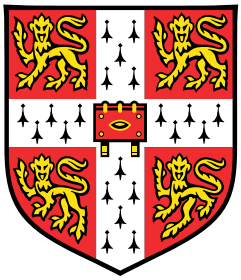
Approaches to developing clinically useful Bayesian risk prediction models



Solon Karapanagiotis

MRC Biostatistics Unit
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

St Catharine's College

April 2021

Σὰ βγεῖς στὸν πηγαιμὸ γιὰ τὴν Ἰθάκη, νὰ εὔχεσαι νά ᾽ναι μακρὺς ὁ δρόμος, γεμάτος περιπέτειες, γεμάτος γνώσεις.

[. . .]

Κι ἂν πτωχικὴ τὴν βρεῖς, ἡ Ἰθάκη δὲν σὲ γέλασε. Ἔτσι σοφὸς ποὺ ἔγινες, μὲ τόση πεῖρα, ἤδη θὰ τὸ κατάλαβες οἱ Ἰθάκες τὶ σημαίνουν.

Κωνσταντίνος Καβάφης, Ιθάκη
Constantine Cavafy, Ithaka

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification except as declared in the preface and specified in the text. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. It does not exceed the prescribed word limit.

Solon Karapanagiotis
April 2021

Abstract

Approaches to developing clinically useful Bayesian risk prediction models

Solon Karapanagiotis

Prediction of the presence of disease (diagnosis) or an event in the future course of disease (prognosis) becomes increasingly important in the current era of personalised medicine. Both tasks (diagnosis and prognosis) are supported using (risk) prediction models. Such models usually combine multiple variables by using different statistical and/or machine learning approaches. Recent advances in prediction models have improved diagnostic and prognostic accuracy, in some cases surpassing the performance of clinicians. However, evidence is lacking that deployment of these models has improved care and patient outcomes. That is, their clinical usefulness is debatable. One barrier to demonstrating such improvement is the basis used to evaluate their performance. In this thesis, we explore methods for developing (building and evaluating) risk prediction models, in an attempt to create clinically useful models.

We start by introducing a few commonly used metrics to evaluate the predictive performance of prediction models. We then show that a model with good predictive performance is not enough to guarantee clinical usefulness. A well performing model can be clinically useless, and a poor model valuable. Following recent line of work, we adopt a decision theoretic approach for model evaluation that allows us to determine whether the model would change medical decisions and, if so, whether the outcome of interest would improve as a result.

We then apply this approach to investigate the clinical usefulness of including information about circulating tumour DNA (ctDNA) when predicting response to treatment in metastatic breast cancer. ctDNA has been proposed as a promising approach to assess response to treatment. We show that incorporating trajectories of circulating tumour DNA results in a clinically useful model and can improve clinical decisions.

However, an inherit limitation to the decision theoretic approach (and related ones) is that model building and evaluation are done independently. During training, the prediction model is agnostic of the clinical consequences from its use. That is, the prediction model is agnostic of its (clinical) purpose, e.g., which type of classification error is more costly (i.e., undesirable). We address this shortcoming by introducing Tailored Bayes (TB), a novel Bayesian inference framework which "tailors" model fitting to

optimise predictive performance with respect to unbalanced misclassification costs. In both simulated and real-world applications, we find our approach to perform favourably in comparison to standard Bayesian methods.

We then move to extend the framework to situations where a large number of (potentially irrelevant) variables are measured. Such high-dimensional settings represent a ubiquitous challenge in modern scientific research. We introduce a sparse TB framework for variable selection and find that TB favours smaller models (with fewer variables) compared to standard Bayesian methods, whilst performing better or no worse. This pattern was seen both in simulated and real data. In addition, we show the relative importance of the variables changes when we consider unbalanced misclassification costs.

Acknowledgements

I would like to thank my supervisors Dr. Paul Newcombe and Dr. Oscar Rueda for all their help and advice with this PhD. I would like to express gratitude to Dr. Paul Kirk who despite not being my official supervisor, has invested his time in discussing parts of this thesis and has contributed to improving it. I would like to thank Dr. Emma Beddowes for provifing the data used in Chapter 2. I am also grateful to Professor Paul Pharoah and Professor Umberto Benedetto who provided me two of the datasets which I use in Chapter 3. I shall also thank Dr. Sach Mukherjee for many engaging discussions over the years, and Dr. Lorenz Wernish for his insightful comments during my first year. I am very grateful to all those listed above for their contributions towards this academic work. This thesis has been substantially enriched by my one-year collaboration with the Alan Turing Institute. Hence, I would like to acknowledge them for their academic and financial support. I also acknowledge the Medical Research Council for the studentship that allowed me to conduct this thesis. A big thank you to those that, willingly or unwillingly, supported me during my studies.

I would like to thank my parents, Georgia and Anestis, for opening all the doors which led me here, and my partner Luce for her unwavering support and encouragement.

Preface

Most of the work presented in this thesis will be published or it is being currently considered for publication. Specifically:

- Parts of Chapter 1 and the vast majority of Chapter 3 are being reviewed and have been made available as a pre-print, see Karapanagiotis et al. (2021).
- Chapter 4, being an extension of Chapter 3, will be subsequently submitted for publication.
- The work in Chapter 2 is self-contained and will be submitted to a scientific journal in due course.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

ESS_t	effective sample size for tailoring
AUROC	area under the ROC curve
AVR	aortic valve replacement
BART	Bayesian Additive Regression Trees
BMA	Bayesian model average
BS	Brier score
BVS	Bayesian variable selection
cfDNA	cell-free DNA
CT	computed tomography
ctDNA	circulating tumour DNA
CV	cross-validation
CVD	cardiovascular disease
DM II	Diabetes Mellitus Type II
ER	oestrogen receptor
EU	expected utility
EuroSCORE	European System for Cardiac Operative Risk Evaluation
HER2	human epidermal growth factor receptor 2
HMC	Hamiltonian Monte Carlo
ICI	integrated calibration index

mBC metastatic breast cancer

MCMC Markov Chain Monte Carlo

METABRIC Molecular Taxonomy of Breast Cancer International Consortium

NB Net Benefit

NICE National Institute for Health and Care Excellence

NUTS No-U-Turn sampler

PAM Predictor analysis of microarray 50

PD progressive disease

PIP posterior inclusion probability

RECIST response evaluation criteria in solid tumours

ROC receiver operating characteristic

SAVR surgical aortic valve replacement

SB Standard Bayes

SUPPORT Study to Understand Prognoses and Preferences for10Outcomes and Risks of Treatments

TAVI transcatheter aortic valve implantation

TB Tailored Bayes

Chapter 1

Introduction

Abstract This work is concerned with model building and evaluation for binary outcomes (risk prediction models). Our principal thesis is to move beyond standard approaches and incorporate information on how a model will be used and the consequences of decisions arising from its use when building and evaluating risk prediction models[1].

In this chapter, we describe a few commonly used performance measures to evaluate risk prediction models and we illustrate their main shortcoming: they are insensitive to the consequences from using a model in clinical practice. They do not provide an answer as to whether a model should be used in clinical practice. More importantly, they do not tell us if using the model will enhance medical decision-making let alone improve health outcomes of the targeted individuals. To overcome this shortcoming, we present an existing model evaluation metric (the Net Benefit), which is sensitive to the clinical consequences. Finally, in this chapter we advocate in favour of and motivate a Bayesian formalism on risk prediction.

Outline We start by motivating our work (Section 1.1) and summarising traditional performance measures for evaluating risk prediction models (Section 1.2). We then present their main shortcoming using two toy examples (Section 1.3). That is, they do not reflect the consequences of taking action based on the model's output. We then rely on decision theory and present a widely used approach to quantify the preferences of different consequences (Section 1.4). Building upon these concepts we introduce a model evaluation metric (the Net Benefit) that is sensitive to the clinical consequences (Section 1.5). In Section 1.6 we motivate the use of the Bayesian paradigm. We finish in Section 1.7 with an overview of the rest of the thesis.

[1] Throughout this dissertation I use "our/we" rather than "my/I" as a stylistic choice.

1.1 Motivation

Risk prediction models are widely used in healthcare. They yield predictions for individuals at risk of having a certain disease or condition (diagnostic setting) or experiencing a certain health event in the future (prognostic setting). In the diagnostic setting, the probability that a particular disease is present can be used, for example, to inform the referral of individuals for further testing, to initiate treatment, or to reassure individuals that a serious cause for their symptoms is unlikely. In the prognostic setting, predictions can be used for planning interventions based on the risk for developing a particular outcome or state of health within a specific period. In both diagnostic and prognostic settings risk prediction models are being developed, validated, implemented, and updated with the aim of assisting clinicians and individuals in estimating probabilities and potentially guide their decision-making (Baumgartner et al., 2017; Down et al., 2014; NICE, 2016, 2018). Risk prediction models are becoming increasingly abundant in the medical literature (Bellou et al., 2019; Carrick et al., 2020; De Kat et al., 2019; de Munter et al., 2017; Dijkland et al., 2020; Hou et al., 2019; Usher-Smith et al., 2016; Volkers et al., 2018), and policymakers are increasingly recommending their use in clinical practice guidelines (Meschia et al., 2014; NICE, 2016, 2018; Rabar et al., 2012). Three illustrative examples of prediction models currently recommenced for use in clinical practice are QRISK2, EuroSCORE, and PREDICT.

QRISK2 estimates the 10-year probability (risk) of developing cardiovascular disease (CVD) (coronary heart disease, stroke, or transient ischaemic attacks) based on age, sex, lifestyle factors (e.g., smoking status), blood-based biomarkers (e.g., total serum cholesterol) and pre-existing conditions (e.g., rheumatoid arthritis) among others (Hippisley-Cox et al., 2008). QRISK2 has been validated in multiple studies (Collins and Altman, 2010, 2012; Hippisley-Cox et al., 2014; Pike et al., 2016) and is currently recommended by the National Institute for Health and Care Excellence (NICE) for use in clinical practice. Specifically, the guidelines recommend that clinicians use QRISK2 to determine whether to prescribe statins for primary prevention of CVD if a person's CVD risk is 10% or more (NICE, 2016).

Another widely used model is the European System for Cardiac Operative Risk Evaluation (EuroSCORE) (Nashef et al., 1999; Roques et al., 2003). EuroSCORE is used for predicting the risk of operative mortality in cardiac surgery patients taking into account risk factors encompassing patient-related, cardiac and operation-related characteristics. It is recommended as a tool for weighing the risk of surgery against the expected natural history of valvular heart disease and as a basis for decision-making (Baumgartner et al., 2017). More specifically, the EuroSCORE is routinely used to guide clinical decisions for cardiac patients undergoing aortic valve replacement (AVR). Cardiac patients with severe symptomatic aortic stenosis are considered for surgical AVR (SAVR) or transcatheter aortic valve implantation (TAVI). Published guidelines recommend TAVI over SAVR if a patient's predicted mortality risk is above 10% (Baumgartner et al., 2017). We revisit EuroSCORE in Chapter 3.

A third example is PREDICT which allows estimation of prognosis and the absolute benefits of adjuvant therapy for women with invasive breast cancer (Wishart et al., 2010). Accurate prognosis is crucial in clinical decision-making around adjuvant therapy in breast cancer. Adjuvant therapies, such as

chemotherapy, are associated with serious side-effects and so are typically only in the patient's interest when there is non-negligible risk of mortality. PREDICT is based on clinical and pathological factors such as age, tumour size, tumour grade, number of positive nodes, estrogen receptor (ER) status, human epidermal growth factor receptor 2 (HER2) status, among others. It has been validated in many cohorts (De Glas et al., 2016; Dos Reis et al., 2017; Karapanagiotis et al., 2018; Wishart et al., 2011), and is currently recommended by the NICE to estimate prognosis and the benefits of adjuvant therapy (NICE, 2018).

A common feature of all these models is that probability estimates are based on combining information from multiple predictors (or risk factors, or covariates) observed or measured from an individual. They are typically constructed as regression models (or machine learning algorithms) by having multiple predictors as inputs and a continuous risk estimate between 0 and 1 as output. In this work, we focus on binary outcomes but some of the concepts can be easily extended to multi-category and time-to-event outcomes[2].

A first step before a model is recommended for use in clinical practice is its validation. The aim of model validation is to evaluate (quantify) the model's predictive performance in either resampled (participant) data of the development data set (often referred to as internal validation) or in other, independent participant data that were not used for developing the model (often referred to as external validation) or a combination of the two (Altman and Royston, 2000; Moons et al., 2012).

Several performance measures (or metrics) have been proposed for model validation (Altman and Royston, 2000; Steyerberg et al., 2011; Steyerberg and Vergouwe, 2014; Steyerberg et al., 2010). Here, we divide them into the traditional (statistical) performance measures (Section 1.2) and (more recent) measures that incorporate the clinical consequences of the predictions. We argue that traditional measures do not capture the consequences of using the model in clinical practice and hence present a distorted image of model validation. Two toy examples illustrate this (Section 1.3). As a result, we (and others) advocate for measures that incorporate the clinical consequences of the predictions (Chatterjee et al., 2016; Jung et al., 2020; Localio and Stack, 2015; Moons et al., 2015; Shah et al., 2019). We present such a measure in Section 1.5, first proposed in the seminal work of Vickers and Elkin (2006). The main concept behind it though had been derived previously by Pauker and Kassirer (1975) and Gail and Pfeiffer (2005). The key idea is based on a decision-theoretic approach to quantify the value provided by a model when considering the likely range of an individual's risk and benefit preferences, without the need for actually measuring these preferences for a particular individual (Section 1.4). As a result, we can incorporate consequences of the predictions and, in theory, the method can tell us whether a model is worth using at all or which of several alternative models should be used. Hence, it can be used for model comparison as well.

Even though such measures allow the evaluation of clinical utility of a risk prediction model they are inevitably and inherently (aggregate) summary measures. In clinical practice we need to obtain

[2] Additionally, sometimes continuous outcomes are often treated as binary for modelling purposes (e.g., blood pressure thresholds, disease severity score thresholds, etc).

individualised predictions and quantify our uncertainty for these predictions. Hence, in this work we advocate in favour of a Bayesian formalism to quantify uncertainty in a principled and coherent way (Section 1.6). Finally, Section 1.7 provides an overview and outline to the rest of the thesis.

1.2 Traditional performance measures

Many performance measures have been developed and proposed to evaluate risk prediction models. Here, we group them based on two aspects that characterise the performance of a risk prediction model: calibration and discrimination (Harrell, 2015; Steyerberg et al., 2019).

Calibration refers to the agreement between observed probabilities and predicted probabilities. A model is well calibrated if, for every 100 individuals given a risk of x%, close to x will indeed have the outcome of interest. For instance, if a 60% probability is predicted that the outcome will occur, then the outcome should on average occur in every 60 out of 100 patients with predicted probabilities of 60%. A few examples of calibration measures are the Hosmer-Lemeshow goodness-of-fit test (Hosmer and Lemesbow, 1980), Cox's intercept and slope approach (Cox, 1958) and the integrated calibration index (ICI) (Austin and Steyerberg, 2019). The ICI, which we will be using later, is a numerical summary of model calibration over the observed range of predicted probabilities (Table 1.1). It quantifies the agreement between observed and predicted probabilities, with values closer to zero indicating better performance. We refer to Huang et al. (2020) for more examples of calibration measures. Calibration concerns the average risk in a population and a well-calibrated model may assist in, for example, prevention decisions, but a miscalibrated model may lead to situations where an individual at high risk is assigned a low predicted probability, and thus forgoes effective (preventive) intervention (Holmberg and Vickers, 2013). However, a model that reliably predicts probabilities that are all between 40% and 60% is not able to distinguish patients with the outcome from patients without the outcome. Such a model is not able to separate risk estimates, and consequently, its discriminative ability is poor.

The performance measure that evaluates how well a model separates risk estimates is discrimination. Discrimination quantifies how well the model can separate those who do and do not have the outcome of interest. If the predicted values for cases are all higher than for non-cases, we say the model can discriminate perfectly, even if the predicted risks do not match the proportion with the outcome (i.e., even if the model has poor calibration). A prediction model can excellently distinguish patients with different outcomes, if it predicts probabilities close to 100% for patients with the outcome and probabilities close to 0% for patients without the outcome. Discrimination for binary outcomes is most often measured by the receiver operating characteristic (ROC) curve (Fawcett, 2006). The ROC curve is a plot of true positive rate (e.g., percentage of individuals correctly classified as having the outcome, or sensitivity) versus false positive rate (e.g., percentage of individuals incorrectly classified as having the outcome, or 1-specificity) evaluated at consecutive threshold values of the predicted probability (Table 1.1). The area under the ROC curve (AUROC) is a summary of the ROC curve. It represents the probability that an individual with the outcome has a higher predicted probability than an individual without the

outcome for a random pair of individuals consisting of one with and one without the outcome (Hanley and McNeil, 1982). A model that perfectly discriminates between individuals with and without an outcome would have an AUROC of 1 (the theoretical maximum), whereas a model with no ability to discriminate between such individuals would have an AUROC of 0.5.

It is recommended that risk prediction models are developed and validated using performance measures that capture both calibration and discrimination (Collins et al., 2015). However, both discrimination and calibration are statistical properties characterizing the performance of a prediction model, but neither captures the clinical consequences of a particular level of discrimination or degree of miscalibration. In other words, a model with good discrimination and calibration does not necessarily result in it being clinically useful. This is because traditional model metrics are insensitive to the benefits and harms of correct and incorrect classifications. The two toy examples that follow illustrate this point.

1.3 Limitations of traditional performance measures

The toy examples that follow present two artificial but realistic scenarios that illustrate the shortcomings of relying on traditional performance measures when evaluating prediction models. That is, we may end-up (confidently) selecting an inadequate model, i.e., a model with sub-optimal clinical utility (Table 1.1).

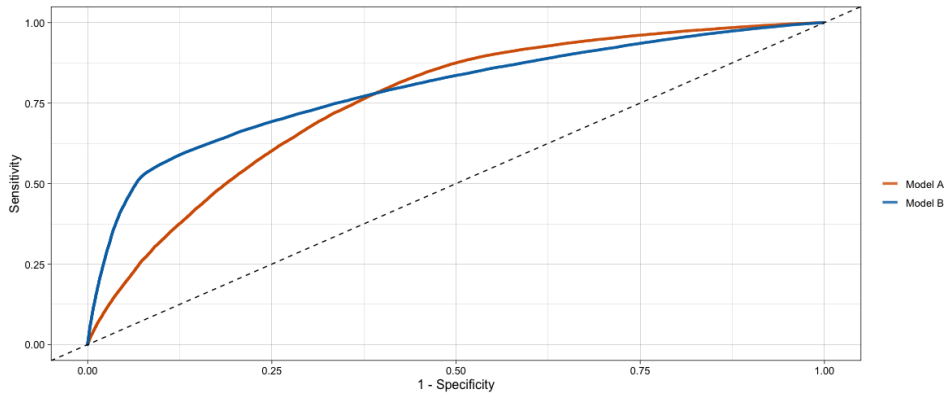
1.3.1 Toy example 1

We consider the following hypothetical scenario for a patient treated with chemotherapy for metastatic non-seminomatous testicular cancer. After chemotherapy, retroperitoneal lymph nodes can either still contain remnants of the metastases (residual disease) or only contain benign necrosis. Surgical lymph node resection (lymphadenectomy) can provide certainty and remove metastatic remnants but should be avoided when lymph nodes only contain benign necrosis. Following Van Calster et al. (2020) we assume that 55% of patients have residual disease (the binary outcome). We compare two logistic regression models to estimate the probability of residual disease. Both include the same continuous covariate but a different binary covariate. These covariates can be thought of as tumour-specific markers (e.g., percentage reduction in residual mass size after chemotherapy) or other biomarkers (e.g., elevated human chorionic gonadotropin levels).

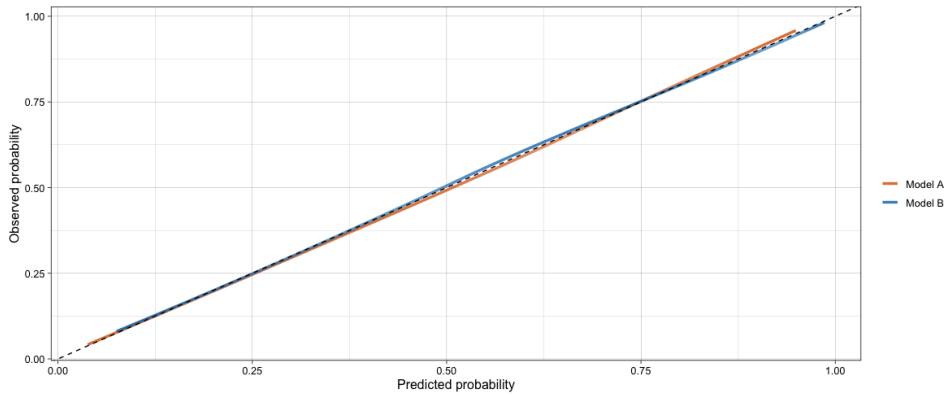
Model A includes a binary covariate with a sensitivity of 88%, and a specificity of 49%. Model B includes a covariate with a sensitivity of 52%, and a specificity of 93%. We refer to Section A.1 for more details on the data generating model. We evaluate the models on an independent dataset of 100,000 patients. Model A and B have AUROCs of 0.754 and 0.791, respectively. The ROC curves have a different shape (Figure 1.1a). For Model A, the curve is higher in the top right and for Model B, in the lower left. Figure 1.1b presents a graphical assessment of calibration by plotting the predicted probabilities (x axis) against the observed probabilities (y axis), calculated using smoothing techniques (Austin and Steyerberg, 2014). Ideally, if predicted and observed probabilities agree (as they do here),

Table 1.1 Common terms

Calibration	Correspondence between predicted and observed probabilities, usually assessed graphically (in calibration plots, e.g., Figure 1.1b). Further, in this work, we are using the ICI as a summary measure of calibration: ICI $= \frac{1}{N}\sum_{i=1}^{N} \lvert\text{predicted}_i - \text{observed}_i\rvert$ where $i = 1,\ldots,N$ indexes datapoints.
Clinical utility of model	Quantification of whether the model will lead to better (clinical) decisions and consequently to an improved health outcome, compared to a default strategy (e.g., not using the model or using another model).
Discrimination	The ability of a model to differentiate between those who do or do not experience the outcome. A model has perfect discrimination if the predicted risks for all individuals who have (diagnostic) or develop (prognosis) the outcome are higher than those for all individuals who do not experience the outcome. In this work, we are using the AUROC as a measure of discrimination. It corresponds to the probability that a randomly selected diseased patient had a higher risk prediction than a randomly selected patient who does not have the disease.
Model evaluation/validation	The terms evaluation and validation are used interchangeably in this work. The aim of model validation is to evaluate (quantify) the model's predictive performance.
Net benefit (NB)	A simple performance measure, with expected benefits and harms put on the same scale so that they can be compared directly. Benefit - (harm $\times$ exchange rate). The expected benefit is represented by the number of patients who have the disease and who will receive treatment (true positives) using the proposed strategy. The expected harm is represented by number of patients without the disease who would be treated in error (false positives) multiplied by a weighting factor (exchange rate) based on the patient's target threshold. See Section 1.5 for details.
Sensitivity	(or true positive rate) The proportion of true positives in individuals with the outcome.
Specificity	(or true negative rate) The proportion of true negatives in individuals without the outcome.
Target threshold (t)	It reflects the probability at which we are indifferent about between two strategies (e.g., administer treatment or not). It captures the relative value the patient places on receiving treatment for the disease, if present, to the value of avoiding treatment if the disease is not present. If the treatment has high efficacy and minimal cost, inconvenience, and adverse effects (e.g., oral antibiotics for community acquired pneumonia), then the target threshold will be low; conversely, if the treatment is minimally effective or associated with substantial morbidity(e.g., radiation for a malignant brain tumour), then the target threshold will be high. For details see Section 1.4.



(a)



(b)

Figure 1.1 Results Toy example 1: (a) ROC curves, and (b) calibration curves for Models A and B.

the plots show a 45-degree line. The ICIs for Models A and B are 0.382 and 0.300, respectively. Based on these results we would conclude Model B demonstrates better predictive performance. But, as we will now show, when we take into consideration the clinical consequences of taking action based on the two models' output, Model A should be preferred.

To take into account the clinical consequences we need to valuate decisions based on using each of the models. To do this, we need information about the benefits and harms of correct and incorrect decisions. Cost-benefit analysis is a widely used approach to valuate decisions (Hunink et al., 2014). Briefly, benefits and harms are summarised using utilities (formally defined in Section 1.4). A utility function assigns a value to each of the four possible decision-outcome combinations stating exactly how beneficial/harmful each decision (surgery or no surgery in our example) is. For each of the four possible outcomes derived from the model output for each patient: true positives (the model correctly flags a patient having residual disease and lymphadenectomy is carried out), false positives (the model incorrectly flags a patient as having residual disease, and lymphadenectomy is carried out), true negatives (the model correctly does not flag a patient having residual disease, and no surgery is carried out), and false negatives (the model incorrectly fails to predict residual disease, and no surgery is carried out), a

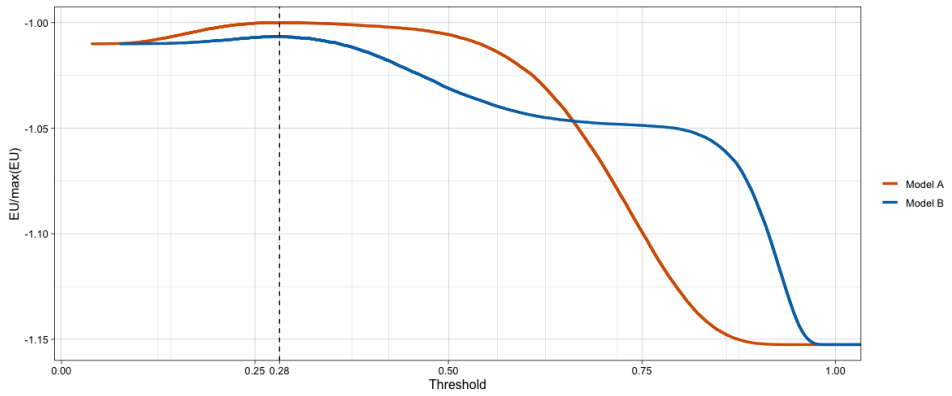


Figure 1.2 Re-scaled expected utility (EU) for different thresholds values for Models A and B. For each threshold value the model with the highest EU should be preferred. max(EU) is the maximum of the EU across both models and therefore does not correspond to the model-specific maxima.

Table 1.2 Utility values from Jung et al. (2020).

Parameter	Description	Value
U_{TP}	Utility for true positives (lymphadenectomy is appropriate and provided)	-28,613
U_{FP}	Utility for false positives (lymphadenectomy is not appropriate but provided)	-14,970
U_{TN}	Utility for true negatives (lymphadenectomy is not appropriate and not provided)	-11,646
U_{FN}	Utility for false negatives (lymphadenectomy is appropriate but not provided)	-37,085

utility value is assigned. Then, we can calculate expected utility (EU) of making treatment decisions (surgery yes or no) based on each of the models.

Figure 1.2 shows a re-scaled version of the EU for each of the models. The graph, allows us to show the relationship between the EU (y axis) for different values of the utilities which are summarised using the threshold concept (x axis) (Pauker and Kassirer, 1980). (We formalise this concept in Section 1.4). For now it is sufficient to state that for each threshold on the x axis the model with the highest EU should be preferred. The vertical dashed line presents the threshold corresponding to utilities values reported by Jung et al. (2020) (Table 1.2), which we assume are the "true" utilities. We see that under these utilities Model A should be preferred. Actually, this is the case for a wide range of threshold values up to 65%. Recall that both the AUROC and ICI, being agnostic to the clinical consequences of using the models, preferred Model B.

Our results are subject to variability due to the stochasticity of the data generation process. To account for this variability we repeat the simulation 1000 times, and we record the percentage of times each model is declared a winner based on each performance measure (Table 1.3). The discrepancy between the performance measures is evident, with the AUROC choosing Model B as the best 100% of the times, whist Model A is the best when we take into consideration the benefits and harms. Interestingly, the ICI is slightly better than a coin flip in choosing the best model.

Table 1.3 Percentage of times each model (A or B) is declared a winner (over 1000 repetitions) based on each performance measure. AUROC: area under the ROC curve, ICI: integrated calibration index, EU: expected utility.

	AUROC	ICI	EU
Model A	0	54	100
Model B	100	46	0

For this example, we used the utilities calculated by Jung et al. (2020) which results in an optimal threshold of 0.28 (vertical line, Figure 1.2). This is very close to the 0.2 threshold previously suggested as clinically reasonable for this scenario (Verbakel et al., 2020). Even though Jung et al. (2020) investigate a different clinical condition and in a different context, our conclusions still hold.

This is because the crucial point in our example is that the ROC curves cross. In many practical applications, it is likely that the ROC curves will cross. One reason for this is that comparisons are likely to be between models with similar performance. In many situations, a model is adjusted a small, incremental step at a time. The result is a series of comparisons between similar models, which are therefore likely to have similar ROC curves. When curves are similar, it is unlikely that one will dominate another – unlikely that one will have a superior sensitivity for all choices of specificity (Hand, 2009). Indeed, empirical evidence supports this hypothesis: Provost et al. (1998) compared a variety of models on ten datasets and found that for only one there was an absolute best performing model. More recently, Shah et al. (2019) presented a similar scenario to ours using real data.

Nevertheless, the problem does not arise only when ROC curves cross. There are situations where both AUROCs and ROCs are almost indistinguishable and yet there is a clear winner between the models when information about the benefits and costs is considered. The next toy example illustrates such a scenario.

1.3.2 Toy example 2

Our second example is a diagnostic scenario inspired by Gail and Pfeiffer (2005). They consider a self-administered questionnaire designed to estimate risk of colorectal cancer in the next 5 years. The decision to be taken is whether an individual should be referred for further evaluation, such as colonoscopy, or not. Gail and Pfeiffer (2005) set the four utilities as: $U_{FN} = -100$ for the possibility of death and morbidity due to failing to detect colorectal cancer, $U_{FP} = -1$ for the risk of bleeding or perforation of the colon, $U_{TP} = -11$ for the risk of bleeding or perforation of the colon and the lowered chance of death or morbidity from colorectal cancer due to early detection, and $U_{TN} = 0$ as a reference value. They further assume prevalence of colorectal cancer is 0.01.

We compare three models to predict the risk of colorectal cancer. Model A includes a covariate with sensitivity 0.9 and specificity 0.2, Model B a covariate with sensitivity 0.2 and specificity 0.6, and Model C includes both covariates. All three models include the same continuous covariate. Details on

Table 1.4 Average performance for each model (A, B or C) over 1000 repetitions. AUROC: area under the ROC curve (AUROC), ICI: integrated calibration index, EU: expected utility. The EU for each model was based on the four utilities specified at the start of Section 1.3.2. Combined these four utilities result in an optimal threshold of 0.01. The calculation follows equation (1.2). Additionally, max(EU) is the maximum of the EU across both models and does not correspond to the model-specific maxima.

	AUROC	ICI*1000	EU/-max(EU)
Model A	0.923	0.907	-1.136
Model B	0.926	0.933	-1.117
Model C	0.928	0.980	-1.105

the data generating model are given in Section A.2. All models are trained and evaluated in independent datasets of 100,000.

Table 1.4 shows the results across 1000 repetitions. The conclusions are subtle. In terms of AUROC Model C is the best, but in terms of ICI Model A has the advantage. Based on the AUROC and ICI together a researcher would probably choose Model B as the best, as the inclusion of the added covariate in Model C does not seem to offer much in terms of AUROC and results in worse ICI. But, when we calculate the expected utility Model C is the best performing model.

These two examples showcase the disadvantage of commonly used performance measures, that is they are insensitive to the clinical implications. Since none of these traditional metrics encapsulate if a model's prediction will result in a favourable change in patient care and outcome, there are recent calls to develop clinical useful models (Chatterjee et al., 2016; Moons et al., 2015; Shah et al., 2019). To achieve this, model evaluation must move beyond traditional metrics and "into clinically meaningful presentations to evaluate improvement in clinical benefits of one model versus another or any model versus none" (Localio and Stack, 2015). In the following section, we introduce the concept of the target threshold that allows us to summarise the different benefits/harms associated with correct/incorrect classifications of a binary outcome into a single number.

1.4 The target threshold

We showed that the traditional metrics have limited value for choosing between models (or treatment decisions) because they do not account for benefits and harms. Here we take on a decision theoretic approach in weighing benefits and harms when using a model to guide clinical decisions (Pauker and Kassirer, 1975, 1980). In the rest of this section, we summarise the benefit and harms of correct and incorrect classifications of a binary outcome into a single number, which we refer to as the target threshold. This approach allows us to achieve two related goals:

1. determine the optimal specified level of risk to use as the target threshold for treatment decisions, and

2. consider the possibility of incorrectly specifying benefits and harms.

Let $Y \in \{0,1\}$ represent a binary outcome of interest. In our previous examples, these were the presence or not of residual disease, and the development or not of colorectal cancer the next 5 years. The observed Y is a realisation of a binary random variable following a Bernoulli distribution with $\pi = P[Y = 1]$. This is the marginal probability of the outcome being present, and consequently, the probability the outcome being absent is $(1-\pi)$.

We introduce utility functions to take into account the benefits or harms of different classifications. A utility function assigns a value to each of the four possible classification-outcome combinations stating exactly how beneficial/costly each action (treat or no treat) is. We assume that people who are classified as positive receive treatment and people who are classified as negative do not receive treatment[3]. We use "treatment" in the generic sense of healthcare intervention which could be a drug, surgery or further testing. Each possible combination of classification (negative and positive) and outcome status (0, 1) is associated with an utility. The four utilities associated with binary classification problems are:

- U_{TP}, the utility of a true positive classification, that is administering treatment to an individual who has the outcome (i.e., treat when necessary),
- U_{FP}, the utility of a false positive classification, that is the utility of administering treatment to an individual who does not have the outcome (i.e., administering unnecessary treatment),
- U_{FN}, the utility of a false negative classification, that is the utility of withholding treatment from an individual that has the outcome (i.e., withholding beneficial treatment), and
- U_{TN}, the utility of a true negative classification, that is the utility of withholding treatment from an individual who does not have the outcome (i.e., withholding unnecessary treatment).

The expected utilities of the two fixed courses of action (or policies) of always treat and never treat are given by

$$EU_{treat} = \pi U_{TP} + (1-\pi)U_{FP}, \tag{1.1a}$$

$$EU_{no\,treat} = \pi U_{FN} + (1-\pi)U_{TN} \tag{1.1b}$$

where EU_{treat} and $EU_{no\,treat}$ are the expected utility of treating and not treating, respectively. In principle, one should choose the course of action with the highest expected utility. When the expected utilities are equal, the decision maker is indifferent on the course of action (Pauker and Kassirer, 1975). Based on classical decision theory, we employ the threshold concept and denote with t the threshold at which the decision maker is indifferent on the course of action (Pauker and Kassirer, 1980). This is the principle of clinical equipoise which exists when all of the available evidence about a course of action does not show that it is more beneficial than an alternative and, equally, does not show that it is less beneficial than the alternative (Turner, 2013). Clinical equipoise is regarded as an "ethically necessary condition in all cases of clinical research" (Freedman, 1987). Based on the threshold concept, an individual should be treated (i.e., classified as positive) if $\pi \geq t$ and should not be treated (i.e., classified

[3] In other words, "classify as positive" and "classify as negative" are equivalent to "treat" and "no treat" actions, respectively.

as negative) otherwise. Having defined t as the value of π of clinical equipoise where the expected benefit of treatment is equal to the expected benefit of avoiding treatment implies $EU_{treat} = EU_{no\ treat}$ or equivalently, $tU_{TP} + (1-t)U_{FP} = tU_{FN} + (1-t)U_{TN}$. Solving for t,

$$\begin{aligned} t &= \frac{U_{TN} - U_{FP}}{U_{TN} - U_{FP} + U_{TP} - U_{FN}} \\ &= \frac{H}{H+B} = \frac{1}{1+\frac{B}{H}}, \end{aligned} \tag{1.2}$$

where $B = U_{TP} - U_{FN}$ is the difference between the utility of administering treatment to individuals who have the outcome and the utility of withholding treatment in those who have the outcome. In other words, B is the benefit for positive prediction, and consequent treatment, among those with the outcome. Similarly, B can be interpreted as the consequence of failing to treat when it would have been of benefit, that is, the harm from a false negative result (compared to a true positive result). Comparably, H is the difference between the utility of avoiding treatment in patients who do not have the outcome and the utility of administering treatment to those who do not have the outcome (i.e., $U_{TN} - U_{FP}$). In other words, H is the consequence of being treated unnecessarily, this is the harm associated with a false positive result (compared to a true negative result).

We henceforth refer to t as the target threshold. Alternative names in the literature are risk threshold (Baker et al., 2009) and threshold probability (Tsalatsanis et al., 2010). It is a scalar function of U_{TP}, U_{FN}, U_{TN} and U_{FP} that determines the cut-off point for calling a result positive that maximizes expected utility. Equation (1.2) therefore tells us that the target threshold at which the decision maker will opt for treatment is informative of how they weigh the relative harms of false positive and false negative results. The main advantage of this decision theoretic approach is there is no need to explicitly specify the relevant utilities, but only the desired target threshold, which in many clinical settings may be more intuitive. Two illustrative examples follow.

> **Example 1:** Assume that for every correctly treated patient (true positive) we are willing to incorrectly treat 9 healthy individuals (false positives). Then we consider the benefit of correctly treating a patient to be nine times larger than the harm of an unnecessary treatment: the harm-to-benefit ratio is 1:9. This ratio has a direct relationship to t: the odds of t equal the harm-to-benefit ratio. That is, $H/B = t/(1-t)$ which is implied by (1.2). For example, t of 10% implies a harm-to-benefit ratio of 1:9 (odds(10%) = 10/90).

> **Example 2:** Assume that not treating an individual with the outcome (false negative) is 9 times worse than treating unnecessarily a healthy individual (false positive). Then we consider the harm of not treating a patient to be nine times larger than the harm of an unnecessary treatment: the harm-to-benefit ratio is again 1:9.

Both statements are equivalent since they result in the same H/B ratio. The difference is how each statement is formulated. In Example 1 we formulated the statement in terms of true/false positives to

showcase that we can derive the target threshold, t, by thinking in terms of "numbers needed" (numbers needed to treat, numbers needed to test, etc). In Example 2 we formulated the statement in terms of false negatives/positives to showcase that we can derive the target threshold, t, by thinking solely in terms of harms of incorrect decisions.

Having derived the target threshold, in the next section we define the expected utility of risk prediction. The expected utility depends on the four basic utilities which is not desirable. We instead use the target threshold and the never treat policy to derive an interpretable simplification of the expected utility which we call the Net Benefit of a risk prediction model. We use the Net Benefit as our model evaluation metric throughout this thesis.

1.5 Net Benefit for risk prediction

In practice, we do not know the probability of the outcome of any given individual. Instead, we need to estimate it, according to a set of covariates. Let $\mathbf{X} \in \mathbb{R}^d$ be a vector of d covariates and define $\pi(\mathbf{x})$ as the conditional class 1 probability given the observed values of the covariates, $\mathbf{x}$: $\pi(\mathbf{x}) = P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$. We are concerned with the problem of classifying future values of Y from the information that the covariates $\mathbf{X}$ contain. Assume we have a prediction model and an estimate of $\pi(\mathbf{x})$, denoted $\hat{\pi}(\mathbf{x})$. We classify an individual as positive if $\hat{\pi}(\mathbf{x}) \geq t$, where t is the target threshold (defined in (1.2)) and as negative otherwise. The expected utility of assigning treatment or not (i.e., classifying positive or negative) at t based on the model's predictions $\hat{\pi}(\mathbf{x})$ can be written as

$$\begin{aligned} EU_{Pred(t)} &= P(\hat{\pi}(\mathbf{x}) \geq t, y = 1)U_{TP} + P(\hat{\pi}(\mathbf{x}) < t, y = 1)U_{FN} + \\ &\quad P(\hat{\pi}(\mathbf{x}) < t, y = 0)U_{TN} + P(\hat{\pi}(\mathbf{x}) \geq t, y = 0)U_{FP} \\ &= \pi TPR_t U_{TP} + \pi(1 - TPR_t)U_{FN} + (1 - \pi)FPR_t U_{FP} + (1 - \pi)(1 - FPR_t)U_{TN} \\ &= \{\pi TPR_t B - (1 - \pi)FPR_t H\} + \{\pi U_{FN} + (1 - \pi)U_{TN}\}, \end{aligned} \tag{1.3}$$

where TPR_t is the true positive rate, i.e., $P(\hat{\pi}(\mathbf{x}) \geq t | y = 1)$ and FPR_t is the false positive rate, i.e., $P(\hat{\pi}(\mathbf{x}) \geq t | y = 0)$. The drawback of this formulation is the need to specify the four utilities. Equation (1.3) can be simplified by considering the expected utility of risk prediction in excess of the expected utility of no treatment. The expected utility of no treatment is given in (1.1b), and so, subtracting this from both sides of (1.3), the expected utility of risk prediction in excess of the expected utility of no treatment is

$$\begin{aligned} EU_{Pred(t)} - EU_{no\ treat} &= \pi TPR_t B - (1 - \pi)FPR_t H \\ &= B\Big\{\pi TPR_t - (1 - \pi)FPR_t \frac{t}{1 - t}\Big\}. \end{aligned} \tag{1.4}$$

This is a Hippocratic utility function because it is motivated by the Hippocratic oath. It incorporates both the principles of *beneficence* (do the best in one's ability) and *non-maleficence* (do no harm). Both principles are considered central notions in bioethics (Childress and Beauchamp, 2001). To be consistent

with the Hippocrates's oath, the modeller chooses the model that has the greatest chance of giving an outcome no worse than the outcome of no treatment. With $B = 1$, (1.4) is defined as the Net Benefit of risk prediction versus treat none (Baker et al., 2009; Vickers and Elkin, 2006). Setting $B = 1$ as the reference level means that Net Benefit is measured in units of true positive predictions. To see this we re-write (1.4) as

$$\mathrm{NB}_{Pred(t)} = \frac{TP_t}{n} - \frac{FP_t}{n}\frac{t}{1-t}, \tag{1.5}$$

where TP_t is number of patients with true positive results, FP_t is number of patients with false positive results, and n is the sample size. To simplify notation we write NB instead of $\mathrm{NB}_{Pred(t)}$. NB gives the proportion of net true positives in the dataset, accounting for the different misclassification costs. In other words, the observed number of true positives is corrected for the observed proportion of false positives weighted by the odds of the target threshold, and the result is divided by the sample size. This net proportion is equivalent to the proportion of true positives in the absence of false positives. For instance, a NB of 0.05 for a given target threshold, can be interpreted as meaning that use of the model, as opposed to simply assuming that all patients are negative, leads to the equivalent of an additional 5 net true positives per 100 patients.

Note that alternative measures, such as the relative utility (Baker, 2009; Baker et al., 2009) and the weighted net reclassification improvement (Pencina et al., 2011), have been suggested. These 3 measures are mathematically interconnected (Van Calster et al., 2013). Consequently, we only focus on NB which will be our main performance measure for model evaluation throughout this thesis. In the above, we have shown NB is defined as a function of the target threshold t, which captures the relative utilities of treatment decisions. It is possible to plot NB for a risk prediction model, across a range of t values, thereby allowing straightforward visual comparison across a range of possible utilities. In addition, it can be used to compare the clinical utility of different models: for example, a basic and extended model fitted on the same data set, or even 2 different models (developed from 2 different data sets) validated on the same independent data set.

Even though NB (and related measures) allow us to evaluate the clinical utility of a prediction model they remain aggregate measures. Ideally, we would like to use models to improve decision-making at the individual level. A crucial step for this is the quantification and communication of uncertainty for individual predictions. As a result, in this work we adopt a Bayesian paradigm for modelling.

1.6 Bayesian modelling in healthcare

Quantification of uncertainty is critically important. This is easily understood as the ability to say, "I don't know" and potentially abstain from providing a diagnosis or prediction when there is a large amount of uncertainty for a given individual. With this ability, additional (human) expertise can be sought or additional data can be collected to reduce the uncertainty to make a more informed decision. To be more precise, in this work we focus on uncertainty quantification for individual-specific probabilities.

We aim to get an answer to the following question: Given the available information, what is the range of plausible values $\pi(\mathbf{x})$ for a specific individual could take?

The Bayesian formalism can provide us with such an answer. Bayesian statistics is an approach to data analysis based on Bayes' theorem (Laplace, 1814), where available knowledge about parameters in a model is updated with the information in observed data (Bernardo and Smith, 2009). The background knowledge is expressed as a prior distribution and is combined with observed data in the form of a likelihood function to determine the posterior distribution. The posterior can also be used for making predictions about future events. The Bayesian paradigm allows for the construction of powerful and flexible statistical models within a rigorous and coherent probability framework. Bayesian analysis has been successfully applied across various research fields such as social sciences, ecology, genetics, medicine and more (see van de Schoot et al. (2021) for a recent review).

Individual-specific quantification of uncertainty is crucial, especially in healthcare applications (Begoli et al., 2019; Kompa et al., 2021). Whilst two (or more) models can perform similarly in terms of aggregate metrics (e.g., AUROC, NB, etc) they can provide very different individual (risk) predictions for the same individual (Li et al., 2020; Pate et al., 2019). This can ultimately lead to different decisions for the individual, with potential detrimental effects. Uncertainty quantification can mitigate this issue since it allows the clinician to abstain from utilising the model's predictions. If there is high predictive uncertainty for an individual, the clinician can discount or even disregard the prediction.

