for stage iv non–small-cell lung cancer: American society of clinical oncology clinical practice guideline update. *Journal of Clinical Oncology*, 2017.

[94] Brady Neal. Lecture 2: Potential outcomes. In *Introduction to Causal Inference*. 2021.

[95] Abraham Wald. Statistical decision functions. 1950.

[96] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[97] Ben Lambert. *A students guide to Bayesian statistics*. Sage, 2018.

[98] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[99] Grant Sanderson. Bayes theorem, the geometry of changing beliefs. In *Mathematics with a distinct visual perspective. Linear algebra, calculus, neural networks, topology, and more.* 2018.

[100] Richard McElreath. Lecture 8: Markov chain monte carlo. In *Statistical Rethinking Winter 2019*. 2019.

[101] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[102] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32, 2017.

[103] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[104] Richard McElreath. Lecture 7: Overfitting. In *Statistical Rethinking Winter 2019*. 2019.

[105] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[106] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[107] Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for bayesian regression models. *The American Statistician*, 2019.

[108] Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

[109] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

[110] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017.

[111] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[112] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[113] Grant Sanderson. Lecture 1: But what is a neural network? In *Mathematics with a distinct visual perspective. Linear algebra, calculus, neural networks, topology, and more.* 2018.

[114] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[115] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[116] Paul D Stolley. When genius errs: Ra fisher and the lung cancer controversy. *American Journal of Epidemiology*, 133(5):416–425, 1991.

[117] Mark Parascandola. Two approaches to etiology: the debate over smoking and lung cancer in the 1950s. *Endeavour*, 28(2):81–86, 2004.

[118] James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[119] David Sontag. Lecture 14: Causal inference, part 2. In *Machine Learning for Healthcare—MIT Course No. 6.871/HST.956*. Cambridge MA, 2021. MIT OpenCourseWare.

[120] John Antonakis and Rafael Lalive. Counterfactuals and causal inference: Methods and principles for social research by stephen l. morgan & christopher winship, 2011.

[121] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[122] Sengwee Toh, Luis A García Rodríguez, and Miguel A Hernán. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiology and drug safety*, 20(8):849–857, 2011.

[123] Brady Neal. Lecture 6: Estimation. In *Introduction to Causal Inference*. 2021.

[124] Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

[125] Charles F Manski. *Partial identification of probability distributions*. Springer Science & Business Media, 2003.

[126] Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[127] Brady Neal. Lecture 8: Unobserved confounding, bounds, and sensitivity analysis. In *Introduction to Causal Inference*. 2021.

[128] Peter W Horby, Lise Estcourt, Leon Peto, Jonathan R Emberson, Natalie Staplin, Enti Spata, Guilherme Pessoa-Amorim, Mark Campbell, Alistair Roddick, Nigel E Brunskill, et al. Convalescent plasma in patients admitted to hospital with covid-19 (recovery): a randomised, controlled, open-label, platform trial. *medRxiv*, 2021.

[129] Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *arXiv preprint arXiv:2003.01747*, 2020.

[130] Olaf Wolkenhauer, Peter Wellstead, Kwang-Hyun Cho, and Brian Ingalls. Sensitivity analysis: from model parameters to system behaviour. *Essays in biochemistry*, 45:177–194, 2008.

[131] Gerko Vink. Towards a standardized evaluation of multiple imputation routines. 2016.

[132] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[133] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[134] Hongshan Li, Jingyu Yu, Chao Liu, Jiang Liu, Sriram Subramaniam, Hong Zhao, Gideon M Blumenthal, David C Turner, Claire Li, Malidi Ahamadi, et al. Time dependent pharmacokinetics of pembrolizumab in patients with solid tumor and its correlation with best overall response. *Journal of pharmacokinetics and pharmacodynamics*, 44(5):403–414, 2017.

[135] Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

[136] Craig K Enders. *Applied missing data analysis*. Guilford press, 2010.

[137] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of statistical software*, 45(7):1–47, 2011.

[138] Lisa L Doove, Stef Van Buuren, and Elise Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014.

[139] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

[140] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[141] Stéphane Dray and Julie Josse. Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, 216(5):657–667, 2015.

