ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Uncertainty Estimation and Natural Language Processing to Identify Patients under Chemotherapy at Risk of Acute Care Use

Master Thesis

Claudio Fanconi
Department of Information Technology and Electrical Engineering (D-ITET)

**A**dvisor: Prof. Dr. Tina Hernandez-Boussard Stanford University
**S**upervisor: Prof. Dr. Ender Konukoglu ETH Zürich

November 6, 2022

# Abstract

Identifying cancer patients at risk of acute care use (ACU), such as emergency department visits and inpatient admissions, once they start chemotherapy is crucial because these events can often be prevented and are expensive. Therefore, this thesis evaluates how uncertainty estimation and natural language processing (NLP) methods can help predict the risk of ACU.

A first experiment explores how uncertainty can be applied in the clinical setting. Consequently, Bayesian logistic LASSO regression (BLLR) models are compared with standard $\ell_1$-penalised logistic regression on high-dimensional structured health data (SHD). This analysis shows that BLLR with a Horseshoe+ prior and a posterior approximated by Metropolis-Hastings sampling is a promising alternative that performs on par with ordinary logistic LASSO in terms of predictive performance and offers the advantage of uncertainty estimation. Furthermore, this work shows how different Bayesian models and their uncertainties can be compared for clinical classification tasks. Additionally, it highlights an interesting phenomenon where predictive uncertainties can be biased across different patient subgroups.

In a second experiment, this thesis explores how NLP can be used to determine the risk of ACU. Risk prediction using SHD is now standard, but prediction using free-text formats is complex. Clinical notes of 6,938 cancer patients are investigated as input for ACU prediction instead of SHD or a combination thereof. Deep learning models are compared with manually engineered language features. The results show that SHD models slightly outperformed NLP models; a logistic LASSO regression with SHD achieves a C-statistic of 0.748 (95%-CI: 0.735, 0.762), while the same model with clinical notes obtains 0.730 (95%-CI: 0.717, 0.745) and a transformer-based model achieves 0.702 (95%-CI: 0.688, 0.717). This experiment demonstrates how language models could be used in clinical applications and draws attention to bias in risk predictions across patient groups, even when using free-text data.

This thesis highlights the importance of estimating uncertainty in medicine to address over-reliance on point estimates and that clinical notes can be used to predict the risk of ACU.

# Acknowledgements

Most importantly, this work would not have been possible without my advisor's invitation and tireless support, Professor Tina Hernandez-Boussard. Tina was an exceptional mentor and ensured I had the best working environment to thrive, grow and learn. I would also like to thank all the other members of the Boussard Lab who always went the extra mile to answer my questions and help me solve complex problems. Additionally, I would especially like to thank three friends from the lab. First, Anne, for all the engaging debates about Bayesian statistics and the laughs during our tea time. Second, Marieke, for all the fruitful discussions on neural network architectures and surfing adventures. Third, Angelo, for all the input from a medical perspective and the endless rapid chess games.

I would also like to thank Professor Ender Konukoglu of the Computer Vision Laboratory at ETH Zürich, who served as my official academic supervisor for this work. I especially appreciate the trust Professor Konukoglu has placed in me to independently pursue research in the Department of Biomedical Informatics at Stanford.

With this final capstone to my studies, I would like to thank my friends, Giovanni, Fabio, Andrea, Leone & Reto, for the good times and fun evenings and for their love and support during the learning process over the past years. Of the connections I made during my Bachelor's and Master's studies, I am especially indebted to Frédéric and Nikola; my friendships with them were invaluable not only for my education in electrical engineering but also for my development as a person.

Finally, I cannot thank my parents, Cinzia and Daniel, my sisters, Linda and Miranda, and my girlfriend, Sara, enough for supporting me so kindly throughout my studies and for their love, warmth and appreciation.

# Contents

# List of Figures

# List of Tables


.

# Nomenclature

| | |
|---|---|
| ACU | Acute Care Use |
| ADVI | Automatic Differentiation Variational Inference |
| AUPRC | Area Under the Precision-Recall Curve |
| AUROC | Area Under the Receiver Operator Characteristic |
| BERT | Bidirectional Encoder Representations from Transformers |
| BLLR | Bayesian Logistic LASSO Regression |
| BNN | Bayesian Neural Network |
| BoW | Bag-of-Words |
| CDF | Cumulative Distribution Function |
| CI | Confidence Interval |
| CMS | Center for Medicare & Medicaid Services |
| DCA | Decision Curve Analysis |
| ECE | Expected Calibration Error |
| EHR(s) | Electronic Health Record(s) |
| GPU | Graphics Processing Unit |
| HS | Horseshoe (prior) |
| KL | Kullback-Leibler |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| MAP | Maximum A Posteriori |
| MH | Metropolis-Hastings |
| ML | Machine Learning |
| MLE | Maximum Likelihood Estimation |
| NLP | Natural Language Processing |
| SD | Standard Deviation |
| SHD | Structured Health Data |
| VI | Variational Inference |

# Chapter 1

# Introduction

This thesis investigates how machine learning (ML) methods can help predict cancer patients' risk of acute care use (ACU) after starting chemotherapy. More specifically, the added value of uncertainty estimation methods for clinical decision support and the use of free-text medical notes to predict ACU are analysed. Taking the use case of ACU prediction as its pivot, this thesis offered two improved solutions: the first focused on uncertainty quantification and the second on natural language processing (NLP). This chapter introduces the clinical problem of ACU. The subsequent sections explain the motivation and scope of this work and the two experiments. Finally, the organisation of the thesis is outlined in section 1.4.