To illustrate this point, we use the standard Bayes posterior from the breast cancer prognostication case study (Section 3.4.1, Chapter 3). The posterior predictive distributions for two patients are displayed in Figure 1.3. The average posterior risk for each patient is indicated by the vertical line at 34 and 35%, respectively. Based solely on these average estimates chemotherapy should be recommended as a treatment option to both patients (see Section 3.4.1 for justification). It is clear, however, that the predictive uncertainty for these two patients is quite different, as the distribution of risk for patient 1 is much more dispersed than the distribution for patient 2. One way to quantify the predictive uncertainty would be to calculate the standard deviation of these distributions, which are 6.9% and 2.8% for patient 1 and patient 2, respectively. Even though both estimates are centred at similar values the predictive uncertainty for patient 1 is more than two times higher than patient 2. Using this information, we could flag patient 1 as needing more information before making a clinical decision.

In addition, for models that predict critical conditions (e.g., sepsis), uncertainty quantification is vital for triaging patients. Clinicians could focus on patients with highly certain estimates, but also further examine patients for whom the model is uncertain with respect to their current condition. For patients with highly uncertain predictions, additional lab values or more frequent monitoring could be requested.

1.7 Thesis overview and outline

In this chapter, we argued that clinical utility needs to be taken into consideration when evaluating risk prediction models. We then showed how traditional performance metrics can provide misleading

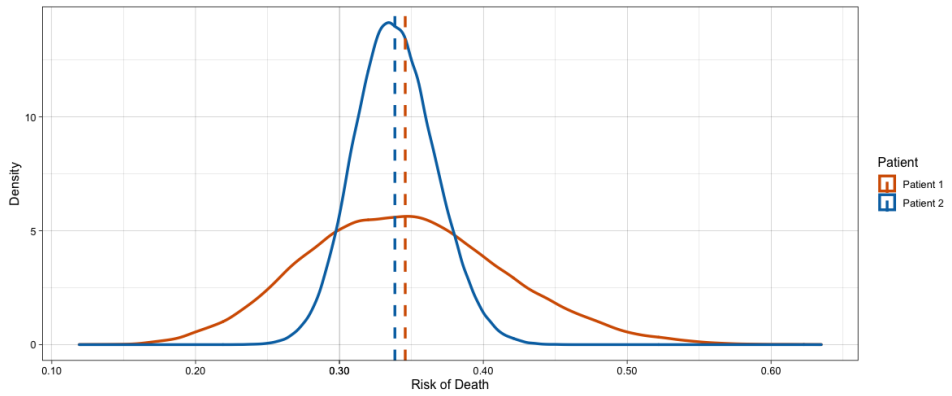


Figure 1.3 Predictive uncertainty for the risk of death in two patients. The posterior predictive distributions reflect the range of risks assigned to these patients, and the mean risk is shown as vertical lines. Despite the fact that both patients have similar mean risks, we may be more inclined to trust the predictions for patient 2 given the lower amount of uncertainty associated with that prediction.

conclusions about a model's utility. We concluded that traditional measures have limited use for (medical) decision-making because they do not incorporate benefits and harms related to treatment decisions. We then presented the work of Pauker and Kassirer (1975, 1980) and introduced the concept of the target threshold, a scalar function of the four basic utilities of prediction, i.e., the optimal level of risk for positive prediction, in the sense of maximizing a person's expected utility given his/her four basic utilities. The main advantage of this decision theoretic approach is there is no need to explicitly specify the relevant utilities, but only the desired target threshold.

As a result, we have a principled and cohesive approach[4], which we call standard Bayes (SB) to evaluate the clinical utility of a risk prediction model:

1. derive an estimate of $\pi(\mathbf{x})$[5],
2. set t, based on acceptable benefits and harms,
3. create classifications based on $\pi(\mathbf{x}) \geq t$, and
4. calculate NB. A model is clinically useful if NB > 0, for the target threshold of interest. In a model comparison setting, Model A should be preferred over Model B if $\text{NB}_{\text{Model A}} > \text{NB}_{\text{Model B}}$.

We use this approach in Chapter 2 to investigate the clinical utility of including information about circulating tumour DNA when predicting response to treatment in metastatic breast cancer. We show that incorporating trajectories of circulating tumour DNA can improve patient decisions.

Nevertheless, the SB approach has an inherit limitation. That is, the harms and benefits of different (mis)classifications are not taken into consideration during model training. The posed model for $\pi(\mathbf{x})$ is

[4]Note, this concept is general. It applies equally to a non-Bayesian as it does to a Bayesian framework. In this work, we focus on a Bayesian paradigm, hence the name, standard Bayes (SB).

[5]this is a point summary of the posterior predictive distribution, such the mean or median.

agnostic on the benefits and harms. In other words, steps 1 and 2 above are independent of each other. Then, a natural question arises: would integrating steps 1 and 2 offer any advantages compared to SB?

In Chapter 3 we propose a novel framework that allows us to integrate steps 1 and 2. The framework allows us to "tailor" model development with the aim of improving performance in the presence of unequal misclassification costs. We call the approach Tailored Bayes (TB). TB allows information about the relative utilities (summarised by the target threshold) to be taken into account whilst training the model. We use simulation studies to showcase when TB is expected to outperform SB. We then apply the methodology to three real-data applications. We show that incorporating this information into the model through our TB approach leads to better treatment decisions.

An interesting observation from these applications of TB was that the relative importance of the covariates in terms of their contribution to the predictions changes. This can have an impact for applications where variable selection is a key (e.g., in many applications in Systems Biology (Vyshemirsky and Girolami, 2008)) or a desirable (e.g., when cost or practicality considerations dictate that wider deployment of the prediction model would require a reduced set of covariates to be used) inferential task. As a result, in Chapter 4 we extend the framework to incorporate a variable selection procedure. Variable selection is a ubiquitous challenge in statistical modelling, especially, with the rise of high-dimensional data. We show that TB favours smaller models (with fewer covariates) compared to SB, whilst performing better or no worse than SB. This pattern is seen both in simulated and real data. In addition, we show that the relative importance of the covariates changes when we consider unequal misclassification costs.

We finish with a concluding chapter that summarises the work that is presented in this thesis and provides an outlook for future work.

Chapter 2

Individualised Predictions of Disease Progression using ctDNA for Metastatic Breast Cancer

Abstract Having previously introduced an approach to evaluate the clinical utility of a risk prediction model (see Section 1.5), in this chapter we apply that approach to assess the clinical utility of circulating tumour DNA (ctDNA) in evaluating response to treatment in metastatic breast cancer (mBC).

Evaluation of response to treatment is essential in the management of metastatic disease. Response to treatment is currently evaluated either using imaging techniques or repeated tumour biopsies. However, both have limitations, such as the increased radiation burden from repeated imaging assessments and the invasiveness of repeated biopsies. To bypass these limitations, ctDNA monitoring has been proposed as an attractive alternative to track response to treatment. In this chapter, we establish the clinical utility of serial ctDNA monitoring in predicting response to treatment in mBC. We found the incorporation of ctDNA allows us to predict radiologic response to treatment which ultimately leads to improved clinical decisions. In addition, the proposed modelling framework allows us to create dynamically updated individual-level predictions. Our results demonstrate the promise of ctDNA monitoring in predicting treatment response and its potential for personalised clinical decision-making. We anticipate our modelling framework will to be a starting point for more sophisticated models in additional cancer types.

Outline This chapter is concerned with a Bayesian modelling framework to predict response to treatment in mBC. We start in Section 2.1 with an introduction to the problem. We then describe the data (Section 2.2.1) and its challenges (Section 2.2.2). To address these challenges, we propose a Bayesian two-stage modelling framework (Sections 2.2.3 and 2.2.4). In Section 2.2.5 we show how we use the model output to create dynamic,

individualised predictions. We then apply the model to our data set and present the results in Section 2.3. Section 2.4 contains a discussion commenting both on the potential impact of the proposed model and on future work directions.

2.1 Introduction

Accurate and reliable evaluation of response to treatment is fundamental in the management of metastatic cancer. Response evaluation during cancer therapy and follow-up of patients with solid malignancies is currently primarily based on radiological assessments according to response evaluation criteria in solid tumours (RECIST) (Eisenhauer et al., 2009). Repeated radiologic assessments are however time consuming, costly, and increase the radiation burden for the patient.

In addition, novel therapies are often aimed at specific mutations. Therefore, for treatment decision-making, up-to-date information about the genomic composition of the tumour lesions is crucial. However, tumour characteristics can change during the course of the disease. Within a single patient, distinct metastatic lesions can be molecularly divergent, and therapeutic stress exerted on tumour cells, particularly by targeted drugs, can dynamically modify the genomic landscape of tumours (Siravegna et al., 2017). Repeated biopsies can provide information about these dynamic adaptations. But obtaining repeated biopsies is not always feasible, given the invasiveness of the procedure. In addition, a tissue biopsy is not always representative of the whole tumour burden due to sampling error, tumour heterogeneity and the dynamic adaptations caused by anticancer treatments (Chung et al., 2007).

To circumvent the above-mentioned limitations regarding radiologic response assessment, as well as the need for up-to-date information about molecular characteristics, there is a requirement for tumour-specific, highly sensitive, non-invasive methods to determine the genomic composition of tumours and to assess response to treatment. A potential method to obtain information about both the genomic composition of tumours and the tumour burden is through detection and quantification of tumour DNA in plasma. Analysis of circulating tumour DNA (ctDNA), commonly referred to as “liquid biopsy”, is a non-invasive way to detect and measure cancer-specific molecular alterations in the blood (Alix-Panabières et al., 2012; Siravegna et al., 2017). The use of ctDNA is emerging as a useful tool in several settings, including cancer diagnosis and prognosis (see Siravegna et al. (2017) for an overview).

Additionally, liquid biopsies can be applied to the monitoring of response and/or resistance to systemic therapy. For example, they can provide temporal measurements of the total tumour burden as well as identify specific mutations that arise during therapy (Bettegowda et al., 2014). Blood-based monitoring of treatment response is particularly attractive because it is minimally invasive, does not involve radiation, and could ultimately be less expensive than current approaches to response assessment (Merker et al., 2018). More importantly, an early prediction of therapeutic response could be useful to distinguish patients most likely to benefit from continued therapy from patients unlikely to benefit, in whom an earlier switch to an alternative therapy may spare toxicity and provide clinical benefit.

At present, however, the clinical value of ctDNA analysis remains controversial, although evidence indicates that the abundance of tumour cells in the blood after treatment can be predictive of response to therapy and, thus, treatment outcomes (Cayrefourcq and Alix-Panabières, 2019; De Rubis et al., 2019). Specifically, data suggests that ctDNA levels within an individual patient correlate with tumour burden over time and that serial assessment of ctDNA may represent a promising approach for monitoring treatment response, with early decreases in ctDNA serving as a predictor of response. Such correlations between changes in ctDNA levels and tumour responses have been demonstrated in small proof-of-principle studies in a variety of cancer types, such as lung cancer (Mok et al., 2015), colorectal cancer (Siravegna et al., 2015), breast cancer (Dawson et al., 2013), and melanoma (Chen et al., 2016).

Specifically, for metastatic breast cancer (mBC), the capacity of serial monitoring of ctDNA to track tumour burden has been previously addressed by several researchers. Dawson et al. (2013) evaluated ctDNA in serially collected blood samples and determined that ctDNA exhibited greater correlation with changes in tumour burden than other biomarkers (circulating tumour cells and CA15-3). Importantly, ctDNA assessment was able to provide the earliest measure of treatment response in 53% of the patients (increased ctDNA levels were detectable on average 5 months before the detection of progressive disease using imaging). In addition, Schiavon et al. (2015) analysed ESR1 mutations in ctDNA to demonstrate the evolution of resistance during therapy in 171 patients with advanced-stage breast cancer. In another study Murtaza et al. (2013) showed that ctDNA could complement current invasive biopsy approaches to identify mutations associated with acquired drug resistance in advanced cancers. These are, however, proof-of-concept studies with small sample sizes, limited follow-up and focused on descriptive statistics.

An additional limitation worth discussing is the specificity of ctDNA quantification. The studies above have used targeted approaches for quantification of ctDNA levels. Targeted approaches involve the detection of previously determined genetic mutations, that is known mutations from the primary tumour which are tracked in plasma. However, tumours are constantly evolving and in advanced stages, genetic alterations in metastases may significantly differ from those identified in the primary tumour (Siravegna et al., 2015). In fact, the driver landscape of metastases may differ from primary cancers (Yates et al., 2017). This highlights the need for untargeted assessment of the tumour load. Untargeted approaches can identify novel changes occurring during tumour treatment and do not require a priori knowledge about the primary tumour's genome (Elazezy and Joosse, 2018). Hence, they offer a more comprehensive evaluation of tumour genomic composition and burden.

Here, we are proposing a probabilistic framework to model both (untargeted) ctDNA levels and the probability of response to treatment in mBC. The model can be used to monitor both the change in ctDNA levels but also to predict radiologic response to treatment. A key feature of the framework is the ability to dynamically update the predictions at the individual level as additional measurements become available. Such a longitudinal assessment, therefore, could enable the clinician to keep track of both the overall tumour burden and disease progression status, ultimately aiding clinical decision-making. For instance, clinicians could identify patients at risk for disease progression and select or adjust systemic

therapy accordingly to improve patient-tailored therapy. We provide evidence of this evaluating the clinical utility of the model using the tools introduced in Chapter 1.

The rest of the chapter is organised as follows. Section 2.2.1 presents the data, alongside the analysis challenges (Section 2.2.2). To overcome these challenges we introduce a two-stage Bayesian modelling framework (Section 2.2.3). In Section 2.3 we present the results and showcase the key feature of our framework, the ability to obtain individualised, dynamically updated predictions. Section 2.4 contains a discussion.

2.2 Methods

In this section we describe the data structure and pre-processing steps (Section 2.2.1) and the associated challenges (Section 2.2.2). To overcome these challenges we propose a two-stage Bayesian modelling framework (Section 2.2.3). In Section 2.2.4 we give details on the priors and computation implementation. We then introduce two approaches to create subject-specific, dynamic predictions (Section 2.2.5). Section 2.2.6 outlines the performance measures we use for model evaluation.

2.2.1 Data collection and pre-processing

Between August 2007 and January 2020, 1023 computed tomography (CT) scans were collected from 135 patients with metastatic breast cancer, recruited as part of the DETECT clinical trial based at Cambridge University Hospital, UK (led by Dr. Emma Beddowes)[1]. The disease status, progressive disease (PD) or not (progression: 1 = yes, 0 = no) at each scan was determined according to the RECIST 1.1 criteria[2](Eisenhauer et al., 2009). More specifically, the non-progressive disease category encompassed all the patients with radiographically determined stable disease, partial or complete response. At baseline 80% of the cohort was in this category. Patients were then followed during standard-of-care (cytotoxic) chemotherapy, targeted or endocrine therapy. For each line of treatment, the CT scan prior to the start of this line of treatment was used as baseline. As a result, the PD outcome at each visit was determined based on the previous baseline assessment. Note that PD is repeatedly measured during follow-up i.e., the PD outcome for each patient is recorded at each visit, and as a result it can switch between 0 and 1 repeatedly over time.

Between August 2010 and October 2019, 992 blood samples were collected from all patients. Blood samples were collected as follows: for patients on chemotherapy samples were taken once per cycle (pre-treatment) and for a minimum of four cycles where applicable. For patients on continuous

[1]The DETECT clinical trial is an ongoing investigator-led pilot study jointly sponsored by Cambridge University Hospital NHS Foundation Trust and the University of Cambridge (IRAS number 214569). The trial is open to any patient with metastatic breast cancer with a clinical performance status of 0-2 which involves consenting patients to give blood samples for research into ctDNA. The main goals are to investigate the clinical utility of ctDNA in predicting treatment response, to identify early treatment failure and to evaluate treatment resistance, and to develop personalised treatment options evaluating biomarkers of response.

[2]The RECIST criteria are used for the assessment of disease progression, and consequently for the assessment of treatment response.

treatments such as endocrine therapy blood samples were taken at routine clinic visits (typically every 3-6 months).This resulted in samples taken up to 24 times per patient (median 6 samples per patient). We used a genome-wide untargeted approach, the ichorCNA algorithm, to quantify the ctDNA fraction (Adalsteinsson et al., 2017). The algorithm quantifies tumour content in cell-free DNA (cfDNA) based on estimation of copy number aberrations in the genome. Blood plasma of patients with cancer contains ctDNA, but this valuable source of information is diluted by much larger quantities of DNA of noncancerous origins, such that ctDNA usually represents only a small fraction of the total cfDNA. The ichorCNA is a widely used tool to quantify the fraction of ctDNA in cfDNA (Ben-David et al., 2018; Manier et al., 2018; Stover et al., 2018; Taylor et al., 2018). Prior to the analysis, the output of the ichorCNA algorithm was arcsine transformed, which is recommended for this type of data (Ahrens et al., 1990). All results are provided in this transformed scale.

For statistical analysis, the first CT scan was designated time zero. Blood samples taken before these days were assigned negative time values. Also, ctDNA data collected after the last CT scan were not used[3]. Further, we excluded 2 patients due to missing covariate values resulting in 133 patients included in the analysis. The covariates include time (years) since baseline visit, ER status (1 = positive; 0 = negative), HER2 status (1 = positive, 0 = negative), and treatment regime and treatment duration (years). Treatment regime was defined based on the base treatment compound or the add-on if the base treatment was missing.

We assessed the performance of the model in predicting PD by applying it to held-out test data corresponding to last CT per patient (for the patients with more than 1 scans). The rest of the data was used for training. This temporal validation approach (Miller et al., 1991), matches how such a model would be used in clinical practice. All results are reported as temporal validation performance metrics, obtained by applying the trained model to the test set. Finally, to illustrate how the model can be used to dynamically update predictions at the individual level we excluded 3 patients from the training dataset.

2.2.2 Data challenges

Here, we outline some of the data challenges which will lead us to introducing our proposed model in the next section.

First, the CT scans are taken at irregular time points. Figure 2.1 illustrates this for a random sample of six patients. We see the variability in both the length of the follow-up for each patient and the actual timing of the CT scans. For example, Patient 1 was measured 12 times within a 4.5-year period whilst Patient 2 was measured only five times within a 1.5-year period. This makes simple approaches such as performing separate t-tests comparing the two groups (i.e., progression yes/no at specific timepoints[4]) infeasible. A more efficient approach that does not require common measurement times across all patients is a random effects model (Laird and Ware, 1982; Raudenbush and Bryk, 2002) (sometimes also

[3]We recognise that ctDNA data collected after the last CT scan could have been used in the analyses (and would in principle improve efficiency of estimation in the training dataset). Their exclusion was done to reflect how in practice the model would be used for prediction.

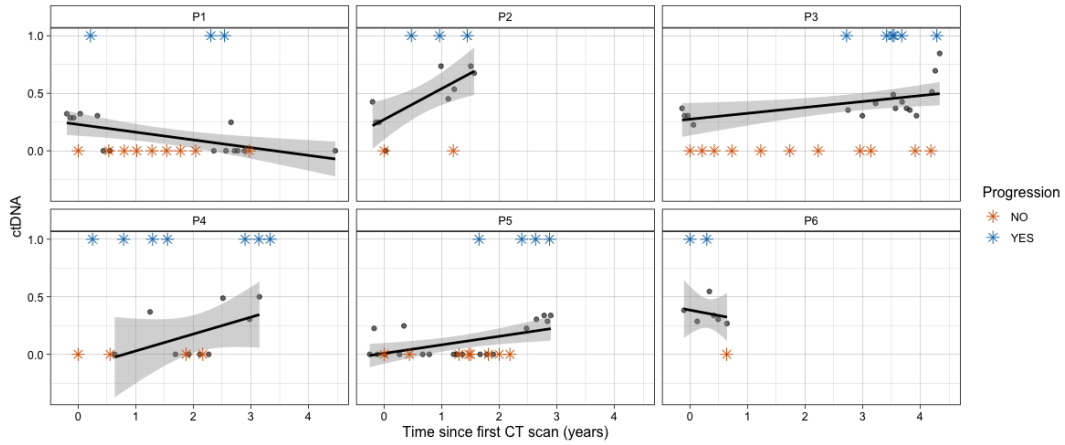


Figure 2.1 Longitudinal ctDNA (grey dots) and PD (asterisks) measurements for six patients (panels). Ordinary least squares (OLS) summaries of the ctDNA profiles (solid lines) with 95% confidence intervals are given per subject. Not only the direction (increase or decrease) of ctDNA varies but the rate as well.

known as mixed effect models, hierarchical models, or multilevel models). This model also captures the correlation between the progression measurements over time from the same patient.

Second, the CT and ctDNA measurements are not necessarily taken at the same timepoint (Figure 2.1). This poses challenges on how to incorporate information about the longitudinal ctDNA measurements in the model for disease progression (our primary outcome). Approaches that collapse the observed longitudinal information over time using a summary measure (e.g., overall average, change score, maximum, minimum, achievement of a threshold, etc.) which then could be used as a covariate in a (logistic) model for PD are inadequate for two reasons. First, it may be difficult to decide which summary measure to use, and as there are many candidate measures, there is perhaps a problem of multiple comparisons. Second, we would be interpolating the chosen summary measure to the timepoint the CT scan was taken.

To address these challenges we implement a two-stage model: first a linear random effects model is fitted to the longitudinal ctDNA data and predicted values of the random coefficients are computed; then a logistic random effects regression model is fitted to the binary outcome, using the random coefficients as covariates. That is, the unobserved underlying parameters (random effects) in the longitudinal ctDNA model are included as covariates to the primary model on disease progression. This two-stage random effects modelling approach has many advantages. First, it allows us to handle the irregularly measured time-varying ctDNA. Second, the model allows to incorporate the inter-subject variability in terms of ctDNA profiles, which is illustrated in Figure 2.1. The ordinary least squares summaries of the ctDNA measurements (grey dots) are given. Not only the direction (increase or decrease) of ctDNA varies but the rate as well. Third, it allows us to deal with the time-gap between the CT and ctDNA

[4]Note that each patient can transition between PD=0 and 1 (or between PD=1 and 0) status at any timepoint during follow-up.

measurements, by "predicting" ctDNA values at the timepoints when CT scans were taken. Fourth, we can take into account the inter-subject association of the successive CT scans. Lastly, the model allows to obtain dynamic personalized prediction of future longitudinal outcome trajectories and risks of disease progression at any time, given the subject-specific outcome profiles up to the time of prediction. A key feature of these dynamic prediction frameworks is that the predictive measures can be dynamically updated as additional longitudinal measurements become available for the target subjects, providing instantaneous risk assessment.

The most common data types for such two-stage models are for continuous longitudinal and time-to-event data (e.g., Rizopoulos (2012); Self and Pawitan (1992); Tsiatis et al. (1995)). For instance, a binary variable indicating whether a patient survives five years after the start of treatment is a common outcome in medicine, but the exact survival time (possibly censored) is more informative. However, in many applications, the exact timing of the event is either unknown or not informative. For instance, for our application, the time of disease progression within a single treatment regime is unknown. Thus, a continuous/binary (ctDNA/PD) model is appropriate.

2.2.3 Modelling framework

We consider the problem of predicting the probability of disease progression at any timepoint during follow-up, in a population of subjects undergoing treatment for mBC, based on longitudinal measurements of ctDNA. Thus, for each subject, we have a set of continuous longitudinal measurements, and a set of longitudinal primary outcomes (a binary variable indicating whether the subject has progressed or not at the specific timepoint). A goal of the analysis is to provide, for each subject, an estimated probability of progression and a quantification of the uncertainty of this estimate. This probability and associated uncertainty estimate could be used for deciding whether the tumour is responding to the treatment and whether the subject could benefit from changing treatment.

The statistical model has two components, one to model the longitudinal ctDNA measurements, and one to model the longitudinal PD status. The data available for each patient i ($i = 1, \ldots, n$) are

$$[\text{ctDNA}_i(s_{i1}), \text{ctDNA}_i(s_{i2}), \ldots, \text{ctDNA}_i(s_{im_i}), \text{PD}_i(t_{i1}), \text{PD}_i(t_{i2}), \ldots, \text{PD}_i(t_{ik_i}), \mathbf{X}_i]$$

where $\text{ctDNA}_i(s_{i1})$ is the value of ctDNA at time s_1, $\text{PD}_i(t_{i1})$ is the progressive disease status at time t_1, and $\mathbf{X}_i$ are the rest of the covariates (see Section 2.2.1). Note that timepoints s and t are indexed by i since patients are measured at different timepoints.

The two-stage approach

In the first stage, the evolutions of the repeated ctDNA measurements are summarised by random effects obtained by fitting a linear random effects model, and in the second stage, the resulting random effects are used as covariates in a logistic regression model to predict the risk of PD.

Let $\text{ctDNA}_i(s_{ij}) = \text{ctDNA}_i(s)$ represent the continuous ctDNA longitudinal measurements for individual i at measurement time j $(j = 1,\ldots,m_i)$. The first stage model can be written as follows

$$\text{ctDNA}_i(s) = \mathbf{X}_{1i}(s)\boldsymbol{\beta}_1 + \mathbf{Z}_{1i}(s)\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{2.1}$$

where vectors $\mathbf{X}_{1i}(s)$ and $\mathbf{Z}_{1i}(s)$ are p and q dimensional covariates corresponding to fixed and random effects, respectively. They include the covariates of interest such as treatment and time. The vector $\boldsymbol{\beta}_1$ contains the fixed effect parameters. The vector $\boldsymbol{b}_i = (b_{i1},\ldots,b_{iq})'$ contains the random effects for patient i and it is distributed as $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_b)$. Finally, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{im_i})'$ is a vector of measurement errors with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0},\sigma_\varepsilon^2\mathbf{I}_{m_i})$. We further assume that $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are independent.

In the second stage, point estimates of the subject-specific random effects, $\boldsymbol{b}_i$ from stage 1 are used as predictors in a random effects logistic regression model with the disease progression as the outcome. Let $\text{PD}_i(t_{ij}) = \text{PD}_i(t)$ represent the binary measurements (progression yes/no) for individual i at time j $(j = 1,\ldots,k_i)$. Then, the second stage model can be written as

$$\text{logit}\{p(\text{PD}_i(t) = 1)\} = \mathbf{X}_{2i}(t)\boldsymbol{\beta}_2 + \mathbf{Z}_{2i}(t)\boldsymbol{u}_i + \boldsymbol{\gamma}'\hat{\boldsymbol{b}}_i \tag{2.2}$$

where $\hat{\boldsymbol{b}}_i$ is the estimated vector of random effects from (2.1), $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters, and $\boldsymbol{u}_i$ is a vector of unknown random effects, distributed as $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_u)$. Hence, the two models are linked through $\boldsymbol{b}_i$. Note that some (or all) covariates in $\mathbf{X}_{2i}$ can overlap with vector $\mathbf{X}_{1i}$ in model (2.1). In addition, we choose to include the random effects themselves from stage 1 as covariates in stage 2 which is similar to previous work (Baghfalaki et al., 2014; Choi et al., 2014; He and Luo, 2016; Wulfsohn and Tsiatis, 1997). We did not consider different representations of the random effects $\boldsymbol{b}_i$ in (2.2), but we recognise other options are available, such as using the trend $(b_{i0} + b_{i1}t)$ as a covariate, which we plan to study in future extensions of the model.

A simplification

Here we implement a convenient and widely used simplification of (2.1) and (2.2). In (2.1), letting the j^{th} row of $\mathbf{Z}_{1i}$ equal $(1, s_{ij})$ for $j = 1,\ldots,m_i$, so that $\boldsymbol{b}_i = (b_{i0}, b_{i1})'$ corresponding to the random intercept and slope for subject i, and with

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma_{11}^2 & \sigma \\ \sigma & \sigma_{12}^2 \end{bmatrix} \tag{2.3}$$

we obtain a random slopes and intercepts model (Laird and Ware, 1982). In (2.2) letting the j^{th} row of $\mathbf{Z}_{2i}$ equal 1, so that $\boldsymbol{u}_i = u'_{i0}$ corresponding to the random intercept for subject i, then

$$\boldsymbol{\Sigma}_u = \sigma_u^2 \tag{2.4}$$

We justify this choice on the random effects in Appendix B. This parameterization postulates that patients who have a lower/higher level for the ctDNA at baseline (i.e., intercept) or who show a steeper increase/decrease in their longitudinal trajectories (i.e., slope) are more likely to experience the event.

Given this simplification the unknown parameter vector is $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \boldsymbol{\Theta}_2')'$, where $\boldsymbol{\Theta}_1'$ and $\boldsymbol{\Theta}_2'$ refer to the parameter vectors for the 1st stage and 2nd stage models, respectively. Namely, $\boldsymbol{\Theta}_1 = (\boldsymbol{\beta}_1, \sigma_\varepsilon, \boldsymbol{\Sigma}_b)'$ and $\boldsymbol{\Theta}_2 = (\boldsymbol{\beta}_2, \boldsymbol{\gamma}, \sigma_u^2)'$.

2.2.4 Bayesian inference

To infer the unknown parameter vector $\boldsymbol{\Theta}$, we use Bayesian inference based on Markov chain Monte Carlo (MCMC) simulations.

Bayesian inference has many advantages. First, MCMC algorithms can be used to estimate the posterior distributions of the parameters, while likelihood-based estimation only produces a point estimate of the parameters. Second, Bayesian inference provides better performance in small samples compared to likelihood-based estimation (Lee and Song, 2004).

In addition, an attractive feature of our two-stage formulation is that the model can be fitted with readily available software, as follows. First, the linear random effects model (equation (2.1) with the simplification (2.3)) is fitted to the longitudinal ctDNA data, and predicted summaries $\hat{\boldsymbol{b}}_i$ for the random effects are computed. We use the means of the posterior distribution of $\hat{\boldsymbol{b}}_i$ as summary measures. Then, the predicted values of the random effects are used as covariates in the generalized logistic model for PD (equation (2.2) with the simplification given in (2.4)).

Fitting the two-stage model under a Bayesian framework requires first to estimate the fixed effects of the Bayesian linear random effects model as well as the means of the posterior distribution of the random effects, followed by estimating the logistic model parameters and random effects in the second stage. This requires specification of the prior distributions for the parameters of the submodels in each stage.

We use vague priors on all elements in $\boldsymbol{\Theta}$. Specifically, except the intercept terms, all other elements in $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ are $\mathcal{N}(0, 100)$. The intercept terms as well as σ_ε are assigned Student-t prior distribution with mean 0, 3 degrees-of-freedom, and scale 2.5 (Gelman et al., 2008). We parametrise the covariance matrix $\boldsymbol{\Sigma}_b$ in terms of a correlation matrix $\boldsymbol{\Omega}_b$ and a vector of standard deviations $\boldsymbol{\sigma}$ through

$$\boldsymbol{\Sigma}_b = \boldsymbol{\sigma}' \boldsymbol{\Omega}_b \boldsymbol{\sigma} \tag{2.5}$$

Priors are then specified for the parameters on the right hand side of the equation. For $\boldsymbol{\Omega}_b$, we use the LKJ-Correlation prior with parameter $\zeta = 1$, i.e.,

$$\boldsymbol{\Omega}_b \sim LKJ(1)$$

which corresponds to a uniform density over correlation matrices of the respective dimension i.e., all correlations matrices are equally likely a priori (Figure 2.2). Following Gelman et al. (2006) we use a

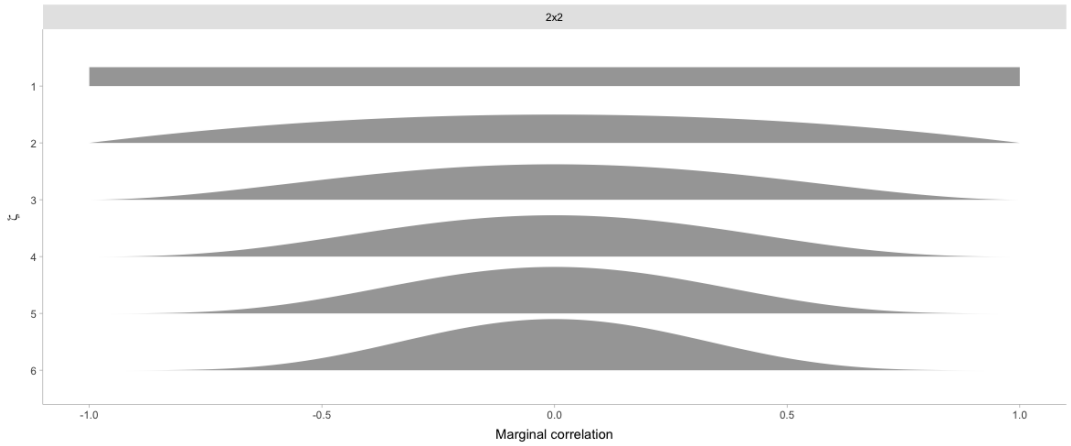


Figure 2.2 Marginal correlation for $LKJ(\zeta)$ prior on a 2×2 matrix size.

half Cauchy prior with scale parameter equal to 2 for every element of $\boldsymbol{\sigma}$. We also use we use a half Cauchy prior for σ_u with scale parameter equal to 2.

The posterior samples are obtained from the full conditional of each unknown parameter using Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) and the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014). HMC algorithms produce samples, which are less autocorrelated than those of other samplers such as the random-walk Metropolis algorithm. Both HMC and NUTS samplers are implemented in `Stan`, which is a probabilistic programming language (Carpenter et al., 2017). We use the `brms` `R` package version 2.14.4 to implement the models (Bürkner, 2017).

To monitor Markov chain convergence, we use the trace plots and view the absence of apparent trends in the plot as evidence of convergence. In addition, we use the Gelman–Rubin diagnostic to ensure the scale reduction $\hat{R}$ of all parameters are smaller than 1.1 (Gelman et al., 2013). After fitting the model to the training dataset (the dataset used to build the model) using Bayesian approaches via MCMC, we obtain M (e.g., $M = 10,000$ after burn-in) samples for the parameter vector $\boldsymbol{\Theta}$.

2.2.5 Dynamic Prediction Framework

We illustrate how to make predictions for a new subject I at time t, based on the available ctDNA history $\text{ctDNA}_I^{\{t\}} = \{\text{ctDNA}_I(s_{Ij}); 0 \leq s_{Ij} \leq t\}$, the PD history $\text{PD}_I^{\{t\}} = \{\text{PD}_I(t_{Ij}); 0 \leq t_{Ij} < t\}$ and the covariate history $\bar{\boldsymbol{X}}_I^{\{t\}} = \{\boldsymbol{X}_{1I}(s_{Ij}), \boldsymbol{X}_{2I}(t_{Ij}), \boldsymbol{Z}_{1I}(s_{Ij}), \boldsymbol{Z}_{2I}(t_{Ij}); 0 \leq s_{Ij}, t_{Ij} \leq t\}$ up to time t, based on the fitted models given in equation (2.1) with the simplification (2.3), and equation (2.2) with the simplification given in (2.4).

We are interested in obtaining predictions for the probability of progressive disease (PD) at time t. We propose two approaches to create these predictions, based on different use of the histories. The first approach relies solely on the available ctDNA history i.e.,

$$p(\text{PD}_I = 1 | \text{ctDNA}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}) \tag{2.6}$$

Equation (2.6) is the probability of progressive disease at time t based on the available ctDNA and covariate histories up to time t. We refer to this approach as "Dynamic Predictions". To do this, the key step is to obtain the mean for patient I's random effects vector $\boldsymbol{b}_I$ from its posterior distribution $p(\boldsymbol{b}_I|\text{ctDNA}_I^{\{t\}}, \boldsymbol{\Theta}_1)$. Likely, for the linear mixed model this has a closed form solution, see next section for details.

We further need samples from u_{I0}. Let its posterior distribution be $p(u_{I0}|\text{ctDNA}_I^{\{t\}}, \text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2)$ which simplifies to $p(u_{I0}|\text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2)$, since the ctDNA history, $\text{ctDNA}_I^{\{t\}}$, is incorporated through the vector $\hat{\boldsymbol{b}}_I$ which is introduced as fixed effects in equation (2.2). In this approach, we do not use information about the PD history so the posterior further simplifies to $p(u_{I0}|\boldsymbol{\Theta}_2)$ i.e., the posterior depends only on the parameters in the $\boldsymbol{\Theta}_2$ vector. Hence, the posterior of u_{I0} reduces to $p(u_{I0}|\sigma_u^2)$, which actually is the prior distribution of the random effects. Then, (2.6) becomes

$$p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}) = \int p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}, \boldsymbol{\Theta}_2, u_{I0}) p(u_{I0}|\sigma_u^2) du_{I0} \tag{2.7}$$

We use the MCMC samples to approximate this expectation. Specifically, conditional on the m^{th} posterior sample $\boldsymbol{\Theta}_2^{(m)}$, we draw the m^{th} sample from

$$p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}, \boldsymbol{\Theta}_2^{(m)}, \hat{u}_{I0}^{(m)}) \tag{2.8}$$

where in each iteration, the predicted random effect $\hat{u}_{I0}^{(m)}$ is drawn from $\mathcal{N} \sim (0, \tilde{\sigma}_u^{2(m)})$, where $\tilde{\sigma}_u^{2(m)}$ is the sampled value of σ_u^2 at the m^{th} iteration. The above steps can be repeated to obtain an updated prediction at any time a new measurement is recorded for this patient, by adjusting the vector $\text{ctDNA}_I^{\{t\}}$ and the corresponding $\boldsymbol{Z}_{1I}$, $\boldsymbol{X}_{1I}$ and $\boldsymbol{X}_{2I}$ matrices.

The second approach creates predictions using both the available ctDNA and PD histories, i.e.,

$$p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \text{PD}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}) \tag{2.9}$$

Equation (2.9) is the probability of progressive disease at time t based on the available ctDNA, PD and covariate histories up to time t. We refer to this approach as "Individualised Dynamic Predictions". Now, we need to obtain samples from both $\boldsymbol{b}_I$ and u_{I0}. For $\boldsymbol{b}_I$ we follow the same procedure as before and outlined in the next section. For u_{I0}, let its posterior distribution be $p(u_{I0}|\text{ctDNA}_I^{\{t\}}, \text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2)$ which simplifies to $p(u_{I0}|\text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2)$ using the same reasoning as above. Then, (2.9) becomes

$$\begin{aligned} &p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \text{PD}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}) \\ &= \int p(\text{PD}_I = 1|\text{ctDNA}_I^{\{t\}}, \text{PD}_I^{\{t\}}, \bar{\boldsymbol{X}}_I^{\{t\}}, \boldsymbol{\Theta}_2, u_{I0}) p(u_{I0}|\text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2) du_{I0} \end{aligned} \tag{2.10}$$

This posterior expectation can be used to make predictions for a new patient, exploiting the information that we already have about the patient. To do this, the key step is to obtain samples from the

posterior

$$p(u_{I0}|\text{PD}_I^{\{t\}}, \boldsymbol{\Theta}_2^{(m)}) = \frac{p(u_{I0}, \text{PD}_I^{\{t\}}|\boldsymbol{\Theta}_2^{(m)})}{p(\text{PD}_I^{\{t\}}|\boldsymbol{\Theta}_2^{(m)})} \propto p(\text{PD}_I^{\{t\}}|u_{I0}, \boldsymbol{\Theta}_2^{(m)}) p(u_{I0}|\boldsymbol{\Theta}_2^{(m)}) \tag{2.11}$$

where the first equality is from Bayes theorem. Following Wang et al. (2017) for each of $\boldsymbol{\Theta}_2^{(m)}$, $m = 1, \ldots, M$, we use adaptive rejection Metropolis sampling (Gilks et al., 1995) to draw 30 samples of u_{I0} and retain the final sample. This process is repeated for the M saved values of $\boldsymbol{\Theta}_2$.

Prediction of $\boldsymbol{b}_{1i}$

Lastly, to perform all the above calculations we need first to estimate the mean of the subject-specific random effects $\hat{\boldsymbol{b}}_I$. Denote the posterior of $\boldsymbol{b}_I$ with $p(\boldsymbol{b}_I|\text{ctDNA}_I^{\{t\}}, \boldsymbol{\Theta}_1)$. For linear models it follows from standard results on conditional multivariate normal densities that the posterior density is multivariate normal. The mean of the posterior has the following form

$$\mathbb{E}[\boldsymbol{b}_I|\text{ctDNA}_I^{\{t\}}, \boldsymbol{\Theta}_1] = \Sigma_b \boldsymbol{Z}_{1I}' (\sigma_\varepsilon^2 \boldsymbol{I}_{m_i} + \boldsymbol{Z}_{1I} \boldsymbol{\Sigma}_b \boldsymbol{Z}_{1I}')^{-1} (\text{ctDNA}_I - \boldsymbol{X}_{1I} \boldsymbol{\beta}_1) \tag{2.12}$$

Then, we set $\hat{\boldsymbol{b}}_I$ as

$$\begin{aligned} \hat{\boldsymbol{b}}_I = \mathbb{E}[\boldsymbol{b}_I|\text{ctDNA}_I^{\{t\}}] &= \int \mathbb{E}[\boldsymbol{b}_I|\text{ctDNA}_I^{\{t\}}, \boldsymbol{\Theta}_1] p(\boldsymbol{\Theta}_1|D) d\boldsymbol{\Theta}_1 \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \Sigma_b^{(m)} \boldsymbol{Z}_{1I}' (\hat{\sigma}_\varepsilon^{2(m)} \boldsymbol{I}_{m_i} + \boldsymbol{Z}_{1I} \Sigma_b^{(m)} \boldsymbol{Z}_{1I}')^{-1} (\text{ctDNA}_I - \boldsymbol{X}_{1I} \boldsymbol{\beta}_1^{(m)}), \end{aligned} \tag{2.13}$$

which is the mean of the posterior distribution of the random effects. D denotes the training data, and $p(\boldsymbol{\Theta}_1|D)$ is the posterior of the fixed effects parameters. Note this is a conditional mean, i.e., it is updated with ctDNA measurements. For instance, suppose for patient I we are interested in the predicted probability at a new timepoint t', then the ctDNA history is updated to $\text{ctDNA}_I^{\{t'\}}$. We can then dynamically update the posterior mean distribution using equation (2.13), i.e., calculate $\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[\boldsymbol{b}_I|\text{ctDNA}_I^{\{t'\}}, \boldsymbol{\Theta}_1^{(m)}]$, after adjusting appropriately the $\boldsymbol{Z}_{1I}$ and $\boldsymbol{X}_{1I}$ matrices. This closed-form solution for the random effects estimates is an advantage of the linear model in (2.1).

To summarise, with the M posterior samples for patient I's random effects $\boldsymbol{b}_I$ and u_{I0}, predictions can be obtained by simply plugging in realisations of the parameter vector and random effects vector $\{\boldsymbol{\Theta}^{(m)}, \boldsymbol{b}_I^{(m)}, u_{I0}^{(m)}, m = 1, \ldots, M\}$. For example, the m^{th} sample of the PD outcome at time t, $\text{PD}_I(t)$, is obtained from equation (2.2):

$$p(\text{PD}_I^{(m)}(t) = 1) = \text{expit}\{\mathbf{X}_{2I}(t) \boldsymbol{\beta}_2^{(m)} + \hat{u}_{I0}^{(m)} + \boldsymbol{\gamma}'^{(m)} \hat{\boldsymbol{b}}_I\}$$

where $\text{expit}(x) = 1/(1+\exp(-x))$, $\hat{\boldsymbol{b}}_I$ is estimated as in (2.13) and $\hat{u}_{I0}^{(m)}$ is estimated depending on whether we are using the Dynamic Predictions or Individualised Dynamic Predictions approach. In the former case, $\hat{u}_{I0}^{(m)}$ is estimated by sampling from the prior, using the procedure (2.8). In the latter case, it is estimated by sampling from the posterior, using the procedure (2.11). All prediction results can then be obtained by calculating simple summaries (e.g., mean, variance, quantiles) of the posterior distributions of M samples $\{p(\text{PD}_I^{(m)}(t) = 1), m = 1, \ldots, M\}$.

2.2.6 Assessing predictive performance

It is essential to assess the performance of the proposed predictive measures. Here, we focus on the average of the posterior PD probability which we denote with $\pi(t|\text{ctDNA}^{\{t\}}, \text{PD}^{\{t\}})$. Specifically, we assess the discrimination (how well the models discriminate between patients who had the event from patients who did not) using the area under the ROC curve (AUROC). We assess the calibration (the agreement between observed and predicted probabilities) using the integrated calibration index (ICI) (Austin and Steyerberg, 2019). We assess the overall performance (how well the models predict the observed data) using the expected Brier score (BS) (Brier, 1950). Finally, we assess clinical utility (how well the models classify at target thresholds) using the Net Benefit (NB).

Area under the ROC curve

First we define the true positive (TPR) and true negative (TNR) rates (see also Table 1.1) at any given cut point $c \in (0,1)$ as $TPR = P\{\pi > c|\text{PD} = 1\}$ and $TNR = P\{\pi \leq c|\text{PD} = 0\}$, where $\pi = \pi(t|\text{ctDNA}^{\{t\}}, \text{PD}^{\{t\}})$ the predicted probability of progressive disease and PD is the observed disease status that equals to 1 if the subject experiences progressive disease at timepoint t, and 0 otherwise[5]. Then, the TPR and TNR can be estimated from the empirical distribution of the predicted probabilities (i.e., $\hat{\pi}$) among either cases or controls. With the estimation of TPR and TNR, the ROC curve can be constructed for all possible cut points $c \in (0,1)$ and the corresponding AUC can be estimated. In general, AUROC = 1 indicates perfect discrimination and AUROC = 0.5 means no better than random guess. We use the pROC package (version 1.16.2) to estimate the AUROC (Robin et al., 2011).