[142] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[143] Sean Khozin, Rebecca A Miksad, Johan Adami, Mariel Boyd, Nicholas R Brown, Anala Gossai, Irene Kaganman, Deborah Kuk, Jillian M Rockland, Richard Pazdur, et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non–small cell lung cancer. *Cancer*, 125(22):4019–4032, 2019.

[144] Gerko Vink, Laurence E Frank, Jeroen Pannekoek, and Stef Van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, 2014.

[145] Michael Mayer. *missRanger: Fast Imputation of Missing Values*, 2018. R package version 1.0.3.

[146] Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

[147] Yves Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[148] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3), 2013.

[149] Carlos Serra Traynor. *simtte: Simulate bespoke time-to-event models*, 2020. R package version 1.0.0.

[150] Irva Hertz-Picciotto and Beverly Rockhill. Validity and efficiency of approximation methods for tied survival times in cox regression. *Biometrics*, pages 1151–1156, 1997.

[151] Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2018. R package version 5.1-2.

[152] John Barnard and Donald B Rubin. Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

[153] John W Graham, Allison E Olchowski, and Tamika D Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, 8(3):206–213, 2007.

[154] Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

[155] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[156] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[157] John W Graham. Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, 10(1):80–100, 2003.

[158] Andrew Gelman and Jennifer Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011.

[159] Rachel M Webster. The immune checkpoint inhibitors: where are we now? *Nature reviews. Drug discovery*, 13(12):883, 2014.

[160] Daniel S Chen and Ira Mellman. Oncology meets immunology: the cancer-immunity cycle. *immunity*, 39(1):1–10, 2013.

[161] Jianda Yuan, Sacha Gnjatic, Hao Li, Sarah Powel, Humilidad F Gallardo, Erika Ritter, Geoffrey Y Ku, Achim A Jungbluth, Neil H Segal, Teresa S Rasalan, et al. Ctla-4 blockade enhances polyfunctional ny-eso-1 specific t cell responses in metastatic melanoma patients with clinical benefit. *Proceedings of the National Academy of Sciences*, 105(51):20410–20415, 2008.

[162] Dawn E Dolan and Shilpa Gupta. Pd-1 pathway inhibitors: changing the landscape of cancer immunotherapy. *Cancer Control*, 21(3):231–237, 2014.

[163] Raju K Vaddepally, Prakash Kharel, Ramesh Pandey, Rohan Garje, and Abhinav B Chandra. Review of indications of fda-approved immune checkpoint inhibitors per nccn guidelines with the level of evidence. *Cancers*, 12(3):738, 2020.

[164] Vamsidhar Velcheti, Pallavi D Patwardhan, Frank Xiaoqing Liu, Xin Chen, Xiting Cao, and Thomas Burke. Real-world pd-l1 testing and distribution of pd-l1 tumor expression by immunohistochemistry assay type among patients with metastatic non-small cell lung cancer in the united states. *PLoS One*, 13(11):e0206370, 2018.

[165] Omar Abdel-Rahman. Correlation between pd-l1 expression and outcome of nsclc patients treated with anti-pd-1/pd-l1 agents: a meta-analysis. *Critical reviews in oncology/hematology*, 101:75–85, 2016.

[166] Martin Reck, Delvys Rodriguez-Abreu, Andrew G Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi,

Sinead Cuffe, et al. Pembrolizumab versus chemotherapy for pd-l1–positive non–small-cell lung cancer. *N engl J med*, 375:1823–1833, 2016.

[167] Tony SK Mok, Yi-Long Wu, Iveta Kudaba, Dariusz M Kowalski, Byoung Chul Cho, Hande Z Turna, Gilberto Castro Jr, Vichien Srimuninnimit, Konstantin K Laktionov, Igor Bondarenko, et al. Pembrolizumab versus chemotherapy for previously untreated, pd-l1-expressing, locally advanced or metastatic non-small-cell lung cancer (keynote-042): a randomised, open-label, controlled, phase 3 trial. *The Lancet*, 393(10183):1819–1830, 2019.