## 1.1 Acute Care for Oncology Patients

After starting chemotherapy, cancer patients undergoing chemotherapy often require acute care, such as emergency room visits and inpatient admissions. Some reasons for ACU include pain, fever, sepsis, vomiting, pneumonia, diarrhoea and nausea [1]. Because chemotherapy often requires several treatment cycles, ACU events can occur at different intervals after treatment begins. Patients may need acute care within the first month and at later stages of their therapy, and different complications might require acute care at different times during a patient's clinical trajectory [2, 3].

## 1.2 Motivation

ACU interventions account for nearly half of the costs associated with oncology care in the United States [4, 5]. Evidence suggests that about 50% of these treatments are preventable with early outpatient interventions [2, 3]. A previous paper by Peterson et al. [1] presented an ML model using structured health data (SHD) from electronic health records (EHR) to identify patients at high risk for ACU after chemotherapy initiation. These and other papers highlight the potential of data-driven models to predict ACU risk [6, 7, 8]. This thesis was inspired by trying to reduce preventable ACU events and the costs associated with them.

The motivation for the first experiment comes from the fact that, traditionally, ML models in healthcare have used point estimates, a single number known as risk probability, to report performance/predictions. However, these models rarely quantify the uncertainty of their predictions and inform their users how likely it is for them to be wrong - a piece of information that bears considerable value for clinical deployment. This information would be necessary for several stakeholders. First, for the data scientist, this information helps to develop robust models and validate them. For the clinician, the information serves to understand better and interpret the risk probability. Finally, it enables healthcare decision-makers to select which prediction tasks can be automated and which should abstain from automatic decision-making

if the estimated uncertainty is too high [9]. We believe that uncertainty estimation can significantly aid in using risk estimates to make more informed, accurate and reliable clinical decisions and to avoid over-reliance on a point estimate of probability to decide whether or not to intervene.
The motivation for the second experiment comes from the observation that most EHRs are not mapped to a common data model and are not necessarily standardised between different facilities. To replicate other hospitals' predictive models based on SHD could require intensive data preparation. On the other hand, 96% of hospitals in the United States [10], and 92% in Switzerland [11], collect digital clinical notes from healthcare workers by the time of this thesis. This information remains mainly unused without NLP but is essential to understand a patient's trajectory. We believe that NLP methods can extract useful information from these unstructured clinical texts and be used to predict the risk of ACU.

## 1.3 Scope of Thesis

To demonstrate the value of quantified uncertainty in clinical decision support, this work first developed a Bayesian $\ell_1$-penalised logistic regression (the penalty is also known as the Least Absolute Shrinkage and Selection Operator - LASSO). The performance of this Bayesian logistic LASSO regression (BLLR) in predicting ACU risk was compared with a traditional (frequentist) logistic LASSO on an existing patient dataset consisting of the patients' high-dimensional tabular health data. This experiment showed how the quantified uncertainty from BLLR is applied to individual risk predictions and the entire patient cohort. Furthermore, the results showed how the estimated uncertainty in the BLLR differs depending on the choice of parameter prior distribution (i.e. Laplace and Horseshoe+ priors) and posterior approximation methods (i.e. variational inference and Metropolis-Hastings sampling). Additionally, disparities in predictive uncertainties across patient groups were analysed.
In a second step, this thesis aimed to replace tabular inputs with features from unstructured clinical notes or combine both modalities to identify patients at risk of needing an ACU. In addition, the goal was to investigate whether novel deep learning language models outperform traditional language feature extraction and linear models. These aspects were investigated by increasing the number of features of the existing cohort of patients with the corresponding medical notes and developing five prediction models. The models were trained to predict ACU with different inputs and compared their predictive performance and utility when applied at the point of care. In addition, risk prediction bias across patient subgroups was investigated.

## 1.4 Thesis Organisation

This thesis is organised as follows: In the following Chapter (2), previous works that focused on predicting the acute care of oncology patients using machine learning, uncertainty estimation in medicine, and NLP methods in healthcare are presented. Chapter 3 outlines the theory and mathematical setting of Bayesian machine learning and NLP methods. Subsequently, the methods and results for the uncertainty experiments are presented in the context of their added value for clinical decision support (Chapter 4). Then, the methods and results for our second experiment were reported, which focused on using free-text clinical notes to predict the risk of ACU in Chapter 5. The results of the two experiments are discussed in Chapter 6. Finally, a conclusion is provided (Chapter 7), and possible directions for future work are presented (Chapter 8).

# Chapter 2

# Related Work

This chapter discusses previous findings from the literature. Initially, it focuses on works on data-driven prediction methods used to estimate the risk of acute care use for cancer patients. In a second step, existing literature on uncertainty quantification and how it is currently utilised in healthcare is pointed out. Finally, previous works that applied natural language processing in healthcare settings are examined.