Integrated Calibration Index

The Integrated Calibration Index (ICI) provides a numerical summary of model calibration over the observed range of predicted probabilities (Austin and Steyerberg, 2019). For each subject in the test sample, a binary outcome (PD) is observed and a predicted probability of the disease progression $\hat{\pi}$ is estimated. In addition, let $\hat{\pi}^c$ denote the smoothed probability based on the loess calibration curve; $\hat{\pi}^c$ is an estimate of the observed probability of the outcome that corresponds to the given predicted probability.

[5]For each subject we are predicting the probability of progressive disease (π) at the last visit (which is not used for model training, see Section 2.2.1). Consequently, we only have one predicted probability for each subject and we do not take into account the actual time this prediction is made for.

A common approach is to estimate $\hat{\pi}^c$ using a locally weighted least squares regression smoother (i.e., the loess algorithm (Cleveland, 1991)). We can then calculate the absolute difference between these two quantities: $f(\hat{\pi}) = |\hat{\pi} - \hat{\pi}^c|$. Let $\phi(\hat{\pi})$ denote the density function of the distribution of predicted probabilities. Then, we define ICI $= \mathbb{E}[f(\hat{\pi})]$, where the expectation is over $\phi(\hat{\pi})$. Hence, the ICI is the weighted difference between smoothed observed proportions (i.e., $\hat{\pi}^c$) and predicted probabilities (i.e., $\hat{\pi}$), in which observations are weighted by the empirical density function of the predicted probabilities (i.e., $\phi(\hat{\pi})$). This is equivalent to integrating $f(\hat{\pi})$ over the distribution of the predicted probabilities. An estimator of ICI is $\widehat{ICI} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\pi}_i - \hat{\pi}_i^c|$, where $i = 1,\ldots,N$ indexes the subjects in the test set. Contrary to the AUROC, the ICI does not have an intuitive interpretation, but given its definition, lower values indicate better performance.

Brier Score

Similarly to the ICI, the expected Brier Score (BS) is defined as BS $= \mathbb{E}[(\text{PD} - \hat{\pi})^2]$, where the observed disease status PD equals to 1 if the subject experiences progressive disease and 0 otherwise. An estimator of BS is $\widehat{BS} = \frac{1}{N}\sum_{i=1}^{N}(\text{PD}_i - \hat{\pi}_i)^2$, where $i = 1,\ldots,N$ indexes the subjects in the test set. A BS = 0 indicates perfect prediction and BS = 0.25 means no better than random guess.

Net Benefit

For convenience we restate the Net Benefit equation:

$$\text{NB} = \frac{TP_t}{N} - \frac{FP_t}{N}\frac{t}{1-t}, \tag{2.14}$$

where TP_t is number of patients with true positive results, FP_t is number of patients with false positive results, and N is the test sample size.

From (2.14) we see the NB of a model is the difference between the proportion of true positives (TP_t) and the proportion of false positives (FP_t) weighted by the target threshold t (expressed on the odds scale) at which an individual is classified as "high risk" (see Section 1.5 for details). The interpretation of the target threshold is given in Section 1.4. Briefly, in this context, it implies the relative value for either changing treatment if PD was likely or avoiding treatment change if PD was not likely. For example, if the clinician views unnecessary treatment change in nine women as an acceptable cost for correctly treating one woman with PD, this is equivalent to changing treatment in all women with $\geq 10\%$ risk. To show how to such information we use a range of target thresholds and plot NB on the y axis against alternative values of t on the x axis.

2.3 Results

In this section, we apply the proposed model and prediction process to our dataset. For all results in this section, we run two parallel MCMC chains and run each chain for 40,000 iterations. The first 2000

Table 2.1 Area under the ROC curve (AUROC), Integrated Calibration index (ICI), and Brier score (BS) for Models A and B evaluated on the test set.

	AUROC	ICI	BS
Model A	**0.750**	**0.153**	**0.222**
Model B	0.711	0.165	0.238

iterations are discarded as burn-in and the inference is based on the remaining 38,000 iterations from each chain. Good mixing properties of the MCMC chains for all model parameters are observed in the trace plots. The scale reduction $\hat{R}$ of all parameters are smaller than 1.1.

Since our aim is to evaluate whether ctDNA helps predicting progressive disease we compare two models: Model A includes information about ctDNA, whilst Model B does not. This is achieved by including (Model A) and excluding (Model B) the $\boldsymbol{b}_i$ term from the 2nd stage model.

Initially, we included treatment duration in both 1st and 2nd stage models, but we excluded it later from the 2nd stage models as it did not offer any predictive performance gain. Also, there is no enough information in the data to estimate the parameters for two treatment regimes, treatment 10 and 12. The parameters corresponding these treatments are not identifiable within the fitted models. Hence, we recommend against using the model for patients treated with either regimes. The posterior parameter estimates are given in Appendix B (Tables B.1 and B.2).

We compare the two candidate models in terms of discrimination, calibration and overall performance in the test set. We present AUROC, ICI, and BS score in Table 2.1. Under all performance measures Model A performs better than Model B.

We further compare the models in terms of their clinical utility and present NB in Figure 2.3. The interpretation of this plot is that the model with the highest NB at a particular target threshold has the highest clinical value. We note Model A has the highest NB across all target thresholds, except for very low or high values of t. NB gives the proportion of "net" true positives in the dataset. This "net" proportion is equivalent to the proportion of true positives in the absence of false positives (i.e., perfect specificity). In fact, this is a direct comparison with the treat none policy, which has zero true positives and zero false positives by default. For example, Model A has an NB of 0.289 at the 25% target threshold, which is equivalent to detecting ≈ 29 PD cases and suggesting zero unnecessary treatment changes per 100 patients. Under Model B, this would be ≈ 25 detected PD cases. We can also calculate the difference in NB between the two models. Consider again the Model A which yields 41 true positives and 19 false positives at the 25% risk threshold (NB = 0.289). At the same threshold, Model B yields 38 true positives and 22 false positives (NB = 0.252). The difference in NB for the Models A and B at the 25% target threshold is 0.289–0.252 = 0.037. Hence, Model A has 3.7 more net detected PD cases per 100 patients. This is equivalent to having 3.7 more detected PD cases per 100 patients for the same number of unnecessary treatment changes.

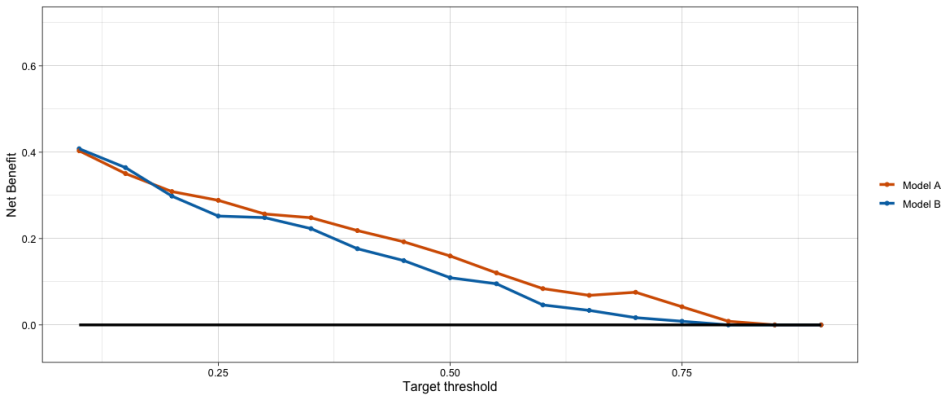


Figure 2.3 Net Benefit for a range of target thresholds. A model is of clinical value if it has the highest NB compared with the simple strategy of no treatment (horizontal black line) across the full range of target thresholds. The model with the highest NB at a particular target threshold enables us to change treatment as many high risk patients as possible while avoiding harm from unnecessarily changing treatment to low risk patients.

2.3.1 Dynamic Predictions

To illustrate the subject-specific predictions, we set aside three patients (not used in model training) and predict the probability of PD at clinically relevant timepoints, conditional on their available measurements. The patients were chosen to represent three distinct types of longitudinal profiles (increasing, decreasing, stable) alongside some variation on the PD outcome, i.e., both positive and negative CT scans. Figure 2.4 presents the profiles of the three patients alongside their PD outcomes. Patients A and B present clinically worsening and improving longitudinal measure profiles, respectively. Patient C has a stable profile. All three patients transition between progressive and non-progressive disease at various rates and timepoints.

Following the methodology of Section 2.2.5, we calculate dynamic predictions using ctDNA history alone (Dynamic Predictions), or combined with the PD history (Individualised Dynamic Predictions), based on Model A. We show the predictions of PD for each patient in Figure 2.5. To create predictions at any given timepoint, t, we only use the available ctDNA measurements up to 0.5 years in the past, i.e., $t - 0.5$. This was for two reasons. First, because more recent measurements are more informative of a patient's clinical status. Second, a lead-time of a few months has been observed between a rise in ctDNA levels and clinical progression (Dawson et al., 2013; Hrebien et al., 2019)[6]. We start by observing that the predictions seem to be sensitive to the ctDNA measurements and generally follow the ctDNA profiles. For instance, the upward trend in ctDNA seen in the second to last panel ($t : 1.37$) of Figure 2.5a is captured by the high predicted risk of PD. Similarly, the third ($t : 0.67$) and fourth panels ($t : 0.94$) are examples where a decreasing ctDNA profile leads to low predicted probability of PD. Second, the two dynamic approaches (Dynamic Predictions and Individualised Dynamic Predictions) have comparable predicted values and prediction intervals in these patients. This indicates the use of only the ctDNA history offers the same conclusions as both ctDNA and PD histories together. Hence, we can potentially

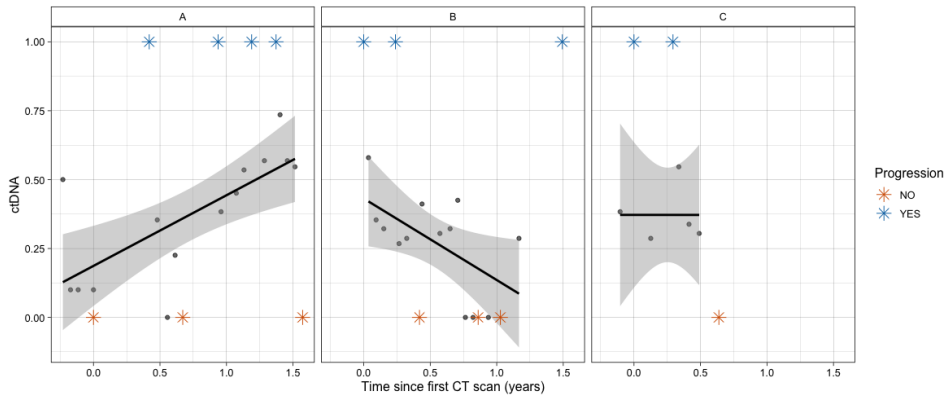


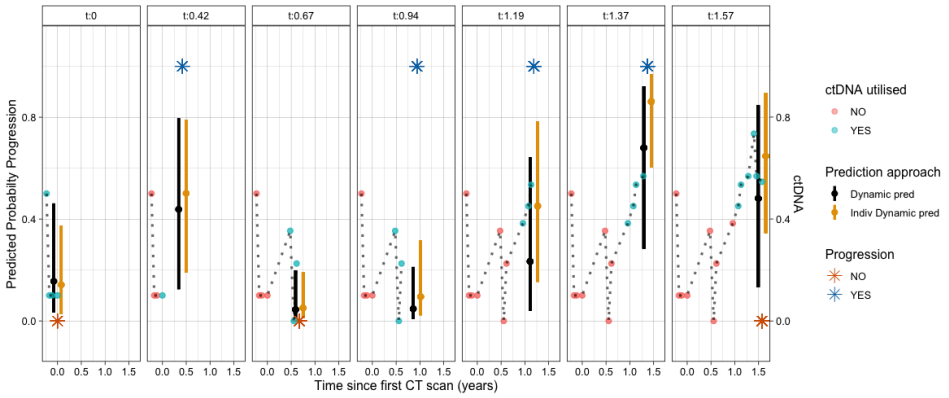
Figure 2.4 Longitudinal ctDNA (grey dots) and PD (asterisks) measurements for patients A, B, and C who have been excluded from the analysis dataset, and for whom we calculate predictions. The solid lines are OLS estimates with 95% confidence intervals.

avoid the extra computational cost of sampling for the posterior of the random effect(s) in 2nd stage model. Third, with such predictions, clinicians are able to precisely track the health condition of each patient and make better informed decisions at the individual level. In particular, predicting the outcome of therapy early would allow adaptive changes to treatment. For example, for Patient A, the ctDNA measurements up to timepoint 0.67 (third panel) exhibit a declining pattern, and based on the last 0.5 years of data, the corresponding median predicted probabilities are 5% and 6% for the two approaches, respectively. This indicates she is responding to treatment. On the other hand, up to timepoint 1.37 (6th panel) the ctDNA measurements exhibit an upward trend which indicates a worsening clinical condition. This is captured by the corresponding median predicted probabilities at 65% and 82%, respectively. This higher risk of progressive disease at timepoint 1.37 indicates the tumour has developed resistance to treatment and clinicians may consider alternative treatment regimes.

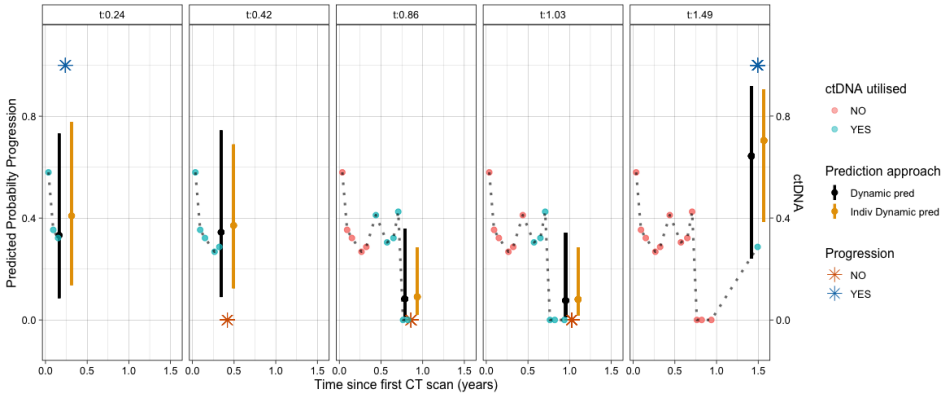
2.4 Discussion

In this work, we proposed a two-stage Bayesian probabilistic model to answer two related questions: (1) whether ctDNA helps predict response to treatment in mBC (2) and, given that patients with metastatic disease are followed-up for considerable time, can we dynamically update those predictions, as additional longitudinal measurements become available? We found that (1) the incorporation of ctDNA to monitor response to treatment offers clinical utility, and (2) the model allows for individualised dynamically updated predictions.

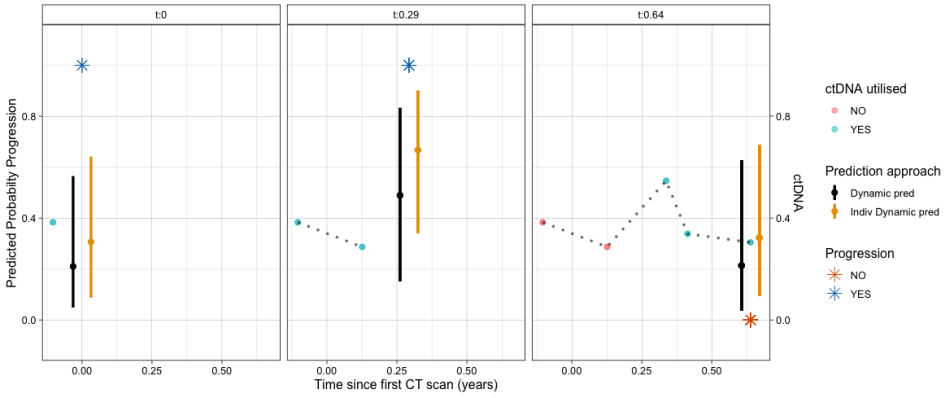
[6]We recognise this strategy for dynamic predictions, using the available ctDNA measurements within 6 months of the timepoint of prediction, is not reflected in how we develop the models on the training data. This is because we wanted to showcase the flexibility of our approach with regards to the dynamic predictions. That is, we can calculate dynamic predictions using the patients' past information in various ways. For instance, the models allow for different past timeframes to be used for different patients, leading to individualised use of a patient's history.



(a)



(b)



(c)

Figure 2.5 Dynamically updated predicted probabilities of PD for patients (a) A, (b) B, and (c) C. Left axis: Median and 90% credible intervals of the predicted probability. Right axis: Observed ctDNA measurements. The observed outcome PD (yes, no) is given as well (asterisks). Each panel shows the corresponding predicted probabilities at timepoint, t, and the longitudinal ctDNA measurements up to that timepoint. These measurements are color-coded depending on whether they were utilised or not in the calculations at the corresponding timepoint.

Our framework provides a potential path to aid personalised decision-making by providing quantitative estimates of clinical outcomes, alongside uncertainty quantification. Our framework considers the probability of disease progression for each patient individually; by doing so, it can identify individual patients with extremely high-risk of non-response to treatment. This information can be used by clinicians, for example to discontinue treatment if the tumour is not responding. An additional desirable element of our framework is uncertainty quantification at the individual level. Thus, a clinician can use the predictions to make treatment decisions or if predictive uncertainty is high, collect more information, e.g., more blood samples, or refer to CT scan for confirmation.

To our knowledge, this is the first attempt to propose a statistical model to predict the probability of disease progression given available past ctDNA information in mBC. Previous studies were limited due to small sample sizes, short follow-up, focus on descriptive relationships and targeted ctDNA quantification. Here, we used a larger sample size of 135 patients who we followed for a median of 1.5 years. In addition, we used an untargeted approach for ctDNA quantification. The main advantage of untargeted approaches is not needing prior information about the primary tumour's genome. In addition, only low depth sequencing is required, keeping the cost of the assay low. On the other hand, a high concentration of ctDNA is needed for reliable reconstruction of tumour-specific genome-wide changes. Furthermore, untargeted approaches show lower sensitivity than targeted ones (Chen and Zhao, 2019). This suggests one of many avenues for future research. First, it will be useful to compare several approaches to estimate the tumour burden and determine the most sensitive option (targeted or untargeted) balancing the minimum cost and prior information about the primary tumour.

Second, more appropriate study designs will allow better use of the available information. An obvious improvement would be to measure both ctDNA and PD at the same timepoints. As we mentioned in Section 2.2.2 some patients have had considerable time gaps between the two measurements. Our model allows us to deal with such cases by "predicting" the ctDNA values but this process in inherently error-prone and model-dependent, especially the longer the time-gap is. This is also supported by the high turnover of the ctDNA, which implies measurements taken further apart provide a distorted picture of the patient clinical condition.

Third, we opted for a two-stage over a joint model to link the two processes, ctDNA and PD. That is, we defined two separate models for the ctDNA and PD outcomes, each of them containing individual-level random effects. We estimated the parameters of the two models separately, as this allowed us to use existing, off-the-shelf software. Additionally, a two-stage approach can achieve comparable predictive performance to joint modelling, although their performances in parameter estimation can be very different (Barrett and Su, 2017; Dandis et al., 2020). However, the two-stage approach ignores the fact that the random effects are not exactly observed but estimated, which may lead to underestimation of the uncertainty as well as imprecision in the regression coefficients (De la Cruz et al., 2011; Sayers et al., 2017; Wang et al., 2000). Hence, a future avenue would be the estimation of the parameters together, in a joint modelling approach.

Fourth, integration of more cancer-patient biomarker data could improve prognostic accuracy. The last decades have seen significant advances in the genomics, proteomics and molecular pathology of biomarkers in cancer. For examples, both DNA- and RNA-based biomarkers have been identified (including mRNA and non-coding RNA) (Rapisuwon et al., 2016). In this work, we only used ctDNA longitudinal measurements to predict disease progression. But our model could be extended to more serially measured biomarkers. This would allow us to consider the between-biomarkers correlation, and estimate the adjusted association of each longitudinal biomarker with the risk of disease progression. Often, such biomarkers vary over time, leading to scenarios where disease progression can be predicted using covariates that are both longitudinal and high-dimensional. To our knowledge, extensions to high-dimensional settings where the covariates are measured longitudinally are currently lacking. This is even more pressing under a joint modelling approach as estimation of joint models becomes computationally prohibitive when more than a handful of longitudinal covariates are included.

Another avenue for future research is to determine optimal number and/or timing of ctDNA monitoring before a clinical decision is made. We showed that tumour burden can be used to predict disease progression and it can be used to change therapeutic regime, if necessary. An open question is how long after treatment initiation we should wait before start measuring ctDNA and how often. This is crucial given recent evidence of time-lag between disease progression determined from ctDNA and CT scans (Bettegowda et al., 2014; Dawson et al., 2013; Ma et al., 2020).

Lastly, we used a wide range of target thresholds to evaluate the clinical utility of our model. Future research can be directed towards identifying a smaller range of reasonable target thresholds. This can be achieved by eliciting patients' and/or clinicians' preferences regarding the relative costs of different classification errors. Numerous techniques that have been proposed in the literature can help with this (e.g., Hunink et al. (2014); Tsalatsanis et al. (2010)). Nevertheless, in practice, eliciting these preferences may be challenging. In such situations, we advocate using a range of plausible t values that reflect general decision preferences. This range can be set by asking for sensible upper and lower bounds on the maximum number of false positives one would tolerate to find one true positive. For example, if a detected progressive disease is worth 10 unnecessary treatment changes, an appropriate target threshold would be $1/(1 + 10) = 9\%$. A risk-averse person would perhaps tolerate more unnecessary treatment changes and motivate a lower bound on the target threshold of 1%. In a clinical context, and with severe illness, the upper bound on the target threshold usually does not exceed 50% – an undetected progressive disease is generally considered more harmful than a false positive case (Wynants et al., 2019).

To conclude, our results suggest the incorporation of ctDNA in clinical practice offers clinical utility and allows for individualised risk predictions throughout a course of therapy. Individualised risk modelling will facilitate personalised therapeutic approaches, with wide applicability in oncology and other areas of medicine.

Chapter 3

Tailored Bayes: a risk modelling framework under unequal misclassification costs

Abstract Risk prediction models are a crucial tool in healthcare (see Chapter 1). However, risk prediction models for a binary outcome are often constructed using methodology which assumes the costs of different classification errors are equal. In many healthcare applications this assumption is not valid, and the differences between misclassification costs can be quite large. In Chapter 1 we introduced a decision analytic approach to take into account unbalanced misclassification costs. We called it standard Bayes (SB). In Chapter 2 we applied that approach to assess the clinical utility of ctDNA in evaluating response to treatment in metastatic breast cancer. Here we show this strategy may still result to sub-optimal solutions. This lead us to present Tailored Bayes (TB), a novel Bayesian inference framework which "tailors" model fitting to optimise predictive performance with respect to unbalanced misclassification costs. We use both simulation and real data to show the advantages of TB compared to SB.

Outline We start in Section 3.1 showing that most commonly used prediction models for a binary outcome implicitly assume equal misclassification costs, which is undesirable in many applications. We then briefly introduce already proposed solutions and their limitations. We then present TB, a novel framework to incorporate misclassification costs into Bayesian modelling (Section 3.2). In Section 3.3 we use simulation studies to showcase when TB is expected to outperform SB. We then apply TB to three real-world applications, and demonstrate the improvement in predictive performance over standard methods (Section 3.4). Finally, Section 3.5 contains a discussion on the methodology and avenues for future work.

3.1 Introduction

In clinical medical research, much effort has been invested in developing decision support systems for diagnostic and prognostic settings. Most of these systems are based on risk prediction models. We saw three examples (QRISK2, EuroSCORE and PREDICT) in Chapter 1 (Section 1.1). Such models are used to guide clinical decision-making, for instance whether an individual should be treated or not. "Treatment" can refer to a wide range of healthcare interventions, such as additional diagnostic workup, referral to specialized care, a procedure, delaying surgery (e.g., in patients at high risk of complications), or lifestyle changes. But the probabilistic nature of risk prediction models complicates significantly the decision process. For example, if a prediction model estimates the probability of an individual having a disease equal to 40%, it is unclear whether this individual should receive treatment or not. We have seen how classical decision theory provides us with a normative approach to answer such questions (Chapter 1).

Relying on classical decision theory, we employ the concept of the target threshold. The target threshold, t, is defined as the probability at which the decision maker is indifferent between two strategies (e.g., administer treatment or not) (Pauker and Kassirer, 1975, 1980). We derived the target threshold from first principles and saw how it summarises the benefits and harms of correct and incorrect decisions through the four basic utilities of a binary classification problem (Section 1.4). The main advantage of the threshold concept is there is no need to explicitly specify these utilities, but only the desired target threshold. Hence, the threshold concept allows us to incorporate decision-makers' preferences when making treatment decisions.

Based on the threshold concept, a consistent approach to decision-making is: having a probability estimate $\hat{\pi}(\mathbf{x})$ we treat based on whether or not the following holds: $\hat{\pi}(\mathbf{x}) \geq t$. We called this paradigm standard Bayes (SB)[1]. That is, we decide to give treatment if $\hat{\pi}(\mathbf{x}) \geq t$, and withhold treatment otherwise. Note that "treat" and "no treat" policies are equivalent to "classify as positive" and "classify as negative", respectively. So, the above statement can be re-worded as: we decide to classify as positive if $\hat{\pi}(\mathbf{x}) \geq t$, and negative otherwise.

Here, we show that despite being an easy to use and widely applicable approach, the SB paradigm may result in sub-optimal classifications (and consequent decisions). Figure 3.1 presents such a setting. We applied SB logistic regression to data generated from a regression model with two covariates, x_1 and x_2 (see Section 3.3.1 for details). The black lines give the optimal model output for various values of t. We focus on $t = 0.3$ which corresponds to $B/H = 2.3/1$ (see equation (1.2)). This implies that we are assuming a false negative classification is 2.3 times worse than a false positive one. The grey line shows the average posterior predictive estimate based on the SB approach described above. The sub-optimality of this solution is apparent, since the posterior does not cover the target $t = 0.3$ line. We revisit this example in more detail in Section 3.3.1.

[1] Although this concept is general, we focus on a Bayesian regression paradigm, hence the name. This implies that $\hat{\pi}(\mathbf{x})$ is a point summary of the posterior predictive distribution (see Section C.2).

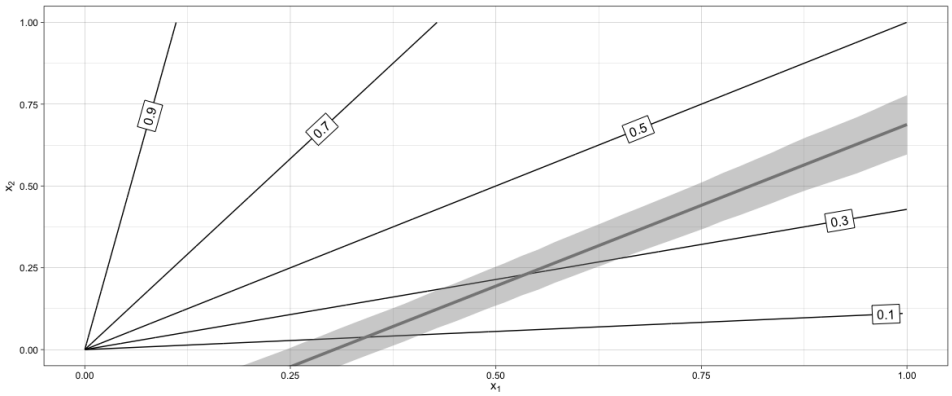


Figure 3.1 Optimal model output (black lines) for target thresholds 0.1, 0.3, 0.5, 0.7, 0.9. Posterior mean boundaries for SB (grey) when targeting $t = 0.3$. Shaded regions represent 90% highest predictive density (HPD) regions. Data simulated from $\theta := p(y = 1|x_1, x_2) = \frac{x_2}{x_1+x_2}$, with $y \sim Bernoulli(\theta)$ and $x_1, x_2 \sim \mathscr{U}(0,1)$ and $n = 5000$ (see Section 3.3.1 for details).

The fundamental cause for this sub-optimality is that most models for binary outcomes are often constructed to minimise the expected classification error; that is the proportion of incorrect classifications. This has been shown in Steinwart (2005); Zhang (2004) and Bartlett et al. (2006). Here, we present a pictorial "proof" using three popular choices of loss (objective) functions used in binary classification models - the exponential loss used in boosting classifiers (Friedman et al., 2000), the hinge loss of support vector machines (Zhang, 2004), and the logistic loss of logistic regression (Friedman et al., 2000; Zhang, 2004).

Figure 3.2 shows these three losses as a function of yf rather than f, because of the symmetry between the $y = +1$ and $y = -1$ case ($f = f(x) = h(x)^T\beta + \beta_0$) [2]. The symmetry between $y = +1$ and $y = -1$ cases implies that the positive part of the x axis, i.e., $yf > 0$ corresponds to correctly classified points while the negative part corresponds to incorrect classifications. Since there is no distinction on the type of misclassification, both errors (i.e., false positives and negatives) are treated equally while training the models with these loss functions. Hence, it is implicitly assumed that all classification errors have equal costs, i.e., the cost of misclassification of a positive label equals the cost of misclassification of a negative label. (Throughout the document we refer to the costs of incorrect classifications as misclassification costs).

However, equal costs may not always be appropriate, and will depend on the scientific or medical context. For example, in cancer diagnosis, a false negative (that is, misdiagnosing a cancer patient as healthy) could have more severe consequences than a false positive (that is, misdiagnosing a healthy individual with cancer); the latter may lead to extra medical costs and unnecessary anxiety for the individual but not result in loss of life. For such applications, a prioritised control of asymmetric misclassification costs is needed.

[2] Note, in contrast to the rest of this work, in Figure 3.2 the outcome, y, is coded as $\{-1, +1\}$, instead of $\{0, 1\}$. Also, f can be any function of the input, x, linear or non-linear.

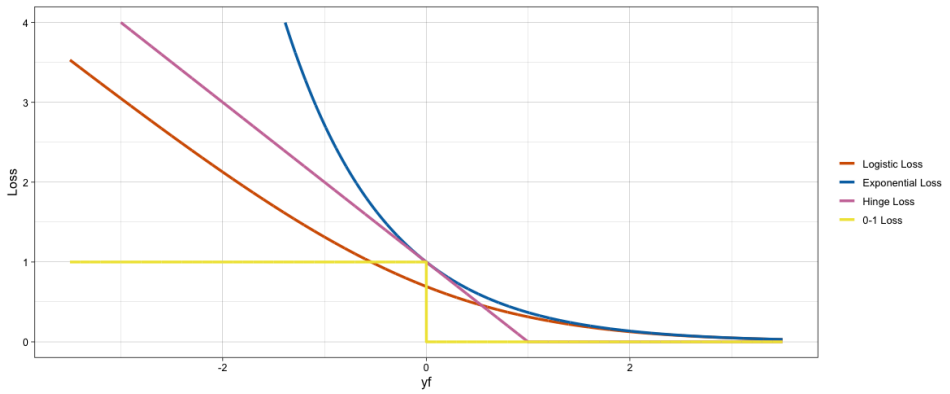


Figure 3.2 The 0-1 (misclassification) loss function and surrogates (hinge loss, logistic loss and exponential loss). All are shown as a function of yf rather than f, because of the symmetry between $y = +1$ and $y = -1$ case ($f = f(x) = h(x)^T\beta + \beta_0$). Note that a classification error is made if and only if yf is negative; thus the 0-1 loss is a step function that is equal to 1 for negative values of the abscissa.

3.1.1 Related work and our proposal

To meet this need, different methods have been developed. In the machine learning literature they are studied under the term cost-sensitive learning (Elkan, 2001). Existing research on cost-sensitive learning can be grouped into two main categories: direct and indirect approaches. Direct approaches aim to make particular classification algorithms cost-sensitive by incorporating different misclassification costs into the training process. This amounts to changing the objective/loss function that is optimised when training the model (e.g., Kukar et al. (1998); Ling et al. (2004); Masnadi-Shirazi and Vasconcelos (2010)). A limitation is that these approaches are designed to be problem-specific, requiring considerable knowledge of the model in conjunction with its theoretical properties, and possibly new computational tools. Conversely, indirect approaches are more general because they achieve cost-sensitivity without any, or with minor modification to existing modelling frameworks. In this work we focus on indirect approaches.

Indirect methods can be further subdivided into thresholding and sampling/weighting. Thresholding refers to the target threshold concept we have already presented. We have already seen that thresholding simply changes the classification threshold of an existing risk prediction model. We can use the target threshold to classify datapoints into positive or negative status if the model can produce probability estimates. This strategy is optimal if the true class probabilities were available. In other words, if the model is based on the logarithm of the ratio of true class probabilities, the threshold should be modified by a value equal to the logarithm of the ratio of misclassification costs (Duda et al., 2012). In practice, however, this strategy may lead to sub-optimal solutions - we already presented such a setting (see Figure 3.1). We further demonstrate this using synthetic (Section 3.3) and real-life data (Section 3.4).

Alternatively, sampling methods modify the distribution of the training data according to misclassification costs (see Elkan (2001) for a theoretical justification). This can be achieved by generating new datapoints from the class with smaller numbers of datapoints i.e., oversampling from the minority

class, or by removing datapoints from the majority class (undersampling). The simplest form is random sampling (over- or under-). However, both come with drawbacks. Duplicating samples from the minority class may cause overfitting (Zadrozny et al., 2003). Similarly, random elimination of samples from the majority class can result in loss of data which might be useful for the learning process. Weighting (e.g., Margineantu and Dietterich (2003); Ting (1998)) can also be conceptually viewed as a sampling method, where weights are assigned proportionally to misclassification costs. For example, datapoints of the minority class, which usually carries a higher misclassification cost, may be assigned higher weights. Datapoints with high weights can be viewed as sample duplication – thus oversampling. In general, random sampling/weighting determine the datapoints to be duplicated or eliminated based on outcome information (whether a datapoint belongs to the majority or the minority class). Notably, they do not take into account critical regions of the covariate space, such as regions that are closer to the target decision boundary. A decision boundary specifies distinct classification regions on the covariate space based on specified misclassification costs (see Section 3.3). This is the goal of the framework presented here.

In this chapter, we build upon the seminal work of Hand and Vinciotti (2003), and present an umbrella framework that allows us to incorporate misclassification costs into commonly used models for binary outcomes. The framework allows us to tailor model development with the aim of improving performance in the presence of unequal misclassification costs. Although the concepts we discuss are general, and allow for relatively simple tailoring of a wide range of models (essentially whenever the objective function can be expressed as a sum over samples), we focus on a Bayesian regression paradigm. Hence, we present Tailored Bayes (TB), a framework for tailored Bayesian inference when different classification errors incur different penalties. We rely on the target threshold concept to quantify the benefits and harms of correct and incorrect classifications. We then build a 2-stage model (Section 3.2). In the first stage, the most informative datapoints are identified. A datapoint is treated as informative if it is close to the target threshold of interest. Each datapoint is assigned a weight proportional to its distance from the target threshold. Intuitively, one would expect improvements in performance to be possible by putting decreasing weights on the class labels of the successively more distant datapoints. In the second stage, these weights are used to downweight each datapoint's likelihood contribution during model training. A key feature is that this changes the estimation output in a way that goes beyond thresholding and we demonstrate this effect in simple examples (Section 3.3).

The rest of this chapter is organised as follows. Section 3.2 presents the TB framework. In Section 3.3 we conduct simulation studies to illustrate the improvement in predictive performance of our proposed TB modelling framework over the SB paradigm. We then apply the methodology to three real-data applications (Section 3.4). We show that incorporating this information about misclassification costs into the model through our TB approach leads to better treatment decisions. We finish with a discussion of our approach, findings and provide some general remarks in Section 3.5.

3.2 Methods

In Section 3.2.1 we incorporate the target threshold in the model formulation which results in the tailored likelihood function (Section 3.2.2) and the tailored posterior (Section 3.2.3). Section 3.2.4 gives the data-splitting strategy we are implementing.

3.2.1 Model formulation

Denote data $D = \{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ where y_i is the binary outcome indicating the class to which the i^{th} datapoint belongs and $\mathbf{x}_i$ is the vector of covariates of size d [3]. The objective is to estimate the posterior probability of belonging to one of the classes given a set of new datapoints. We use D to fit a model $p(y_i \mid \mathbf{x}_i)$ and use it to obtain $\pi(\mathbf{x}_*)$ for a future datapoint y_* with covariates $\mathbf{x}_*$. We simplify the structure using $p(y_i \mid f(\mathbf{x}_i))$, where $f : \mathscr{X} \to \mathbb{R}$ is a function that maps the vector of the covariates to the real line i.e., the linear predictor used in generalised linear models. To develop the complete model, we need to specify $p(y_i \mid f(\mathbf{x}_i))$ and f.

In the machine learning literature, most of the binary classification procedures use a loss-function-based approach. In the same spirit, we model $p(y_i \mid f(\mathbf{x}_i))$ according to a loss function $\ell(y_i, f(\mathbf{x}_i))$ which measures the loss for reporting f when the truth is y. Mathematically, minimizing this loss function can be equivalent to maximizing $-\ell(y, f)$, where $\exp\{-\ell(y, f)\}$ is proportional to the likelihood function. This duality between "likelihood" and "loss", that is viewing the loss as the negative of the log-likelihood is referred to in the Bayesian literature as a logarithmic score (or loss) function (Bernardo and Smith, 2009; Bissiri et al., 2016). In the introduction we saw a few popular choices of loss functions such as the exponential loss (Friedman et al., 2000), the hinge loss (Zhang, 2004), and the logistic loss (Friedman et al., 2000; Zhang, 2004). In this work, we focus on the following loss,

$$\ell_{w_i}(y_i, f(\mathbf{x}_i)) = -w_i y_i \log \pi(f(\mathbf{x}_i)) - w_i(1 - y_i)\log(1 - \pi(f(\mathbf{x}_i))), \text{ for } i = 1, \ldots, n \tag{3.1}$$

and we define $\pi_{w_i}(f(\mathbf{x}_i)) := \pi(f(\mathbf{x}_i))^{w_i} = (\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}/1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\})^{w_i}$, where $w_i \in [0, 1]$ are datapoint-specific weights. This is a generalised version of the logistic loss, first introduced by Hand and Vinciotti (2003). We recover the standard logistic loss by setting $w_i = 1$ for all $i = 1, \ldots, n$. Note that we specify f as a linear function, i.e., $f(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $(d+1)$-dimensional vector of regression coefficients. Hence, our objective is to learn $\boldsymbol{\beta}$. We make this explicit by replacing $\pi_{w_i}(f(\mathbf{x}_i))$ with $\pi_{w_i}(\mathbf{x}_i; \boldsymbol{\beta})$ for the rest of this work.

The datapoint-specific weights, w_i, allow us to tailor the standard logistic model. We wish to weigh observations based on their vicinity to the target threshold, t, upweighting observations close to t (the most informative) and downweighting those that are further away. To accomplish this we set the weights as

$$w_i = \exp\big\{-\lambda h(\pi_u(\mathbf{x}_i), t)\big\} = \exp\big\{-\lambda(\pi_u(\mathbf{x}_i) - t)^2\big\}, \tag{3.2}$$

[3] In this chapter we follow the notation introduced in Chapter 1. As a result, $Y \in \{0, 1\}$ and $\mathbf{X} \in \mathbb{R}^d$, and $\pi(\mathbf{x}) = P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ is the conditional class 1 probability given the observed values of the covariates.

where h is the squared distance (see Section C.6 for other options) and $\pi_u(\mathbf{x}_i)$ is a base/reference prediction model. Of course, in practice we do not know $\pi_u(\mathbf{x}_i)$ so we cannot measure the distance between t and each datapoint's predicted probability, $\pi_u(\mathbf{x}_i)$, in order to derive these weights. To overcome this, we propose a two-stage procedure. First, the distance is measured according to an estimate of $\pi_u(\mathbf{x}_i)$, $\hat{\pi}_u(\mathbf{x}_i)$, which can be compared with t to yield the weights. This estimate could be based on any classification method: we use standard Bayesian logistic regression in the analysis below. If a well-established model of $\pi_u(\mathbf{x}_i)$ already exists in the literature that could be used (as in our cardiac surgery case study, see Section 3.4.2) this task would not be necessary. After deriving the weights, they are then used to estimate $\pi_{w_i}(\mathbf{x}_i;\boldsymbol{\beta})$. Finally, under the formulation in (3.2) the weights decrease with increasing distance from the target threshold t. The tuning parameter $\lambda \geq 0$ controls the rate of that decrease. For $\lambda = 0$ we recover the standard logistic regression model. We use cross-validation to choose λ, see later for details.

3.2.2 Tailored likelihood function

To gain a better insight into the model we define the tailored likelihood function as

$$L(D \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \exp\{-\ell_{w_i}(y_i, \mathbf{x}_i^T\boldsymbol{\beta})\} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right)^{y_i w_i} \left(1 - \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right)^{w_i(1-y_i)} \tag{3.3}$$

Strictly speaking, this quantity is not the standard logistic likelihood function. Nevertheless, it is instinctive to see its correspondence with the standard likelihood function. Thus, we rewrite (3.3) (after taking the log in both sides) as

$$\begin{aligned}
\log(L(D \mid \boldsymbol{\beta})) &= -\textstyle\sum_{i=1}^{n} \ell_{w_i}(y_i, \mathbf{x}_i^T\boldsymbol{\beta}) \\
&= \textstyle\sum_{i=1}^{n} y_i w_i \log\left(\frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right) + w_i(1-y_i)\log\left(1 - \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right) \\
&= \textstyle\sum_{i=1}^{n} w_i \left[y_i \log\left(\frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right) + (1-y_i)\log\left(1 - \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}\right)\right] \\
&= \textstyle\sum_{i=1}^{n} w_i l_i(D \mid \boldsymbol{\beta})
\end{aligned} \tag{3.4}$$

where $l_i(D \mid \boldsymbol{\beta})$ is the standard logistic log-likelihood function. We can further replace (3.2) into (3.4)

$$\log(L(D \mid \boldsymbol{\beta})) = \sum_{i=1}^{n} \exp\left\{-\lambda(\pi_u(\mathbf{x}_i) - t)^2\right\} l_i(D \mid \boldsymbol{\beta})$$

to see that each datapoint contributes exponentially proportional to its distance from the target threshold t, which summarises the four utilities associated with binary classification problems (see Section 1.4). One option to proceed is by optimising the tailored likelihood function with respect to the coefficients in an empirical risk minimisation approach (Vapnik, 1998). An attractive feature of (3.4) is

that this optimisation is computationally efficient since we can rely on existing algorithmic tools, e.g., (stochastic) gradient optimisation. However, here we learn the coefficients in a Bayesian formalism.

3.2.3 Bayesian tailoring

Following Bayes Theorem, the TB posterior is

$$p(\boldsymbol{\beta} \mid D) = \frac{L(D \mid \boldsymbol{\beta}) p(\boldsymbol{\beta})}{p(D)}, \tag{3.5}$$

where $L(D \mid \boldsymbol{\beta})$ is the tailored likelihood function given in (3.3), $p(\boldsymbol{\beta})$ is the prior on the coefficients, and $p(D) = \int L(D|\tilde{\boldsymbol{\beta}}) p(\tilde{\boldsymbol{\beta}}) d\tilde{\boldsymbol{\beta}}$, is the normalising constant. In this work we assume a normal prior distribution for each element of $\boldsymbol{\beta}$, i.e., $p(\beta_j) = \mathcal{N}(\mu_j, \sigma_j^2)$, where μ_j and σ_j are the mean and standard deviation respectively for the j^{th} element of $\boldsymbol{\beta}$ ($j = 1, \ldots, d+1$). For all analysis below we use vague priors with $\mu_j = 0$ and $\sigma_j = 100$, for all j.

Conveniently, we can interpret the choice of prior as a regularizer on a per-datapoint influence (or importance) (see Section C.1). Crucially, this allows us to view the TB posterior as combining a standard likelihood function with a data-dependent prior (Section C.1). Hence, even though the tailored likelihood function does not have a probabilistic interpretation the TB posterior is a proper posterior.

In Appendix C we provide details on the model inference and predictions steps (Section C.2), the cross-validation scheme for choosing λ (Section C.3), and the Markov chain Monte Carlo (MCMC) algorithm we are implementing (Section C.4).

3.2.4 Data splitting strategy

To avoid overfitting due to the estimation of both $\pi_u(\mathbf{x}_i)$ and $\pi_{w_i}(\mathbf{x}_i; \boldsymbol{\beta})$ from the same dataset we use the following data splitting process (Figure 3.3). First, the data is split into training and testing sets. This step is avoided if we already have an independent test set (Section 3.4.1) or if data is simulated (Section 3.3). The train set is subsequently split again into design (20%) and development (80%). The design part is used to estimate $\pi_u(\mathbf{x}_i)$. The development part is used to choose λ (with 5-fold stratified CV, see Section C.3 for details). After choosing a λ value the model is fit to the entire development part, and it is with respect to the posterior from this final fit credible intervals are generated in the analyses below. Finally, the test set is used to validate the performance of the model. We opted for this three-way splitting strategy, design-development-test set, because we have large datasets available but any other method such as leave-one-out, (nested) cross-validation, or bootstrap methods could be used. In addition, note that compared to SB, this data splitting strategy will increase the posterior variance of our TB estimates. This is because contrary to SB, the TB posterior is based on the development part of the training set rather than the entire training set.