[168] S Peters, SS Ramalingam, L Paz-Ares, R Bernabe Caro, B Zurawski, S-W Kim, A Alexandru, L Lupinacci, E de la Mora Jimenez, H Sakai, et al. Nivolumab (nivo)+ low-dose ipilimumab (ipi) vs platinum-doublet chemotherapy (chemo) as first-line (1l) treatment (tx) for advanced non-small cell lung cancer (nsclc): Checkmate 227 part 1 final analysis. *Annals of Oncology*, 30:v913–v914, 2019.

[169] Scott J Antonia, Augusto Villegas, Davey Daniel, David Vicente, Shuji Murakami, Rina Hui, Takayasu Kurata, Alberto Chiappori, Ki H Lee, Maike De Wit, et al. Overall survival with durvalumab after chemoradiotherapy in stage iii nsclc. *New England Journal of Medicine*, 379(24):2342–2350, 2018.

[170] Xinran Ma, Lawrence Bellomo, Kelly Magee, Caroline S Bennette, Olga Tymejczyk, Meghna Samant, Melisa Tucker, Nathan Nussbaum, Bryan E Bowser, Joshua S Kraut, et al. Characterization of a real-world response variable and comparison with recist-based response rates from clinical trials in advanced nsclc. *Advances in therapy*, 38(4):1843–1859, 2021.

[171] Lihui Zhao, Brian Claggett, Lu Tian, Hajime Uno, Marc A Pfeffer, Scott D Solomon, Lorenzo Trippa, and LJ Wei. On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221, 2016.

[172] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[173] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[174] John A List, Azeem M Shaikh, and Yang Xu. Multiple hypothesis testing in experimental economics (no. w21875), 2016.

[175] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

[176] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[177] Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

[178] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[179] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[180] Corwin M Zigler, Francesca Dominici, and Yun Wang. Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics*, 13(2):289–302, 2012.

[181] Stan Development Team. RStan: the R interface to Stan, 2019. R package version 2.19.2.

[182] Andreas Bender, Andreas Groll, and Fabian Scheipl. A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321, 2018.

[183] Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pages 517–523. Elsevier, 2012.

[184] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[185] F. Le Borgne and Y. Foucher. *IPWsurvival: Propensity Score Based Adjusted Survival Curves and Corresponding Log-Rank Statistic*, 2017. R package version 0.5.

[186] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[187] Pierre JM Verweij and Hans C Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

[188] Kyle T Baron. *mrgsolve: Simulate from ODE-Based Models*, 2021. R package version 0.10.9.

[189] Jeffrey Crawford, David C Dale, and Gary H Lyman. Chemotherapy-induced neutropenia: risks, consequences, and new directions for its management. *Cancer*, 100(2):228–237, 2004.

[190] Jean-David Fumet, Caroline Truntzer, Mark Yarchoan, and Francois Ghiringhelli. Tumour mutational burden as a biomarker for immunotherapy: current data and emerging concepts. *European Journal of Cancer*, 131:40–50, 2020.

[191] Leena Gandhi and Marina C Garassino. Pembrolizumab plus chemotherapy in lung cancer. *N Engl J Med*, 379(11):e18, 2018.

[192] Patrick Royston and Mahesh KB 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. *BMC medical research methodology*, 13(1):1–15, 2013.

[193] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[194] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[195] Barbra A Dickerman, Xabier García-Albéniz, Roger W Logan, Spiros Denaxas, and Miguel A Hernán. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature medicine*, 25(10):1601–1606, 2019.

[196] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[197] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Appendix A

Result Reproduction

A.1 R code-simulation of missing data

The following *R* code was used for generating the synthetic data in the simulation experiments in chapter 4.

Load the *simstandard*, *mice*, and *simtte* package:

```
library(simstandard)
library(mice)
library(simtte) # https://github.com/csetraynor/simtte
```

Define the function that generates datasets from an SCM with a high correlation between manifest variables:

```
# Simulation of high correlated datasets
# s: sample number (1 to 200)
# N: number of individuals to simulate
sample_data_sim_high_cor <- function(s, N) {
  # lavaan syntax for model (high correlation)
  m <- "
  A =~ 0.7 * A1 + 0.9 * A2 + 0.9 * A3 + 0.3 * B1
  B =~ 0.7 * B1 + 0.9 * B2 + 0.9 * B3
  B ~ 0.9 * A
  "
  # Simulate data
  d <- sim_standardized(m,
                        n = N,
                        latent = FALSE,
                        errors = FALSE)
```

```
}
```