## 2.1 Machine Learning to Predict Acute Care

Previous papers focused on machine learning and statistics to predict chemotherapy-related acute care utilisation. Brooks et al. [6] studied palliative cancer patients with malignant solid tumours who required hospitalisation within 30 days of their last chemotherapy administration. In their work, they developed a multivariable logistic regression model based on demographic, clinical, and laboratory variables. Their results showed that seven variables were significantly correlated with higher risk of chemotherapy-related hospitalisations: age, Charlson comorbidity score, creatinine clearance, calcium level, low white blood cell count, polychemotherapy (compared to monotherapy) and receipt of camptothecin chemotherapy [6].
Grant et al. [7] analysed the risk of ACU within 30 days of starting chemotherapy in a cohort of 28,010 cancer patients. Using backwards characteristic selection, they introduced the Prediction of Acute Care Use During Cancer Treatment (PROACCT) score, which combines four variables: cancer type and treatment regimen, age, and emergency department visits in the previous year. They discuss that the PROACCT score can be used directly in a univariate regression to estimate the risk of ACU.
Brooks et al. [12] investigated the risk of 30-day hospitalisation in patients with an advanced cancer. They developed a risk stratification model that divided a cohort of patients into high-risk and low-risk patients, based solely on albumin and sodium values. They selected these input variables using a logistic regression LASSO.
Daly et al. [8] developed a model that identifies patients at high risk for a potentially avoidable acute care visit within the first six months of chemotherapy. Compared to the work above, this research focused on high-dimensional clinical data and used 270 features extracted from the EHR system.
Finally, Peterson et al. [1] analysed a cohort of patients for risk of preventable ACU within 30, 180 and 365 days of initiating chemotherapy. In their paper, they selected a study population subject to the OP-35 metric [13], a quality metric implemented by the Centers for Medicare & Medicaid Services (CMS) that penalises healthcare providers for preventable ACU events. Using dense electronic medical records originally with 760 variables (including demographic factors, laboratory values, medication orders and procedure codes), they showed that these can be brought into play to predict the risk of inpatient admissions and emergency department stays. Their studies analysed several ML models and proposed a logistic regression LASSO as the most appropriate solution to the problem. By using the $\ell_1$-penalty in their logistic regression model, they reduced the number of predictive features to 125.

## 2.2 Uncertainty Quantification in Medicine

There are several approaches to quantifying uncertainty in ML models [9]. One of the most common approaches to estimating predictive uncertainty in statistics is the Bayesian framework. Compared to the currently more common frequentist statistics framework, where the risk predictions are just single-point probabilities, Bayesian approaches produce predictive distributions that can quantify uncertainty (more is explained in Chapter 3.1).
Carlin et al. [14] analysed Bayesian and frequentist statistical methods for comparing multiple treatments in the context of pharmacological treatments for female urinary incontinence. Their results showed that Bayesian methods are more flexible than frequentist ones, and their results are more clinically interpretable but more challenging to develop.
Dagliati et al. [15] investigated a Bayesian logistic regression to forecast metabolic control in type II diabethes patients. The Bayesian model outperformed a classic logistic regression significantly on the Matthews Correlation Coefficient on the test set. The hierarchical structure allowed them to take into account population and individual variability.
Beker et al. [16] developed a Bayesian neural network (BNN [17]) for predicting the similarity to drugs of molecules. They improved their classification by selecting predictions with low predictive uncertainty from the BNN.
Similarly, BNNs are exploited by Joshi and Dhar [18] for filtering and correcting uncertainties in cancer classification. The authors show how quantified uncertainty (defined here as the variance of the predictive distribution) improves classification accuracy.
To classify medical images, Syrykh et al. [19] used a convolutional neural network with Monte Carlo dropout [20] to diagnose lymphoma on histopathological images. The authors argue that in addition to achieving strong prediction results, predictive uncertainty (quantified by the variance of the predictive distribution) helped detect unknown data. Furthermore, they demonstrated that the area under the receiver operating curve increased when uncertain slides were removed from the images.
For semantic segmentation of medical images, Baumgartner et al. [21] proposed a conditional variational autoencoder to model segmentation at different resolutions. They quantify the pixel-wise uncertainty by taking the expected cross-entropy between the mean segmentation mask and the samples. Their results show that their model can produce realistic uncertainty segmentation maps.
Meijerink et al. [22] researched uncertainty estimates for out-of-distribution detection based on structured EHR data. In their work, they used ensemble models to quantify uncertainty by calculating the entropy of the predicted distributions. They also proposed four clinical scenarios in which out-of-distribution can be useful: detecting rare and new diseases, finding underrepresented patient groups, detecting false admissions and finding corrupted data.
Kang et al. [23] demonstrated how uncertainty can be exploited in a "human in the loop" approach to testing uncertain sleep stages on an epoch basis with neural network models. To estimate uncertainty, they used the Shannon entropy [24]. The authors argued that their uncertainty-based clinician-in-the-loop framework improves both classification accuracy and trustworthiness in a cost-effective and economically resourceful manner.
More recently, generalised additive models were used to show how quantified uncertainty (using the variance of the predictive distribution) can be applied to visualise a risk distribution for patient mortality, such as in the work of Mathiszig-Lee et al. [25]. The authors illustrated a case study in which the clinical risk model has high predictive uncertainty when missing laboratory values.

## 2.3 Natural Language Processing in Healthcare

NLP methods have already proven useful in clinical applications. Currently, the most popular methods can be divided into the manual engineering of language features and using neural networks to learn the