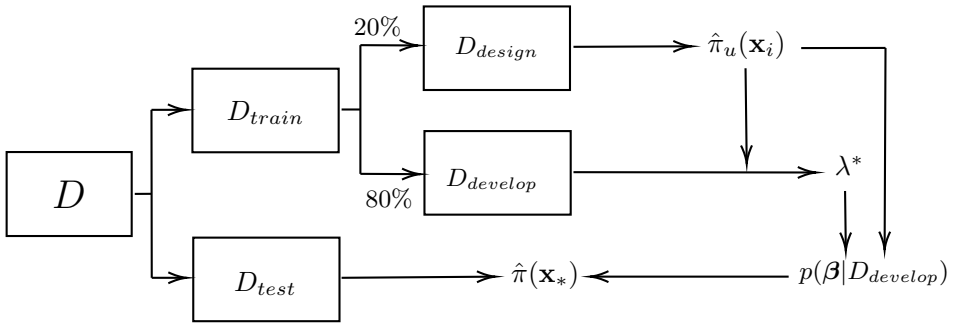


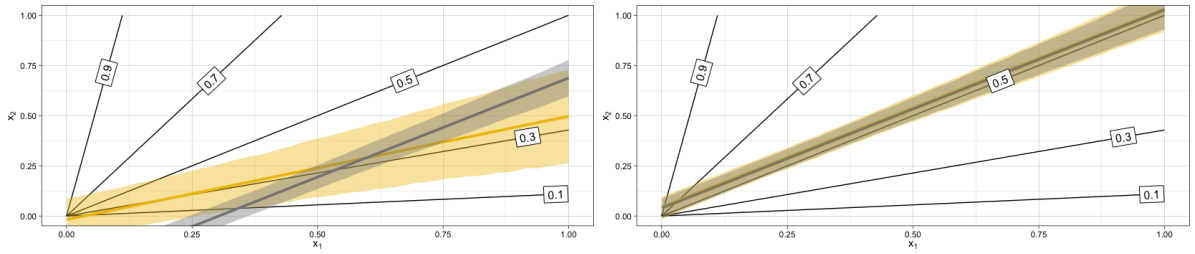
Figure 3.3 The data splitting strategy. The dataset, D, is split into train (D_{train}) and test (D_{test}) sets. The train set is subsequently split again into design (D_{design}) (20%) and development ($D_{develop}$) (80%). The design part is used to estimate $\hat{\pi}_u(\mathbf{x}_i)$. The development part is used to choose λ^* (5-fold CV, see Section C.3 for details). After choosing a λ^* value the model is fit to the entire development part, obtaining the posterior, $p(\boldsymbol{\beta}|D_{develop})$. Finally, the test set is used to create predictions, $\hat{\pi}(\mathbf{x}_*)$ (this is the posterior predictive mean defined in Section C.2.)

3.3 Simulations

The simulations are designed to provide insight into when TB can be advantageous compared to the standard Bayesian paradigm. Two scenarios where TB is expected to outperform standard Bayes (SB) are the absence of parallelism of the optimal decision boundaries and data contamination. A decision boundary determines distinct classification regions in the covariate space. It provides a rule to classify datapoints based on whether the datapoint's covariate vector falls inside or outside the classification region. If a datapoint falls inside the classification region it will be labelled as belonging to class 1 (e.g., positive), if it falls outside it will be labelled as belonging to class 0 (e.g., negative). According to Bayesian decision theory the optimal decision boundaries determine the classification regions where the expected reward is maximised given pre-specified misclassification costs (Duda et al., 2012). More specifically, we classify as positive if $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} > \frac{t}{1-t}$, where $\pi(\mathbf{x})$ denotes the true class 1 probability, as in Section 1.5. Simulations 1 and 2 present two settings where the optimal decision boundaries are not parallel with their orientation changing as a function of the target threshold. Simulation 3 is an example of data contamination.

3.3.1 Simulation 1: Linear Decision Boundaries

We first evaluate the performance of tailoring by extending a simulation from Hand and Vinciotti (2003). We simulate n data points according to two covariates, x_1 and x_2, and assign label 1 with probability: $\theta := p(y=1|x_1,x_2) = \frac{qx_2}{x_1+qx_2}$ with $y \sim Bernoulli(\theta)$, $x_1, x_2 \sim \mathscr{U}[0,1]$ and where q is a scalar. The parameter q determines the relative prevalence of the two classes, when $q > 1$ there are more class 1 than class 0, otherwise there are more class 0 than class 1. Figure 3.4 shows the optimal decision boundaries in the covariate space for a range of target thresholds using $n = 5000$ and $q = 1$ (which



(a) (b)

Figure 3.4 Optimal decision boundaries (black lines) for target thresholds 0.1, 0.3, 0.5, 0.7, 0.9. Posterior mean boundaries for SB (grey) and TB (yellow) when targeting the (a) 0.3, and (b) 0.5 boundary. Shaded regions represent 90% highest predictive density (HPD) regions.

leads to a prevalence of 0.5)[4]. A key feature is that these boundaries are linear, but not parallel. The absence of parallelism renders any linear model unsuitable as a global fit, but the linearity of the decision boundaries allows linear models to describe these boundaries sufficiently.

We use the decision boundaries corresponding to 0.3 and 0.5 target thresholds as exemplars. SB results in a sub-optimal estimated decision boundary for $t = 0.3$ (Figure 3.4a). The estimated 0.3 boundary from SB is parallel to the 0.5-optimal boundary. This is expected because under this simulation setting logistic regression is bound to find a compromise model which should be linear with level lines roughly parallel to the true 0.5 boundary (where misclassification costs are equal). On the other hand, TB allows derivation of a decision boundary which is far closer to the optimum. Note the wider predictive regions of the tailoring. This is an expected consequence of our framework which we comment on in Section C.6. When deriving decision boundaries under the equal costs implied by a 0.5 target threshold (Figure 3.4b), the two models are almost indistinguishable.

To systematically investigate the performance of tailoring across a wide range of settings, we set-up different scenarios by varying: (1) the sample size, (2) the prevalence of the outcome, (3) and the target threshold. Model performance is evaluated in an independently sampled dataset of size 2000. Under most scenarios TB outperforms SB (Figure 3.5). The performance gains are evident even for small sample sizes. With a few exceptions (most notably $t = 0.7$ and 0.9) the advantage of tailoring is relatively stable across sample sizes. The advantage of tailoring persists even when varying the prevalence of the outcome. In fact, we see that under certain scenarios TB is superior to SB even for the 0.5 boundary. Figure 3.6 illustrates such a scenario for $q = 0.1$, which corresponds to prevalence of 0.15. Under such class imbalance, which is common in medical applications, even when targeting the 0.5 boundary, one might want to use tailoring over standard modelling approaches.

[4]This is the simulation scenario in Figure 3.1.

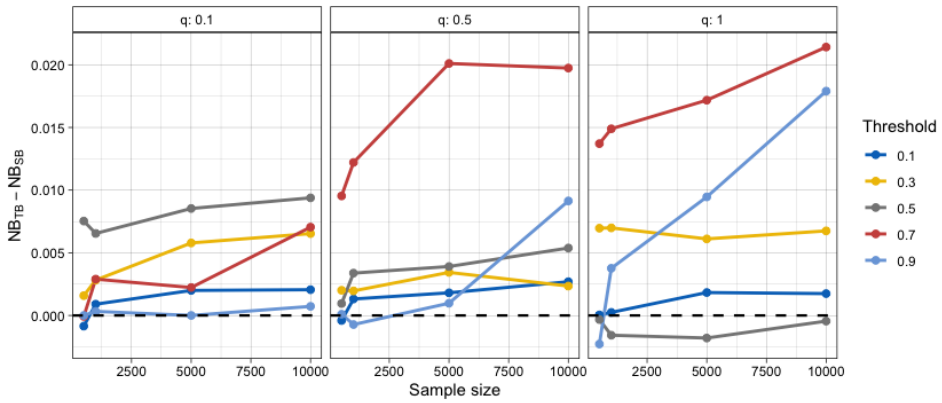


Figure 3.5 Difference in Net Benefit for samples sizes of 500, 1000, 5000, 10000 averaged over 20 repetitions. A positive difference means TB outperforms SB. The values of 0.1, 0.5, 1, for the q parameter correspond to prevalence of around 0.15, 0.36, 0.50, respectively.

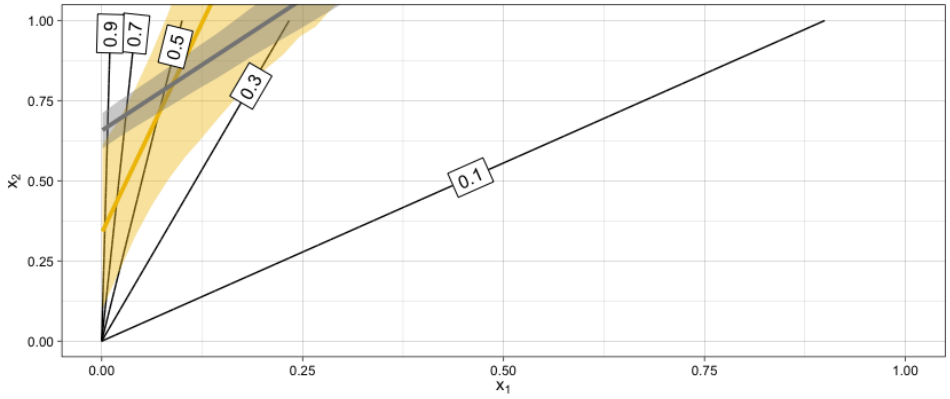


Figure 3.6 Single realisation with $q = 0.1$ corresponding to prevalence of around 0.15. Optimal decision boundaries (black lines) for target thresholds 0.1, 0.3, 0.5, 0.7, 0.9. Posterior mean boundaries for SB (grey) and TB (yellow) when targeting the 0.5 boundary. Shaded regions represent 90% highest predictive density (HPD) regions.

3.3.2 Simulation 2: Quadratic Decision Boundaries

Our second simulation is a more pragmatic scenario where the optimal decision boundaries a quadratic rather that linear function of the covariates. The model is of the form

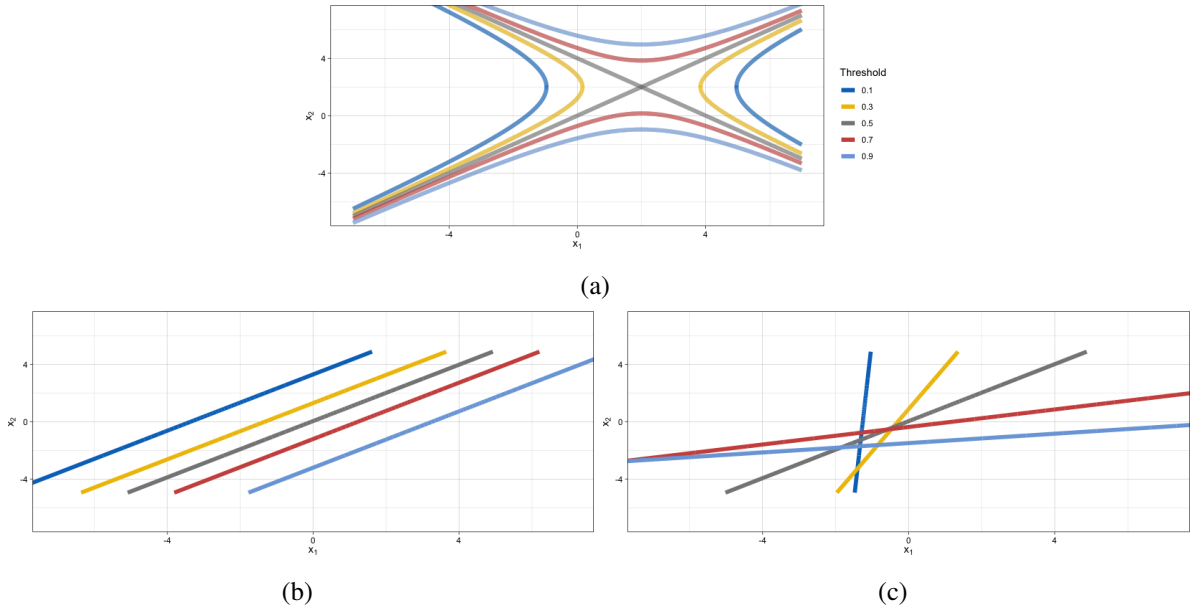
$$\mathbf{x}|y=1 \quad \sim \quad \mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1,0 \\ 0,2 \end{bmatrix}\right)$$

$$\mathbf{x}|y=0 \quad \sim \quad \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2,0 \\ 0,1 \end{bmatrix}\right)$$

where $\mathbf{x} = (x_1, x_2)^T$ contains the two continuous-valued predictors. The marginal probabilities of the outcome are equal, i.e., $p(y=0) = p(y=1) = 0.5$. In this case of unequal covariance matrices, the optimal decision boundaries are a quadratic function of $\mathbf{x}$ (Figure 3.7a) (Duda et al. (2012), Chapter 2). The model we implement is sub-optimal. Nevertheless, this example allows us to demonstrate in an analytically tractable way the advantage of tailoring and it allows us to explore a broader array of generic simulation examples, since arbitrary Gaussian distributions lead to decision boundaries that are general hyperquadrics.

Figures 3.7b and 3.7c show the posterior median decision boundaries for SB and TB using $n = 5000$ under the data generating model described above, and for a range of target thresholds. It is clear that the direction of the optimal decision plane is a function of the costs. The parallel decision boundaries obtained by applying different thresholds to the standard logistic predictions are clearly not an optimal solution when comparing against the optimal boundaries depicted in Figure 3.7a. Although limited to estimation of linear boundaries, tailoring is able to adapt the angle of the boundary to better approximate the optimal curves. One exception in comparative performance is the 0.5 threshold which is estimated perfectly for both models. This is expected, since the standard logistic model targets the 0.5 boundary.

As before, we investigate the performance of tailoring across a wide range of settings, by varying: (1) the sample size, (2) the prevalence of the outcome, (3) and the target threshold. Performance is evaluated in an independently sampled test set of size 2000. Figure 3.8 shows the difference in NB between TB and SB. Tailoring performs similarly or better than standard regression across all target thresholds for prevalence scenarios 0.3 and 0.5. For 0.1 the two models are closely matched. A further comparison with a non-linear model, namely Bayesian Additive Regression Trees (BART) (Sparapani et al., 2021) is detailed in Section C.5. Briefly, TB demonstrated equivalent or better performance than BART at the clinically relevant lower prevalences of 0.1 and 0.3, indicating that the benefits offered by TB cannot be matched simply by switching to a non-linear modelling framework.



(a)

(b) (c)

Figure 3.7 (a) Optimal decision boundaries for target thresholds 0.1, 0.3, 0.5, 0.7, 0.9. Posterior median boundaries for (b) SB, and (c) TB.

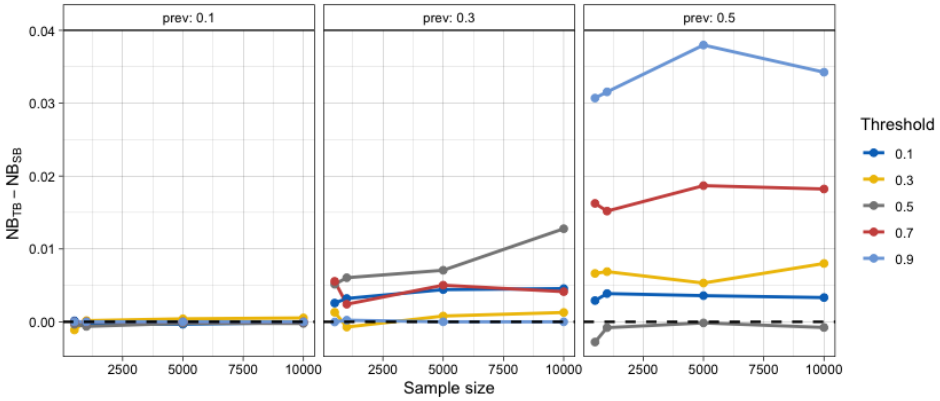


Figure 3.8 Difference in Net Benefit for samples sizes of 500, 1000, 5000, 10000 averaged over 20 repetitions. A positive difference means TB outperforms SB. Each grid corresponds to a different prevalence setting.

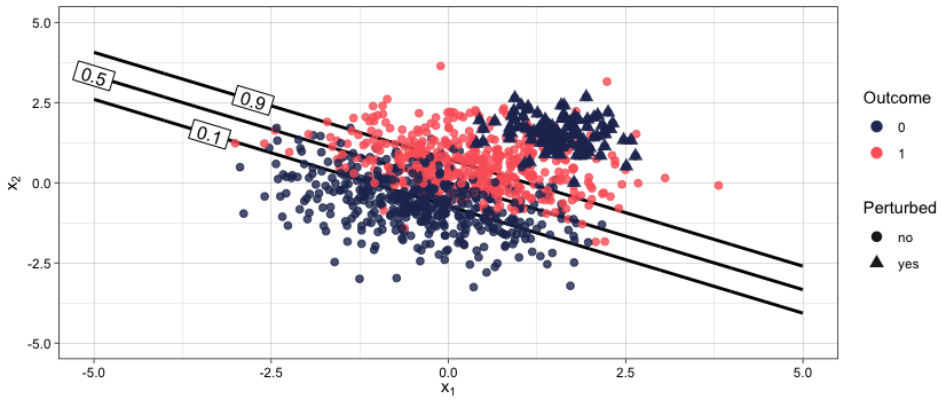


Figure 3.9 Single realisation from contaminated distribution with 10% corrupted datapoints. Data ($n = 1000$) with labels 0 and 1 are shown in blue and red, respectively. The corrupted data points are depicted with triangles on the upper right-hand corner of the data cloud. The lines corresponds to target thresholds 0.1, 0.5, and 0.9.

3.3.3 Simulation 3: Data contamination

Our third simulation scenario demonstrates the robustness of tailoring to data contamination i.e., the situation in which a fraction of the data have been mislabelled. The data generating model is a logistic regression with a large fraction of mislabelled datapoints. We simulate $d = 2$ covariates and $n = 1000$ datapoints. Figure 3.9 depicts a scenario with 10% of datapoints mislabelled among those with high values of both covariates, i.e., among the upper right hand side of the data cloud. For each covariate, 1000 values are independently drawn from a standard Gaussian distribution. Denoting the coefficient vector by $\boldsymbol{\beta} \in \mathbb{R}^3$ with values $\boldsymbol{\beta} = (0,2,3)$ (the first value corresponds to the intercept term) we simulate the outcome vector as $y \sim Bernoulli\left(\frac{\exp\{\mathbf{x}^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}^T\boldsymbol{\beta}\}}\right)$, where $\mathbf{x} = (1, x_1, x_2)^T$. We then corrupt the data with class 0 datapoints, i.e., we set $y := 0$ for ψn datapoints where ψ is the fraction of contamination taking values $5\%, 10\%, 15\%, 20\%$ and 30%. The covariates are generated from equivalent and independent normal distributions, specifically $x_1, x_2 \sim \mathcal{N}(1.5, 0.5)$. This type of contamination framework has been popularised by Huber (1964, 1965) and used extensively to study the robustness of learning algorithms to adversarial attacks in general (Balakrishnan et al., 2017; Diakonikolas et al., 2018; Osama et al., 2019; Prasad et al., 2018) and medical applications (Paschali et al., 2018).

We derive the optimal NB based on the true probability score in an independent non-contaminated test dataset of size $n = 2000$[5]. Figure 3.10 shows the results for various contamination fractions. For most fractions TB outperforms SB. As the contamination fraction gets larger the performance of both models degrades, but standard regression degrades at a faster rate. Tailoring can accommodate various degrees of contamination better than standard logistic regression, while generally never resulting in poorer performance.

[5]Here, we chose to use a non-contaminated test dataset. As part of future work we could consider situations where (a) both train and test sets are contaminated and (b) only the test set is contaminated.

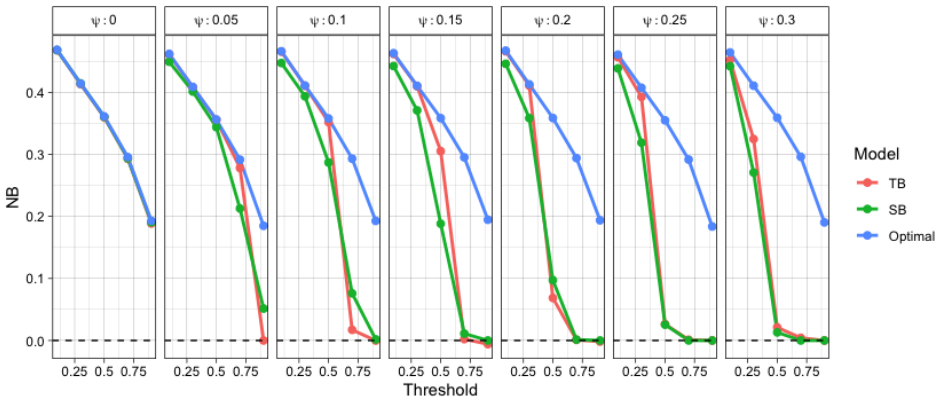


Figure 3.10 Net Benefit of tailoring (red) and standard regression (green) compared to optimal classification (blue) averaged over 20 repetitions. Each grid corresponds to different contamination fraction.

Note that under no contamination (i.e., $\psi = 0$, first panel Figure 3.10) SB is an optimal classifier, since the optimal decision boundaries are parallel straight lines (Figure 3.9). However, for all other scenarios even a data corruption as small as 5% results in poor performance under SB for target thresholds > 0.5. On the contrary, tailoring maintains stable performance and close to the optimal for $t < 0.5$, for up to 15% of mislabelled datapoints.

3.4 Real data applications

We evaluate the performance of TB on three real-data applications involving a breast cancer prognostication task (Section 3.4.1), a cardiac surgery prognostication task (Section 3.4.2) and a breast cancer tumour classification task (Section 3.4.3). Overall, our empirical results demonstrate the improvement in predictive performance when taking into consideration misclassification costs during model training.

3.4.1 Real data application 1: Breast cancer prognostication

Here, we apply the TB methodology to predict mortality after diagnosis with invasive breast cancer. The training data is based on 4718 oestrogen receptor positive subjects diagnosed in East Anglia, UK between 1999 and 2003. The outcome modelled is 10-year mortality. The covariates are age at diagnosis, tumour grade, tumour size, number of positive lymph nodes, presentation (screening vs. clinical), and type of adjuvant therapy (chemotherapy, endocrine therapy, or both). We use 20% of the data as design and the rest as development set (see Figure 3.3), repeating the design/development set split $m = 5$ times. The entire training dataset is used to fit SB. Both models are evaluated in an independent test set consisting of 3810 subjects. Detailed information on the datasets can be found in Karapanagiotis et al. (2018).

An important part of the TB methodology is the choice of t. In breast cancer, accurate predictions are decisive because they guide treatment. In clinical practice, treatment is given if it is expected to

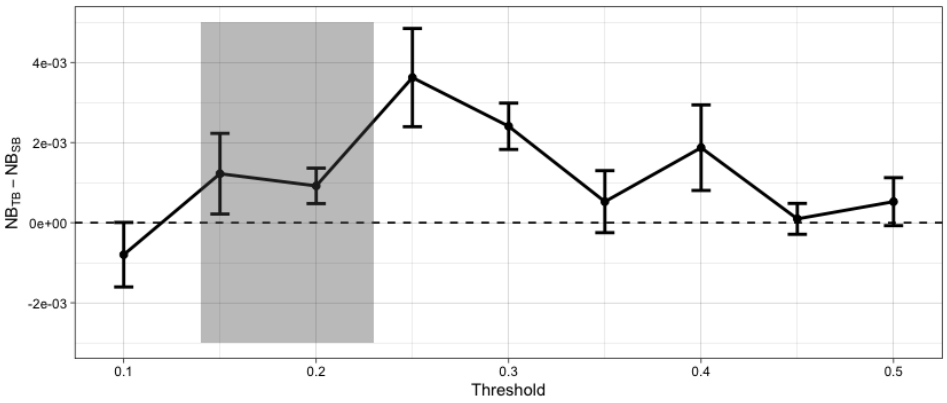


Figure 3.11 Difference in Net Benefit for various t values evaluated on the test set. Error bars correspond to one standard error of the difference. That is, denoting the difference in Net Benefit $D^i = \text{NB}^i_{TB} - \text{NB}^i_{SB}$ with $i = 1, \ldots, m = 5$ for each t then the standard error of the difference is $SE_D = \sqrt{\frac{\sum(D^i - \bar{D})^2}{m(m-1)}}$, where $\bar{D} = \sum_i D^i / m$. This accounts for the fact that both models have been evaluated on the same data. The units on the y axis may be interpreted as the difference in benefit associated with one patient who would die without treatment and who receives therapy. The 0.14 to 0.23 shaded area on the x axis corresponds to 3%–5% absolute risk of death reduction with and without chemotherapy. These are the risk ranges where chemotherapy is discussed as a treatment option.

reduce the predicted risk by at least some pre-specified magnitude. For instance, clinicians in the Cambridge Breast Unit (Addenbrooke's Hospital, Cambridge, UK) currently use the absolute 10-year survival benefit from chemotherapy to guide decision-making for adjuvant chemotherapy as follows: $< 3\%$ no chemotherapy; $3\% - 5\%$ chemotherapy discussed as a possible option; $> 5\%$ chemotherapy recommended (Down et al., 2014). Following previous work (Karapanagiotis et al., 2018), we assume that chemotherapy reduces the 10-year risk of death by 22% (Peto et al., 2012). Then, a risk reduction between 3% and 5%, corresponds to target thresholds between 14% and 23%. Hence, we explore misclassification cost ratios corresponding to t in the range between 0.1 and 0.5.

Figure 3.11 shows the difference in NB between the two models averaged over the 5 splits. We see TB outperforms SB for most target thresholds, especially where decisions about adjuvant chemotherapy are more crucial. Compared to SB, tailoring achieves up to 3.6 more true positives per 1000 patients (when $t = 0.15$), which is equivalent to having 3.6 more true positives per 1000 patients for the same number of unnecessary treatments.

Next, we examine the effect of tailoring on the posterior distributions of the coefficients. As an exemplar, we use the posterior samples for the model corresponding to $t = 0.15$ (Figure 3.12). We see that tailoring affects both the location and spread of the estimates compared to standard modelling. First, note the wider spread of tailoring compared to the standard models. Second, the tailored posteriors are centred on different values. The most extreme example is the coefficient for the number of nodes. Under tailoring it has a stronger positive association with the risk of death. To quantify the discrepancy between the posteriors of the two models table 3.1 shows estimates of the overlapping area between

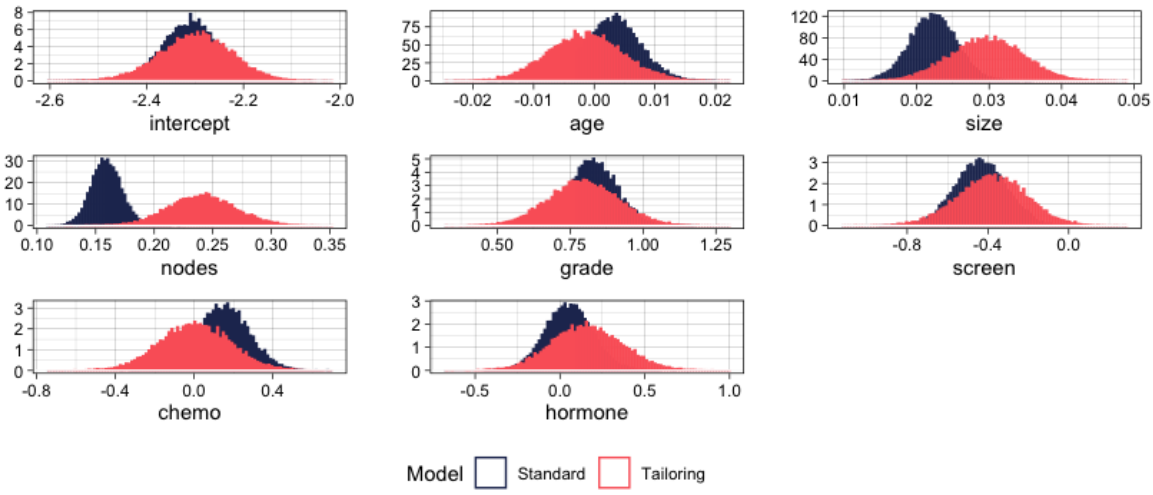


Figure 3.12 Marginal density plots of posterior parameters for $t = 0.15$ for SB (blue) and TB (red).

Covariate	Posterior Overlap (%)
nodes	3.05
size	23.46
chemo	41.92
age	48.78
hormone	57.76
grade	62.66
screen	69.94

Table 3.1 Overlapping area of posterior distributions for each coefficient based Gaussian kernel density estimations (Pastore and Calcagnì, 2019).

the posteriors for each covariate. These range from 3% to 70%. The relative shifts in magnitude of the effect sizes indicates different relative importance of the covariates in terms of their contribution to the predictions from the two models.

Additionally, we explore the sensitivity of our results to the choice of the prior standard deviation. Our main analysis is based on a prior with standard deviation, $\sigma_j = 100$. Since this prior puts a large proportion of the prior density on regions that may not be supported by the data (especially for standardised coefficients), we repeated the analysis using $\sigma_j = 10$ and $\sigma_j = 1$. Figure 3.13 shows the TB posterior distributions of the coefficients under $t = 0.15$. We see the posteriors between different choices of prior standard deviations are almost indistinguishable. As a result, we conclude that for this application the choice of prior standard deviation does not seem to have any quantifiable effect on the posterior.

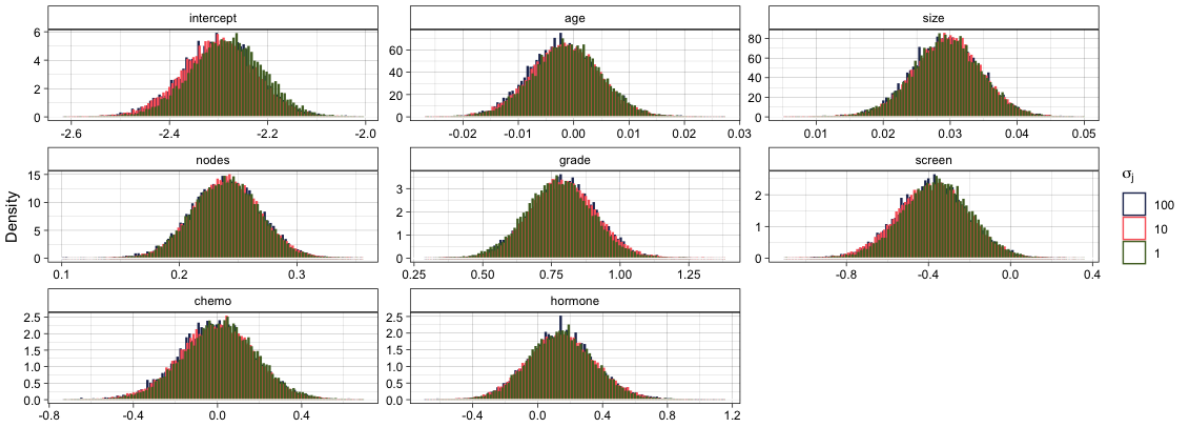


Figure 3.13 Marginal density plots of posterior parameters for $t = 0.15$ for different σ_j values.

3.4.2 Real data application 2: Cardiac surgery prognostication

For our second case study we investigate whether TB allows for better predictions, and consequently improved clinical decisions for patients undergoing aortic valve replacement (AVR). Cardiac patients with severe symptomatic aortic stenosis are considered for surgical AVR (SAVR). Given that SAVR is typically a high-risk procedure, transcatheter aortic valve implantation (TAVI) is recommended as a lower risk alternative but it is associated with higher rates of complications (Baumgartner et al., 2017). The European System for Cardiac Operative Risk Evaluation (EuroSCORE) is routinely used as a criterion to choose between SAVR and TAVI (Roques et al., 2003). EuroSCORE is an operative mortality risk prediction model which takes into account 17 covariates encompassing patient-related, cardiac and operation-related characteristics. It was first introduced by Nashef et al. (1999) and it has been updated in 2003 (Roques et al., 2003) and 2012 (Nashef et al., 2012). Published guidelines recommend TAVI over SAVR if a patient's predicted mortality risk is above 10% (Baumgartner et al., 2017) or 20% (Vahanian et al., 2008). Here, we compare the performance of TB with EuroSCORE and SB given these target thresholds.

We use data (n = 9031) from the National Adult Cardiac Surgery Audit (UK) collected between 2011 and 2018. We use 80% of the data for training and the rest for testing, repeating the train/test set split $m = 5$ times. For this data a design set to estimate $\pi_u(\mathbf{x}_i)$ is not necessary (see Figure 3.3) but instead we use the predictions from EuroSCORE (Roques et al., 2003). We add an extra step of re-calibration to account for the population/time drift (Cox, 1958; Miller et al., 1993). Figure 3.14 presents the results. We see TB outperforms both EuroSCORE and SB when targeting the 0.1 threshold, and only EuroSCORE at $t = 0.2$.

We further investigate the effect of tailoring to individual parameters. Figure 3.15 shows the highest posterior density (HPD) regions for a subset of the covariates under SB and TB for $t = 0.1$ and 0.2. As in the previous case study, under tailoring the regions are generally wider and are centred on different values. For instance, compared to SB under both $t = 0.1$ and 0.2 the posteriors of critical operative

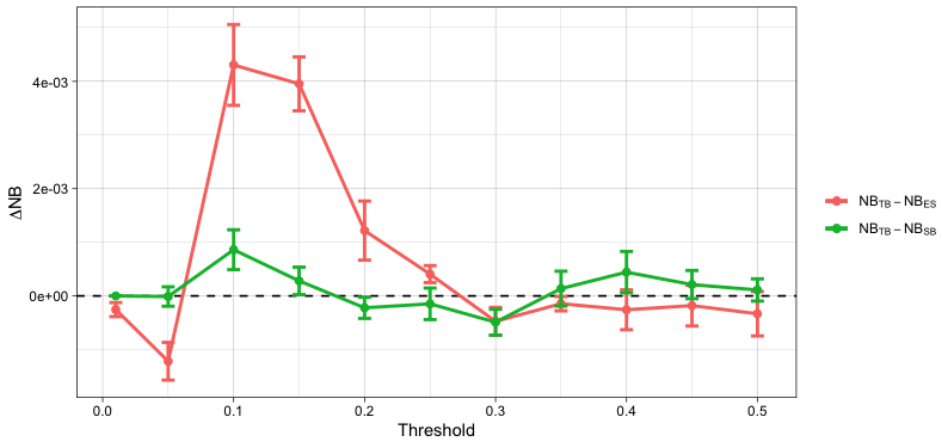


Figure 3.14 Difference in Net Benefit (ΔNB) between TB and EuroSCORE (ES) (red), and between TB and SB (green) for various target thresholds evaluated on the test set. Error bars correspond to one standard error of the difference (see caption of Figure 3.11 for details).

state and unstable angina are shifted towards the same direction (positive for critical operative state and negative for unstable angina). Contrast these with the posteriors of emergency and active endocarditis that compared to SB they are centred on more positive values under $t = 0.1$ and more negative under $t = 0.2$. On the contrary, extracardiac arteriopathy, recent myocardial infarct, and sex are centred on similar values across the three models. This once more exemplifies the change in the contribution of some covariates towards the predicted risks when taking into account misclassification costs.

3.4.3 Real data application 3: Breast cancer tumour classification

For our third case study we use the Wisconsin breast cancer tumour dataset from the UCI repository (Dua and Graff, 2017). The dataset consists of $n = 699$ points, with covariates $\mathbf{x} \in \mathbb{R}^9$, which describe characteristics of the cell nuclei present in digitized images of a breast mass, and labels $y \in \{0, 1\}$. The class labels 0 and 1 correspond to 'benign' and 'malignant' cancers, respectively. To validate the results of the simulation in Section 3.3.3 we artificially contaminate the dataset. More precisely, we use 70% of the data for training, which is corrupted by flipping the labels of 49 class 1 datapoints to 0 (10% contamination). In clinical practice, such data contamination may arise due to the manual nature of breast cancer detection and classification. Breast cancer detection is commonly performed through medical imaging modalities by one or more experts (usually pathologists) (Murtaza et al., 2019). The procedure is time-consuming and dependent on the professional experience and domain knowledge of the pathologists, thus making it prone to errors. This is highlighted by the significant inter- and intra-variability between pathologists (Hong et al., 2012; Li et al., 2009; Warfield et al., 2008). We use 20% of the training data as design and the rest as development set. We assume that missing a malignant cancer is more severe than misdiagnosing a benign as malignant, and so we focus on target thresholds $t < 0.5$, which correspond to a larger weight placed on false negatives vs false positives. Figure 3.16

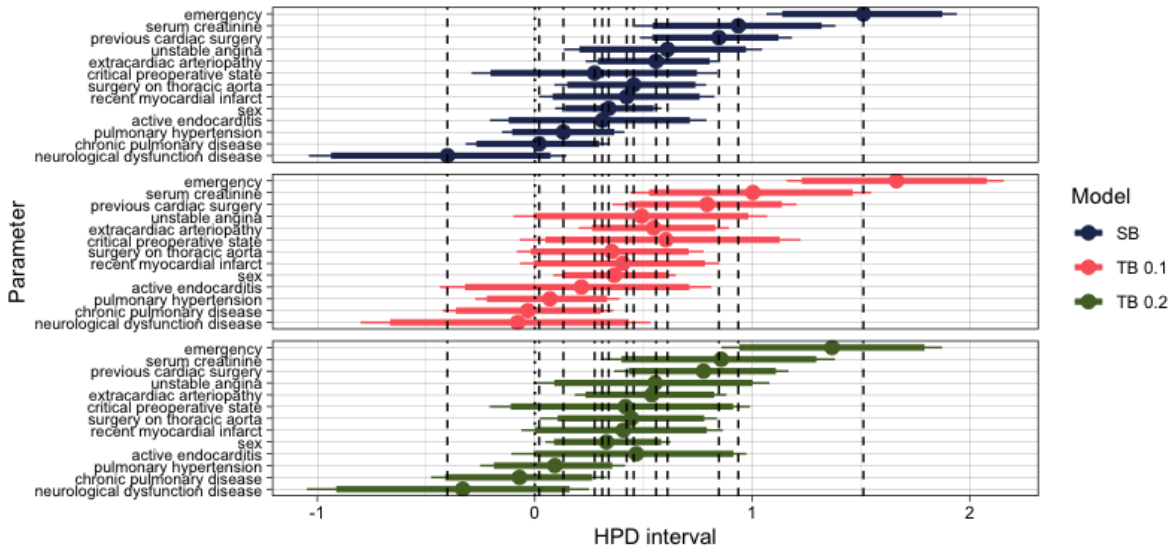


Figure 3.15 Highest posterior density (HPD) regions for the parameters. Dots represent medians, and thick and thin lines represent 90 and the 95% of the HPD regions, respectively. The dashed vertical lines pass through the posterior median values of the SB parameters.

presents the difference in NB for various t values over 5 splits of the training data into design and development. We see tailoring outperforms standard regression for most target thresholds.

We further investigate the effect of tailoring on individual parameter values. Figure 3.17 shows the HPD regions under SB and TB for $t = 0.3$ and 0.5. As in the previous case studies, under tailoring the regions are generally wider and are centred on different values. For instance, under $t = 0.3$ all posteriors are shifted towards more positives values. The only two exceptions are the coefficients of clump thickness, cell shape and mitosis which are pulled towards zero. Similar conclusions, but less pronounced are seen under $t = 0.5$. This again indicates that the relative importance of different covariates changes when using our tailored modelling approach.

3.5 Discussion

In this chapter, we presented Tailored Bayes, a framework to incorporate misclassification costs into Bayesian modelling. We demonstrated that our framework improves predictive performance compared to standard Bayesian modelling over a wide range of scenarios in which the costs of different classification errors are unbalanced.

The methodology relies solely on the construction of the datapoint-specific weights (see equation (3.2)). In particular, we need to specify t, the grid of λ values for the CV, a model to estimate $\pi_u(\mathbf{x}_i)$ and the weighting function, h. For some applications there may be a recommended target threshold, t. For instance, UK national guidelines recommend that clinicians use a risk prediction model (QRISK2; (Hippisley-Cox et al., 2008)) to determine whether to prescribe statins for primary prevention of

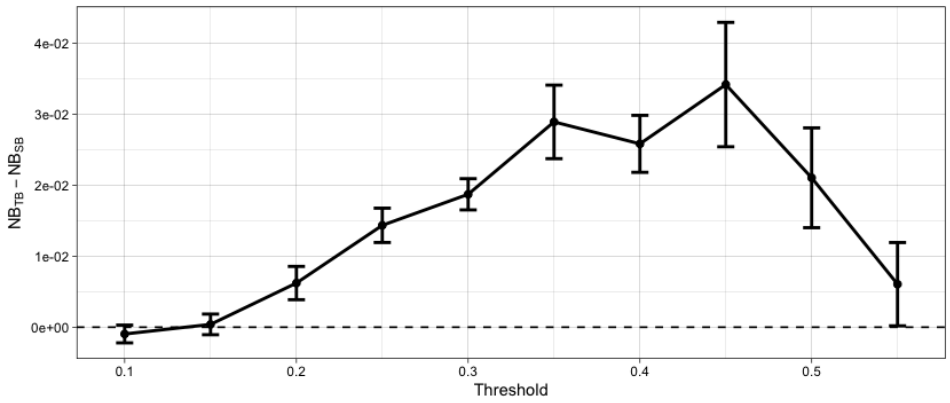


Figure 3.16 Difference in Net Benefit for various t values evaluated on the test set. Error bars correspond to one standard error of the difference (see caption of Figure 3.11 for details).

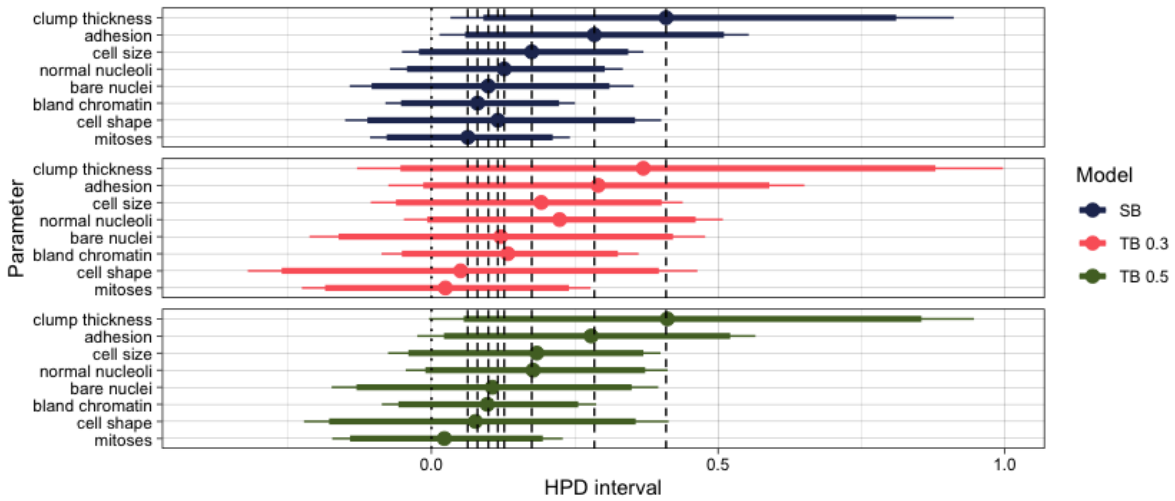


Figure 3.17 Highest posterior density (HPD) regions for the parameters. Dots represent medians, and thick and thin lines represent 90 and the 95% of the HPD regions, respectively. The dashed vertical lines pass through the posterior median values of the SB parameters.

cardiovascular disease (CVD) if a person's CVD risk is 10% or more (NICE, 2016). When guidelines are not available, the specification of t is inevitably subjective, since it reflects the decision maker's preferences regarding the relative costs of different classification errors. In practice, eliciting these preferences may be challenging, despite the numerous techniques that have been proposed in the literature to help with this (e.g., Hunink et al. (2014); Tsalatsanis et al. (2010)). In such situations, we advocate fitting the model for a range of plausible t values that reflect general decision preferences. For example, research in both mammographic (Schwartz et al., 2000) and colorectal cancer screening (Boone et al., 2013) has shown that healthcare professionals and patients alike greatly value gains in sensitivity over loss of specificity. For additional examples on setting t see Vickers et al. (2016) and Wynants et al. (2019). Further examples in which benefits and costs associated with an intervention (as well as with patients' preferences) are taken into account, are provided by Le et al. (2017); Manchanda et al. (2016); Watson et al. (2020).

We discuss the remaining elements for the construction of the weights in Section C.6. There we define the effective sample size for tailoring, ESS_t, and showcase how to use it to set the upper limit for the grid of λ values. In addition, we show our framework is robust to miscalibration of $\pi_u(\mathbf{x}_i)$ and the choice of h. The framework is therefore flexible, allowing many ways for the user to specify the weights.