Define the function that generates datasets from an SCM with a low correlation between manifest variables:

```
# Simulation of low correlated datasets
# s: sample number (1 to 200)
# N: number of individuals to simulate
sample_data_sim_low_cor <- function(s, N) {
  # lavaan syntax for model (low correlation)
  m <- "
  A =~ 0.4 * A1 + 0.4 * A2 + 0.4 * A3 + 0.3 * B1
  B =~ 0.4 * B1 + 0.5 * B2 + 0.4 * B3
  B ~ 0.5 * A
  "
  # Simulate data
  d <- sim_standardized(m,
                        n = N,
                        latent = FALSE,
                        errors = FALSE)
}
```

Define the function to amputate datasets:

```
# Amputation of datasets
# d : dataset object
# prop : proportion of missingness
# cor_level :  correlation id
# mech : mechanism of missingness (MCAR, MAR, MNAR)

ampute_sim <- function(d, prop, mech) {
  d_miss <- mice::ampute(d, prop = prop, mech = mech)
  # writeLines(paste0("Percentage of newly generated missing values: ",
  100*sum(is.na(d_miss$amp))/prod(dim(d_miss$amp)), " %"))
}
```

Define the function to simulate event times:

```
# Simulation of event times
# simdata : dataset object
simTteData <- function(simdata) {
  # set true coefficients
  beta <- c(-2, -1, 0, 1, 2, 3)
  # compute linear predictor
  lp <- as.matrix(simdata) %*% beta

  res <- sim_tte(pi = lp,
                 mu = -1,
                 coefs = 1.1,
                 type = "weibull",
                 obs.only = F,
                 obs.aug = T,
                 delta = 0.05,
                 end_time = 500)
  return(loo::nlist(res, simdata))
}
```

A.2 CPP code - proportional hazards

The following *CPP* code was used for the Weibull ODE proportional hazard model used in chapter 4:

```
Model file:  weibull.txt
[PROB]
# Model: ‘Simulate Weibull parametric proportional hazard model‘
  - Forward Kolmogorov differential equation
  - Author: Carlos Traynor
  - Date: ‘r Sys.Date()‘
  - Version: ‘r packageVersion("mrgsolve")‘
```

Define the time-to-event parameters:

```
[PARAM] @annotated
  lp   :  0.15   : linear predictor
  mu   :  0.1    : intercept
```

```
  shape :  1      : shape parameter
```

List of state vector with initial conditions:

```
[INIT]
  p11 = 1
```

Define the link function:

```
[MAIN]
  double eta = exp(mu + lp);
```

Define the ODE system:

```
[ODE]
  if(SOLVERTIME >= 1E-3) {
    dxdt_p11 =  - p11 * shape * pow (SOLVERTIME, shape - 1) * eta;
  } else {
    dxdt_p11 = - p11 * eta; // approximation
  }
```

A.3 Python code - MITABNET

The following *Python* code was used for MITABNET in chapter 4.

Load necessary packages:

```
# Necessary packages
from __future__ import absolute_import
from __future__ import division
from __future__ import print_function
import argparse
import sys
import os
from pathlib import Path
```

```
from pytorch_tabnet_dropout.tab_model import TabNetRegressor
import torch
import pandas as pd
import numpy as np
from pytorch_tabnet.pretraining import TabNetPretrainer
```

Define the main function and read features:

```
def main(args):

  # call parameters
  data_dir = args.data_dir
  perc = args.perc
  dropout = args.dropout
  data_dir = os.path.join("..", data_dir)

  x  = pd.read_csv( os.path.join(data_dir, 'x.csv') )
  ry = pd.read_csv( os.path.join(data_dir, 'ry.csv') )
  wy = pd.read_csv( os.path.join(data_dir, 'wy.csv') )
  y  = pd.read_csv( os.path.join(data_dir, 'y.csv') )

  y = y.values
  x = x.values
  ry = (ry.values)[:,0]
  wy = (wy.values)[:,0]

  xobs = x[ry, :]
  xmis = x[wy, :]
  yobs = y[ry]
```