features directly from the data [26]. The theory of both methods is presented in chapter 3.2.
For predicting intensive care unit (ICU) outcomes, Marafino et al. [27] demonstrated that NLP-derived terms associated with mortality could significantly increase the performance of their model using laboratory test results or vital signs. They extracted NLP terms by filtering the 1,000 most frequently occurring words and weighting them using the TF-IDF algorithm [28].
In another study, Marafino et al. [29] developed a support vector machine (SVM) to identify various procedures and diagnoses in ICU clinical notes for use in risk adjustment. In their work, they compared unigram features with bigram features and features negated with the NegEx [30] algorithm. They argued that SVM-based classifiers can accurately identify ICU patients' procedural status and diagnoses and that the use of n-gram features improves performance.
Heo et al. [31] investigated whether NLP methods can predict poor outcomes in patients with acute ischaemic stroke based on MRI free-text reports of the brain. They compared sequence-agnostic features, such as frequency of terms, with sequence-specific deep learning models and showed that deep learning models outperformed traditional methods.
Alsentzer et al. [32] trained a Bidirectional Encoder Representations from Transformers (BERT [33, 34]) for generic clinical text and discharge reports and named it ClinicalBERT. They demonstrated that using a domain-specific language model leads to performance improvements in common clinical NLP tasks.
Similarly, Huang et al. [35] also worked with ClinicalBERT mentioned above after fine-tuning it for 30-day hospital readmissions. They use the free-text discharge reports of patients in the ICU. Their results show that ClinicalBERT can achieve better risk prediction than other neural network architectures and manually generated features.
Sarraju et al. [36] developed a fine-tuned ClinicalBERT to classify non-use of statins in high-risk populations. In addition, they showed that the model identifies reasons for non-use, such as side effects and patient preferences.
Recently, Gatto et al. [37] evaluated the effectiveness of transfer learning methods for telemedical triage. They compared a TF-IDF features with word embeddings on different models, including BERT. Their results showed that the transformer based models outperformed other models.

## 2.4 Developing the Research Questions

Based on the previous works summarised in the sections above, the research questions for this thesis could be asked. How can uncertainty estimation be used in a clinical risk prediction case, and what models are suited best to quantify it? While the past works mainly focused on providing either a point-estimate models [1, 6, 7, 8, 12] or a single uncertainty model [15, 16, 23, 25], this thesis aimed to provide a comparison of multiple uncertainty methods.
Furthermore, the question was asked: can NLP be used to predict ACU instead of SHD? Or can it be used to improve the predictions if used in combination? The literature on this matter demonstrated that NLP has already been successfully applied to numerous other clinical problems [27, 27, 29, 35, 36, 37], and inspired an investigation of its utility for ACU prediction.

# Chapter 3

# Theory

This chapter explains the theory and mathematics of Bayesian machine learning and how it can quantify the uncertainty of its predictions. Compared to the currently more common frequentist statistics framework, where the model parameters are single-point estimates, the Bayesian approach aims to estimate the distribution of the model parameters. Consequently, the risk predictions are not just single-point probabilities but distributions of probabilities that can quantify uncertainty. This section provides the theoretical foundation necessary to understand the first experiment (chapter 4).
In a second step, this chapter presents the mathematics to manually engineer language features from free text and use transformer models with an attention mechanism to extract features automatically. This theoretical foundation motivates the methods of the second experiment in chapter 5.

## 3.1 Bayesian Machine Learning

Bayesian machine learning is a systematic approach that can approximate the *posterior predictive distribution* of new incoming data points. With this distribution, uncertainty properties of the prediction may be derived. It is obtained via a specific likelihood estimation based on Bayes' Theorem. Bayes' Theorem on two random variables $A$ and $B$ is:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{3.1}$$

To fit a binary risk model, there is feature input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ (where $n \in \mathbb{N}$ is the number of inputs, and $d \in \mathbb{N}$ is the number of features), a label vector $\mathbf{y} \in \{0, 1\}^n$, and $\boldsymbol{\theta} \in \mathbb{R}^d$ a $d$-dimensional vector of the model parameters. The Bayes' Theorem of Equation 3.1, conditioned on the already existing input $\mathbf{X}$, is written as follows:

$$P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}|\mathbf{X})}{P(\mathbf{y}|\mathbf{X})} \tag{3.2}$$

$P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$ is the estimated *posterior* distribution of the model parameters, based on the inputs and labels. $P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is the *likelihood* distribution of the labels, given the weights and input data. $P(\boldsymbol{\theta}|\mathbf{X})$ is the *prior* distribution on the model weights, which are assumed to be independent of the inputs, and therefore reduced to $P(\boldsymbol{\theta})$. Finally, $P(\mathbf{y}|\mathbf{X})$ is the *evidence*, which is a normalisation constant of the posterior probability density. In the literature, Equation 3.2 is therefore often written as:

$$P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{1}{Z} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}) \tag{3.3}$$

$$Z = P(\mathbf{y}|\mathbf{X}) \tag{3.4}$$

### 3.1.1 Maximum Likelihood Estimation vs Maximum a Posteriori Estimation

In traditional frequentist machine learning, the goal is to train model parameters to maximise the likelihood of the labels. Therefore successive iterations of Maximum Likelihood Estimation (MLE) are performed on training, which can mathematically be described as:

$$\boldsymbol{\theta}^*_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \tag{3.5}$$

The main goal in Bayesian ML is to estimate the posterior distribution, given the likelihood and the prior distribution. The parameters that maximise the posterior probability yield the so-called Maximum A Posteriori (MAP) estimate. The estimation problem can mathematically be written as:

$$\begin{aligned}\boldsymbol{\theta}^*_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) && (3.6)\\ &= \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}) && (3.7)\end{aligned}$$

The evidence is dropped in the maximisation equation, as it does not depend on the model parameters $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta}^*_{\text{MLE}}$ and $\boldsymbol{\theta}^*_{\text{MAP}}$ are the same, when the parameter prior distribution $P(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. Therefore, classical ML approaches perform MAP estimation by assuming a uniform prior distribution.