In contrast to the work of Hand and Vinciotti (2003) our approach is framed within the Bayesian formalism. Consequently, the tailored posterior integrates the attractive features of Bayesian inference - such as flexible hierarchical modelling, the use of prior information and quantification of uncertainty - while also allowing for tailored inference. Quantification of uncertainty is critically important, especially in healthcare applications (Begoli et al., 2019; Kompa et al., 2021). We illustrated this point in Section 1.6. In fact, Figure 1.3 presented the SB posterior for two patients from the breast cancer prognostication case study (Section 3.4.1). We see the predictive distributions for the two patients are centred on similar values. As we discussed in Section 3.4.1, based on these average estimates, chemotherapy should be recommended as a treatment option to both patients. However, the predictive uncertainty for patient 1 is considerable. Using this information we may not be inclined to recommend chemotherapy to patient 1, and instead flag her as needing more information before making a clinical decision.

A few additional comments are in order. In this work we used vague Gaussian priors, but they could be replaced with other application-specific distribution choices. For instance, in the case of high-dimensional data another option could be the sparsity-inducing prior used by Bayesian lasso regression (Park and Casella, 2008). Furthermore, we can easily incorporate external information in a flexible manner, through $\pi_u(\mathbf{x})$, in addition to the prior on the coefficients. If a well-established model exists then it is natural to consider using it to improve the performance of an expanded model. We have implemented such an approach in Section 3.4.2. Cheng et al. (2019) propose several approaches for incorporating published summary associations as prior information when building risk models. A limitation of their approaches is the requirement for a parametric model, i.e., information on regression coefficients. Our method does not have any restriction on the form of $\pi_u(\mathbf{x})$, it can arise from a parametric or non-parametric model.

We note that we opted to use the same set of covariates, $\mathbf{x}$, to estimate both $\pi_{w_i}(\mathbf{x};\boldsymbol{\beta})$ and $\pi_u(\mathbf{x})$. This does not need to be the case. If available, we could instead use another set of covariates, say $\mathbf{Z}$ to estimate $\pi_u(\mathbf{z})$. The set $\mathbf{Z}$ could be a superset or a subset of $\mathbf{X}$ or the two sets could be completely disjoint. We implement this approach in the next chapter (Chapter 4). We also note that in this work we focused on logistic regression based on linear combinations of the covariates to showcase the methodology. This is because it is widely utilised and allows analytical and computational tractability. Nevertheless, we would stress that our framework is generic, and not restricted to either linear combinations or logistic regression. It can accommodate a wide range of modelling frameworks, from linear to non-linear and from classical statistical approaches to state-of-the-art machine learning algorithms. As a result, future work could consider such extensions to other models. Also, future work could consider the advantages of a joint estimation, i.e., both steps, stage 1 (estimation of weights) and stage 2 (estimation of weighted prediction probabilities) jointly. A further direction is the extension of the framework to high-dimensional settings. We propose such an extension in the next chapter.

To conclude, in response to recent calls for building clinically useful models (Chatterjee et al., 2016; Shah et al., 2019) we presented an overarching Bayesian learning framework for binary classification where we take into account the different benefits/costs associated with correct and incorrect classifications. The framework requires the modellers to first think of how the model will be used and the consequences of decisions arising from its use - which we would argue should be a prerequisite for any modelling task. Instead of fitting a global, agnostic model and then deploying the result in a clinical setting we propose a Bayesian framework to build towards models tailored to the clinical application under consideration.

3.6 Software

The `R` code used for the experiments in this chapter has been made available as an `R` package, `TailoredBayes`: https://github.com/solonkarapa/TailoredBayes. The implementation is constructed with the scope of being versatile. The user can customise every aspect of the algorithm, if needed, or simply rely on the default options.

Chapter 4

Tailored Bayesian variable selection under unequal misclassification costs

Abstract Risk prediction models for binary outcomes are often constructed using methodology which assumes the costs of different classification errors are equal (see Chapter 3). However, in many healthcare applications this assumption is not realistic. To address this issue, in the previous chapter, we presented Tailored Bayes (TB) a principled, simple and widely applicable umbrella framework to incorporate misclassification costs into Bayesian modelling. Using both simulated and real data we showed that the TB approach allows us to "tailor" model development with the aim of improving performance in the presence of unequal misclassification costs. However, we only considered low-dimensional data. Here, we extend the TB to deal with high(er)-dimensional data, given their widespread use in modern healthcare applications. Consequently, we incorporate the TB approach into a hierarchical sparse regression framework for variable selection. Our aim is to compare TB and standard Bayes (SB) under a wide range of settings both using simulated and real data. Overall, we show that TB favours smaller models (with fewer covariates) compared to SB, whilst performing better or no worse than SB. We thus conclude that TB allows for more parsimonious explanations for the data. In addition, we show the ranking of covariates changes when we take misclassification costs into consideration. This may result in lower data collection costs and different covariates used in further downstream analysis, for instance in genetic fine-mapping and related applications.

Outline In light of our results in the previous chapter, we start by introducing and motivating Bayesian variable selection (BVS) (Section 4.1). We then outline some key concepts of Bayesian variable selection which we then integrate into our TB approach (Section 4.2). In Section 4.3 we use a wide range of simulation studies to compare TB and SB. We then apply TB to three real-world applications, corresponding to a wide range of sample sizes, covariate dimensions and outcome prevalences, and we comment on the

advantages and disadvantages of the two approaches (Section 4.4). Finally, Section 4.5 contains a discussion.

4.1 Introduction

In the previous chapter we showed that the most commonly used models for binary classification seek to minimise the percentage of incorrect classifications (see Section 3.1). This implies the costs of different classification errors are equal. We then argued that in many healthcare applications this assumption is inappropriate. In fact, our two main case studies exemplified this.

In Section 3.4.1 we saw how clinicians make treatment decisions about adjuvant therapy in breast cancer. We derived the target thresholds, between 14% and 23%, where chemotherapy is discussed as a treatment option. These thresholds imply that clinicians consider approximately 3 to 6 times worse to fail to administer treatment that should have been given (a false negative) than to give unnecessary treatment (a false positive)[1].

Similarly, in Section 3.4.2 we presented the guidelines on choosing between two procedures (TAVI and SAVI) in cardiac surgery patients. Again, the recommended target thresholds of 0.1 and 0.2 imply clinicians consider approximately 4 to 9 times worse a false negative than a false positive prediction.

To address this issue of incorporating information about different misclassification costs during model training we proposed Tailored Bayes (TB), a framework for Bayesian inference that allows us to incorporate misclassification costs into commonly used models for binary outcomes. The framework allows us to tailor model development with the aim of improving performance in the presence of unequal costs. We used both simulations and real-data applications to demonstrate the improvement in predictive performance over standard modelling approaches. An interesting finding was that the relative importance of the covariates changes considerably under the TB framework. This phenomenon was present in all case studies (Section 3.3). We concluded TB can change the covariates' contribution towards the predicted risks. An interesting extension is to consider what might happen in high(er)-dimensional settings. We hypothesise the change in the covariates' contribution may lead us to prioritise different covariates to include in the final model (or selection of models) under the TB framework.

Hence, in this chapter, we extend the framework to incorporate a variable (covariate) selection procedure.

4.1.1 Motivation for Bayesian variable selection

Prediction models are increasingly important in the current era of precision medicine (Kattan et al., 2016). Prediction models are often used to describe the association of an outcome of interest with several covariates (or risk factors). In some applications, one may be interested in predicting the outcome using

[1]Recall $t = \frac{1}{1+B/H}$ (see Section 1.4). Setting $t = 0.14$ (and 0.23) and solving for B/H gives approximately 6 and 3, respectively. We saw in Section 1.4, B is the harm from a false negative result and H is the harm associated with a false positive result.

the covariates. For examples of such applications, see chapters 1, 2, and 3. In other applications, one may want to quantify the effects of covariates on the expected value of an outcome, for example, when assessing the predictive value of a cardiovascular risk factor. One particular application where this is paramount is genetic fine mapping (Huang et al., 2017; Schaid et al., 2018). The goal of fine mapping is to identify the most likely genetic variants that causally affect some trait(s) of interest. In other words, the main goal of fine mapping is to increase our mechanistic understanding, rather than to build a better predictive model.

In either case, at the beginning of the analysis many covariates may be available for inclusion in a model, but it is not always clear upfront if all or just some of them should be included in the final model. This is especially important in situations where a large number of potential covariates are available. The inclusion of unnecessary covariates in a model has several disadvantages, such as increased risk of multicollinearity, insufficient samples to estimate all model parameters, overfitting the current data leading to poor predictive performance on new data and making model interpretation more difficult (van de Schoot et al., 2021).

A common approach is to select a subset from a large set of candidate covariates based on statistical significance. The poor performance of models derived this way has been demonstrated (Steyerberg et al., 2018). Similar algorithmic approaches (such as stepwise and best-subset regression) have also been found to perform poorly in simulations and case studies (see Rothman et al., 2008; Viallefont et al., 2001, and references therein).

In contrast, Bayesian model selection provides a systematic way to compare and assess the relevance and the performance of many models or covariates (Mattei, 2019). Bayesian model comparison is theoretically attractive because it produces a distribution over a set of models. This (a) allows comparison of non-nested models, (b) it is consistent provided the data generating model is among the compared ones (Bernardo and Smith, 2009), (c) is connected with cross-validation (Fong and Holmes, 2020), (d) provides a comprehensive quantification of uncertainty in the important covariates, and subsets of covariates.

As variable selection is just a subset of model selection, all the above are directly applicable. More specifically, attractive features of Bayesian variable selection (BVS) include inference of posterior probabilities for each covariate, posterior inference on competing combinations, and, potentially most importantly, the possibility of incorporating prior information into the analysis. That is, Bayesian approaches allow for prior knowledge about correlations among the covariates to be incorporated into the analysis. For example, in models with gene expression data, spike-and-slab variable selection priors incorporating knowledge of gene-to-gene interaction networks have been employed to aid the identification of predictive genes (Li and Zhang, 2010), as well as the identification of both relevant pathways and subsets of genes (Stingo et al., 2011). In addition, BVS has been successfully applied in many related areas such as genome-wide association studies (e.g., Guan and Stephens (2011); Zhao et al. (2019)), genetic fine mapping (e.g., Wallace et al. (2015)) and other biomedical applications (Ge et al., 2019; Gu et al., 2020; Hill et al., 2012; Theorell and Nöh, 2020).

Risk prediction models can be combined with BVS to provide a powerful mechanism to select covariates, enforce sparsity and quantify uncertainty on the selected covariates. In this chapter, armed with the many attractive features of BVS we provide a comprehensive comparison of TB and SB in both simulated and real-data.

The rest of the chapter is organised as follows. Section 4.2 provides an overview of standard Bayesian model/variable selection concepts which we then incorporate into our TB approach. Section 4.3 contains a series of simulation studies aimed to compare TB and SB in wide range of scenarios. In Section 4.4, we further compare TB and SB on three broad-based real-world data applications involving (1) the prediction of hospital mortality, (2) the diagnosis of diabetes, and (3) prediction of breast cancer relapse. Overall, our results are summarised as follows. TB generally favours sparser models (with fewer covariates) compared to SB. At the same time, it performs better or no worse than SB. In addition, the relative importance of the covariates changes under the TB approach, implying that different covariates may be chosen for further downstream analysis. We end with a discussion highlighting avenues for further work (Section 4.5).

4.2 Methods

In Section 4.2.1 we introduce some generic concepts on Bayesian model selection which we then specialise to variable selection (Section 4.2.2). In Section 4.2.3 we present the TB model formulation. We finish this section describing the metrics we use when comparing TB and SB (Section 4.2.4)

4.2.1 Preliminaries

We consider a prediction problem with covariates $\mathbf{x}$ and an outcome y. The observed data is denoted by $D = \{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ and a future observation by $(\mathbf{x}_*, y_*)$. Under an assumed model, M, y has density $p(y|\mathbf{x}, \boldsymbol{\theta}, M)$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Conditioning on observed data D by Bayes' theorem we get the *posterior distribution* $p(\boldsymbol{\theta}|D, M)$ for the model parameters,

$$p(\boldsymbol{\theta}|D, M) = \frac{p(D|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{\int p(D|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)} \tag{4.1}$$

where $p(D|\boldsymbol{\theta}, M)$ is the likelihood and $p(\boldsymbol{\theta}|M)$ is the prior density of $\boldsymbol{\theta}$ under model M. Equation (4.1) can in turn be used to determine the *posterior predictive distribution*,

$$p(y_*|\mathbf{x}_*, D, M) = \int p(y_*|\mathbf{x}_*, \boldsymbol{\theta}, M)p(\boldsymbol{\theta}|D, M)d\boldsymbol{\theta} \tag{4.2}$$

An underlying assumption so far is that the chosen model M is adequate for its designed purpose. This may not be the case. For example, a grossly misspecified model may describe the actual problem very poorly. Hence, we seek to find a model or collection of models that is useful. We evaluate the usefulness of a model by its ability to make predictions about future (i.e., yet unseen) observations.

Hence, after specifying a set of alternative models $\{M_k\}_{k=1}^K$ and a corresponding prior $p(M_k)$ on each model, one can integrate over the models and thereby arrive at the *Bayesian model averaging* (BMA) predictive distribution (Hoeting et al., 1999),

$$p_{BMA} = \sum_{k=1}^{K} p(y_*|\mathbf{x}_*, D, M_k) p(M_k|D) \tag{4.3}$$

where $p(M_k|D)$ are the posterior probability of the model M_k. Equation (4.3) is an average of the posterior predictive distributions under each of the models considered, weighted by their posteriors probabilities. The term $p(M_k|D)$ is given by

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=1}^{K} p(D|M_l)p(M_l)} \tag{4.4}$$

where

$$p(D|M_k) = \int p(D|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k \tag{4.5}$$

is the integrated likelihood[2] of model M_k, $\boldsymbol{\theta}_k$ is the vector of parameters of model M_k, $p(\boldsymbol{\theta}_k|M_k)$ is the prior of $\boldsymbol{\theta}_k$ under model M_k, $p(D|\boldsymbol{\theta}_k, M_k)$ is the likelihood and $p(M_k)$ is the prior probability of model M_k.

The model posterior distribution $p(M_1|D), \ldots, p(M_K|D)$ is the fundamental object of interest for model selection. In the following section we focus on a special case of the model selection problem: variable selection.

4.2.2 Bayesian Variable Selection

The problem of variable selection arises when one wants to model the relationship between y and a subset of $\mathbf{x} = (x_1, \ldots, x_P)$ [3], but there is uncertainty about which subset to use. Such a situation is particularly of interest when P is large and $\mathbf{x}$ is thought to contain many redundant or irrelevant variables. Each model under consideration corresponds to a distinct subset of $\mathbf{x}$. The linear predictor may be written as

$$\eta = \sum_{j=1}^{P} \gamma_j x_j \beta_j \tag{4.6}$$

where $x_j (j = 1, \ldots, P)$ are the covariates and $\beta_j \in \mathbb{R}$, are the regression coefficients. It will be convenient throughout to index each of these 2^P possible subset choices by the vector

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)' \in \{0,1\}^P$$

[2] This quantity is also often called the "marginal likelihood" or "evidence" of model M_k.

[3] to simplify notation we drop the subscript i.

where $\gamma_j = 0$ or 1 according to whether β_j is zero or not, respectively. The vector $\boldsymbol{\gamma}$ uniquely defines a specific model. Note that here, $\boldsymbol{\gamma}$ plays the role of model identifier M_k described above. Hence, in the following we substitute $\boldsymbol{\gamma}$ for the model indicator k. For instance, in the following section we talk about model space prior denoted by $p(\boldsymbol{\gamma})$ which is conceptually equivalent to $p(M_k)$.

Model Space Priors for Variable Selection

For the specification of the model space prior, most BVS implementations (e.g., George and McCulloch (1993, 1997)) have used independence priors of the form

$$p(\boldsymbol{\gamma}) = \prod_j w_j^{\gamma_j}(1-w_j)^{(1-\gamma_j)} \tag{4.7}$$

Under this prior, each x_j enters the model independently of the other covariates, with probability $p(\gamma_j = 1) = 1 - p(\gamma_j = 0) = w_j$. A useful reduction of (4.7) has been to set $w_j := w$, meaning all covariates have equal inclusion probability, w, of being included. This reduction yields

$$p(\boldsymbol{\gamma}) = \prod_j w^{q_\gamma}(1-w)^{(1-q_\gamma)} \tag{4.8}$$

where $q_\gamma := \boldsymbol{\gamma}'1$ denotes the size of the γ^{th} subset. The hyperparameter w is the a priori expected proportion of variables in the model. Setting $w = 1/2$, yields the popular uniform prior

$$p(\boldsymbol{\gamma}) = 1/2^P$$

This prior puts most of its weight near models of size $P/2$ because there are more of them. An alternative is to assign a prior on w. For instance, using a beta prior, i.e., $w \sim Beta(a,b)$, (4.8) becomes,

$$p(\boldsymbol{\gamma}) = \frac{B(a+q_\gamma, b+p-q_\gamma)}{Beta(a,b)} \tag{4.9}$$

where $B(a,b)$ is the beta function. Thus, by setting a and b, one could represent prior belief about the proportion of the covariates that should be included in the model. Under such a prior, the components of $\boldsymbol{\gamma}$ are exchangeable but not independent (except for the special case (4.7)).

As in Ley and Steel (2009) and Scott and Berger (2010), we set $a = 1$, so that the prior distribution on model size is nonincreasing in q_γ. The hyperparameter b can then be chosen to reflect the expected model size, the global prior probability of at least one association or the marginal prior odds that any covariate is associated with the outcome. Following Wilson et al. (2010) we choose $b = \psi P$ i.e., b is proportional to total number of covariates P. Under this formulation the expected model size is $\frac{P}{\psi P+1}$, which approaches a limit of $1/\psi$ as P approaches infinity, the global prior odds of any effect is constant at $1/\psi$, and the marginal prior odds of any single variable having an effect is $1/\psi P$, and therefore decreases with the total number of covariates. As a result, ψ can be chosen to induce more or less

sparsity, depending on prior beliefs. In practice, we recommend trying a range of several ψ and choosing the value which optimizes predictive performance in the validation data. Here, we use $\psi = 0.1$ (unless otherwise stated) to ensure a fair comparison between TB and SB.

Parameter Priors for Selection of Nonzero coefficients

We now consider parameter prior formulations for variable selection where the goal is to ignore only those x_j for which $\beta_j = 0$ in (4.6). In effect, the problem then becomes that of selecting a submodel of (4.6) of the form

$$\eta = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma \tag{4.10}$$

where $\mathbf{X}_\gamma$ is the $n \times q_\gamma$ matrix whose columns correspond to the γ^{th} subset of $x_1, \ldots, x_P$, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of unknown regression coefficients. Here, $\boldsymbol{\gamma}$ plays the role of the model parameter $\boldsymbol{\theta}_k$ described in Section 4.2.1. The most common option is to assign independent normal priors to the non-zero coefficients centred on zero, with a common variance σ^2_β,

$$p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2_\beta) \tag{4.11}$$

Rather than fixing σ^2_β, which controls the magnitude of included effects, we use a flexible hyper-prior to allow adaption to the data at hand. We assign the following vague hyper-prior for σ_β

$$\sigma_\beta \sim \mathscr{U}(0.01, \sigma_u) \tag{4.12}$$

which is recommended by Gelman et al. (2006) and has been used previously (Newcombe et al., 2019; O'Hara et al., 2009). We use $\sigma_u = 5$ for all the analysis below.

Posterior Calculation and Exploration for Variable Selection

An exhaustive evaluation of all possible combinations of covariates is computationally prohibitive even for a moderate number of covariates. Hence, we rely on MCMC methods (Brooks et al., 2011) which have become a principal tool for posterior evaluation and exploration in Bayesian variable selection problems (see Section D.1 for details). Such methods are used to simulate a sequence of models

$$\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots \tag{4.13}$$

that converges (in distribution) to $p(\boldsymbol{\gamma}|D)$ (the model posterior probability), i.e.,

$$\tilde{p}(\boldsymbol{\gamma}|D) = \frac{\sum_{b=1}^{B} I[\boldsymbol{\gamma}^{(b)} = \boldsymbol{\gamma}]}{B} \xrightarrow{d} p(\boldsymbol{\gamma}|D) \tag{4.14}$$

where B is the number of MCMC iterations and $I[.]$ is the indicator function. The marginal posterior inclusion probability (PIP) $p(\gamma_j = 1|D)$ can be approximated by

$$\tilde{p}(\gamma_j = 1|D) = \sum_{\gamma \in U} \gamma_j \tilde{p}(\boldsymbol{\gamma}|D) \tag{4.15}$$

where U is the set of unique models that were sampled. The simulated models (4.13) can also play an important role in model averaging. For instance, (4.3) can be approximated by

$$\tilde{p}_{BMA} = \sum_{\boldsymbol{\gamma} \in S} \tilde{p}(y_*|\mathbf{x}_*, D, \boldsymbol{\gamma}) \tilde{p}(\boldsymbol{\gamma}|D, S) \tag{4.16}$$

where S is a manageable subset of models, $\tilde{p}(\boldsymbol{\gamma}|D,S)$ is a probability distribution over S (given in (4.14)), and $\tilde{p}(y_*|\mathbf{x}_*, D, \boldsymbol{\gamma})$ is the MCMC approximation to $p(y_*|\mathbf{x}_*, D, \boldsymbol{\gamma})$ given by

$$\tilde{p}(y_*|\mathbf{x}_*, D, \boldsymbol{\gamma}) = \frac{1}{B} \sum_{b=1}^{B} p(y_*|\mathbf{x}_*, D, \boldsymbol{\beta}^{(b)}, \boldsymbol{\gamma}) \tag{4.17}$$

4.2.3 Model formulation

Finally, we need to specify the likelihood function. The tailored likelihood function was given in equation (3.2.2). We re-state it here,

$$L(D \mid \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \exp\{-\ell_{w_i}(y_i, \mathbf{x}_i^T \boldsymbol{\beta}_{\boldsymbol{\gamma}})\} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\}}{1+\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\}} \right)^{y_i w_i} \left(1 - \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\}}{1+\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\}} \right)^{w_i(1-y_i)} \tag{4.18}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_\beta^2)^T$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the element-wise product of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and $w_i \in [0,1]$ are the datapoint-specific weights, which are set as

$$w_i = \exp\big\{ -\lambda (\pi_u(\mathbf{x}_i) - t)^2 \big\} \tag{4.19}$$

We refer to Section 3.2.1 for detailed explanation of (4.19). Briefly, datapoints are weighted based on their vicinity to the target threshold, t. More precisely, datapoints are downweighed at an exponential rate. This rate is controlled by λ. $\pi_u(\mathbf{x}_i)$ is a base/reference prediction model[4] and needs to be estimated. To do this, we use the same procedure as in the previous chapter, i.e., we estimate $\pi_u(\mathbf{x}_i)$ using standard Bayesian logistic regression. In the simulations, we use all the available covariates to estimate $\pi_u(\mathbf{x}_i)$. In the real-data applications, we use either all the available covariates or a subset[5], but we do not consider variable selection when estimating $\pi_u(\mathbf{x}_i)$. Another alternative, is to use a well-established model of $\pi_u(\mathbf{x}_i)$. We implement this approach in our third case study (see Section 4.4.3). To avoid overfitting when we estimate both $\pi_u(\mathbf{x}_i)$ and $\pi_{w_i}(\mathbf{x}_i; \boldsymbol{\beta})$ from the same dataset we use the same data splitting process as

[4] Recall from Section 3.2.1, $\pi_{w_i}(\mathbf{x}_i; \boldsymbol{\beta}) = (\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}/1+\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\})^{w_i}$

in the previous chapter (see Section 3.2.4). The only modification is we repeat the data splitting process 10 times in order to assess the reliability of the methods.

Combining the tailored likelihood in (4.18) with the prior we derive the TB posterior as

$$p(\boldsymbol{\theta}|D) \propto L(D \mid \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \tag{4.20}$$

where $p(\boldsymbol{\gamma})$ is the model space prior defined in equation (4.9) and $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is the prior on the parameters conditional on (i.e., included in) the model (see equations (4.11) and (4.12)).

4.2.4 Performance evaluation

We aim to compare TB and SB under a wide range of simulated and real-data scenarios. Our comparison is focused on three aspects:

1. Predictive performance. As in the rest of this thesis our performance measure for model evaluation is the Net Benefit (NB) (see Section 1.5). Recall NB is a suitable model evaluation metric since it allows us to account for the unbalanced misclassification costs. The observed number of true positives is corrected for the observed proportion of false positives weighted by the odds of the target threshold, and the result is divided by the sample size. This net proportion is equivalent to the proportion of true positives in the absence of false positives. To calculate the predictions we use the average of the BMA predictive distribution (see equation (4.16)).

2. Ranking covariates. We are interested in how the covariates' rank changes between TB and SB. For the real-data we use to (marginal) posterior inclusion probability (PIP) (see equation (4.15)). For the simulated data we use the precision-recall plot (see later for details) which allows us to compare the two models with the ground truth.

3. Model size. We are interested in how the posterior model mass is distributed across model sizes between TB and SB. More specifically, we interested in the difference in the preferred models under TB and SB in terms of the number of covariates chosen. The posterior model mass is calculated using the posterior model probabilities (see equation (4.14)).

The three model comparison aspects are motivated as follows. Imagine two modelling frameworks giving rise to Models A and B, using three covariates. Two interesting scenarios can arise. First, the two models may have similar predictive performance but rank differently the three covariates. We argued in the introduction in some biological or other applications covariate ranking may be of (primary) interest. Second, another alternative is both the covariate ranking and the predictive performance for the two models is the same but the latter is based on different number of covariates, e.g., the performance of Model A is based on all three covariates, but the performance of Model B is based only on two covariates. The simulations that follow illustrate both scenarios.

[5] this allows us to illustrate the flexibility of the framework, since it can used in applications where $P \gg n$.

4.3 Simulations

Here we use synthetic and semi-synthetic data to investigate the strengths and weaknesses of TB compared to SB. Simulation 1 is a simple scenario with independent covariates that allows to visualise the effect of TB, that is to show that the decision boundaries actually change. Simulation 2 is a more realistic scenario with correlated covariates. Simulation 3 is an example of a mispecified model, where interaction terms between covariates are not included in the model. Finally, Simulation 4 is a semi-synthetic scenario (real covariates and simulated outcome) which allows us to study the two methods under a more realistic covariate correlation structure.

4.3.1 Simulation 1: Independent Covariates

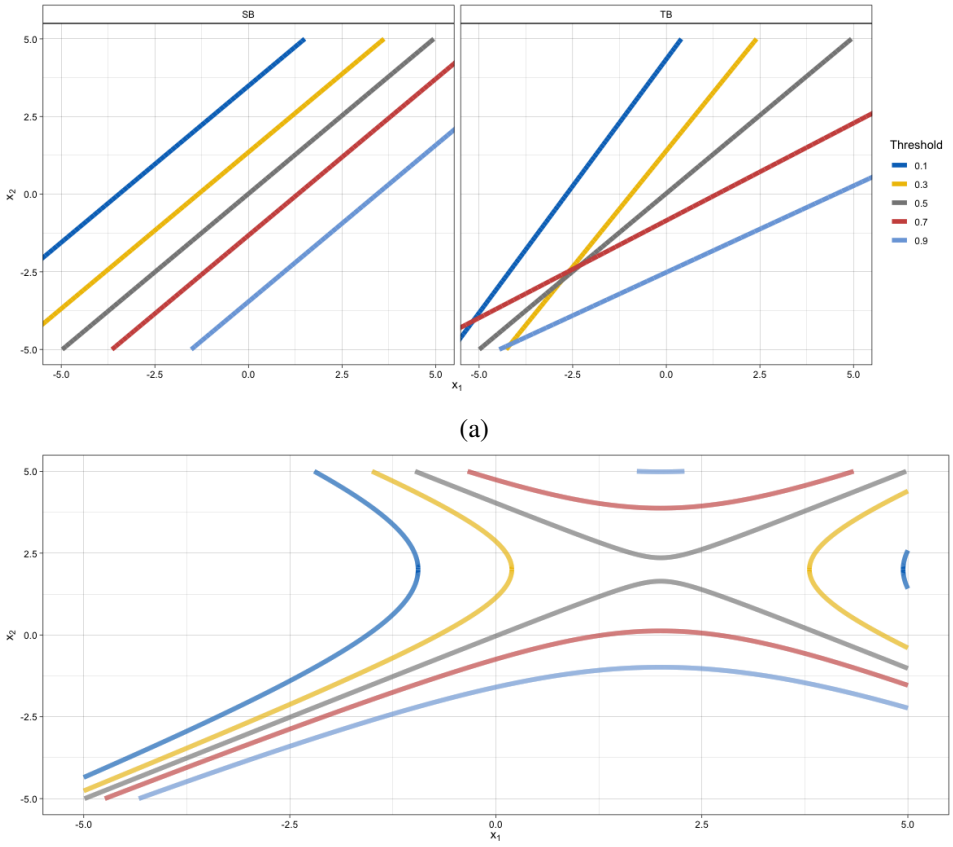
We consider data sets of $n = 1000$ observations with $P \in \{20, 100\}$ continuous covariates. To generate the data, first, n realizations of the binary outcome variable are drawn from the Bernoulli distribution, $y \sim Bernoulli(p)$. Typically the two classes are imbalanced, hence, we set the prevalence, $p \in \{0.1, 0.3, 0.5\}$. After that, the corresponding covariates to each outcome value are drawn from the P-dimensional normal distributions,

$$\begin{aligned} \mathbf{x}|y = 0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \mathbf{x}|y = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \end{aligned} \tag{4.21}$$

with the mean vectors set to $\boldsymbol{\mu}_0^T = (0, 1, \mathbf{0})$ and $\boldsymbol{\mu}_1^T = (1, 0, \mathbf{0})$ where $\mathbf{0}$ is a vector of dimension $P-2$. The covariance matrices are diagonal, $\boldsymbol{\Sigma}_0 = \text{diag}(2, 1, \mathbf{1})$ and $\boldsymbol{\Sigma}_1 = \text{diag}(1, 2, \mathbf{1})$, where $\mathbf{1}$ is a $P-2$ dimensional vector of 1s. This choice of covariance matrices represents a scenario of independent covariates with only univariate relations to the outcome. More importantly, under this setting, only the first two covariates (x_1 and x_2) are associated with the outcome which allows us to visualise the results.

Figure 4.1a shows the posterior median boundaries for SB and TB under the data generating model described above, and for a range of target thresholds. The parallel decision boundaries obtained by applying different thresholds to the standard logistic predictions are clearly not an optimal solution when comparing against the theoretical boundaries depicted in Figure 4.1b. Although limited to estimation of linear boundaries, TB is able to adapt the angle of the boundary to mimic the optimal curves.

Next, we investigate the performance of tailoring across a wide range of settings, by varying: (1) the prevalence of the outcome, p, (2) the target threshold, t, (3) and the covariate dimension, P. Performance is evaluated in independently sampled test sets of size $n = 2000$. Figure 4.2 shows the difference in NB between TB and SB across 100 simulated datasets. Tailoring performs better and generally no worse than SB across all target thresholds for prevalence scenarios 0.3 and 0.5. For $p = 0.1$ the two models are closely matched.



(a)

(b)

Figure 4.1 (a) Posterior median boundaries for SB and TB under $P = 20$ and $p = 0.5$; (b) optimal decision boundaries for target thresholds 0.1, 0.3, 0.5, 0.7, 0.9.

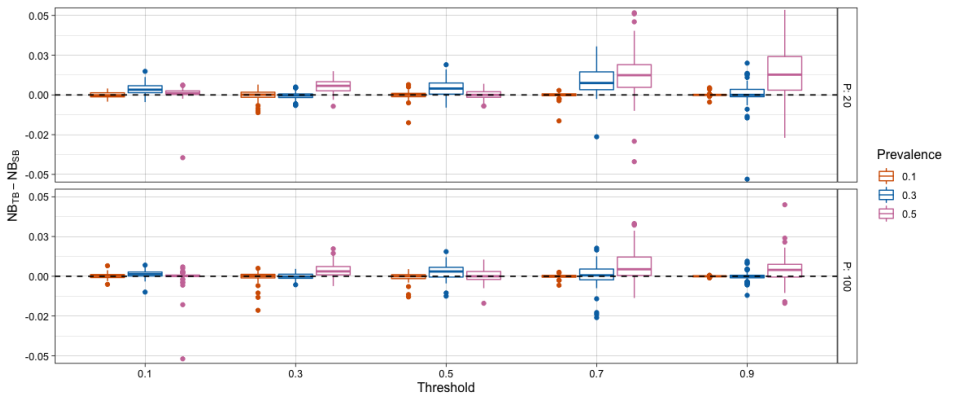


Figure 4.2 Distribution of difference in NB across 100 replications. A positive difference means TB outperforms SB. Each grid corresponds to a different P setting. The lower and upper hinges of the boxplots correspond to the first and third quartiles (25th and 75th percentiles). The upper/lower whisker extends from the hinge to the largest/smallest value no further than/at most 1.5 * IQR from the hinge (where IQR is the inter-quartile range).

Model Size and Detection of Relevant Covariates

Next we investigate the composition of the resulting covariate combinations regarding relevant and irrelevant ones between the two approaches. For our 100 replications, we compute the precision (i.e., positive predictive value) as the ratio of selected relevant covariates among the total number of selected covariates, which answers the question: "what portion of the selected covariates is actually relevant?". Formally,

$$\text{precision} = \frac{TP}{TP+FP}$$

This measure focuses exclusively on the composition of the selected combination. Hence, a model selecting only 1 out of 100 relevant covariates, but no irrelevant covariate would still optimize this measure. To take also into account the total information available in the covariate space we can compute the recall (i.e., true positive rate) as the ratio of selected relevant covariates among all existing relevant covariates, answering the question "what portion of all relevant covariates was selected?". Formally,

$$\text{recall} = \frac{TP}{TP+FN}$$

These two quantities are calculated as follows: in each iteration of the MCMC sampler, TP, FP and FN for TB and SB were counted by comparing the selected model with the true one. Then, TP, FP and FN rates were defined as the average true, false positive and false negative values over the 100 replications. Figure 4.3 shows the results. In all cases, TB achieves the same or marginally better precision than SB, while the recall is the same or slightly worse. This indicates that TB might miss relevant covariates. This is manifested by assigning more posterior mass into smaller models (in terms of number of covariates) (Figure 4.4). But, interestingly, this does not negatively affect predictive performance (see Figure 4.2).

Overall, TB correctly identified the relevant covariates albeit being more conservative (spreading the posterior mass across smaller model sizes) than SB, whilst maintaining the same or improving predictive performance.

Benefits of Model Averaging

Here we investigate the difference in performance between fitting the full model and the BMA solution, as implemented above. This is because, given the dimensionality of the problems investigated, we could, in practice, fit the full model (i.e., include all covariates). This selection of a single model ignores model uncertainty and may have a negative effect on predictive performance. In fact, our results show that the BMA solution over the candidate models yields the best results or at least no worse on expectation (Figure 4.5). This agrees with what is known about the good performance of the BMA (Hoeting et al., 1999; Piironen and Vehtari, 2017). In addition, with model selection we estimate the posterior probability of all models within the considered class of models (or in practice, of all models with non-negligible probability). In our case, this question is asked in variable-specific form: i.e., the task is to estimate the

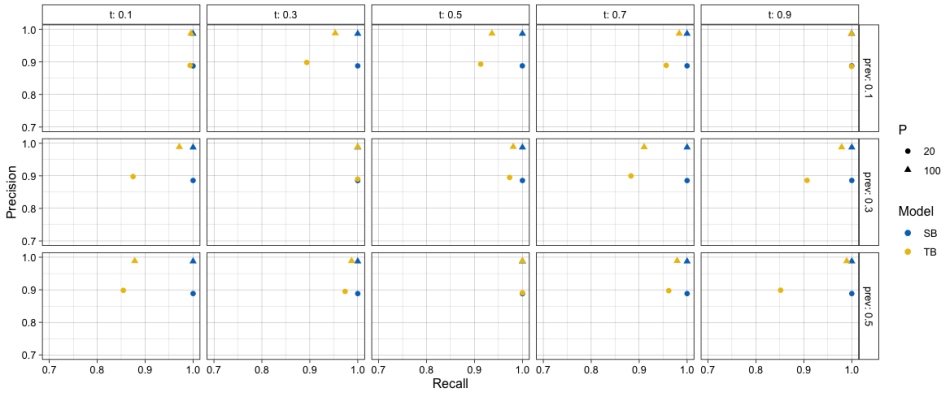
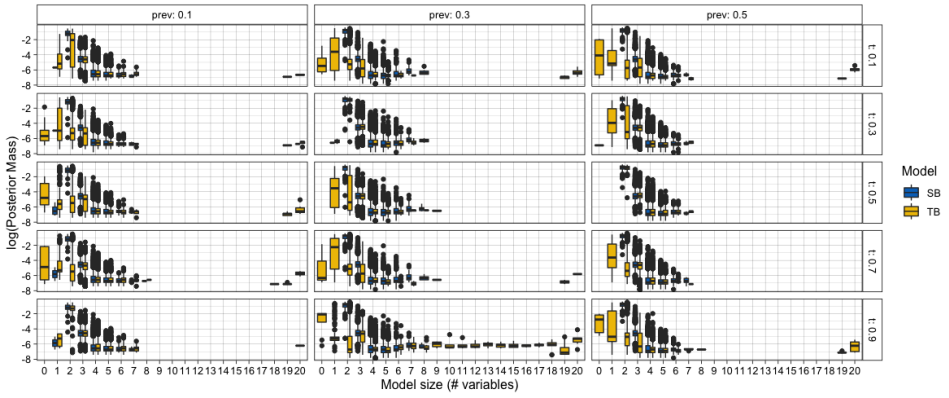
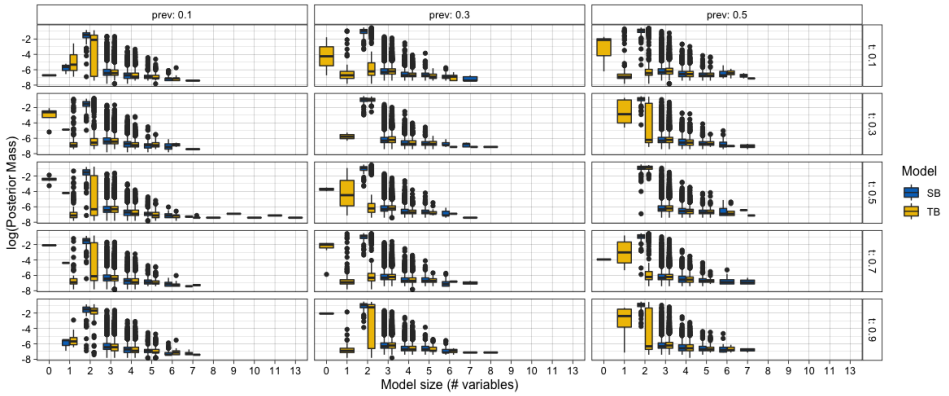


Figure 4.3 Precision-recall plot comparing SB and TB. Precision corresponds to the ratio of relevant detected covariates divided by total amount of covariates in the model. Recall shows the ratio of relevant detected covariates divided by the total existing number of relevant covariates.

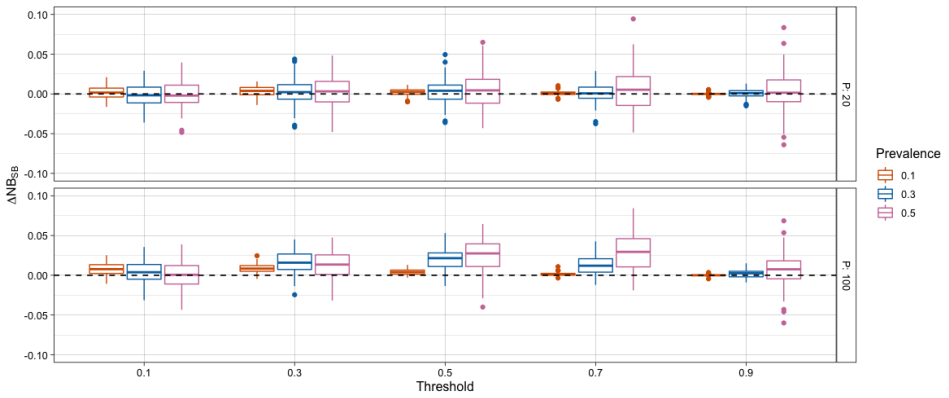


(a)

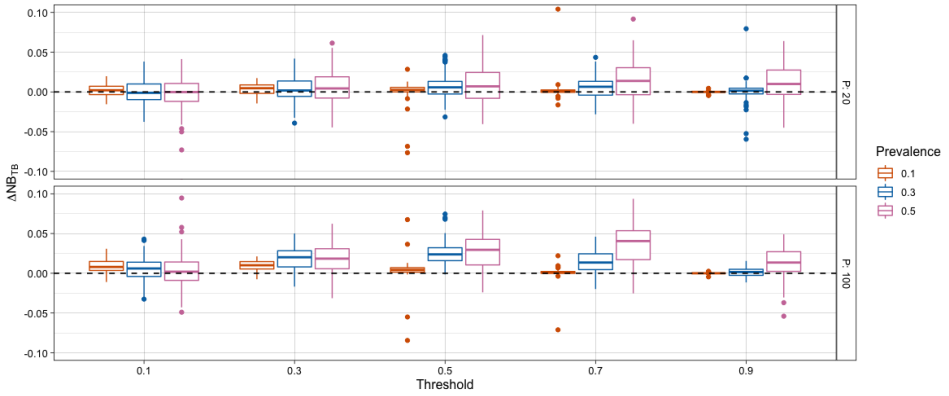


(b)

Figure 4.4 Distribution of Posterior Model Mass (log scale) across model size among top 100 visited models, when (a) $P = 20$ and (b) $P = 100$.



(a)



(b)

Figure 4.5 (a) Difference in NB (ΔNB) for SB, i.e., $SB_{BMA} - SB_{no\ selection}$, (b) Difference in NB (ΔNB) for TB, i.e., $TB_{BMA} - TB_{no\ selection}$. Positive difference means the BMA solution performs better.

marginal posterior probability that a variable should be in the model. Given we may have some a priori expectation that only a small proportion of candidates are truly affecting the outcome, this information should be taken into account in the variable selection.

4.3.2 Simulation 2: Correlated Covariates

Our choice of covariance matrices in the previous section results in independent covariates. However, in most real-world applications, the assumption of completely independent covariates does not necessarily hold true. A typical example is fine mapping applications where the covariates (genetic variants) can be very highly correlated, owing to a phenomenon called linkage disequilibrium (Ott, 1999; Schaid et al., 2018). Therefore, we implement a more widely utilised framework were covariates are correlated (Nikooienejad et al., 2016). We draw the covariates from a multivariate normal distribution centred at zero, i.e.,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

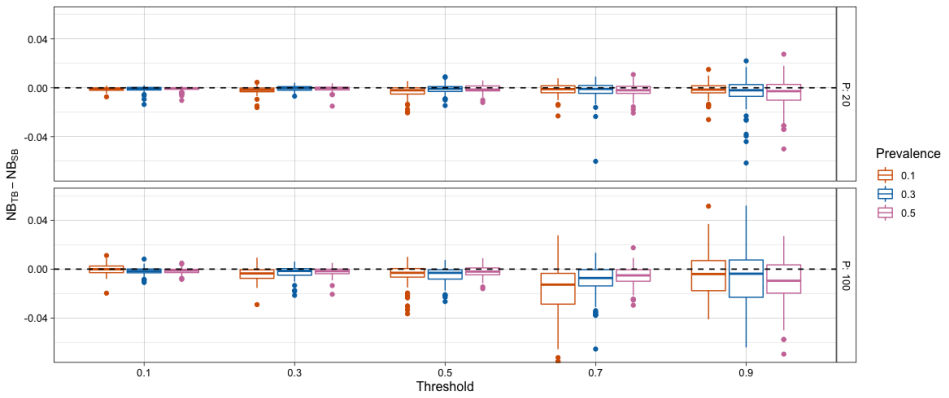


Figure 4.6 Distribution of difference in NB across 100 replications. A positive difference means TB outperforms SB.

For the covariance matrix $\mathbf{\Sigma}$, the diagonal elements of were 1 and off-diagonal elements were $\rho = 0.5$. We then draw realizations of the outcome variable y_i ($i = 1, \ldots, n = 1000$) from the Bernoulli distribution, $y_i \sim Bernoulli(p_i)$ with parameter p_i modelled by a logistic relation to the linear predictor of relevant covariates,

$$p_i = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})}$$

The structure of $\boldsymbol{\beta}^T$ is $(\beta_0, \beta_1, \ldots, \beta_{rel}, \ldots, \mathbf{0})$, where $\dim(\boldsymbol{\beta}) = P$, and $\beta_j = 1, j = 1, \ldots, rel = \lceil P/3 \rceil$. This set-up results in 7 and 34 relevant covariates for $P = 20$ and 100, respectively. The intercept term β_0 is chosen so we obtain approximately the desired outcome prevalence as in the previous section.

The results are given in Figure 4.6. TB provides no benefit under this scenario which is expected because the fitted model is correctly specified. Figure 4.7 plots the precision-recall results. Under $P = 100$, TB achieves better precision than SB. But, under $P = 20$ TB achieves the same or marginally worse precision than SB. In general, the recall is worse for TB. This again indicates that TB might miss relevant covariates which is manifested by assigning more posterior mass into smaller models (Figure 4.8).