Pre-train MITABNET:

```
  # TabNetPretrainer
  unsupervised_model = TabNetPretrainer(
      optimizer_fn=torch.optim.Adam,
      optimizer_params=dict(lr=2e-2),
      mask_type='entmax' # "sparsemax"
```

```
)

max_epochs = 1000 if not os.getenv("CI", False) else 2

Set = np.random.choice(["train", "valid"], p =[perc, 1-perc], size=len(yobs))
X_train = xobs[Set == "train"]
X_valid = xobs[Set == "valid"]
y_train = yobs[Set == "train"]
y_valid = yobs[Set == "valid"]

unsupervised_model.fit(
  X_train=X_train,
  eval_set=[X_valid],
  max_epochs=max_epochs , patience=20,
  batch_size=512, virtual_batch_size=64,
  num_workers=0,
  drop_last=False,
  pretraining_ratio=0.8)
```

Specify MITABNET options and train the model:

```
# Training
clf = TabNetRegressor(optimizer_fn=torch.optim.Adam,
                      optimizer_params=dict(lr=2e-2),
                      scheduler_params={"step_size":10,
                                        "gamma":0.9},
                      scheduler_fn=torch.optim.lr_scheduler.StepLR,
                      mask_type='sparsemax',
                      dropout = dropout
                      )

clf.fit(X_train=X_train, y_train=y_train,
  eval_set=[(X_train, y_train), (X_valid, y_valid)],
  eval_name=['train', 'valid'],
  eval_metric=['rmse'],
  max_epochs=max_epochs,
  patience=50,
  batch_size=256, virtual_batch_size=64,
  num_workers=0,
  drop_last=False,
```

```
    from_unsupervised=unsupervised_model)
```

Issue imputations for missing values:

```
  preds = clf.predict(xmis)
  preds = pd.DataFrame(preds)
  preds.to_csv(os.path.join(data_dir, 'y_pred.csv') , index = False)
```

Set up arguments for main function:

```
if __name__ == '__main__':
  # Inputs for the main function
  parser = argparse.ArgumentParser()

  parser.add_argument(
  '--data_dir',
    default=".",
    type=str)
  parser.add_argument(
  '--perc',
    default=0.8,
    type=float)
  parser.add_argument(
  '--dropout',
    default=0.1,
    type=float)

  args = parser.parse_args()
  # Call main function
  main(args)
```

A.4 CPP code - non-proportional hazards

The following *CPP* code was used for the Weibull ODE non-proportional hazard model used in chapter 5:

```
Model file:  weibull_tve.txt
```

```
[PROB]
# Model: ‘Simulate Weibull parametric non proportional
           hazard model with two tve‘
- Forward Kolmogorov differential equation
  - Author: Carlos Traynor
  - Date: ‘r Sys.Date()‘
  - Version: ‘r packageVersion("mrgsolve")‘
```

Define the time-to-event parameters:

```
[PARAM] @annotated
  X1      :   1     : First variable with tve
  X2      :   1     : Second variable with tve
  coeftv1 : 0.5 : Coefficient tve 1
  coeftv2 : 0.5 : Coefficient tve 1
  lp      :  0.15   : linear predictor
  mu      :  0.1    : intercept
  shape   :  1     : shape parameter
```

List of state vector with initial conditions:

```
[INIT]
p11 = 1
```

Define the link function for the proportional hazard component:

```
[MAIN]
double eta = exp(mu + lp);
```

Define the ODE system:

```
[ODE]
if(SOLVERTIME >= 1E-1) {
  dxdt_p11 =  - p11 * shape * pow (SOLVERTIME, shape - 1) * eta *
```

```
  exp( X1 * coeftv1 * SOLVERTIME) * exp(X2 * coeftv2 * SOLVERTIME);
} else {
  dxdt_p11 = - p11 * eta; // approximate via exponential model
}
```

A.5 Stan code-Weibull GP

The following *Stan* code [11] was used for the Weibull GP hazard model used in chapter 5.