### 3.1.2 Making Predictions

To obtain a point estimate risk prediction of a new data point $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$, the MAP estimate of the parameters and a link function $f : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow [0, 1]$ between the model parameters and the input is used. In the case of binary risk classification, the link function corresponds to a sigmoid (inverse logit) function:

$$\hat{y} = f(\mathbf{x}_{\text{new}}, \boldsymbol{\theta}_{\text{MAP}}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top_{\text{MAP}} \mathbf{x}_{\text{new}})} \tag{3.8}$$

However, an advantage of Bayesian ML is that it is not limited to only obtaining point estimates of the risk probability. By marginalising the model weights of the posterior distribution, one can obtain a distribution of risk probabilities, which account for the uncertainty in $\boldsymbol{\theta}$:

$$\begin{aligned}P(\hat{y}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) &= \int P(\hat{y}, \boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) d\boldsymbol{\theta} && (3.9)\\ &= \int P(\hat{y}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) d\boldsymbol{\theta} && (3.10)\\ &= \int P(\hat{y}|\boldsymbol{\theta}, \mathbf{x}_{\text{new}}) P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\theta} && (3.11)\end{aligned}$$

This distribution is known as posterior predictive distribution and referred to in this thesis only as predictive distribution[1] for brevity. From Equation 3.10 to Equation 3.11, the notation is simplified for the likelihood of $\hat{y}$ (as it only depends on the input vector $\mathbf{x}_{\text{new}}$ and model parameters $\boldsymbol{\theta}$) and the posterior (as it does not depend on the unseen $\mathbf{x}_{\text{new}}$).
Unfortunately, in many cases, the posterior itself is not tractable and cannot be calculated analytically, especially for high-dimensional feature space. Subsections 3.1.5 demonstrates two computationally feasible methods to approximate the posterior distribution.

[1] This applies only to this work and shall not be confused with the *prior predictive distribution*.

### 3.1.3 Quantifying Uncertainty

There are two main types of uncertainty in a prediction model important to machine learning scientists [38]: *aleatoric* uncertainty and *epistemic* uncertainty. Aleatoric uncertainty refers to the uncertainty within the observed data. As it is inherent to the measurements, it cannot be reduced. On the other hand, epistemic uncertainty (also known as model uncertainty) refers to imperfections of the model and its ability to understand the underlying process of the data. As more data becomes available, epistemic uncertainty is reduced.
The uncertainty in a prediction is the sum of epistemic and aleatoric uncertainty [39]. It is represented in the dispersion of the predictive distribution, and it can be quantified in numerous ways. Naturally, variance and standard deviation of the predictive distribution are well-suited metrics for this task, as they are measures of dispersion from the mean of the distribution. High dispersion indicates high uncertainty; low dispersion is for low uncertainty. Let us assume a predictive distribution $P(\hat{y}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}})$ from which $T \in \mathbb{N}$ risk probabilities $\hat{y}^{(t)}$ can be sampled:

$$\hat{y}^{(t)} \sim P(\hat{y}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) \qquad t \in \{1, ...T\}, \tag{3.12}$$

then the standard deviation and the mean can be calculated by

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}^{(t)} - \bar{y})^2} \qquad \bar{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}^{(t)} \tag{3.13}$$

However, one is not limited to these statistics to quantify the predictive uncertainty. Another option is to use ranges in the (empirical) inverse cumulative distribution function. The cumulative distribution function (CDF) of the posterior predictive distribution is defined as:

$$F(\alpha) = P(\hat{y} \leq \alpha|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}}) = \int_0^{\alpha} P(\hat{y}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{\text{new}})d\hat{y} \tag{3.14}$$

where $\alpha \in [0, 1]$ (this is because the risk probabilities $\hat{y}$ cannot be smaller than zero or larger than one), and $F : [0, 1] \to [0, 1]$ is the probability that $\hat{y}$ will have a risk score less then $\alpha$. By taking the inverse of the CDF (also known as the quantile function) $F^{-1} : [0, 1] \to [0, 1]$, the risk score $\hat{y}$ with a probability of less or equal to $\alpha$ can be determined. To demonstrate this with an example: to quantify the predictive uncertainty as the range where 95% of predictions lay, it is calculated by:

$$\gamma = \left[F^{-1}(0.025), F^{-1}(0.975)\right] \tag{3.15}$$

In Bayesian statistics, this is referred to as the 95%-*credible interval*[2] of the predictive distribution.
Other statistics can quantify uncertainty, such as expected cross-entropy, Shannon entropy [24], or mean absolute deviation, which will not be discussed further in this thesis.

### 3.1.4 Likelihood and Priors

Depending on the task, choosing the right distribution for the likelihood and prior is necessary. To achieve the Bayesian equivalent of a classical logistic regression for the case of binary risk classification, the labels are sampled from a Bernoulli distribution (coin flip distribution). Thus, the likelihood term is as follows:

$$y_i \sim P(y_i|\mathbf{X}_i, \boldsymbol{\theta}) = Bernoulli(p_i) \tag{3.16}$$

$$= p_i^{y_i} \cdot (1 - p_i)^{1-y_i} \tag{3.17}$$

[2]This shall not be confused with the 95%-*confidence interval*.

In this case, the probability parameter $p_i$ is the risk probability of the $i$-th ($i \in \{0, 1, ..., n\}$) input of the feature matrix $\mathbf{X}$, denoted here as $\mathbf{X}_i$. The probability parameter $p_i$ is calculated through the sigmoid link function

$$p_i = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{X}_i)}. \tag{3.18}$$