4.3.3 Simulation 3: Interaction Simulation

TB is more likely to outperform SB when the model is misspecified. This was the case in Section 4.3.1. Here, the data generating model is the same as in the previous section, but now the including the following relevant interactions, i.e., $x_1 * x_2$, $x_1 * x_3$, $x_5 * x_6$. The main effects of these covariates are included in the model. In total the data generating model is composed of 5 main effects and 3 interactions, all with coefficients equal to $\log(3)$. The fitted model is misspecified as the interaction term is not included.

Figure 4.9 shows the results. TB performs better or no worse than SB. The precision-recall plot is given in Figure 4.10. Again, both models perform equally well. The only exception is for $t = 0.1$

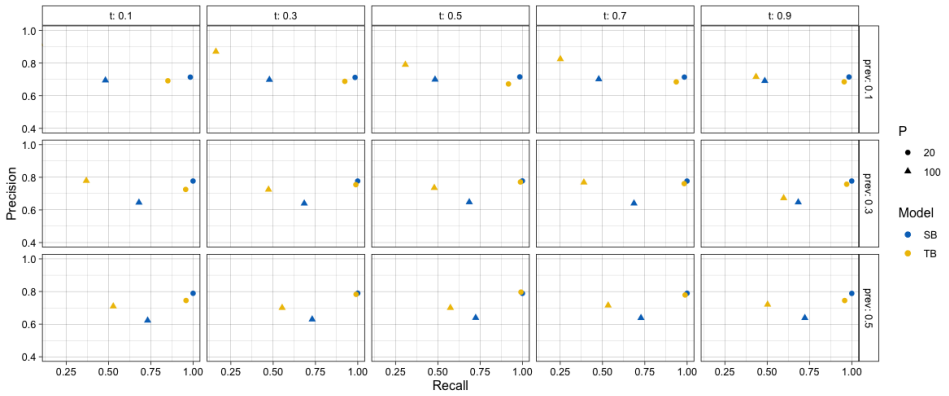


Figure 4.7 Precision-recall plot comparing TB and SB. Precision corresponds to the ratio of relevant detected covariates divided by total amount of covariates in the model. Recall shows the ratio of relevant detected covariates divided by the total existing number of relevant covariates.

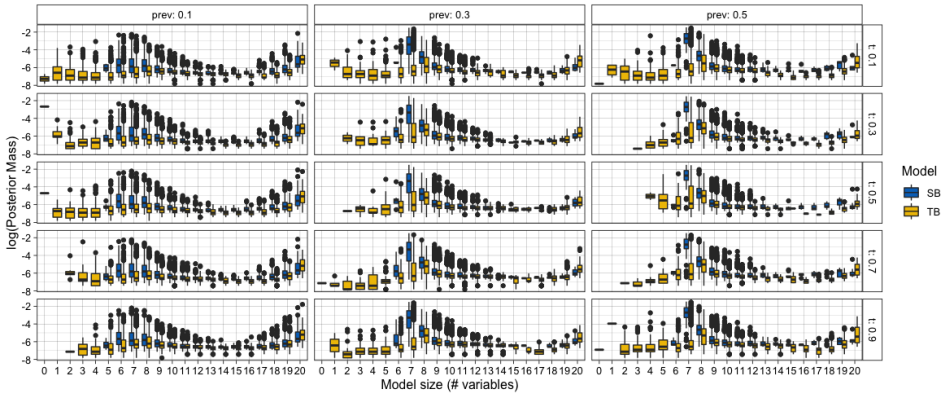


Figure 4.8 Distribution of Posterior Model Mass (log scale) across model size among top 100 models, under $P = 20$.

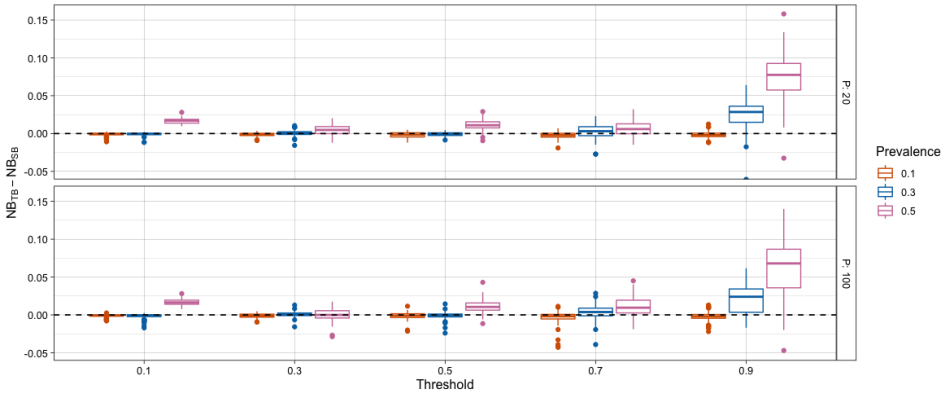


Figure 4.9 Distribution of difference in NB across 100 replications. Each grid corresponds to a different P setting.

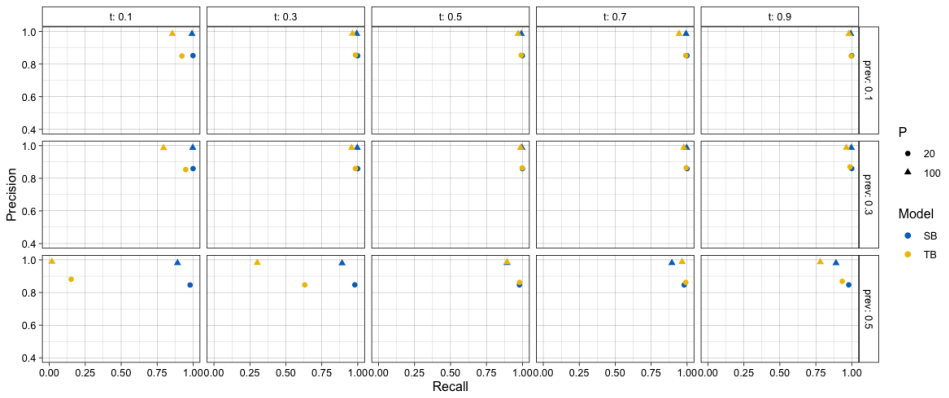


Figure 4.10 Precision-recall plot comparing TB and SB. Precision corresponds to the ratio of relevant detected covariates divided by total amount of covariates in the model. Recall shows the ratio of relevant detected covariates divided by the total existing number of relevant covariates.

and 0.3 and prevalence = 0.5 where TB performs notably worse than SB. Lastly, we investigate how the posterior model mass is distributed across model sizes (Figure 4.11). We see that TB is more conservative, spreading the posterior mass across smaller model sizes.

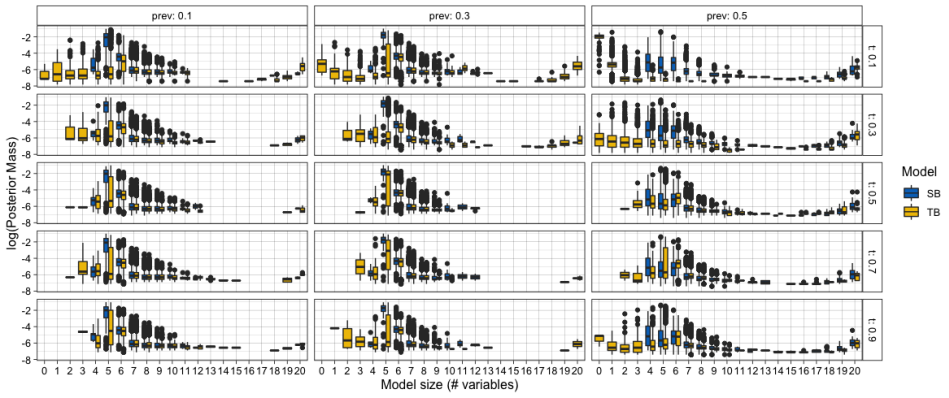
Overall, the conclusions are similar to Section 4.3.1. TB and SB perform equally in terms of selecting the relevant covariates without missing the truly relevant ones. But TB generally favours smaller models while being competitive in terms of predictive performance.

4.3.4 Simulation 4: Semi-Synthetic data

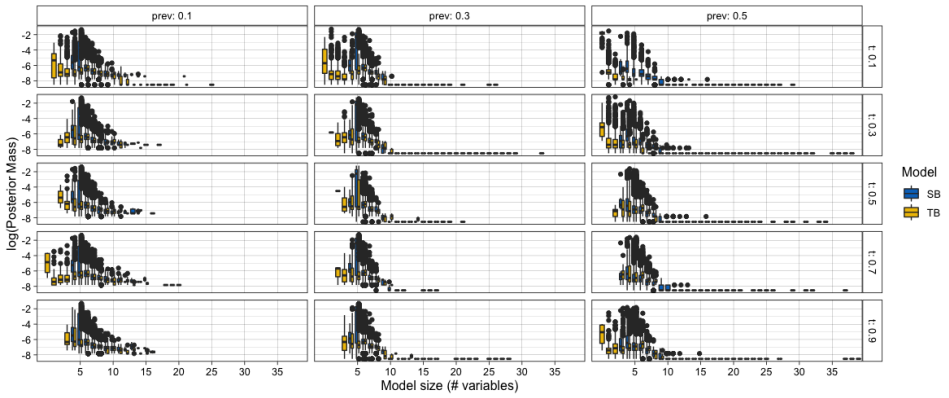
Here we compare TB and SB using data arising from a real-life case study. The advantage of this semi-synthetic scenario is to control the relationship between covariates and outcome whilst based on real data. The semi-synthetic scenario is based on the SUPPORT dataset (see Section 4.4.1 for details). We draw realisations of the outcome variable y_i from a Bernoulli distribution, $y_i \sim Bernoulli(p_i)$ ($i = 1, \ldots, n = 1000$). The parameter p_i is based on the original data analysis and is provided alongside the dataset. The model used to estimate p_i is available online on the Vanderbilt Biostatistics website[6]. Briefly, the model is a Cox proportional hazards regression, from which we extract the p_i corresponding to 180-day risk of mortality. Five out of the 20 covariates are the irrelevant ones, since they were not included in the original model. These were urine, race, diabetes, income and sex. Here, we explore how TB and SB perform in terms of predictive performance, the selected model sizes and identifying the relevant covariates.

Figure 4.12 shows the difference in NB over 100 simulated datasets. There is no difference in average performance between the two models. The precision-recall plot is given in Figure 4.13. TB achieves higher precision and lower recall across all thresholds, except when $t = 0.1$. Under TB if a covariate is selected it is more likely to be relevant (higher precision, and consequently lower false

[6] https://biostat.app.vumc.org/wiki/Main/SupportDesc



(a)



(b)

Figure 4.11 Distribution of Posterior Model Mass (log scale) across model size among top 100 models, when (a) $P = 20$ and (b) $P = 100$.

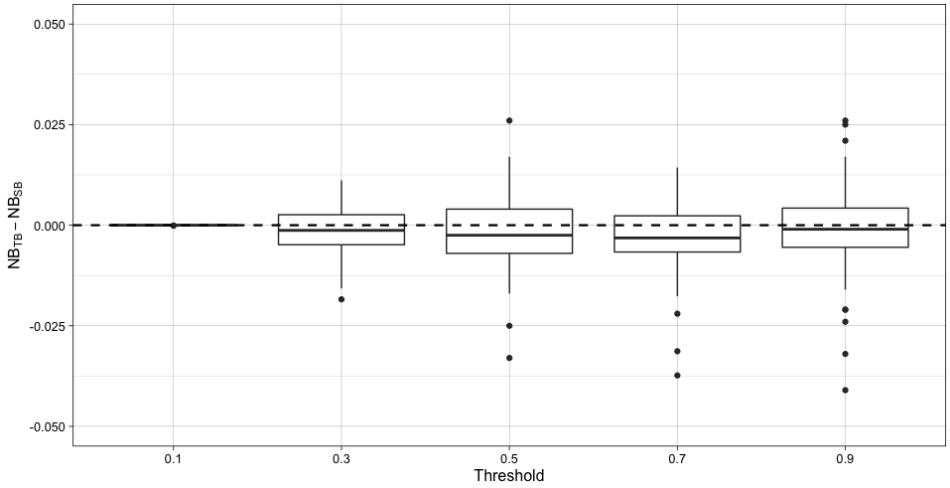


Figure 4.12 Distribution of difference in NB across 100 replications. A positive difference means TB outperforms SB.

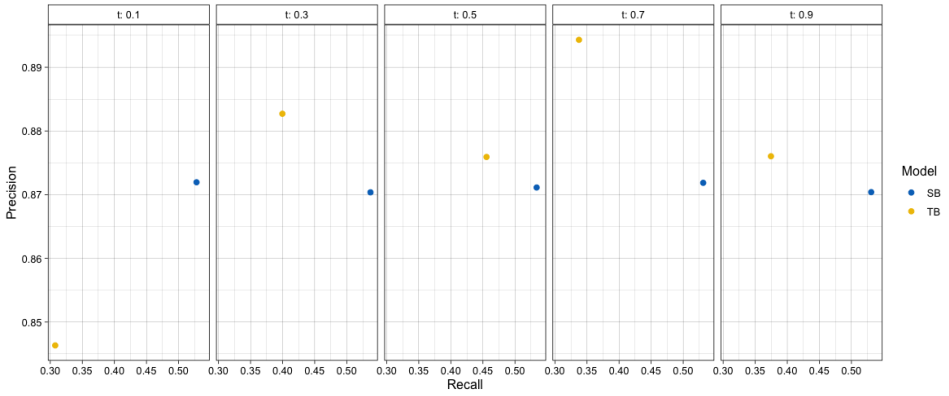


Figure 4.13 Precision-recall plot comparing TB and SB.

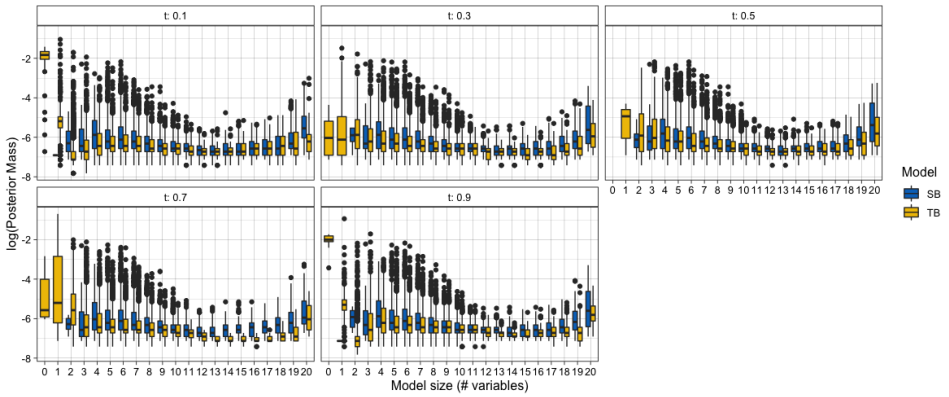


Figure 4.14 Distribution of Posterior Model Mass (log scale) across model size among top 100 models.

discovery rate) but might miss some of the relevant covariates (lower recall). The latter point is illustrated with posterior mass spreading across smaller models under TB compared to SB (Figure 4.14).

4.4 Real data applications

We evaluate the performance of TB and SB on three real-data applications corresponding to a wide range of sample sizes, covariate dimensions and outcome prevalences (see Table 4.1). As before, all the results are presented in terms of predictive performance, covariate ranking and model size. Overall, our empirical results corroborate the findings from the simulations. That is, TB performs better or no worse than SB while favouring sparser models, and prioritising different covariates.

4.4.1 Real data application 1: SUPPORT

For our first case study, we use the data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) (Knaus et al., 1995). In this multicentre study, 9103 hospitalized patients had data recorded including diagnoses, laboratory and vital status (Table 4.2). The

Table 4.1 Real-world datasets used. N: sample size, P: covariate dimension, *Prev*: outcome prevalence.

Name	N	P	*Prev*
SUPPORT	9103	20	0.46
Diabetes	375	10	0.15
METABRIC	1787	1504	0.30

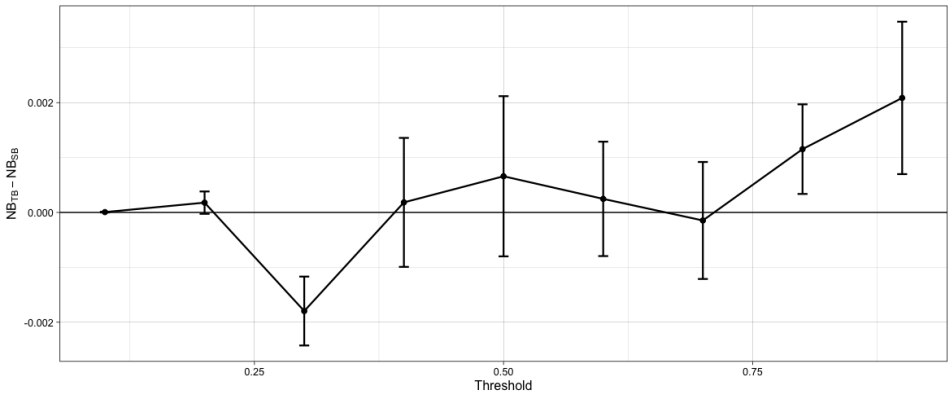


Figure 4.15 Difference in NB for various t values; a positive difference means TB outperforms SB. Error bars correspond to one standard error of the difference.

outcome of interest is 180-day mortality. The dataset is publicly available on the Vanderbilt Biostatistics website[7]. All covariates were standardised prior to the analysis. Some patients had missing values for one or more covariates; in this case we imputed the missing data using the recommended default values listed on the Vanderbilt web site. Data were divided with a 80%/20% split into training and test sets. When estimating $\pi_u(\mathbf{x})$ we use the 13 covariates that form the physiologic score (for details see Knaus et al. (1995)) because they are major prognostic factors (Table 4.2).

As before, we are comparing TB and SB in three aspects: predictive performance, covariate ranking and model sizes. First, we evaluate the out of sample predictive performance for each training-test split using NB as performance measure. TB outperforms SB for $t = 0.8$ and 0.9, performs worse for $t = 0.3$ and there is no difference for the rest of the target thresholds (Figure 4.15).

Second, we evaluate the ranking of covariates under TB and SB. Figure 4.16 shows the posterior inclusion probability (PIP) for each covariate across target thresholds. Notably the variable selection decisions under TB are quite different compared to SB. Even though both models confidently choose *dzgroup, scoma, age, hday* and *ca* as important predictors they disagree on *sod*. Under SB the median PIP for *sod* is 0.8 across thresholds, compared with 0.6 for TB. This implies under standard modelling we would confidently pick *sod* as an important predictor, but under tailoring this would depend on the target threshold. This is exemplified in Figure 4.17 that shows the median PIPs for the two models. The 45-degree line reflects perfect agreement between TB and SB. Most PIPs change between the two

[7] https://biostat.app.vumc.org/wiki/Main/DataSets

Table 4.2 Names and description for each covariate. Covariates with an asterisk are used to form the physiologic score. For more details, we refer to the Vanderbilt website. PaO2: partial pressure oxygen in arterial blood, FiO2: fraction of inspired oxygen.

Covariate	Description
meanbp*	Mean arterial blood pressure
wblc*	White blood cell count in thousands
hrt*	Heart rate per minute
resp*	Respiration rate
temp*	Temperature
pafi*	PaO2/(.01*FiO2)
alb*	Serum albumin
bili*	Serum bilirubin
crea*	Serum creatinine
sod*	Serum sodium
dzgroup*	Disease group
scoma*	SUPPORT Coma Score
age*	(years)
hday	Day in hospital when qualify for study
ca	Cancer by comorbidity or primary disease category (no, present)
urine	Urine output
race	Asian, black, hispanic, white, or other
diabetes	(No or present)
income	One of the following: under \$11k, \$11-\$25k, \$25-\$50k, $>$\$50k
sex	(Male, Female)

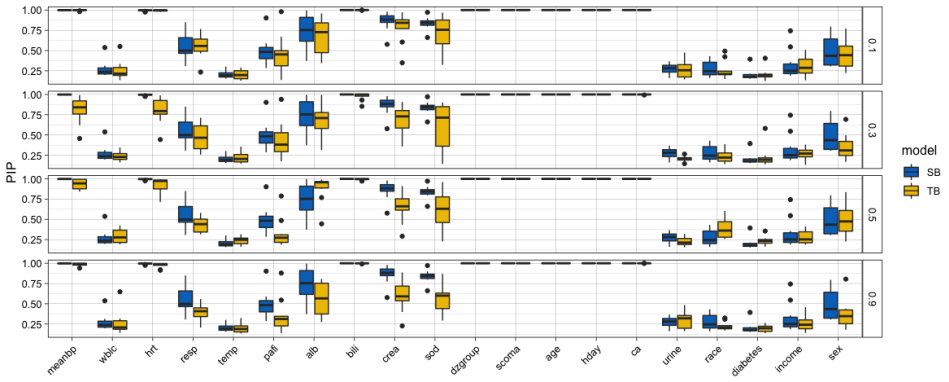


Figure 4.16 Posterior inclusion probability (PIP) for TB and SB for each target threshold (panels).

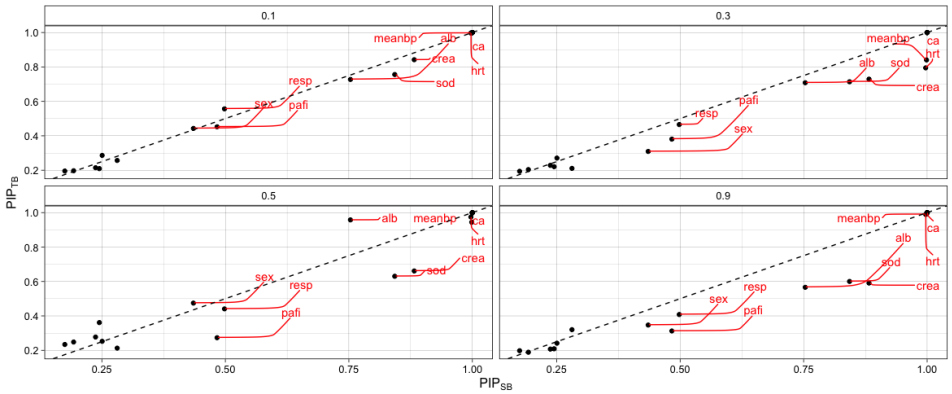


Figure 4.17 Median posterior inclusion probability (PIP) for TB and SB for each target threshold (panels).

models as the target threshold changes. An illustrative example is *alb*: for low thresholds ($t = 0.1$ and 0.3) it has a smaller PIP under TB than SB. This is reversed for $t = 0.5$, where the PIP under TB is higher. Moving to $t = 0.9$ the PIP under TB is smaller again. This is because under TB the relative importance of each covariate changes across different thresholds (Figure 4.18). The depicted covariates exhibit different patterns. For example, *resp* exhibits a decreasing relationship between PIP and threshold, and *alb* exhibits an inverse U-shaped relationship.

Third, we investigate how the posterior model mass is distributed across model sizes. Figure 4.19 shows the posterior probability for each model (dots) under TB and SB. We see that across thresholds tailoring favour smaller (i.e., sparser) models, spreading the posterior mass across smaller model sizes. The top 5 models (in terms of posterior probability) are given in black. We see that they generally tend to correspond to smaller model sizes under TB. Based on the above, we can conclude that TB favours more parsimonious solutions (smaller model sizes) without loss in predictive performance.

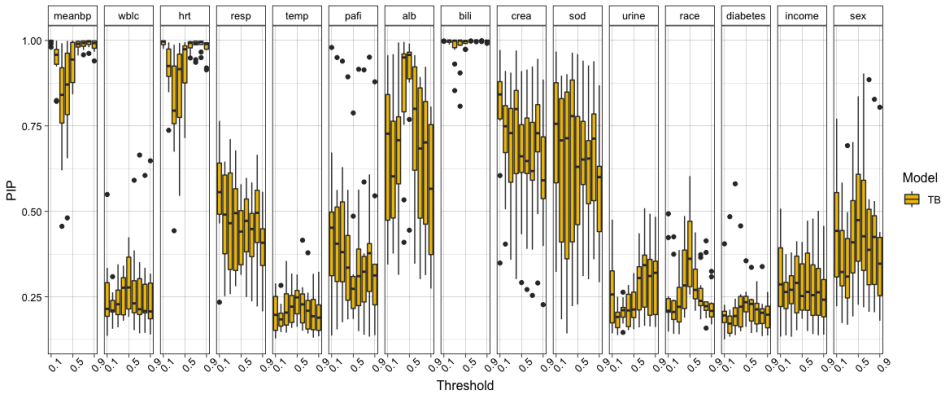


Figure 4.18 Posterior inclusion probability (PIP) of each covariate across target thresholds.

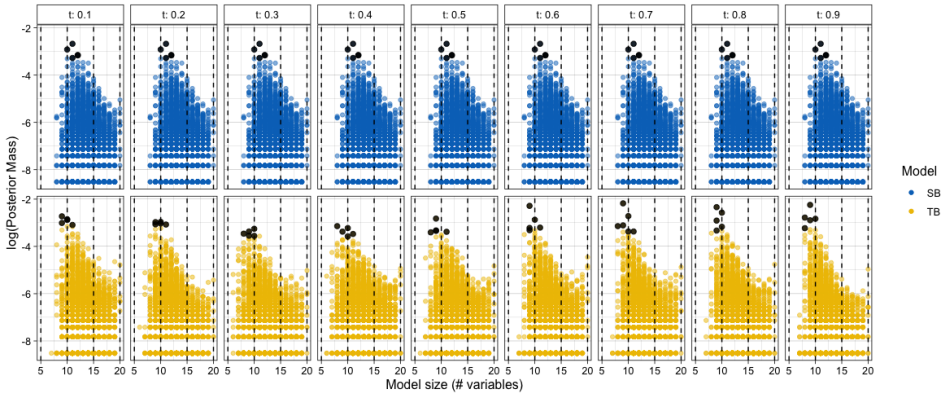


Figure 4.19 Posterior Mass (log scale) across model size. The black dots correspond to the top 5 models (in terms of posterior model probability).

Table 4.3 Names and Description for each covariate.

Covariate	Description
chol	Total cholesterol
stab.glu	Stabilized glucose
hdl	High density lipoprotein
age	(years)
gender	(Male, Female)
bmi	Body mass index
bp.1s	First systolic blood pressure
bp.1d	First diastolic blood pressure
whr	Waist-to-hip ratio
time.ppn	Postprandial time when labs were drawn (minutes)

4.4.2 Real data application 2: Diabetes

We perform a variable selection strategy for indicators of Diabetes Mellitus Type II (DM II) in African Americans. The data (403 subjects) were again obtained from the Department of Biostatistics, Vanderbilt University website. We model the presence of diabetes based on covariates such as total cholesterol, stabilized glucose, age, body mass index, systolic and diastolic blood pressure, waist-to-hip ratio and postprandial time indicator (see Table 4.3). To create the outcome we discretised glycosolated hemoglobin at 7mg/dL. Values of glycosolated hemoglobin $>$ 7mg/dL are usually taken as a positive diagnosis of diabetes (Schorling et al., 1997; Willems et al., 1997). After excluding subjects with missing values, the data consisted of 375 subjects which was split 10 times into training and test samples of sizes 337 and 38, respectively. Furthermore, we use 67 out of the 337 subjects as the design dataset to estimate $\pi_u(\mathbf{x})$, based on all the covariates. To avoid the problem of nonidentifiability, due to collinearity between the covariates, we use weakly informative priors as proposed by Gelman et al. (2008). More specifically, we choose normal priors for the coefficients with variance, $\sigma^2 = 2.5$, after standardising the data to zero mean and unit variance. Recent work has focused on target thresholds $t < 0.5$ when developing/evaluating risk prediction models for DM II (Hippisley-Cox and Coupland, 2017; Mars et al., 2020). Following that line of work we use $t < 0.5$.

First, we evaluate the out of sample predictive performance for each training-test split (Figure 4.20). TB performs equally or slightly better on average for across all thresholds.

Next, we evaluate the ranking of covariates between TB and SB. Figures 4.21 and 4.22 show the PIPs across different thresholds. It is interesting to note the variable selections under tailoring are quite different compared to the standard logistic. In particular, while both the models successfully identify stabilized glucose as an important predictor, age is identified as an important predictor under standard (median PIP = 0.65 across thresholds), but not under tailoring (median PIP = 0.32 across thresholds). Furthermore, the relative importance of each variable seems to change across different thresholds (Figure 4.23).

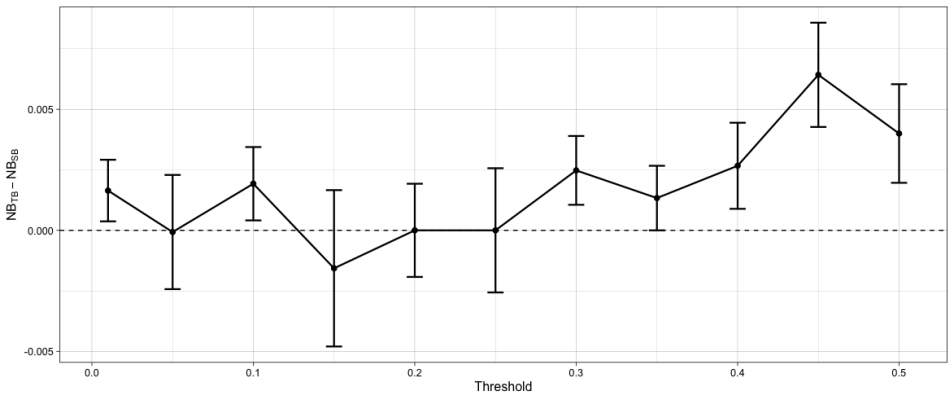


Figure 4.20 Difference in NB for various t values; a positive difference means TB outperforms SB. Error bars correspond to one standard error of the difference.

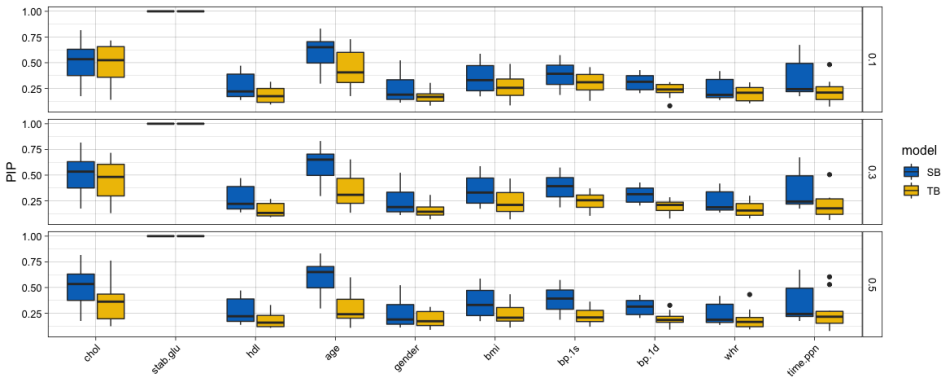


Figure 4.21 Posterior inclusion probability (PIP) for TB and SB for each target threshold (panels).

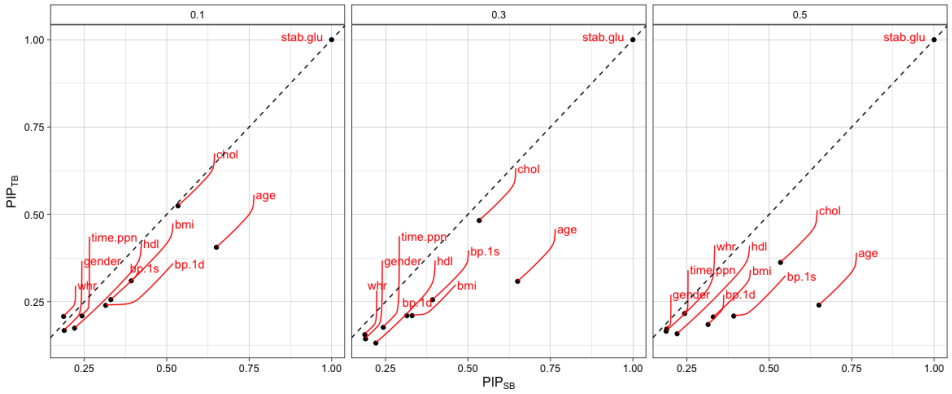


Figure 4.22 Median Posterior inclusion probability (PIP) for TB and SB for each target threshold (panels).

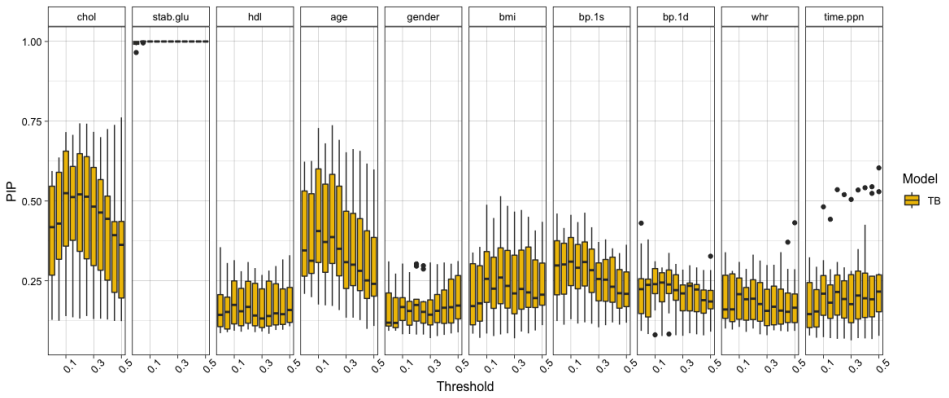


Figure 4.23 Posterior inclusion probability (PIP) of each covariate across target thresholds.

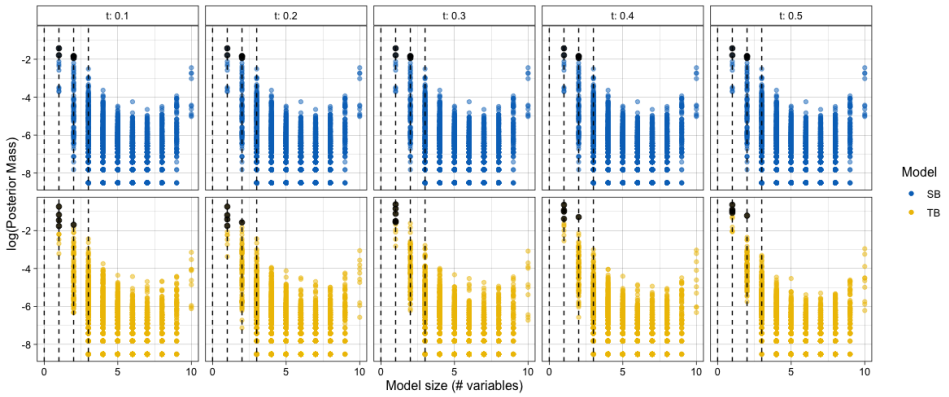


Figure 4.24 Distribution of Posterior Model Mass (log scale) across model sizes. The black dots correspond to the top 5 models (in terms of posterior model probability).

Finally, we investigate how the posterior model mass is distributed across model sizes (Figure 4.24). We see that across thresholds TB favours models with the same or fewer covariates than SB. The top 5 models are given in black. We see that they generally tend to correspond to equal or smaller model sizes under TB. Hence, we again conclude the tailored modelling provides more parsimonious solutions with improvement or no loss in predictive performance. Note that under tailoring smaller models get higher weights in the BMA.

4.4.3 Real data application 3: METABRIC

For our third case study, we aim to identify gene signatures associated with the risk of relapse in breast cancer and to establish their clinical utility. Multi-gene signatures have been extensively studied to provide prognostic and predictive information for breast cancer treatment (Harris et al., 2016). Such molecular assays provide risk prediction with good prognostic value allowing clinically valid risk groups to be defined (Harris et al., 2016; Richman and Dowsett, 2019). Hence, as a secondary goal we investigate whether we can show clinical utility of the already existing risk stratification thresholds.

We use data from $n = 1787$ patients from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort (Curtis et al., 2012; Rueda et al., 2019). The outcome modelled is 5-year risk of relapse. For this data we opted not to estimate $\pi_u(\mathbf{x})$ but instead we use the predictions from an existing model, PREDICT (see Section 1.1). We add an extra step of re-calibration to account for the population/time drift. To choose λ we formulate a working model using common clinicopathological predictors of relapse: age at state entry (diagnosis), tumour grade, tumour size, ER and HER2 status, and the number of positive lymph nodes. Using this working model we perform 10-fold cross-validation.

Here we do not aim to find new gene signatures but instead to establish the clinical utility of the genes that have already been reported in the literature. To this end, we include in our variable selection set-up all the genes that have been appeared in published signatures. Recently, Huang et al. (2018) reported 33 published breast cancer gene signatures with 1895 unique genes out of which 1493 are found in our dataset. We further included 5 genes selected at random as a sanity check. These are CK904829, DB338332, TPH2, B3GAT1, and IER5L. Due to colinearity problems we excluded three genes (HLA-DPA1, HLA-F, HLA-DOB) ending up with a total $P = 1504$ covariates to search over (1498 genes + 6 clinicopathological covariates). We use $\psi = 0.014$ to match the average number of genes (67.8) across existing signatures ($\psi = 0.014$ results in expected model size of 68, a priori).

An important part of the TB methodology is the choice of t. Güler (2017) provides a summary of the proposed thresholds for different assays. For example, using the Predictor analysis of microarray 50 (PAM50) score patients are divided into high (>20%), intermediate (10-20%) and low (<10%) risk groups. Cardoso et al. (2016) defined low clinical risk as the 10-year probability of breast-cancer–specific survival without systemic therapy of more than 88%. Similar, ranges are reported for Oncotype DX. Hence, we set our target threshold, $t < 0.5$, to cover these values.

We start by evaluating the out of sample predictive performance for each training-test split (Figure 4.25). Both models perform equally well across all thresholds of interest. We further compare our models with two commonly used signatures[8], PAM50 (Parker et al., 2009) and Oncotype (Paik et al., 2004) (Figure 4.26). Both signatures feature in the 2016, American Society of Clinical Oncology Breast Cancer Guidelines which provide guidelines for the use of gene-expression assays to help guide treatment decisions regarding the use of adjuvant systemic therapy (Harris et al., 2016). For low thresholds ($t \leq 0.12$), all models demonstrate comparable performance. For higher thresholds TB and SB have a clear advantage over both PAM50 (Figure 4.26a) and Oncotype (Figure 4.26b). Most importantly, TB and SB demonstrate higher clinical utility in the target thresholds where treatment decisions are made (see above).

Next, we evaluate the ranking of variables between TB and SB (Figure 4.27). Once more, the relative importance of each variable changes between models. In particular, while both models identify the positive lymph nodes as an important predictor, they slightly disagree on tumour size and ENC1. Both are identified as less important under tailoring. Furthermore, the relative importance of each variable seems to change across different thresholds (Figure 4.28). For example, ENC1 is more important for low

[8]The algorithms are implemented using the `genefu` package (Gendoo et al., 2020).

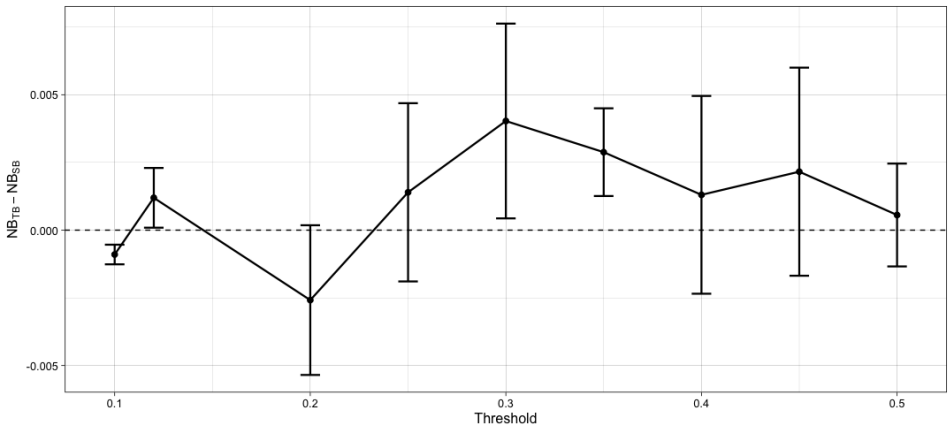
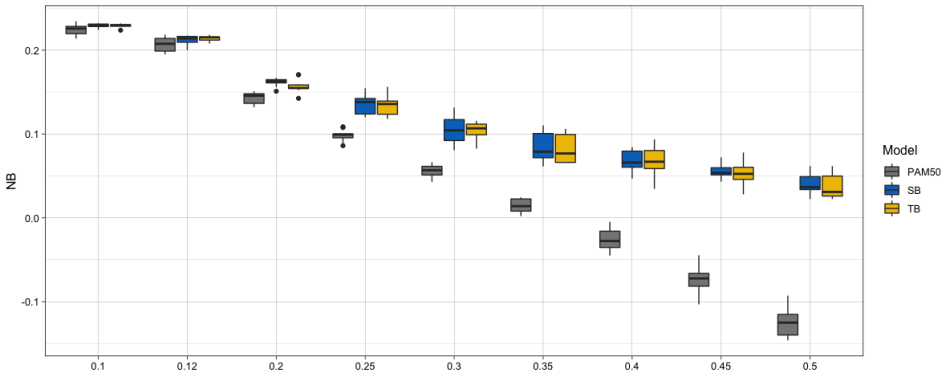
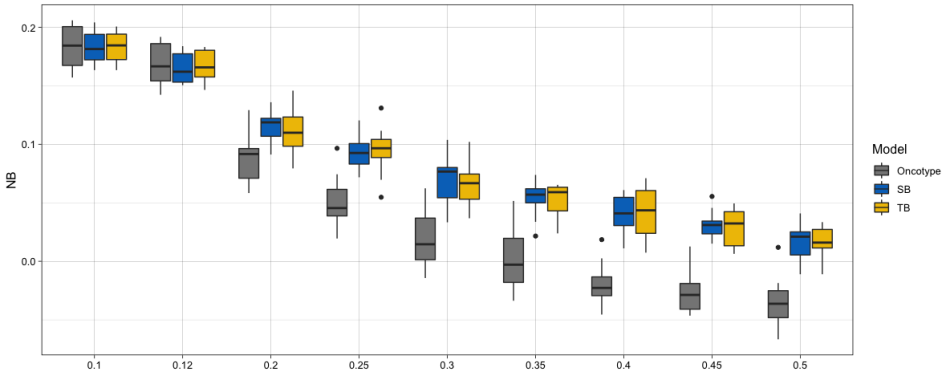


Figure 4.25 Difference in NB for various t values; a positive difference means TB outperforms SB. Error bars correspond to one standard error of the difference.



(a)



(b)

Figure 4.26 NB for various t values comparing TB, SB with (a) PAM50, and (b) Oncotype. Panel (b) shows results only for the ER positive subjects.

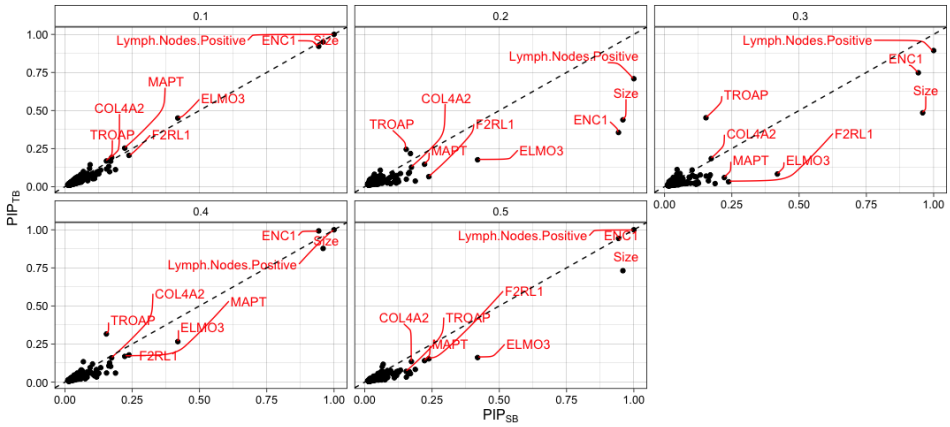


Figure 4.27 Median posterior inclusion probability (PIP) for TB and SB for each target threshold (panels).

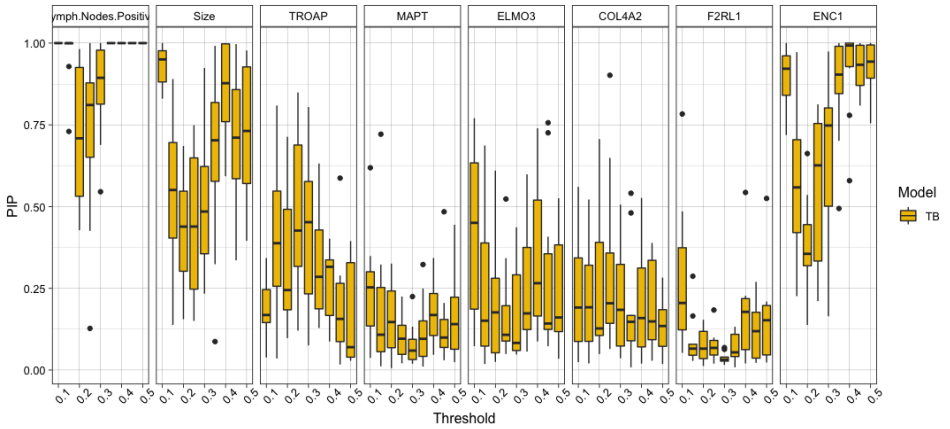


Figure 4.28 Posterior inclusion probability (PIP) for a subset of the covariates across target thresholds.

and high thresholds. Subsequently, we compare the PIPs for the five randomly selected genes (Figure 4.29). Across all thresholds TB assigns the same or lower PIPs to all genes. Even though under both models all genes have very low PIPs, the percentage change (shown in colour) is up to 70%. Additionally, we compare the median PIPs for SB and TB across all the genes included in the 33 published signatures (Figure 4.30). The three genes that stand-out are TROAP, ELMO3 and ENC1. Across thresholds the absolute percentage change for TROAP ranges from 2% to 190%, for ELMO3 from 7% to 80% and for ENC1 from 0% to 64%.