Define the log-survival and log-hazard function for the Weibull model:

```
functions {
  /**
  * Log survival for Weibull distribution
  *
  * @param eta real, prognostic index, aka linear predictor
  * @param t real, event or censoring time
  * @param shape real, Weibull shape
  * @return A real
  */
  real weibull_log_surv(real eta, real t, real shape) {
    real res;
    res = - pow(t, shape) * exp(eta);
    return res;
  }

  /**
  * Log hazard for Weibull distribution
  *
  * @param eta real, prognostic index, aka linear predictor
  * @param t real, event or censoring time
  * @param shape Real, Weibull shape
  * @return A real
  */
  real weibull_log_haz(real eta, real t, real shape) {
    return log(shape) + (shape - 1) * log(t) + eta;
  }
```

Compute the GP latent state:

```
  /**
  * GP: computes sigma_yless Gaussian Process
  * @param volatility volatility of gaussian process
  * @param amplitude
  * @param GP_z_scores
  * @param n_x number of x
  * @param x
  * @return A vector
  */
  vector GP(real volatility, real amplitude, vector GP_z_scores,
  int n_x, real[] x ) {
    matrix[n_x,n_x] cov_mat ;
    real amplitude_sq_plus_jitter ;
    amplitude_sq_plus_jitter = amplitude^2 + 0.001 ;
    cov_mat = gp_exp_quad_cov(x, amplitude, 1/volatility) ;
    cov_mat = add_diag(cov_mat, amplitude_sq_plus_jitter);
    return(cholesky_decompose(cov_mat) * GP_z_scores ) ;
  }
}
```

Define the data inputs:

```
data {

  int<lower=1> N; // N: number of individuals
  vector<lower = 0>[N] time_at_risk; // time:
  int<lower=0, upper = 1> event_status[N]; // status indicator

  int K;
  matrix[N, K] x;            // set of individual covariates
  real log_crude_event_rate; // helps center the intercept

  /*** GP inputs ***/

  // n_w: number of groups in predictor array
  int n_w;
  // data matrix fore regression should include intercept
```

```
  matrix[N, n_w] w; // array of GP weights

  // n_gp_x: number of unique x values
  int<lower=1> n_gp_x ;
  // x: unique values of x
  // should be scaled to min=0,max=1
  real gp_x[n_gp_x] ;
  // x_index: vector indicating which x is associated with each y
  int gp_x_index[N] ;
}
```

Define the model parameters:

```
parameters {
  real<lower=0> shape;
  vector[K] beta;

  // volatility_helper: helper for cauchy-distributed volatility
  vector<lower=0,upper=pi()/2>[n_w] volatility_helper ;

  // amplitude: amplitude of GPs
  vector<lower=0>[n_w] amplitude ;

  // f_GP_z_scores: helper variable for GPs
  matrix[n_gp_x,n_w] f_GP_z_scores;

}
```

Define operations on model parameters:

```
transformed parameters {
  // volatility: volatility of GPs (a.k.a. inverse-lengthscale)
  vector[n_w] volatility ;
  // f: GPs
  matrix[n_gp_x,n_w] f ;

  //next line implies volatility ~ cauchy(0,1)
  volatility = tan (volatility_helper) ;
```

```
  // loop over predictors, computing GPs for each treatment
  for(wi in 1:n_w){
    f[,wi] = GP (volatility[wi] , amplitude[wi] ,
      f_GP_z_scores[ ,wi] , n_gp_x , gp_x);
  }

}
```

Define the model likelihood and the prior distributions for the model parameters:

```
model {

  vector[N] eta = x * beta;

  for ( i in 1:N ) {
    real mu = 0;
    mu += log_crude_event_rate; //helps center the intercept
    mu += eta[i];
    mu += sum(w[i,] .* f[gp_x_index[i], ]);

    // weibull survival likelihood
    if(event_status[i] == 1) {
      target += weibull_log_haz (mu, time_at_risk[i], shape);
      target += weibull_log_surv(mu, time_at_risk[i], shape);
    } else {
      target += weibull_log_surv(mu, time_at_risk[i], shape);
    }
  } // close for loop of individuals!

  /*** priors for survival model ***/
  shape  ~ normal(1, 1); // Half-normal peaks at 1

  /*** priors for GP stuff ***/
  // normal(0,1) priors on GP_z_scores
  target += std_normal_lpdf( to_vector(f_GP_z_scores));
  // amplitude prior
  target += exponential_lpdf (amplitude | 1);
}
```