On the other hand, the prior $P(\boldsymbol{\theta})$ is equivalent to choosing the regularisation term of logistic regression. An $\ell_1$-regularisation penalty (LASSO) is equivalent to finding the MAP by assuming a centered Laplace distribution on the $j$-th model parameter ($j \in \{0, 1, ..., d\}$) with a predefined scale parameter $b \in \mathbb{R}^+$

$$\theta_j \sim P(\theta_j) = Laplace(0, b) \tag{3.19}$$

$$= \frac{1}{2b}\exp\left( - \frac{|\theta_j|}{b} \right) \tag{3.20}$$

This can be shown as follows:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \tag{3.21}$$

$$= \arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \tag{3.22}$$

$$= \arg\max_{\boldsymbol{\theta}} \log\left( \frac{1}{Z} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}) \right) \tag{3.23}$$

$$= \arg\max_{\boldsymbol{\theta}} \log P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) + \log P(\boldsymbol{\theta}) \tag{3.24}$$

$$= \arg\max_{\boldsymbol{\theta}} \log\left( \prod_{i=1}^{n} P(y_i|\mathbf{X}, \boldsymbol{\theta}) \right) + \log\left( \prod_{j=1}^{d} P(\theta_j) \right) \tag{3.25}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log P(y_i|\mathbf{X}, \boldsymbol{\theta}) + \sum_{j=1}^{d} \log P(\theta_j) \tag{3.26}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left( p_i^{y_i} \cdot (1 - p_i)^{1 - y_i} \right) + \sum_{j=1}^{d} \log\left( \frac{1}{2b}\exp\left( - \frac{|\theta_j|}{b} \right) \right) \tag{3.27}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i) + \sum_{j=1}^{d} \log\frac{1}{2b} - \frac{1}{b}\sum_{j=1}^{d} |\theta_j| \tag{3.28}$$

$$= \arg\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i) + \frac{1}{b}\sum_{j=1}^{d} |\theta_j| \tag{3.29}$$

where the first sum is the binary cross entropy loss that minimises logistic regression. The second sum is the $\ell_1$-penalty on all the model parameters weighted by the regularisation parameter $\frac{1}{b}$. The factorisation of the joint distributions of the labels and model parameters from Equation 3.24 to Equation 3.25 comes from the independence assumption of the labels and weights.
Bhadra et al. [40] introduced the Horseshoe+ prior to inducing stronger sparsity in Bayesian generalised linear models. The Horseshoe+ prior is a hierarchical prior (using hyperpriors: priors on priors) of the following distributions:

$$\theta_j \sim P(\theta_j|\lambda_j, \tau) = \mathcal{N}(0, \lambda_j^2\tau^2) \tag{3.30}$$

$$\lambda_j \sim t^+(0, 1) \tag{3.31}$$

$$\tau \sim t^+(0, 1) \tag{3.32}$$

where $t^+$ is a half-$t$ distribution (only the positive support of the $t$ distribution) and $\mathcal{N}$ is the Gaussian distribution.

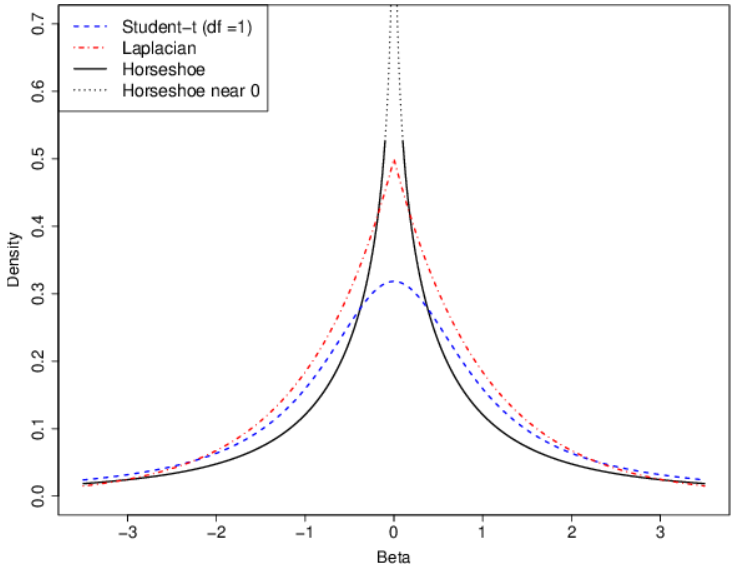



Figure 3.1: The probability density functions of the univariate Laplace, Student-t, and Horseshoe distributions are visualised in the image. The Horseshoe prior has a higher probability mass on 0 than the Laplace distribution. This induces higher sparsity in the features. The figure is taken directly from [41].

### 3.1.5 Posterior Approximation Methods

There is often no analytical solution for a posterior distribution. As these distributions are intractable and hard to sample from, they require methods to approximate them. Variational inference (VI) and Markov-Chain Monte Carlo (MCMC) are popular approximation methods.

#### 3.1.5.1 Variational Inference

Variational inference is a method that seeks to approximate an intractable distribution by a simple one, as closely as possible [42, 43]. Here, the goal is to approximate the posterior with a simpler distribution $q(\boldsymbol{\theta})$:

$$q(\boldsymbol{\theta}) \approx P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \tag{3.33}$$

It does so by finding the distribution in the variational family of Gaussian distributions that minimises the Kullback-Leibler (KL) divergence with the posterior distribution. The KL-divergence measures how one probability distribution differs from a second reference probability distribution. Mathematically, $\mathcal{Q} = \{\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,|\, \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}\}$ is the family of Gaussian distributions with mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, and the optimisation goal is:

$$q^*(\boldsymbol{\theta}) \in \arg\min_{q\in\mathcal{Q}} KL\Big(q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\Big) \tag{3.34}$$

where $q^*$ is a Gaussian distribution, with a specific mean and covariance matrix that minimises the KL-divergence $KL(\cdot||\cdot)$ with the posterior. As this is now a Gaussian, it is simple to sample from this approximated distribution directly. The VI posterior approximation method is not asymptotically exact [43].