Finally, we investigate how the posterior model mass is distributed across model size (Figure 4.31). Again, we see that across thresholds TB favours models with fewer covariates than SB. This supports the notion that TB can lead to sparser models, which are cheaper and easier to deploy, while retaining predictive performance that is at least as good.

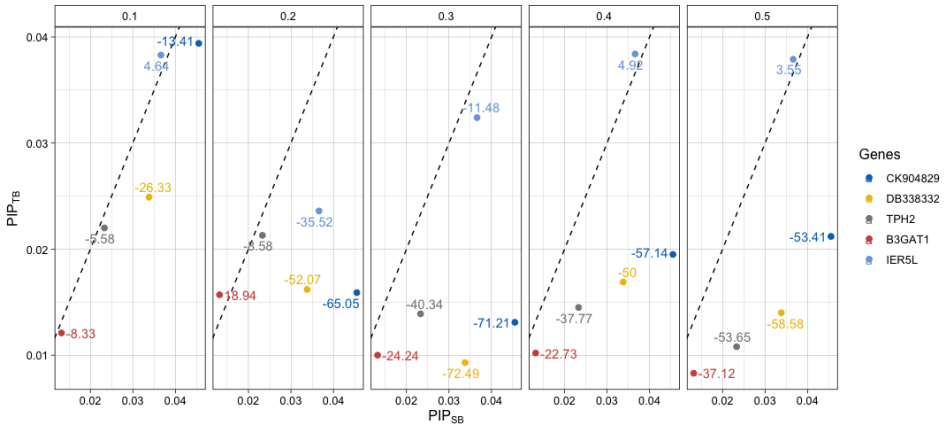


Figure 4.29 Median posterior inclusion probabilities (PIPs) for TB and SB for the five randomly selected genes. Percentage change $\frac{tailor-standard}{standard} * 100$ shown in color.

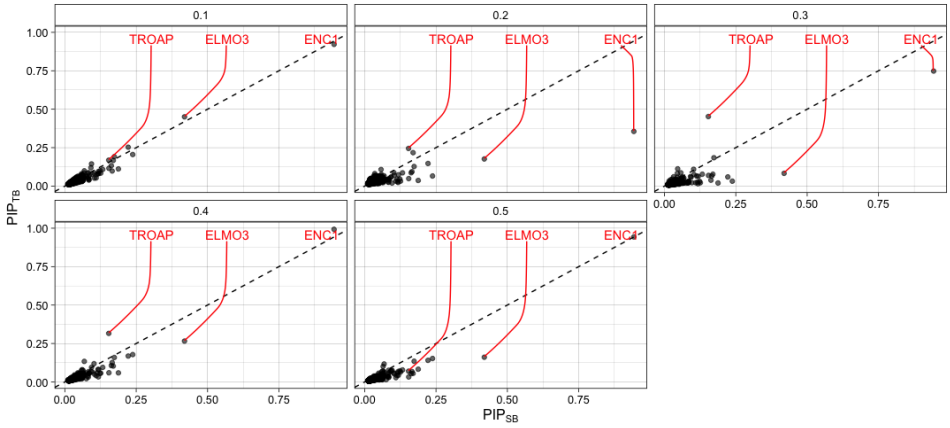


Figure 4.30 Median posterior inclusion probabilities (PIPs) for TB and SB for all the genes included in the 33 published signatures.

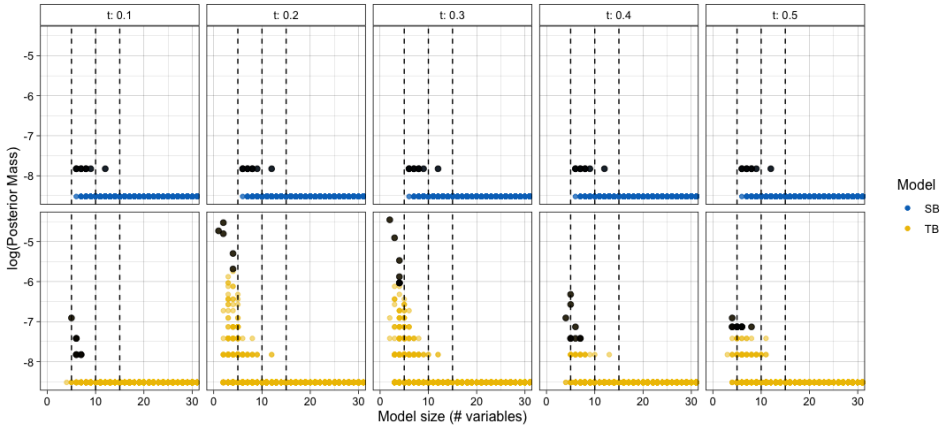


Figure 4.31 Distribution of Posterior Model Mass (log scale) across model sizes. Each panel corresponds to a different choice of t.

4.5 Discussion

Summarising, in this chapter we extended the TB framework to incorporate a variable selection procedure. This allows us to make use of the TB ideas in high-dimensional settings. That is, we can take misclassification costs into consideration while training a binary classification model. The aim of this chapter was to present a comprehensive comparison between TB and SB in settings where variable selection may be adopted. To this end, we simulated data corresponding to a wide range of data generating settings and used three real-word datasets with different sample sizes, number of covariates and outcome prevalences. We found that TB favours smaller models (fewer covariates) compared to SB, whilst performing better or no worse than SB. This pattern was seen both in simulated and real data. This allows more parsimonious explanations for the data at hand. The price to pay is likely to be slightly higher false negative rate (lower recall) while maintaining the same false discovery rate. In addition, we showed the relative importance of the covariates changes when we consider unequal misclassification costs. This has implications for risk prediction models since smaller models may result in lower data collection costs and different covariates being selected for further downstream analysis, for instance in genetic fine-mapping and related applications. In fact, in other related fields, such as Systems Biology model/covariate ranking is important challenge as well (Vyshemirsky and Girolami, 2008).

We opted to introduce sparsity using a model-space prior approach. That is, we viewed the model as a whole, placed priors on $\boldsymbol{\gamma}$, the covariates selected in the model and their coefficients $\boldsymbol{\beta}_{\gamma}$, which then makes the choice of which covariate should be included in the model to be a secondary problem. This is a natural way to introduce sparseness as the model-space prior approach allows us to describe the uncertainty with respect to the model specification given an exhaustive list of candidate models. An alternative approach is to place priors on the individual parameters. Typical examples are the lasso-type (Park and Casella, 2008) and horseshoe-type priors (Carvalho et al., 2010). Such examples are grouped under the continuous shrinkage priors umbrella (see e.g., Polson and Scott, 2010, and references therein). For an overview of the two approaches see O'Hara et al. (2009). Future work could consider the comparing TB and SB when using continuous shrinkage priors.

In the previous chapter we mentioned the flexibility of the TB approach when constructing $\pi_u(\mathbf{x})$. Namely, we have the option to use another set of covariates, say $\mathbf{Z}$ to estimate $\pi_u(\mathbf{z})$. The set $\mathbf{Z}$ could be a superset or a subset of $\mathbf{X}$ or the two sets could be completely disjoint. We used this approach in this chapter. For example, in Section 4.4.1 we used only a subset of the covariates to estimate $\pi_u(\mathbf{x})$. These corresponded to covariates strongly predictive of the outcome. In Section 4.4.3 we opted not to estimate $\pi_u(\mathbf{x})$ but used the predictions from an already existing model (similarly to Section 3.4.2 in previous chapter). In addition, we used a subset of the covariates (the clinicopathological risk factors) to choose a good λ value. The above points showcase the flexibility in combining prior information and the data at hand when constructing the weights.

Finally, both simulation examples and real-data applications have larger sample sizes than number of covariates ($n > P$). We note this is not a requirement for our TB approach and it could also be applied to high(er)-dimensional settings ($P > n$). But we need to point out that since our approach is based

on MCMC methods (to be precise, reversible jump MCMC (RJMCMC), see Section D.1), settings where $P > n$ may be algorithmically challenging. For high-dimensional covariate spaces, the traditional posterior summary statistics of counting the occurrence of each particular posterior model may be infeasible because any model is most likely to be sampled only once in a MCMC with workable length. Of course, this is not an issue inherent to TB but it also affects other models relying on (RJ)MCMC methods. A potential solution would be to construct a RJMCMC sampler for a specific problem of interest, focusing on careful specification of the proposal mechanism and then tuned using pilot runs. Examples for tuning include blocking and re-parameterisation. We refer to the review by Hastie and Green (2012) for such solutions. Another potential avenue is to focus on alternatives to RJMCMC. For instance, both variational inference (Carbonetto and Stephens, 2012) and expectation-maximization (EM) algorithms have been used in BVS settings (Duan et al., 2018; Hayashi and Iwata, 2010).

To conclude, variable selection (and more broadly model selection) is a ubiquitous challenge in statistical modelling, especially, with the rise of high dimensional data. Modern scientific research often involves the simultaneous measurement of a large number of (potentially irrelevant) variables. In this context, it appears of paramount importance to be able to compare and assess the relevance and the performance of these many models, and to identify the most clinically useful ones. Bayesian model uncertainty combined with TB provides a coherent way of answering some of these questions. Here we presented overarching Bayesian model selection framework where we take into account the different benefits/costs associated with correct and incorrect classifications, resulting in selections of variables that are better tailored to the specific clinical application.

4.6 Software

The `R` code used for the TB implementation in this chapter is available as an `R` package, `R2BGLiMS`: https://github.com/pjnewcombe/R2BGLiMS. Section D.2 presents several convergence tests.

Chapter 5

Discussion

5.1 Summary

Throughout this thesis, we have been concerned with approaches to (1) evaluate and (2) build clinically useful risk prediction models for binary outcomes.

In Chapter 2 we focused on model evaluation. We sought to investigate whether ctDNA can be used to predict response to treatment in metastatic breast cancer (mBC). ctDNA has been proposed as a promising approach to assess response to treatment. This is because quantification of ctDNA is less costly, minimally invasive and can be more informative than currently used techniques as it can provide up-to-date information about the genomic composition of the tumour lesions. To assess the usefulness of ctDNA, we proposed a two-stage Bayesian probabilistic model of treatment response. The model allowed us to address the main challenges presented in the data. Namely, the unbalanced study design and the fact that the two outcomes, ctDNA and treatment response are not measured at the same timepoints. We found that ctDNA is useful for predicting response to treatment offering improved clinical utility. In addition, we showed how we can dynamically update individualised predictions, as additional longitudinal measurements become available.

In Chapters 3 and 4 we focused on model building. We sought to incorporate information about the costs of different misclassification errors during model training. Our work was motivated by the main shortcoming of commonly used risk prediction models. That is, models for binary outcomes are often constructed to minimise the expected classification error; that is the proportion of incorrect classifications. This is not desirable in many healthcare applications. To overcome this shortcoming, we proposed a novel Bayesian framework which we called Tailored Bayes (TB) (Chapter 3). We demonstrated TB has favourable properties compared to standard Bayesian (SB) paradigm. In particular, we used toy simulations to showcase scenarios where TB is expected to outperform SB. We then applied TB to three real-world case studies, and showed that incorporating information about misclassification costs into the model leads to better (treatment) decisions. In Chapter 4 we extended TB in high-dimensional settings. We thus proposed a sparse TB model and used extensive simulations and real data to compare TB and SB. We found that TB favours smaller models (with fewer covariates) compared to SB, whilst

performing better or no worse than SB. In addition, TB actually changes the rank order of the covariates compared to SB.

5.2 Conclusions

Evaluating and building clinically useful models is clearly very important, particularly given the widespread adoption of prediction models in healthcare (see Chapter 1). In Chapter 1 we showed that a model with good predictive performance (in terms of traditional performance metrics) is not enough to guarantee clinical usefulness. A well performing model, by traditional metrics, can be clinically useless, and a poorly performing model valuable.

In Chapter 2, we demonstrated the utility of serial ctDNA measurements in predicting response to treatment in mBC. This has important clinical implications for decision-making since clinicians can use the predictions to guide treatment choices. One of the main contributions of this chapter was therefore to provide quantitative estimates of response to treatment. This is the first work to offer individualised predictions of disease progression. Our proposal considers the probability of disease progression for each patient individually; by doing so, it can identify individual patients with extremely high-risk of non-response to treatment. Clinicians can use this information to discontinue treatment if the tumour is not responding. A further attractive feature of our model is uncertainty quantification at the individual level. This allows for more targeted decisions. For instance, if predictive uncertainty is high clinicians can collect more information, e.g., more blood samples, or refer to CT scan for confirmation.

In Chapter 3, we proposed an umbrella framework for Bayesian inference under unbalanced misclassification costs, TB. The framework allows us to take into account misclassification costs whist training the model, thus addressing an inherent limitation of commonly used models. We demonstrated the impact of this approach to the model output (the posterior distribution) and the improvement in predictive performance compared to the SB approach. As far as we know this is the first work to address the issue of incorporating misclassification costs into Bayesian modelling. Hence, one of our contributions is to motivate the approach from a Bayesian perspective. As we discussed in Section 3.5 the tailored posterior (being a proper posterior) integrates the attractive features of Bayesian inference - such as flexible hierarchical modelling, the use of prior information and quantification of uncertainty. Another important contribution was the use of a wide range of simulated scenarios which allowed to us gain some intuition on the underpinnings of the method and insights into when TB can be advantageous compared to the standard Bayesian paradigm. Two such scenarios are the absence of parallelism of the optimal decision boundaries and data contamination. Further, we applied the method to three real-data applications and showed that it results in clinical useful models. An interesting finding from these applications was that under the TB approach the relative importance of the covariates in terms of their contribution to the prediction's changes. We thus hypothesised this could lead us to prioritise different covariates under a variable selection setting.

Hence, in Chapter 4 we extended the TB framework to high-dimensional settings. Based on our comprehensive empirical comparison between TB and SB we concluded that TB favours more parsimonious solutions which may result in lower data collection costs and different covariates being selected (or prioritised) for further downstream analysis. Even though this was an empirical observation based on simulation and real-data analysis we can gain some intuition of this behaviour if we view the TB posterior as combining a standard likelihood function with a data-dependent prior. As we mentioned in Chapter 3 and showed in Appendix C.1 we can interpret the TB prior as a data-dependent regularizer. Thus, compared to SB, TB can provide additional regularisation which is driven by the data. This could explain the results in Chapter 4, i.e., that TB favours smaller models (with fewer covariates) compared to SB. We believe a comprehensive study of the theoretical properties of TB is a promising avenue for future work. In this work, we focused on extensive simulations and real-life scenarios to evaluate the two approaches, which constitutes our main contribution. An attractive feature was the consistency of the conclusions in both the simulated and real datasets. This allowed us to gain better insights on what to expect when applying TB to a real-word setting and how to explain the results.

5.3 Future work

Several possible directions for further work have been suggested throughout this thesis. Here we outline two further directions selected as likely the most promising to prioritise: (1) extension of TB to non-linear models and (2) to multiple health states and more interventions.

In this work, we focused on logistic regression based on linear combinations of the covariates to develop the TB framework. The choice to employ this modeling framework brings many advantages such as ease of implementation, analytical and computational tractability and a smaller number of parameters to estimate. Nevertheless, a potential promising future avenue could be to develop non-linear TB implementations. This is motivated by the popularity and empirically demonstrated good performance of many non-linear models. We believe it would be interesting to study under what circumstances non-linear TB would be expected to outperform non-linear SB and compare their performance in real-world applications.

Throughout this thesis we have been concerned with problems involving two health states and two intervention actions/policies. Many important problems in clinical practice can be represented like this. For example, the breast cancer prognostication case study (Section 3.4.1) involved the health states, dead or alive, and the intervention actions, treat with chemotherapy or not. In screening for disease, a person either has detectable disease or not, and an intervention choice is whether to apply a screening test to that person, perhaps based on an estimate of the probability that the person is diseased. In predicting whether a person will have a myocardial infarction in the next ten years, the person either will or will not be so diagnosed, and an intervention might be whether to recommend taking statins to prevent cardiovascular events. Although such formulations for two health states and two intervention actions

have many applications and have been well studied, there may be multiple health states to consider, and the intervention actions may be more elaborate.

An example of more than two intervention actions is given in Richman and Dowsett (2019). They propose three intervention actions for women with breast cancer upon completion of 5 years of endocrine therapy. After calculating the risk of late recurrence ($\pi(\mathbf{x})$) in women who have not had distant disease recurrence 5 years after diagnosis, they recommend one of three possible actions: stop endocrine therapy, offer genomic test for late recurrence or extend endocrine therapy. For each of these three possible actions, the disease is either present or absent, resulting in 6 possible combinations of intervention actions and disease states. By assigning utilities to each of these 6 conditions we can then define two target thresholds for clinical management. First, we need a target threshold for stopping endocrine therapy versus offering genomic testing. We denote this target threshold by t_{stop}. Second, we need another target threshold for continuing endocrine therapy versus offering genomic testing. We denote this target threshold by t_{cont}. Based on these thresholds one stops therapy if $\pi(\mathbf{x}) < t_{\text{stop}}$. If $\pi(\mathbf{x}) > t_{cont}$ one continues therapy and if $t_{\text{stop}} \leq \pi(\mathbf{x}) \leq t_{\text{cont}}$ one offers further genomic testing. Future work could be directed towards developing evaluation metrics for such examples. For instance, it is unclear if an interpretable metric such as the Net Benefit (see Section 1.5) can be derived for these types of problems. Another direction would be to study how to incorporate these new target thresholds in the TB framework.

Sometimes an intervention affects multiple health states. For example, Hippisley-Cox and Coupland (2010) studied the unintended effects of prescribing statins to reduce the risk of cardiovascular disease (CVD) among high-risk patients. Risk prediction models, such as QRISK2, are used to identify high risk patients most likely to benefit from interventions, including statins (see Section 1.1). However, the use of statins has negative effects such as myopathy, cataract, acute renal failure, oesophageal cancer, and moderate or serious liver dysfunction. A few challenges arise when trying to derive a net benefit function for these types of health states. A first challenge is to assign costs to all the possible outcomes. Not everyone would agree to the costs used, but let's take this step as given. Let the cost c_k to each outcome, $k = 1, 2, \ldots, K$, where $k = 1$ corresponds a CVD event and $K = 7$ to liver dysfunction, then we can determine whether there is a net benefit from prescribing statins as

$$\text{Net Benefit} = \sum_k c_k \pi_{1k}(\mathbf{x}) + \sum_k c_k \pi_{0k}(\mathbf{x}) \tag{5.1}$$

where $\pi_{0k}(\mathbf{x})$ is the probability of the outcome in the absence of statins and $\pi_{1k}(\mathbf{x})$ is the probability in the presence of statins. The net benefit is the expected cost in the absence of statins minus the expected cost in the presence of statins. Relative risks estimates RR_k can be used to compute $\pi_{1k}(\mathbf{x}) = RR_k \pi_{0k}(\mathbf{x})$. The above calculation has a drawback though. It ignores the possibility that a person would develop more than one of these conditions. This assumption means that one does not need to assign costs to multiple simultaneous outcomes. Moreover, one does not need to know the joint distribution of these various outcomes with and without statins; the marginal estimates like $\pi_{0k}(\mathbf{x})$ and $\pi_{1k}(\mathbf{x})$ are sufficient. But in many settings, this assumption is unrealistic. We then need to estimate the joint distributions of

the outcomes. We believe that model evaluation metrics taking into account all these settings are an interesting avenue for future work.

Bibliography

Adalsteinsson, V. A., Ha, G., Freeman, S. S., Choudhury, A. D., Stover, D. G., Parsons, H. A., Gydush, G., Reed, S. C., Rotem, D., Rhoades, J., et al. (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1):1–13.

Ahrens, W. H., Cox, D. J., and Budhwar, G. (1990). Use of the arcsine and square root transformations for subjectively determined percentage data. *Weed Science*, pages 452–458.

Alix-Panabières, C., Schwarzenbach, H., and Pantel, K. (2012). Circulating tumor cells and circulating tumor DNA. *Annual Review of Medicine*, 63:199–215.

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19(4):453–473.

Austin, P. C. and Steyerberg, E. W. (2014). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33(3):517–535.

Austin, P. C. and Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38(21):4051–4065.

Baghfalaki, T., Ganjali, M., and Berridge, D. (2014). Joint modeling of multivariate longitudinal mixed measurements and time to event data using a bayesian approach. *Journal of Applied Statistics*, 41(9):1934–1955.

Baker, S. G. (2009). Putting risk prediction in perspective: relative utility curves. *Journal of the National Cancer Institute*, 101(22):1538–1542.

Baker, S. G., Cook, N. R., Vickers, A., and Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A*, 172(4):729–748.

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212.

Barrett, J. and Su, L. (2017). Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Statistics in medicine*, 36(9):1447–1460.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bassett Jr, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622.

Baumgartner, H., Falk, V., Bax, J. J., De Bonis, M., Hamm, C., Holm, P. J., Iung, B., Lancellotti, P., Lansac, E., Rodriguez Munoz, D., et al. (2017). 2017 ESC/EACTS guidelines for the management of valvular heart disease. *European Heart Journal*, 38(36):2739–2791.

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

Bellou, V., Belbasis, L., Konstantinidis, A. K., Tzoulaki, I., and Evangelou, E. (2019). Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ*, 367.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C. A., Dempster, J., Lyons, N. J., Burns, R., et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian Theory*. John Wiley & Sons.

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., et al. (2014). Detection of circulating tumor DNA in early-and late-stage human malignancies. *Science Translational Medicine*, 6(224):224ra24–224ra24.

Bissiri, P. G., Holmes, C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130.

Boone, D., Mallett, S., Zhu, S., Yao, G. L., Bell, N., Ghanouni, A., von Wagner, C., Taylor, S. A., Altman, D. G., Lilford, R., et al. (2013). Patients' healthcare professionals' values regarding true-& false-positive diagnosis when colorectal cancer screening by CT colonography: discrete choice experiment. *PLoS ONE*, 8(12).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., et al. (2016). 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine*, 375(8):717–729.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Carrick, R. T., Park, J. G., McGinnes, H. L., Lundquist, C., Brown, K. D., Janes, W. A., Wessler, B. S., and Kent, D. M. (2020). Clinical Predictive Models of Sudden Cardiac Arrest: A Survey of the Current Science and Analysis of Model Performances. *Journal of the American Heart Association*, 9(16):e017625.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Cayrefourcq, L. and Alix-Panabières, C. (2019). CTCs as liquid biopsy: Where are we now? *Molecular Medicine*.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392.

Chen, G., McQuade, J. L., Panka, D. J., Hudgens, C. W., Amin-Mansour, A., Mu, X. J., Bahl, S., Jané-Valbuena, J., Wani, K. M., Reuben, A., et al. (2016). Clinical, molecular, and immune analysis of dabrafenib-trametinib combination treatment for BRAF inhibitor–refractory metastatic melanoma: a phase 2 clinical trial. *JAMA Oncology*, 2(8):1056–1064.

Chen, M. and Zhao, H. (2019). Next-generation sequencing in liquid biopsy: cancer screening and early detection. *Human Genomics*, 13(1):1–10.

Cheng, W., Taylor, J. M., Gu, T., Tomlins, S. A., and Mukherjee, B. (2019). Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C*, 68(1):121–139.

Childress, J. F. and Beauchamp, T. L. (2001). *Principles of Biomedical Ethics*. Oxford University Press.

Choi, J., Anderson, S. J., Richards, T. J., and Thompson, W. K. (2014). Prediction of transplant-free survival in idiopathic pulmonary fibrosis patients using joint models for event times and mixed multivariate longitudinal data. *Journal of applied statistics*, 41(10):2192–2205.

Chung, G. G., Zerkowski, M. P., Ghosh, S., Camp, R. L., and Rimm, D. L. (2007). Quantitative analysis of estrogen receptor heterogeneity in breast cancer. *Laboratory Investigation*, 87(7):662–669.

Cleveland, W. S. (1991). Local regression models. *Statistical models in S*.

Collins, G. S. and Altman, D. G. (2010). An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*, 340.

Collins, G. S. and Altman, D. G. (2012). Predicting the 10 year risk of cardiovascular disease in the united kingdom: independent and external validation of an updated version of QRISK2. *BMJ*, 344.

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*, 131(2):211–219.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Dandis, R., Teerenstra, S., Massuger, L., Sweep, F., Eysbouts, Y., and IntHout, J. (2020). A tutorial on dynamic risk prediction of a binary outcome based on a longitudinal biomarker. *Biometrical Journal*, 62(2):398–413.

Dawson, S.-J., Tsui, D. W., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S.-F., Dunning, M. J., Gale, D., Forshew, T., Mahler-Araujo, B., et al. (2013). Analysis of circulating tumor DNA to monitor metastatic breast cancer. *New England Journal of Medicine*, 368(13):1199–1209.

De Glas, N., Bastiaannet, E., Engels, C., De Craen, A., Putter, H., Van De Velde, C., Hurria, A., Liefers, G., and Portielje, J. (2016). Validity of the online PREDICT tool in older patients with breast cancer: a population-based study. *British Journal of Cancer*, 114(4):395–400.

De Kat, A. C., Hirst, J., Woodward, M., Kennedy, S., and Peters, S. A. (2019). Prediction models for preeclampsia: a systematic review. *Pregnancy Hypertension*, 16:48–66.

De la Cruz, R., Marshall, G., and Quintana, F. A. (2011). Logistic regression when covariates are random effects from a non-linear mixed model. *Biometrical journal*, 53(5):735–749.

de Munter, L., Polinder, S., Lansink, K. W., Cnossen, M. C., Steyerberg, E. W., and de Jongh, M. A. (2017). Mortality prediction models in the general trauma population: A systematic review. *Injury*, 48(2):221–229.

De Rubis, G., Krishnan, S. R., and Bebawy, M. (2019). Liquid biopsies in cancer diagnosis, monitoring, and prognosis. *Trends in Pharmacological Sciences*, 40(3):172–186.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. (2018). Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*.

Dijkland, S. A., Foks, K. A., Polinder, S., Dippel, D. W., Maas, A. I., Lingsma, H. F., and Steyerberg, E. W. (2020). Prognosis in moderate and severe traumatic brain injury: a systematic review of contemporary models and validation studies. *Journal of Neurotrauma*, 37(1):1–13.

Dos Reis, F. J. C., Wishart, G. C., Dicks, E. M., Greenberg, D., Rashbass, J., Schmidt, M. K., van den Broek, A. J., Ellis, I. O., Green, A., Rakha, E., et al. (2017). An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Research*, 19(1):1–13.

Down, S. K., Lucas, O., Benson, J. R., and Wishart, G. C. (2014). Effect of PREDICT on chemotherapy/trastuzumab recommendations in HER2-positive patients with early-stage breast cancer. *Oncology Letters*, 8(6):2757–2761.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository.

Duan, W., Zhang, R., Zhao, Y., Shen, S., Wei, Y., Chen, F., and Christiani, D. C. (2018). Bayesian variable selection for parametric survival model with applications to cancer omics data. *Human genomics*, 12(1):1–15.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247.

Elazezy, M. and Joosse, S. A. (2018). Techniques of using circulating tumor DNA as a liquid biopsy component in cancer management. *Computational and Structural Biotechnology Journal*, 16:370–378.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, page 973–978.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fong, E. and Holmes, C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.

Freedman, B. (1987). Equipoise and the Ethics of Clinical Research. *New England Journal of Medicine*, 317(3):141–145.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1–10.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383.

Gendoo, D. M., Ratanasirigulchai, N., Schroeder, M. S., Pare, L., Parker, J. S., Prat, A., and Haibe-Kains, B. (2020). *genefu: Computation of Gene Expression-Based Signatures in Breast Cancer*. R package version 2.22.0.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373.

Gilks, W. R., Best, N. G., and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society: Series C*, 44(4):455–472.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Gu, X., Tadesse, M. G., Foulkes, A. S., Ma, Y., and Balasubramanian, R. (2020). Bayesian variable selection for high dimensional predictors and self-reported outcomes. *BMC Medical Informatics and Decision making*, 20(1):1–11.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815.

Güler, E. N. (2017). Gene expression profiling in breast cancer and its effect on therapy selection in early-stage breast cancer. *European Journal of Breast Health*, 13(4):168.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123.

Hand, D. J. and Vinciotti, V. (2003). Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2):124–131.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., Hammond, E. H., Kuderer, N. M., Liu, M. C., Mennel, R. G., et al. (2016). Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. *Journal of Clinical Oncology*, 34(10):1134.

Hastie, D. I. and Green, P. J. (2012). Model choice using reversible jump markov chain monte carlo. *Statistica Neerlandica*, 66(3):309–338.

Hayashi, T. and Iwata, H. (2010). EM algorithm for bayesian estimation of genomic breeding values. *BMC Genetics*, 11(1):1–9.

He, B. and Luo, S. (2016). Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. *Statistical methods in medical research*, 25(4):1346–1358.

Hill, S. M., Neve, R. M., Bayani, N., Kuo, W.-L., Ziyad, S., Spellman, P. T., Gray, J. W., and Mukherjee, S. (2012). Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. *BMC Bioinformatics*, 13(1):1–15.

Hippisley-Cox, J. and Coupland, C. (2010). Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. *BMJ*, 340:c2197.

Hippisley-Cox, J. and Coupland, C. (2017). Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ*, 359:j5019.

Hippisley-Cox, J., Coupland, C., and Brindle, P. (2014). The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open*, 4(8).

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., and Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336(7659):1475–1482.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Holmberg, L. and Vickers, A. (2013). Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS Med*, 10(7):e1001491.

Hong, T. S., Tomé, W. A., and Harari, P. M. (2012). Heterogeneity in head and neck IMRT target design and clinical practice. *Radiotherapy and Oncology*, 103(1):92–98.

Hosmer, D. W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069.

Hou, X.-H., Feng, L., Zhang, C., Cao, X.-P., Tan, L., and Yu, J.-T. (2019). Models for predicting risk of dementia: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(4):373–379.

Hrebien, S., Citi, V., Garcia-Murillas, I., Cutts, R., Fenwick, K., Kozarewa, I., McEwen, R., Ratnayake, J., Maudsley, R., Carr, T., et al. (2019). Early ctDNA dynamics as a surrogate for progression-free survival in advanced breast cancer in the BEECH trial. *Annals of Oncology*, 30(6):945–952.

Huang, H., Fang, M., Jostins, L., Mirkov, M. U., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662):173–178.

Huang, S., Murphy, L., and Xu, W. (2018). Genes and functions from breast cancer signatures. *BMC Cancer*, 18(1):473.

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., and Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1965). A Robust Version of the Probability Ratio Test. *The Annals of Mathematical Statistics*, 36(6):1753–1758.

Hunink, M. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., and Glasziou, P. P. (2014). *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press.

Janssen, K., Moons, K., Kalkman, C., Grobbee, D., and Vergouwe, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, 61(1):76–86.

Jung, K., Kashyap, S., Avati, A., Harman, S., Shaw, H., Li, R., Smith, M. A., Shum, K. F. K., Javitz, J., Vetteth, Y., et al. (2020). A framework for making predictive models useful in practice. *medRxiv*.

Karapanagiotis, S., Benedetto, U., Mukherjee, S., Kirk, P. D. W., and Newcombe, P. J. (2021). Tailored Bayes: a risk modelling framework under unequal misclassification costs. *arXiv preprint arXiv:2104.01822*.

Karapanagiotis, S., Pharoah, P. D., Jackson, C. H., and Newcombe, P. J. (2018). Development and external validation of prediction models for 10-year survival of invasive breast cancer. Comparison with PREDICT and CancerMath. *Clinical Cancer Research*, 24(9):2110–2115.

Kattan, M. W., Hess, K. R., Amin, M. B., Lu, Y., Moons, K. G., Gershenwald, J. E., Gimotty, P. A., Guinney, J. H., Halabi, S., Lazar, A. J., et al. (2016). American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA: A Cancer Journal for Clinicians*, 66(5):370–374.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., et al. (1995). The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203.

Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):1–6.

Kukar, M., Kononenko, I., et al. (1998). Cost-sensitive learning with neural networks. In *ECAI*, volume 98, pages 445–449.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Laplace, P. (1814). Essai Philosophique sur les Probabilities, Courcier Imprimeur, Paris; reprints of this work and of Laplace's much larger Theorie Analytic des Probabilities are available from Editions Culture et Civilisation, 115 Ave. *Cabriel Lebron*, 1160.

Le, P., Martinez, K., Pappas, M., and Rothberg, M. (2017). A decision model to estimate a risk threshold for venous thromboembolism prophylaxis in hospitalized medical patients. *Journal of Thrombosis and Haemostasis*, 15(6):1132–1141.

Lee, S.-Y. and Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4):653–686.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Li, X. A., Tai, A., Arthur, D. W., Buchholz, T. A., Macdonald, S., Marks, L. B., Moran, J. M., Pierce, L. J., Rabinovitch, R., Taghian, A., et al. (2009). Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG Multi-Institutional and Multiobserver Study. *International Journal of Radiation Oncology*Biology* Physics*, 73(3):944–951.

Li, Y., Sperrin, M., Ashcroft, D. M., and van Staa, T. P. (2020). Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ*, 371.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B*, 80(5):1087–1110.

Ling, C. X., Yang, Q., Wang, J., and Zhang, S. (2004). Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, page 69.

Localio, A. R. and Stack, C. B. (2015). TRIPOD: a new reporting baseline for developing and interpreting prediction models. *Annals of Internal Medicine*, 162(1):73–74.

Ma, F., Guan, Y., Yi, Z., Chang, L., Li, Q., Chen, S., Zhu, W., Guan, X., Li, C., Qian, H., et al. (2020). Assessing tumor heterogeneity using ctDNA to predict and monitor therapeutic response in metastatic breast cancer. *International Journal of Cancer*, 146(5):1359–1368.

Manchanda, R., Legood, R., Antoniou, A. C., Gordeev, V. S., and Menon, U. (2016). Specifying the ovarian cancer risk threshold of 'premenopausal risk-reducing salpingo-oophorectomy' for ovarian cancer prevention: a cost-effectiveness analysis. *Journal of Medical Genetics*, 53(9):591–599.

Manier, S., Park, J., Capelletti, M., Bustoros, M., Freeman, S., Ha, G., Rhoades, J., Liu, C., Huynh, D., Reed, S., et al. (2018). Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nature Communications*, 9(1):1–11.

Margineantu, D. and Dietterich, T. (2003). A wrapper method for cost-sensitive learning via stratification. [Online; cited Decemeber 2019] Available at: :http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.1102.

Mars, N., Koskela, J. T., Ripatti, P., Kiiskinen, T. T., Havulinna, A. S., Lindbohm, J. V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(4):549–557.

Masnadi-Shirazi, H. and Vasconcelos, N. (2010). Risk minimization, probability elicitation, and cost-sensitive svms. In *ICML*, pages 759–766.

Mattei, P.-A. (2019). A Parsimonious Tour of Bayesian Model Uncertainty. *arXiv preprint arXiv:1902.05539*.

Merker, J. D., Oxnard, G. R., Compton, C., Diehn, M., Hurley, P., Lazar, A. J., Lindeman, N., Lockwood, C. M., Rai, A. J., Schilsky, R. L., Tsimberidou, A. M., Vasalos, P., Billman, B. L., Oliver, T. K., Bruinooge, S. S., Hayes, D. F., and Turner, N. C. (2018). Circulating Tumor DNA Analysis in Patients With Cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. *Journal of Clinical Oncology*, 142(10):1242–1253.

Meschia, J. F., Bushnell, C., Boden-Albala, B., Braun, L. T., Bravata, D. M., Chaturvedi, S., Creager, M. A., Eckel, R. H., Elkind, M. S., Fornage, M., et al. (2014). Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, 45(12):3754–3832.

Miller, M. E., Hui, S. L., and Tierney, W. M. (1991). Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–1226.

Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., and McDonald, C. J. (1993). Validation of probabilistic predictions. *Medical Decision Making*, 13(1):49–57.

Mok, T., Wu, Y.-L., Lee, J. S., Yu, C.-J., Sriuranpong, V., Sandoval-Tan, J., Ladrera, G., Thongprasert, S., Srimuninnimit, V., Liao, M., et al. (2015). Detection and dynamic changes of EGFR mutations from circulating tumor DNA as a predictor of survival outcomes in NSCLC patients treated with first-line intercalated erlotinib and chemotherapy. *Clinical Cancer Research*, 21(14):3196–3203.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., and Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1–W73.

Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., and Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, 98(9):691–698.

Murtaza, G., Shuib, L., Wahab, A. W. A., Mujtaba, G., Nweke, H. F., Al-garadi, M. A., Zulfiqar, F., Raza, G., and Azmi, N. A. (2019). Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, pages 1–66.

Murtaza, M., Dawson, S.-J., Tsui, D. W., Gale, D., Forshew, T., Piskorz, A. M., Parkinson, C., Chin, S.-F., Kingsbury, Z., Wong, A. S., et al. (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108–112.

Nashef, S. A., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., Salamon, R., and Group, E. S. (1999). European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, 16(1):9–13.

Nashef, S. A., Roques, F., Sharples, L. D., Nilsson, J., Smith, C., Goldstone, A. R., and Lockowandt, U. (2012). EuroSCORE II. *European Journal of Cardio-Thoracic Surgery*, 41(4):734–745.

Newcombe, P. J., Nelson, C. P., Samani, N. J., and Dudbridge, F. (2019). A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genetic Epidemiology*, 43(7):730–741.

Newcombe, P. J., Raza Ali, H., Blows, F. M., Provenzano, E., Pharoah, P. D., Caldas, C., and Richardson, S. (2017). Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical Methods in Medical Research*, 26(1):414–436.

NICE (2016). Cardiovascular disease: risk assessment and reduction, including lipid modification. [Online; cited December 2019] Available from: https://www.nice.org.uk/guidance/cg181/chapter/1-recommendations.

NICE (2018). Early and locally advanced breast cancer: diagnosis and management. [Online; cited November 2020] Available from: https://www.ncbi.nlm.nih.gov/books/NBK541671/.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, 32(9):1338–1345.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117.

Osama, M., Zachariah, D., and Stoica, P. (2019). Robust Risk Minimization for Statistical Learning. *arXiv preprint arXiv:1910.01544*.

Ott, J. (1999). *Analysis of Human Genetic Linkage*. Johns Hopkins University Press.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160.

Paschali, M., Conjeti, S., Navarro, F., and Navab, N. (2018). Generalizability vs. robustness: adversarial examples for medical imaging. *arXiv preprint arXiv:1804.00504*.

Pastore, M. and Calcagnì, A. (2019). Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. *Frontiers in Psychology*, 10:1089.

Pate, A., Emsley, R., Ashcroft, D. M., Brown, B., and van Staa, T. (2019). The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Medicine*, 17(1):1–16.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine*, 293(5):229–234.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

Pencina, M. J., D'Agostino Sr, R. B., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, 30(1):11–21.

Peto, R., Davies, C., Godwin, J., Gray, R., Pan, H., Clarke, M., Cutter, D., Darby, S., McGale, P., Taylor, C., et al. (2012). Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*, 379(9814):432–444.

Piironen, J. and Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.

Pike, M. M., Decker, P. A., Larson, N. B., Sauver, J. L. S., Takahashi, P. Y., Roger, V. L., Rocca, W. A., Miller, V. M., Olson, J. E., Pathak, J., et al. (2016). Improvement in cardiovascular risk prediction with electronic health records. *Journal of Cardiovascular Translational Research*, 9(3):214–222.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9(501-538):105.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th international conference on machine learning ICML-98 Morgan Kaufmann. San Mateo, CA*.

Rabar, S., Lau, R., O'Flynn, N., Li, L., and Barry, P. (2012). Risk assessment of fragility fractures: summary of NICE guidance. *BMJ*, 345.

Rapisuwon, S., Vietsch, E. E., and Wellstein, A. (2016). Circulating biomarkers to monitor cancer progression and treatment. *Computational and Structural Biotechnology Journal*, 14:211–222.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data Analysis Methods*, volume 1. Sage Publications.

Richman, J. and Dowsett, M. (2019). Beyond 5 years: enduring risk of recurrence in oestrogen receptor-positive breast cancer. *Nature Reviews Clinical Oncology*, 16(5):296–311.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-event Data: With Applications in R*. Chapman and Hall/CRC.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):1–8.

Roques, F., Michel, P., Goldstone, A., and Nashef, S. (2003). The logistic EuroSCORE. *European Heart Journal*, 24(9):882–883.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins.

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., Batra, R., Pereira, B., Bruna, A., Ali, H. R., et al. (2019). Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature*, 567(7748):399–404.

Sayers, A., Heron, J., Smith, A. D., Macdonald-Wallis, C., Gilthorpe, M., Steele, F., and Tilling, K. (2017). Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: a simulation study of childhood growth and BP. *Statistical Methods in Medical Research*, 26(1):437–452.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

Schiavon, G., Hrebien, S., Garcia-Murillas, I., Cutts, R. J., Pearson, A., Tarazona, N., Fenwick, K., Kozarewa, I., Lopez-Knowles, E., Ribas, R., et al. (2015). Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Science Translational Medicine*, 7(313):313ra182–313ra182.

Schorling, J. B., Roach, J., Siegel, M., Baturka, N., Hunt, D. E., Guterbock, T. M., and Stewart, H. L. (1997). A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine*, 26(1):92–101.

Schwartz, L. M., Woloshin, S., Sox, H. C., Fischhoff, B., and Welch, H. G. (2000). US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *BMJ*, 320(7250):1635–1640.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS epidemiology*, pages 231–255. Springer.

Shah, N. H., Milstein, A., and Bagley, S. C. (2019). Making machine learning models clinically useful. *JAMA*, 322(14):1351–1352.

Siravegna, G., Marsoni, S., Siena, S., and Bardelli, A. (2017). Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical Oncology*, 14(9):531–548.

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., Ponzetti, A., Cremolini, C., Amatu, A., Lauricella, C., et al. (2015). Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nature Medicine*, 21(7):795–801.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software*, 97(1):1–66.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142.

Steyerberg, E. W., Borsboom, G. J., van Houwelingen, H. C., Eijkemans, M. J., and Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586.

Steyerberg, E. W. et al. (2019). *Clinical Prediction Models*. Springer.

Steyerberg, E. W., Uno, H., Ioannidis, J. P., Van Calster, B., Ukaegbu, C., Dhingra, T., Syngal, S., and Kastrinos, F. (2018). Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*, 98:133–143.

Steyerberg, E. W., Van Calster, B., and Pencina, M. J. (2011). Performance measures for prediction models and markers: evaluation of predictions and classifications. *Revista Española de Cardiología (English Edition)*, 64(9):788–794.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, 35(29):1925–1931.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5(3).

Stover, D. G., Parsons, H. A., Ha, G., Freeman, S. S., Barry, W. T., Guo, H., Choudhury, A. D., Gydush, G., Reed, S. C., Rhoades, J., et al. (2018). Association of cell-free DNA tumor fraction and somatic copy number alterations with survival in metastatic triple-negative breast cancer. *Journal of Clinical Oncology*, 36(6):543.

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., Schumacher, S. E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*, 33(4):676–689.

Theorell, A. and Nöh, K. (2020). Reversible jump MCMC for multi-model inference in Metabolic Flux Analysis. *Bioinformatics*, 36(1):232–240.

Ting, K. M. (1998). Inducing cost-sensitive trees via instance weighting. In *Principles of Data Mining and Knowledge Discovery*, pages 139–147. Springer Berlin Heidelberg.

Tsalatsanis, A., Hozo, I., Vickers, A., and Djulbegovic, B. (2010). A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making*, 10(1):51.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Turner, J. R. (2013). *Encyclopedia of Behavioral Medicine*, chapter Principle of Equipoise, pages 1537–1538. Springer New York.

Usher-Smith, J. A., Walter, F. M., Emery, J. D., Win, A. K., and Griffin, S. J. (2016). Risk prediction models for colorectal cancer: a systematic review. *Cancer Prevention Research*, 9(1):13–26.

Vahanian, A., Alfieri, O. R., Al-Attar, N., Antunes, M. J., Bax, J., Cormier, B., Cribier, A., De Jaegere, P., Fournial, G., Kappetein, A. P., et al. (2008). Transcatheter valve implantation for patients with aortic stenosis: a position statement from the European Association of Cardio-Thoracic Surgery (EACTS) and the European Society of Cardiology (ESC), in collaboration with the European Association of Percutaneous Cardiovascular Interventions (EAPCI). *European Journal of Cardio-Thoracic Surgery*, 34(1):1–8.

Van Calster, B., Vickers, A. J., Pencina, M. J., Baker, S. G., Timmerman, D., and Steyerberg, E. W. (2013). Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Medical Decision Making*, 33(4):490–501.