#### 3.1.5.2 Metropolis-Hastings Sampling

Metropolis-Hastings (MH) is an MCMC method that seeks to approximate an intractable distribution (i.e. the posterior) by obtaining a sequence of random samples from that distribution by simulating a Markov

Chain. The key idea is to create a sequence of sample values that are iteratively produced, with the distribution of the following sample being dependent only on the current sample value. If the Markov Chain is simulated sufficiently long, the resulting samples are drawn from a distribution that is very close to the intractable distribution, e.g. the posterior. The Metropolis-Hastings algorithm picks at each iteration a candidate for the following sample. With some probability, the sample is either accepted (set as the following sample) or rejected (the current sample is kept for the next iteration) [44, 45]. The MH algorithm is asymptotically exact [43]. Algortihm 1 demonstrates how Metropolis-Hastings sampling works in pseudo-code:

**Algorithm 1** Metropolis-Hastings Sampling for Posterior Approximation

**Input:** Proposal distribution $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, (unnormalised) posterior $\pi(\boldsymbol{\theta}) = P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta})$
**Output:** Sequence of samples $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$
  Pick an initial state $\boldsymbol{\theta}^{(0)}$
  **for** $t \in \{0, ..., T-1\}$ **do**
    Sample a candidate $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})$
    Calculate acceptance rate $A(\boldsymbol{\theta}', \boldsymbol{\theta}^{(t)}) = \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t)})} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})}$
    **if** $A(\boldsymbol{\theta}', \boldsymbol{\theta}^{(t)}) \geq 1$ **then**
      $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}'$
    **else**
      $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)}$
    **end if**
  **end for**
  **return** $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$

## 3.2 Natural Language Processing

NLP is a branch of computer science and artificial intelligence that aims to understand and model the natural language of humans. Due to ambiguities and context-specific language, this is particularly challenging. In ACU risk prediction, the aim is to extract useful features from free text that machine learning models can process. A particular challenge is that machine learning models, in most cases, require numerical input and not strings. The most relevant approaches to process free-text data are presented in the following sections.

### 3.2.1 Text Preprocessing

The first step in most NLP tasks is to preprocess the text to extract the most useful features, as the following examples show. *Tokenisation* refers to the process of dividing a text into smaller units (i.e. tokens). These tokens can be single characters, words, numbers or $n$-grams (combinations of $n$ words). *Stopword Removal* removes predefined words from the text that are potentially useless for risk prediction (e.g. "a", "an", "the", "what" etc...). *Lemmatisation* is the process of grouping words into their base dictionary form so that they can be exploited as a single element ("are" $\rightarrow$ "to be", "is" $\rightarrow$ "to be"). This requires morphological analysis and dictionaries such as WordNet [46]. *Part-of-speech tagging* is the process of classifying words into parts of speech, such as nouns, verbs, and adjectives. Stochastic and rule-based algorithms are often exploited for this, and it is useful to remove potentially useless tags. Finally, *negation* is the process of algorithmically [30] tagging words or parts of sentences that are negated in the text. This procedure is useful to avoid missing context when tokenising the text.

### 3.2.2 Bag-of-Words

Since ML models often require numerical inputs for their predictions, assigning meaningful numerical values to the tokens extracted from the text is important. The so-called *Bag-of-Words* (BoW) approach is a simple and popular method. Each word is represented in a one-hot vector $\left\{\mathbf{x}_i \in \{0,1\}^V : \sum_{j=1}^{V} x_{i,j} = 1\right\}$ where $V \in \mathbb{N}$ is the size of the available vocabulary (i.e. the number of unique tokens) and the $i$-th entry destined for that token is 1 while the others are 0. To represent a complete text document, the one-hot vectors of the different tokens can be pooled, e.g. by summation. The vector representing the text document can be described mathematically as follows:

$$\mathbf{x}_{\text{doc}} = \sum_{i \in W} \mathbf{x_i} \tag{3.35}$$

where $W$ is the set of token indices contained in the document. In short, when pooling through summation, the vector $\mathbf{x}_{\text{doc}}$ contains the number of occurrences of each token of the vocabulary in the document. This strategy is also called *Term Frequency* (TF). To avoid common words in the vocabulary having too high values (in this thesis e.g. "patient", "doctor" or "hospital"), the Term Frequency Inverse Document Frequency (TF-IDF) algorithm [28] is applied. The value increases proportionally to the frequency of occurrence of a word in a document but is reweighted according to its occurrence in all documents.

### 3.2.3 Word Embeddings

One problem with the BoW method mentioned above is its large dimensionality. Especially for $n$-grams, the extracted features can become computationally infeasible. In an unsupervised approach, Mikolov et al. [47] proposed an efficient representation of words in vector space that can be learned directly from large text corpora such as Wikipedia, Quora, and Reddit. Instead of high-dimensional sparse vectors, words are now represented in a dense mathematical vector space $\mathbf{e} \in \mathbb{R}^q$, where $q \in \mathbb{N}$ is the dimension. These dense vectors representing tokens are called *embeddings*. Their position in space is optimised using the unsupervised learning approach. A lookup table is used to find an embedding that corresponds to a token.