Van Calster, B., Wynants, L., Collins, G. S., Verbakel, J. Y., and Steyerberg, E. W. (2020). ROC curves for clinical prediction models part 3. The ROC plot: a picture that needs a 1000 words. *Journal of Clinical Epidemiology*, 126:220–223.

van de Schoot, R., Depaoli, S., King, R., et al. (2021). Bayesian statistics and modelling. *Nat Rev Methods Primers*, 1(1).

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for bayesian models. *R package version*, 2(4):1003.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., and Van Calster, B. (2020). ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, 126:207–216.

Viallefont, V., Raftery, A. E., and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, 20(21):3215–3230.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Vickers, A. J., Van Calster, B., and Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352:i6.

Volkers, E. J., Algra, A., Kappelle, L. J., Jansen, O., Howard, G., Hendrikse, J., Halliday, A., Gregson, J., Fraedrich, G., Eckstein, H.-H., et al. (2018). Prediction Models for Clinical Outcome After a Carotid Revascularization Procedure: An External Validation Study. *Stroke*, 49(8):1880–1885.

Vyshemirsky, V. and Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839.

Walker, S. and Hjort, N. L. (2001). On Bayesian Consistency. *Journal of the Royal Statistical Society: Series B*, 63(4):811–821.

Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D., García, A. R., Ferreira, R. C., Guo, H., Walker, N. M., et al. (2015). Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS Genet*, 11(6):e1005272.

Wang, C., Wang, N., and Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56(2):487–495.

Wang, J., Luo, S., and Li, L. (2017). Dynamic prediction for multiple repeated measures and event time data: an application to Parkinson's disease. *The Annals of Applied Statistics*, 11(3):1787.

Warfield, S. K., Zou, K. H., and Wells, W. M. (2008). Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1874):2361–2375.

Watson, V., McCartan, N., Krucien, N., Abu, V., Ikenwilo, D., Emberton, M., and Ahmed, H. U. (2020). Evaluating the Trade-offs Men with Localised Prostate Cancer Make Between the Risks and Benefits of Treatments: The COMPARE Study. *The Journal of Urology*, pages 10–1097.

Willems, J. P., Saunders, J. T., Hunt, D. E., and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: a community-based study. *Southern Medical Journal*, 90(8):814–820.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., and Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *The Annals of Applied Statistics*, 4(3):1342.

Wishart, G., Bajdik, C., Azzato, E., Dicks, E., Greenberg, D., Rashbass, J., Caldas, C., and Pharoah, P. (2011). A population-based validation of the prognostic model PREDICT for early breast cancer. *European Journal of Surgical Oncology*, 37(5):411–417.

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., and Pharoah, P. D. (2010). PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12(1):1–10.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., Van Calster, B., et al. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(1):192.

Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H., Gonzalez, S., Martincorena, I., Alexandrov, L. B., Van Loo, P., Haugland, H. K., Lilleng, P. K., et al. (2017). Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell*, 32(2):169–184.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pages 435–442. IEEE.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85.

Zhao, Y., Zhu, H., Lu, Z., Knickmeyer, R. C., and Zou, F. (2019). Structured genome-wide association studies with Bayesian hierarchical variable selection. *Genetics*, 212(2):397–415.

Appendix A

Appendix to Chapter 1

A.1 Details to Toy example 1

Here we give details for the simulation presented in Section 1.3.1. Let $y \in \{0,1\}$ be the binary outcome (residual disease: {no, yes}) and x the covariates. Both models include the same continuous covariate, x_1, simulated from

$$\begin{aligned} x_1|y=1 &\sim \mathcal{N}(0.5,1) \\ x_1|y=0 &\sim \mathcal{N}(0,1) \end{aligned} \tag{A.1}$$

The binary covariate, x_A, in Model A is simulated as

$$\begin{aligned} x_A|y=1 &\sim Bernoulli(sens_A) \\ x_A|y=0 &\sim Bernoulli(1-spec_A) \end{aligned} \tag{A.2}$$

where $sens_A = 0.88$ and $spec_A = 0.49$ are the sensitivity and specificity of x_A, respectively. Similarly, the binary covariate, x_B, in Model B is simulated as

$$\begin{aligned} x_B|y=1 &\sim Bernoulli(sens_B) \\ x_B|y=0 &\sim Bernoulli(1-spec_B) \end{aligned} \tag{A.3}$$

where $sens_B = 0.52$ and $spec_B = 0.93$ are the sensitivity and specificity of x_B, respectively. Finally, the outcome is simulated as

$$y \sim Bernoulli(prev) \tag{A.4}$$

where $prev = 0.55$ is the prevalence of y.

A.2 Details to Toy example 2

The simulation in Section 1.3.2 follows the same structure as above. That is, the continuous covariate, x_1, is simulated from

$$\begin{aligned} x_1|y=1 &\sim \mathcal{N}(-1,1) \\ x_1|y=0 &\sim \mathcal{N}(1,1) \end{aligned} \tag{A.5}$$

The binary covariate, x_A, in Models A and C is simulated as

$$\begin{aligned} x_A|y=1 &\sim Bernoulli(sens_A) \\ x_A|y=0 &\sim Bernoulli(1-spec_A) \end{aligned} \tag{A.6}$$

where $sens_A = 0.9$ and $spec_A = 0.2$ are the sensitivity and specificity of x_A, respectively. The binary covariate, x_B, in Models B and C is simulated as

$$\begin{aligned} x_B|y=1 &\sim Bernoulli(sens_B) \\ x_B|y=0 &\sim Bernoulli(1-spec_B) \end{aligned} \tag{A.7}$$

where $sens_B = 0.2$ and $spec_B = 0.6$ are the sensitivity and specificity of x_B, respectively. Finally, the outcome is simulated as

$$y \sim Bernoulli(prev) \tag{A.8}$$

where $prev = 0.01$ is the prevalence of y.

Appendix B

Appendix to Chapter 2

To compare between including or not a random slope parameter in the 2nd stage model we used the expected log pointwise predictive density to assess model performance (Vehtari et al., 2017). More specifically, we used the Bayesian leave-one-out cross-validation estimate of the expected log pointwise predictive density which is a sum of n individual pointwise log predictive densities. It is defined as,

$$\widehat{\text{elpd}}_{loo} = \sum_{i=1}^{n} \log \hat{p}(y_i|y_{-i}),$$

where

$$\hat{p}(y_i|y_{-i}) = \frac{1}{M}\sum_{m=1}^{M} p(y_i|\theta^{(m)})$$

is the leave-one-out predictive density given the data without the i^{th} data point, θ refers to all the model parameters, and $\theta^{(m)}$, $m = 1,\ldots,M$ refers to draws from the posterior, with $p(y_i|\theta)$ being the likelihood. $\widehat{\text{elpd}}_{loo}$ was estimated as described in Vehtari et al. (2017) and Vehtari et al. (2015) and implemented using the `loo` package (Vehtari et al., 2020) in `R`. To compare between the two models, including (Model 1) and excluding (Model 2) the random slope, we calculate the difference (and standard error) in $\widehat{\text{elpd}}_{loo}$, $\Delta\widehat{\text{elpd}}_{loo}$. Comparing the models reveals an estimated $\Delta\widehat{\text{elpd}}_{loo}$ of -0.7 (with a standard error of 2.0) in favour of Model 1. But, given the large standard error compared to $\Delta\widehat{\text{elpd}}_{loo}$ we opted to proceed with the simpler (with fewer parameters) Model 2 at the 2nd stage.

Parameter estimates based on Model A are presented as follows:

1. Table B.1 presents the parameters of the 1st stage model, corresponding to the $\boldsymbol{\Theta}_1$ parameter vector.
2. Table B.2 presents the parameters of the 2nd stage model, corresponding to the $\boldsymbol{\Theta}_2$ parameter vector.

Table B.1 Parameter estimates 1st stage model. ρ denotes the correlation between σ_{11} and σ_{22}.

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	0.22	0.10	0.03	0.41
time	0.02	0.01	-0.00	0.04
ER status (pos)	-0.05	0.04	-0.13	0.03
Her2 status (pos)	-0.02	0.02	-0.07	0.03
Treatment 1	-0.05	0.09	-0.22	0.12
Treatment 2	0.00	0.09	-0.18	0.19
Treatment 3	-0.06	0.09	-0.24	0.11
Treatment 4	0.01	0.12	-0.23	0.24
Treatment 5	-0.01	0.09	-0.18	0.17
Treatment 6	0.00	0.10	-0.19	0.20
Treatment 7	0.01	0.09	-0.17	0.20
Treatment 8	-0.13	0.12	-0.37	0.11
Treatment 9	0.07	0.10	-0.13	0.26
Treatment 10	-0.10	0.13	-0.35	0.15
Treatment 11	-0.04	0.09	-0.21	0.14
Treatment 12	0.14	0.18	-0.21	0.49
Treatment 13	0.11	0.09	-0.06	0.29
Treatment 14	-0.01	0.09	-0.20	0.17
Treatment 15	-0.03	0.09	-0.21	0.15
Treatment 16	0.01	0.09	-0.17	0.20
Treatment 17	0.02	0.09	-0.16	0.21
Treatment 18	-0.08	0.11	-0.29	0.14
Treatment 19	0.07	0.10	-0.11	0.26
Treatment duration	0.00	0.01	-0.01	0.01
σ_ε	0.15	0.00	0.14	0.16
σ_{11}	0.09	0.01	0.07	0.12
σ_{22}	0.05	0.01	0.03	0.07
ρ	0.40	0.32	-0.23	0.96

Table B.2 Parameter estimates 2nd stage model. γ_1 and γ_2 are two elements of the $\boldsymbol{\gamma}$ vector.

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	-3.85	1.70	-7.56	-0.88
time	0.30	0.10	0.12	0.50
ER status (pos)	-0.83	0.56	-1.96	0.24
Her2 status (pos)	-0.29	0.31	-0.90	0.30
Treatment 1	2.79	1.61	0.03	6.37
Treatment 2	2.37	1.71	-0.66	6.10
Treatment 3	1.21	1.66	-1.70	4.84
Treatment 4	2.54	2.23	-1.84	7.00
Treatment 5	2.95	1.64	0.10	6.59
Treatment 6	2.54	1.75	-0.61	6.32
Treatment 7	3.19	1.65	0.36	6.83
Treatment 8	2.50	2.23	-1.89	6.93
Treatment 9	3.19	1.67	0.28	6.86
Treatment 10	-74.68	59.72	-217.33	0.83
Treatment 11	2.76	1.62	-0.01	6.37
Treatment 12	-76.16	60.06	-220.33	0.38
Treatment 13	4.14	1.63	1.33	7.75
Treatment 14	3.95	1.66	1.11	7.61
Treatment 15	2.70	1.60	-0.06	6.26
Treatment 16	3.24	1.67	0.36	6.91
Treatment 17	2.72	1.67	-0.19	6.41
Treatment 18	3.31	1.89	-0.12	7.34
Treatment 19	4.29	1.71	1.27	8.00
γ_1	7.03	3.76	-0.01	14.81
γ_2	4.42	7.21	-9.86	18.59
σ_u	0.90	0.23	0.47	1.37

Appendix C

Appendix to Chapter 3

C.1 Interpretation of the TB prior (and posterior)

Here we show the prior in TB can be interpreted as a regularizer on a per-datapoint influence/importance. First, we slightly modify notation and consider data $D_i = (X_i, Y_i)$ as a copy of a random variable $D = (X, Y) \in \mathbb{R}^d \times \{0, 1\}$. Then, let $\mathscr{L}(D_i|\boldsymbol{\beta})$ be the standard likelihood contribution of datapoint i ($i = 1, \ldots, n$). The TB posterior, up to a normalising constant, is

$$p(\boldsymbol{\beta}|D) \propto \prod_{i=1}^{n} \mathscr{L}(D_i|\boldsymbol{\beta})^{w_i} p(\boldsymbol{\beta}). \tag{C.1}$$

Following Walker and Hjort (2001) we can view (C.1) as combining the original likelihood function with a *data-dependent prior* that is divided by a portion of the likelihood. To see this, we first define the data-dependent prior as

$$\frac{p(\boldsymbol{\beta})}{\prod_{i=1}^{n} \mathscr{L}(D_i|\boldsymbol{\beta})^{1-w_i}}$$

which corresponds to

$$p(\boldsymbol{\beta}|D) \propto \prod_{i=1}^{n} \mathscr{L}(D_i|\boldsymbol{\beta}) \frac{p(\boldsymbol{\beta})}{\prod_{i=1}^{n} \mathscr{L}(D_i|\boldsymbol{\beta})^{1-w_i}} = \prod_{i=1}^{n} \mathscr{L}(D_i|\boldsymbol{\beta})^{w_i} p(\boldsymbol{\beta}),$$

which is seen to coincide with (C.1). This data-dependent downweighting of the prior reduces the weights of those parameter values that "track the data too closely" (Linero and Yang, 2018).

C.2 Model inference and prediction

To sample from the TB posterior we use Markov Chain Monte Carlo (MCMC) (see Section C.4 for details on the computational scheme). We obtain S posterior samples $\{\boldsymbol{\beta}^s\}_{s=1}^S$, where $\boldsymbol{\beta}^s = (\beta_1^s, \ldots, \beta_{d+1}^s)$. We

use the posterior samples to approximate the predictive density for test data $\mathbf{x}_*$

$$\begin{aligned} p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, D) &= \int p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid D) d\boldsymbol{\beta} \\ &\approx \frac{1}{S} \sum_{s=1}^{S} p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, \boldsymbol{\beta}^s). \end{aligned} \tag{C.2}$$

To calculate point predictions we summarise (C.2) by using the posterior predictive mean,

$$\hat{\pi}(\mathbf{x}_*) = \int \pi(\mathbf{x}_*) p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, D) d\pi(\mathbf{x}_*), \tag{C.3}$$

which is used as a plug-in into the NB function (equation (1.5)). We also use Bayesian inference for the estimation of $\pi_u(\mathbf{x})$ so $\pi(\mathbf{x}_*)$ can be conceptually replaced by $\pi_u(\mathbf{x})$ in equations (C.2) and (C.3) with the caveat that they are estimated in different subsets of the data (see Section 3.2.4 for the data splitting strategy we are implementing).

C.3 Cross-validation to choose λ

We use stratified K-fold cross-validation (CV) to choose λ in equation (3.2). The stratification ensures the prevalence of the outcome is the same in each fold. In K-fold CV, the data is partitioned into K subsets $D_{(k)}$, for $k = 1, \ldots, K$ and then the model is fit separately to each training set $D_{(-k)}$ thus yielding a posterior distribution $p(\boldsymbol{\beta} \mid D_{(-k)})$. When calculating the predictive performance of the model the data of the k^{th} fold is used as test data. The predictive density for $\mathbf{x}_*$, if it is in subset k, is

$$\begin{aligned} p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, D_{(-k)}) &= \int p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid D_{(-k)}) d\boldsymbol{\beta} \\ &\approx \frac{1}{S} \sum_{s=1}^{S} p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, \boldsymbol{\beta}^s) \end{aligned} \tag{C.4}$$

and the posterior predictive expectation is $\hat{\pi}(\mathbf{x}_*) = \int \pi(\mathbf{x}_*) p(\pi(\mathbf{x}_*) \mid \mathbf{x}_*, D_{(-k)}) d\pi(\mathbf{x}_*)$ which is used as a plug-in into equation (1.5) to calculate the K-fold CV estimate of NB in the k^{th} fold, $\text{NB}_{(k)}$. We choose λ as

$$\lambda^* = \arg\max_{\lambda} \frac{1}{K} \sum_{k=1}^{K} NB_{(k)}$$

We use $K = 5$ for the all analysis. In practice, we have seen the results are insensitive to the choice of K.

C.4 Computational scheme

For all analysis in this report we use MCMC which has become a very important computational tool in Bayesian statistics since it allows for Monte Carlo approximation of complex posterior distributions where analytical or numerical integration techniques are not applicable. The Markov chain is constructed using random walk Metropolis-Hastings updates (Brooks et al., 2011). We give a brief overview of the algorithm.

The target distribution is $p(\boldsymbol{\beta}|D)$ (equation (3.5)). The sampling scheme starts at an initial set of parameter values, denote these $\boldsymbol{\beta}^0$. To sample the next set of parameters, which we denote $\boldsymbol{\beta}^1$, we propose moving from the current state to another set of parameter values, $\boldsymbol{\beta}^{new}$, by using a proposal function $q(\boldsymbol{\beta}^{new}|\boldsymbol{\beta})$. We then accept these proposed values as the next sample with probability equal to the Metropolis-Hastings ratio:

$$\text{MHR}(\boldsymbol{\beta}, \boldsymbol{\beta}^{new}) = \frac{L(D|\boldsymbol{\beta}^{new})p(\boldsymbol{\beta}^{new})}{L(D|\boldsymbol{\beta})p(\boldsymbol{\beta})} \times \frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^{new})}{q(\boldsymbol{\beta}^{new}|\boldsymbol{\beta})}, \tag{C.5}$$

where $L(D|\boldsymbol{\beta})$ is the tailored likelihood and $p(\boldsymbol{\beta})$ the prior, given in Section 3.2.2 and 3.2.3. The proposed move is accepted with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^{new}) = min(1, \text{MHR}(\boldsymbol{\beta}, \boldsymbol{\beta}^{new})).$$

If this new set of values is accepted, the proposed set is accepted as $\boldsymbol{\beta}^1$; otherwise, the sample value remains equal to the current sample value, i.e., $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0$. The proposal function is Gaussian, i.e., $q \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{I}sd)$ where sd is chosen to yield an acceptance rate ≈ 0.24 (Brooks et al., 2011). In the current version of the algorithm all parameters are updated jointly.

C.5 Comparison with BART

Given the non-linear decision boundaries of the simulation scenario in Section 3.3.2, we further compare TB with a standard non-linear Bayesian model. We use logistic Bayesian Additive Regression Trees (BART) as implemented in the `BART` package version 2.9 (`lbart()` function) (Sparapani et al., 2021).

Figure C.1 shows the difference in NB between TB and BART. Under the 0.5 prevalence scenario BART performs better than TB except at $t = 0.9$. On the other hand, TB performs better or no worse than BART under prevalence scenarios 0.1 and 0.3. This is noteworthy as these prevalence scenarios are common in medical applications. Together with the results from Figure 3.8, we conclude that for this simulation scenario, TB, albeit implemented as a linear model, mitigates some of the advantages of a non-linear one, such as BART. Note that an additional comparison of interest would be BART with a tailored BART implementation. We leave this for future work.

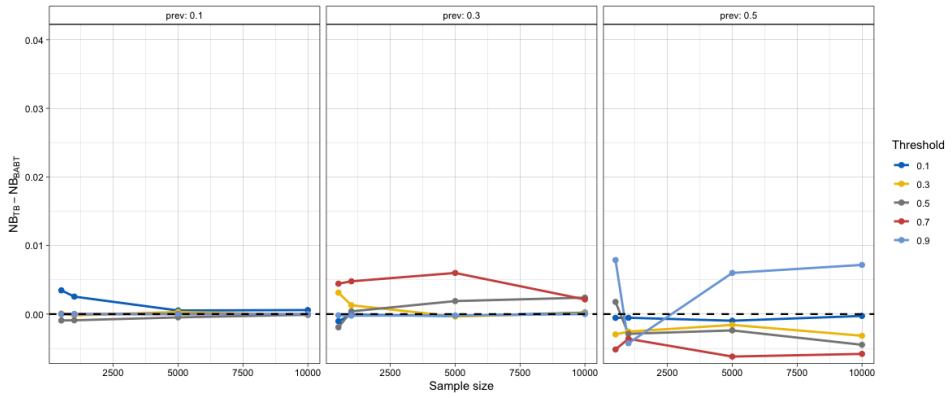


Figure C.1 Difference in Net Benefit for samples sizes of 500, 1000, 5000, 10000 averaged over 20 repetitions. A positive difference means TB outperforms BART. Each grid corresponds to a different prevalence setting.

C.6 Additional experiments and implementation considerations

The method presented in this paper relies on the construction of the datapoint-specific weights (see equation (3.2)). Here we discuss each element in turn.

C.6.1 On choosing λ

We have opted to use cross-validation to choose λ. An open question is how to choose the range of λ values to consider. Our proposal is to consider values of the form $\lambda \in \{0, \ldots, m\}$. When $\lambda = 0$, the model reduces to standard logistic regression, a sensible choice for the lower limit. To choose the upper limit, m, note that as λ increases the rate with which the datapoints are downweighted increases exponentially (Figure C.2). This in turn decreases the effective number of datapoints that are used when training the model. We call this the effective sample size for tailoring, ESS_T. Formally, we define ESS_T as

$$ESS_T = \sum_{i=1}^{n} w_i$$

Under standard modelling, $ESS_T = n$, since $w_i = 1, \forall i$. Under tailoring $ESS_T \leq n$, which is why tailoring results in wider posteriors[1]. This is demonstrated in Figure 3.4. In addition, Figure C.3 shows the precision (as measured by the width) of the HPD credible intervals produced by each model under the simulation setting in Section 3.3.2. The figure suggests that the width of the credible intervals increases under tailoring compared to standard modelling. This is expected due to the downweighting of the likelihood contributions.

As a result, we can use the ESS_T as a guide to choose m. Figure C.4 shows $\frac{ESS_T}{n}$ for various λ values and target thresholds for the breast cancer prognostication case study (Section 3.4.1). Based on a target

[1] recall, $w_i \in [0, 1]$

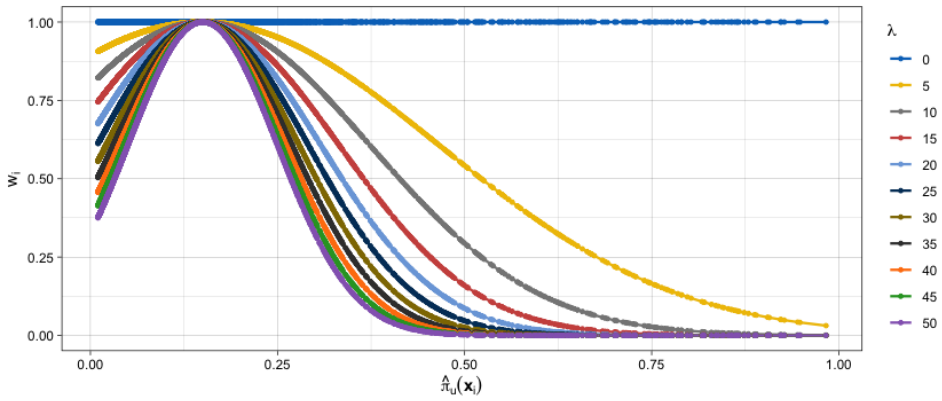


Figure C.2 Distribution of weights, w_i, against $\hat{\pi}_u(\mathbf{x}_i)$ for breast cancer prognostication case study (Section 3.4.1) for $t = 0.15$.

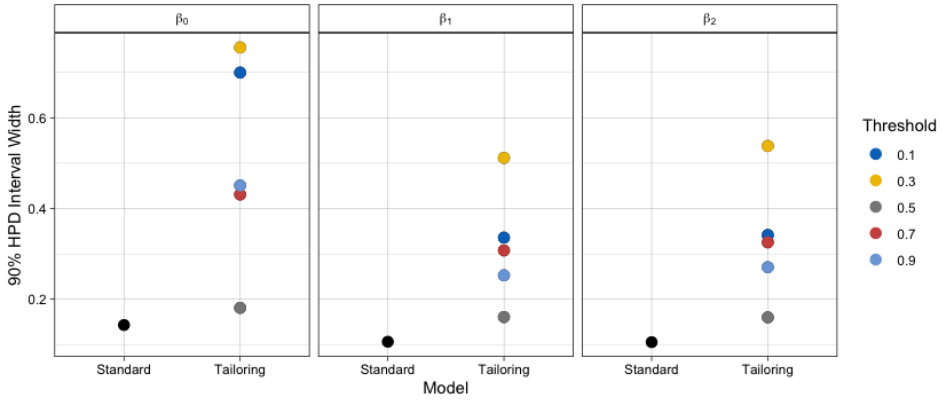


Figure C.3 90% HPD Interval width for each parameter as a function of the model.

threshold we can choose m so the ESS_T does not drop below a pre-specified threshold. Importantly, this plot can be produced before fitting the model since we only need estimates of $\pi_u(\mathbf{x}_i)$.

In a similar fashion, we can have an indication whether TB will outperform SB before fitting the model. This can be achieved by plotting NB as λ increases (Figure C.5). If NB remains stable or decreases as λ increases (Figure C.5, $t = 0.5$ orange points) then TB will probably not offer any performance improvement compared to SB (see Figure 3.11, $t = 0.5$). This is because, as discussed above, when $\lambda = 0$ TB reduces to SB (i.e., all weights are equal to one).

C.6.2 On calibration

Accurate estimation of $\pi_u(\mathbf{x}_i)$ at the first step of our framework is important for the construction of the weights. Ideally, we would like the estimated probabilities, $\hat{\pi}_u(\mathbf{x}_i)$ to be well calibrated. Calibration refers to the degree of agreement between observed and estimated probabilities (Section 1.2). Probabilities are well calibrated if, for every 100 patients given a risk of x%, close to x have the event.

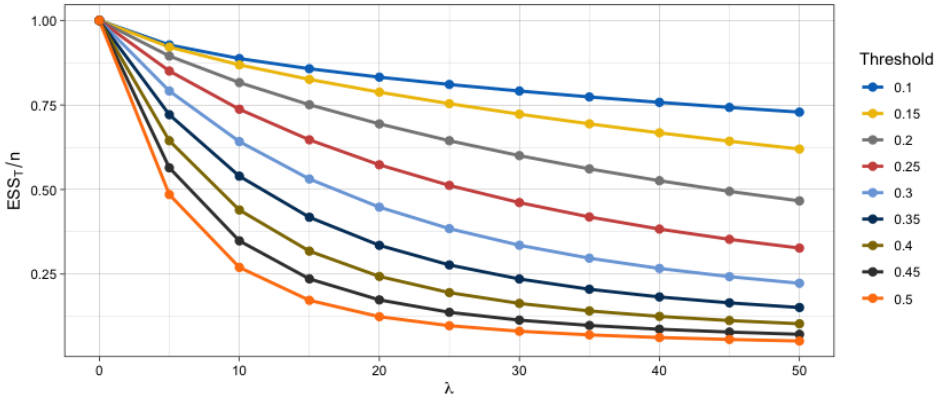


Figure C.4 $\frac{ESS_T}{n}$ for various λ values per target threshold.

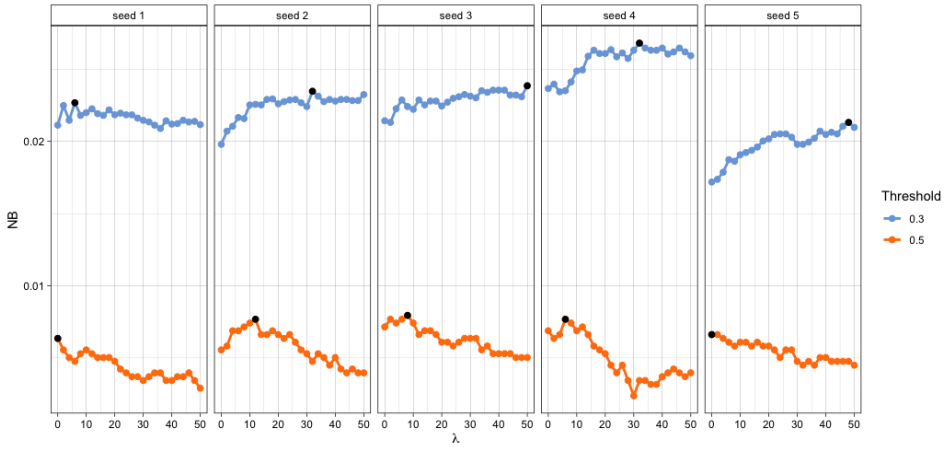
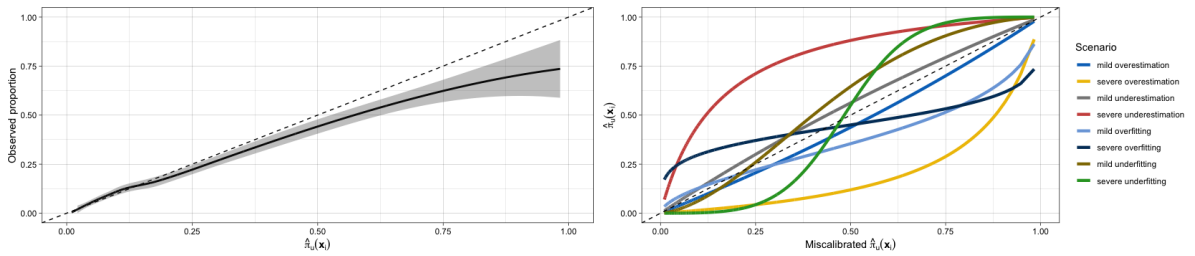


Figure C.5 Average 5-fold CV estimate of Net Benefit for the breast cancer prognostication dataset. Black points correspond to the chosen lambda values, λ^* (defined in Section C.3).



(a) (b)

Figure C.6 (a) Calibration plot of $\hat{\pi}_u(\mathbf{x}_i)$ on the train data using loess smoother. The 45-degree line represents the perfect calibration. (b) Illustrations of different miscalibration scenarios. The y axis shows $\hat{\pi}_u(\mathbf{x}_i)$ and the x axis the miscalibrated $\hat{\pi}(\mathbf{x}_i)$ used in model fitting.

We use the breast cancer prognostication case study (Section 3.4.1) to investigate the effect of miscalibration on the model performance. First, we assess the calibration of $\hat{\pi}_u(\mathbf{x}_i)$. Figure C.6a presents a graphical evaluation of calibration. It is based on loess-based smoothing method (Austin and Steyerberg, 2014) where the estimated (i.e., $\hat{\pi}_u(\mathbf{x}_i)$) and observed probabilities (from the development data) are plotted against each other; good models are close to the 45-degree line. We see the probabilities are well calibrated for the lower thresholds, and tend to be underestimated for the higher risk thresholds. To explore sensitivity of the tailored model to the accuracy of the step 1 probabilities, we deliberately perturbed $\hat{\pi}_u(\mathbf{x}_i)$ generating four miscalibration types: (1) overestimation; (2) underestimation, when probabilities are systematically overestimated or underestimated, respectively; (3) overfitting, when small probabilities are underestimated whereas large ones are overestimated; (4) underfitting, when small probabilities are overestimated whereas large ones are underestimated. We further allowed for two degrees of miscalibration (mild and severe) for each type giving us a total of eight scenarios (Figure C.6b).

Figure C.7 shows the difference in NB between the original tailored (calibrated) model and the tailored miscalibrated ones for the different scenarios. Comparing across miscalibration types we see that the decline in performance depends on the type of miscalibration, with overfitting and underfitting more robust than over- and underestimation. Comparing within each type we note a drop in performance from mild to severe degrees, especially for over- and underestimation.

These results show that the model performance (in terms of NB) depends on the type of miscalibration and is robust to mild miscalibration. In practice, the calibration of the estimated probabilities can be readily evaluated graphically as done here or using statistical tests (Austin and Steyerberg, 2014). If the results show poor calibration we recommend re-calibrating $\hat{\pi}_u(\mathbf{x}_i)$ before calculating the datapoint-specific weights (Janssen et al., 2008; Steyerberg et al., 2004).

C.6.3 On the weighting function

In Section 3.2.1 we defined the weights using the squared distance function, h. Here we investigate the sensitivity of the framework to the choice of the distance function. We choose the family of ε-insensitive

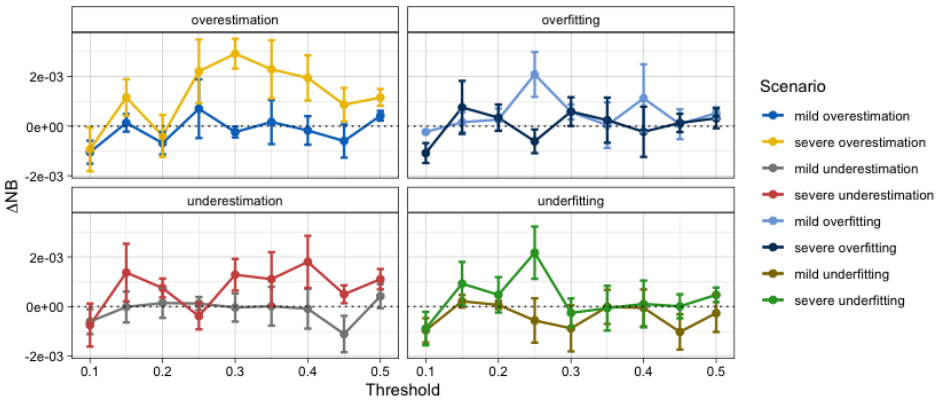


Figure C.7 Difference in NB, ΔNB (on the test set) between original, calibrated TB and miscalibrated TB under different scenarios. A positive difference means the calibrated TB outperforms the miscalibrated one.

functions (Vapnik, 1998), which is defined as

$$h(\pi_u(\mathbf{x}),t) = |\pi_u(\mathbf{x}) - t|_\varepsilon$$

where we denote

$$|\pi_u(\mathbf{x}) - t|_\varepsilon = \begin{cases} 0 & \text{if } |\pi_u(\mathbf{x}) - t| \leq \varepsilon \\ |\pi_u(\mathbf{x}) - t| - \varepsilon & \text{otherwise} \end{cases} \tag{C.6}$$

The ε-insensitivity arises from the fact that the function value is equal to 0 if the discrepancy between the predicted probability $\pi_u(\mathbf{x})$ and the target threshold t is less than ε. In other words, we do not care about the distance as long as it is less than ε, but will not accept any deviation larger than this. As a result, observations with predicted probability within ε of the target threshold will not be downweighted. For $\varepsilon = 0$ we recover the absolute distance, which is the objective function in median regression (Bassett Jr and Koenker, 1978). Both the squared distance and the family of ε-insensitive functions are symmetric, i.e., they downweight equally observations based only on their distance from the target threshold, not taking into account the direction. This is a reasonable requirement for our weighting function. Figure C.8 presents the results for various ε values for the breast cancer case study (Section 3.4.1). The conclusions are qualitatively unchanged when compared within different ε values and between ε-insensitive functions and the squared distance (first panel in Figure C.8). Hence, we conclude that for this dataset the results are also robust to the choice of the weighting function.

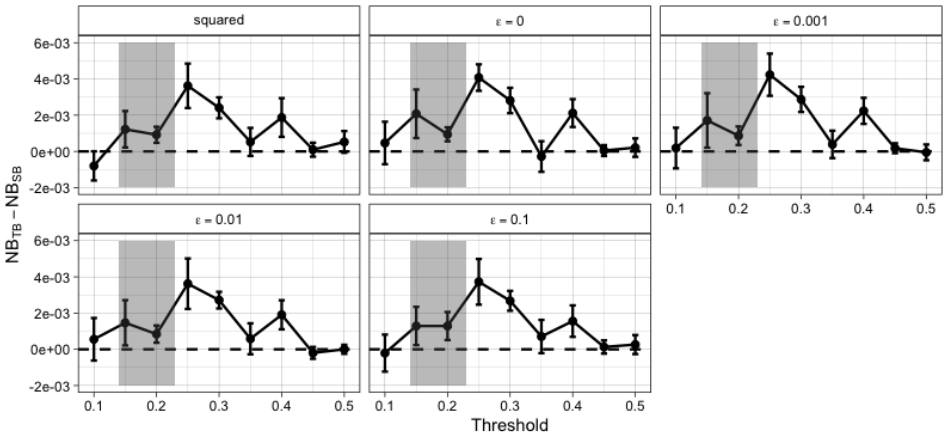


Figure C.8 Difference in NB (breast cancer prognostication case study) between TB and SB under the squared distance and ε-insensitive functions for various ε values. Note the first panel corresponds to Figure 3.11.

Appendix D

Appendix to Chapter 4

D.1 Computational scheme

An exhaustive evaluation of all possible combinations of the covariates is computationally prohibitive even for a moderate number of covariates. To alleviate this issue we have implemented an iterative MCMC sampling algorithm, reversible jump MCMC (RJMCMC) (Green, 1995).

RJMCMC is a class of MCMC methods tailored for drawing posterior samples spanning multiple parametric models together with their model-specific parameters, by jumping between models as part of the sampling. The algorithm is conceptually similar with the one described in Appendix C.4, with the caveat that we now need to sample the model, $\boldsymbol{\gamma}$, alongside the other parameters, $\boldsymbol{\theta}$. The algorithm starts at an initial model, $\boldsymbol{\gamma}^{(0)}$, which is a selection of covariates, and corresponding set of parameter values, $\boldsymbol{\theta}^{(0)}$. To sample the next model and set of parameters, which we denote by $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\theta}^{(1)}$, we propose moving from the current state to another model and/or parameter values, $\boldsymbol{\gamma}^{(new)}$ and $\boldsymbol{\theta}^{(new)}$, using a proposal function $q(\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma})$ (see later for details). We then accept these proposed values as the next sample with probability equal to the Metropolis-Hastings ratio:

$$\text{MHR} = \frac{L(D|\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(new)}) p(\boldsymbol{\theta}^{(new)}|\boldsymbol{\gamma}^{(new)}) p(\boldsymbol{\gamma}^{(new)})}{L(D|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma})} \times \frac{q(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(new)})}{q(\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(new)}|\boldsymbol{\theta}, \boldsymbol{\gamma})}, \tag{D.1}$$

where $L(D|\boldsymbol{\theta}, \boldsymbol{\gamma})$ is the tailored likelihood given in equation (4.18), $p(\boldsymbol{\gamma})$ is the model space prior defined in equation (4.9) and $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is the prior on the parameters conditional on (i.e., included in) the model (see equations (4.11) and (4.12)). The proposed move is accepted with probability $min(1, \text{MHR})$. If this new set of values is accepted, we set $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\gamma}^{(new)}$ and $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(new)}$, otherwise the current values are retained $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$. This produces results in a sequence of parameter/model samples, which converge to the target posterior distribution, $p(\boldsymbol{\theta}|D)$.

We briefly describe how the proposal function, $q(\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma})$, is chosen. For details we refer to Newcombe et al. (2017). The proposal is done in 2 steps with first a proposal for $\boldsymbol{\gamma}^{(new)}$ (the model

proposal) and then a proposal for $\boldsymbol{\theta}^{(new)}$ (the parameter update) which implies that

$$q(\boldsymbol{\theta}^{(new)}, \boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma}) = q(\boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma}) q(\boldsymbol{\theta}^{(new)} | \boldsymbol{\gamma}^{(\text{new})}, \boldsymbol{\theta}, \boldsymbol{\gamma})$$

The model proposal $q(\boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma})$ is done as follows: First, the type of move is determined from four possibilities: (1) adding a covariate, (2) removing a covariate, (3) swapping the presence of one covariate for another, or (4) a 'null' move where no change to the model is made. These move types are assigned conditional probabilities according to the number of covariates currently included in the model; an addition can only occur when there are $< P$ covariates present, a removal can only occur when there are > 0 covariates present, and a swap can only occur when there are ≥ 1 covariates present. Next, if an addition, removal or swap move was selected, the covariates to be involved in the move are picked from the covariates available for the move (e.g., an addition can only involve covariates that are currently excluded) with equal probability. Therefore, $q(\boldsymbol{\gamma}^{(\text{new})} | \boldsymbol{\theta}, \boldsymbol{\gamma})$ is determined by multiplying the probability of the move type and, with the exception of a 'null' move, the probability of selecting the particular covariates involved in the move.

The parameter updates $q(\boldsymbol{\theta}^{(new)} | \boldsymbol{\gamma}^{(\text{new})}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ involve standard sampling methods. For parameters remaining in the model during a move, proposals are drawn from Gaussian distributions centred on the current value. For an addition move, values for the new parameter are drawn from Gaussian distributions, centred on zero. Therefore, $q(\boldsymbol{\theta}^{(new)} | \boldsymbol{\gamma}^{(\text{new})}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is calculated as a Gaussian density (product, where more than one parameter is involved in the move).

D.2 Convergence Diagnostics

To assess the computational performance of the RJMCMC sampling algorithm we performed several convergence tests. These include 1) trace plots, 2) reproducibility of space exploration, and 3) the Potential Scale Reduction Factor (PSRF). The convergence tests are demonstrated using the Diabetes data set (Section 4.4.2) and the TB model for $t = 0.5$. All evaluated after running the algorithm five times with random starting points. All tests indicate convergence of the RJMCMC sampler.

D.2.1 Trace plots

Figure D.1 displays the mixing of the regression coefficients. The relatively constant distribution of parameter values over the course of the Markov Chain indicates good mixing.

D.2.2 Reproducibility of space exploration

To investigate whether final results correspond to the mathematical model and are not a product of poor convergence of the Markov Chain (such as getting stuck in a local optimum) we run the algorithm five times with random starting points. The results are displayed in Figure D.2, but with standard deviations

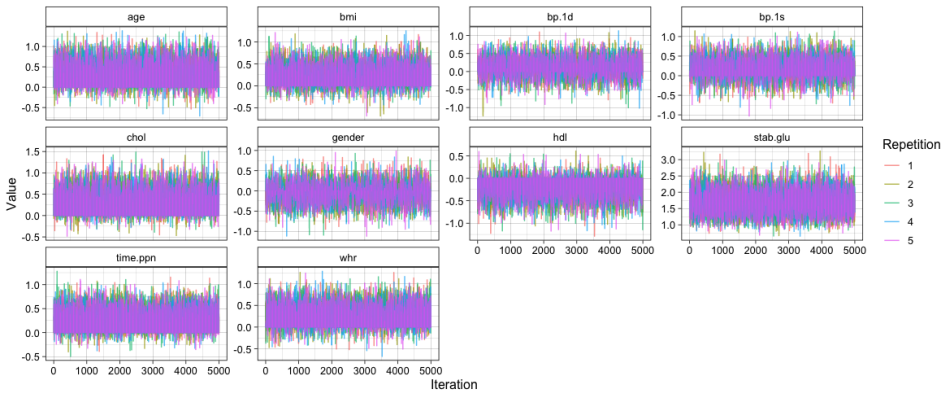


Figure D.1 Trace plots of $\boldsymbol{\beta}$ parameters for each of the five repetitions. The sampler was run for 1 million iterations in total, with 50% as burn-in, after which every 100th sample is stored.

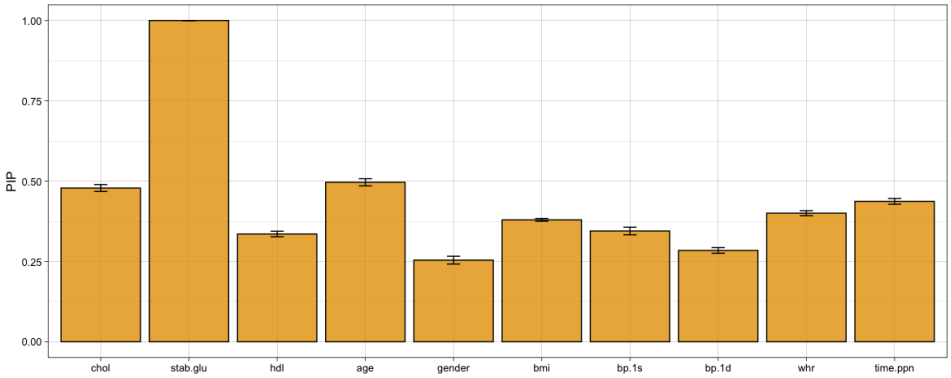


Figure D.2 Posterior inclusion probability (PIP) with standard deviations, based on five independent RJMCMC runs with random starting points.

added, based on the five independent runs. The low standard deviations indicate a good reproducibility of the RJMCMC sampler.

D.2.3 Potential Scale Reduction Factor

The Potential Scale Reduction Factor (PSRF) is a common quantifier of Markov Chain convergence, which takes values in the interval $[1, \infty)$ and quantifies the difference in the obtained distributions from a number of independent executions, here five (Brooks and Gelman, 1998; Gelman et al., 2013). Conventionally, convergence is stated when the PSRF is lower than 1.1. Figure D.3 displays the PSRF averaged over the 5 runs of the RJMCMC sampler. All values are well below 1.1, thereby indicating good convergence.

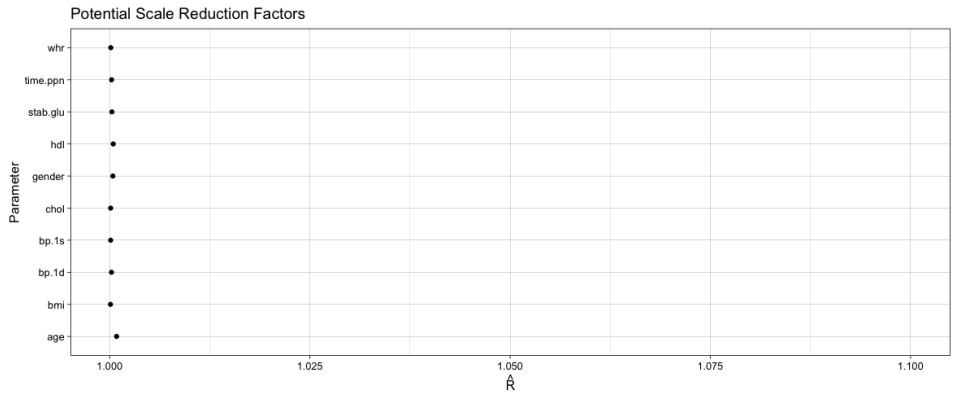


Figure D.3 PSRF values for each covariate, averaged across five independent runs.