### 3.2.4 Attention Layers

In recent years, neural network architectures that can handle embeddings have become increasingly popular for language modelling. To understand transformers in more detail, one needs to analyse their underlying building blocks: *self-attention* [48] layers. These map a sequence of vectors $\mathbf{x}_1, ..., \mathbf{x}_L$ (with $\mathbf{x}_i \in \mathbb{R}^q$, length $L \in \mathbb{N}$) onto a sequence of vectors $\mathbf{y}_1, ...., \mathbf{y}_L$ (with $\mathbf{y}_i \in \mathbb{R}^q$) by taking a weighted average of the input:

$$\mathbf{y}_i = \sum_{j=1}^{L} \mathbf{w}_{ij} \mathbf{x}_j \tag{3.36}$$

In this case, $\mathbf{w}_{ij}$ captures the interaction between each input vector $\mathbf{x}_i$ and $\mathbf{x}_j$. Vaswani et al. [33] proposed to quantify this interaction using the normalised inner product, i.e.

$$\mathbf{w}'_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\sqrt{q}} \tag{3.37}$$

$$\mathbf{w}_{ij} = \text{softmax}(\mathbf{w}'_{ij}) = \frac{\exp(\mathbf{w}'_{ij})}{\sum_{j=0}^{L} \exp(\mathbf{w}'_{ij})} \tag{3.38}$$

With $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{L \times q}$ as the matrices representing the stacked input and output vectors, respectively, the Equation 3.36 can be rewritten in matrix notation:

$$\mathbf{Y} = \text{softmax}\left(\frac{\mathbf{X}\mathbf{X}^\top}{\sqrt{q}}\right)\mathbf{X} \tag{3.39}$$

The above equation can be viewed as a hash function generalised to a scoring function between a query and a set of keys. To transform the inputs $\mathbf{X}$ into queries, keys and values, three linear transformation matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{q \times q}$ are multiplied with $\mathbf{X}$. The inputs to these matrices are learnable model parameters that can be fitted using the backpropagation algorithm [49]. The Equation 3.39 can therefore be extended as follows.

$$\mathbf{Y} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{q}}\right)\mathbf{V} \tag{3.40}$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{3.41}$$

This is the mathematics behind the basic self-observation layer. In practice, several of these attention heads, resulting in a *multi-head attention block*, are connected in parallel to process input and output vectors that form an attention layer in a transformer.

### 3.2.5 Transformers

A transformer model usually consists of several attention heads followed by layer normalisation [50], skip connection [51] and feed-forward networks [52]. Due to its architecture, a transformer, unlike recurrent neural networks, is independent of sequence order and requires position coding [33]. A popular transformer architecture is currently BERT, proposed by Devlin et al. [34]. It uses bidirectional self-attention blocks, where each token can observe the context to its left and right. The input tokens are created with WordPiece embeddings [53], and the first token of each sentence is always a unique classification token (CLS). The last hidden state, corresponding to the CLS token $C \in \mathbb{R}^{786}$, aggregates the sequence representation for classification tasks. BERT models are straightforward because they are trained in two steps: *pretraining* and *fine-tuning*. While pretraining for masked language modelling and next sentence prediction is done in an unsupervised way on large text corpora, fine-tuning for any downstream language task (such as classification) is helpful on smaller niche datasets.

# Chapter 4

# Experiment I: Predictive Uncertainty with Bayesian Logistic LASSO Regression

This chapter presents the methods and results for the first experiment. The aim is to evaluate the predictive performance of BLLRs compared to standard $\ell_1$-penalised logistic regression and to highlight the potential of estimating predictive uncertainty in the clinical setting, rather than over-relying on point estimates.

## 4.1 Methods

### 4.1.1 Dataset

In 2019, the CMS introduced the Chemotherapy Quality Measure (also referred to as OP-35). This quality measure monitors adult patients' inpatient admissions or emergency department visits related to potentially preventable diagnoses within 30 days of starting outpatient chemotherapy [13]. Based on this measure, Peterson et al. [1] created a study population at Stanford Hospital, including a sizeable tertiary practice, for risk prediction over 30, 180 and 365 days after chemotherapy initiation. In this experiment, we focused only on the 30-day ACU prediction, as this is the time frame suggested by CMS for assessing the quality of care. The OP-35 diagnostic codes were the supervised learning labels and defined a positive event. As in [1], we used the 760 features extracted from the EHR (Epic System Corp), such as demographic, social, vital signs, procedural, diagnostic, medication, laboratory, health care utilisation, and cancer-specific data generated prior to the first date of chemotherapy. A detailed description of how the patient cohort was extracted, the inclusion and exclusion criteria for the OP-35 metric, and a complete list of features can be found in the original paper [1]. The cohort was previously randomly divided into a training set (80%) and a test set (20%) for modelling, and we, therefore, kept exactly these sets to obtain comparable results. The resulting input feature matrix was $\mathbf{X} \in \mathbb{R}^{n \times 760}$, where $n$ is the number of patients in the corresponding dataset.

### 4.1.2 Model Development

We compared four prediction models: *Frequentist LASSO*, *Laplace-VI*, *Laplace-MH*, and *Horseshoe-MH*. All were modelled with the Bernoulli likelihood probability distribution.

#### 4.1.2.1 Frequentist LASSO

This was a traditional logistic regression with $\ell_1$-penalty on the model parameters, which we used for baseline comparison, as proposed in the original paper by Peterson et al. [1].

#### 4.1.2.2 Laplace-VI

The first Bayesian model used a Laplace prior (Distributions 3.19-3.20) with Variational Inference to approximate the posterior distribution. The Laplace prior was chosen because of the mathematical equivalence of the optimisation problem between the MAP estimate and a frequentist LASSO (Theory Chapter 3, Equation 3.21-3.29). The Laplace prior had a scaling factor $b = \frac{1}{\sqrt{2}}$ (resulting in a unit variance of the prior). We chose VI approximation because it handles high-dimensional data computationally well [54]. We used the Automatic Differentiation Variational Inference (ADVI) [54] algorithm to solve the optimisation problem in Equation 3.34 via stochastic gradient descent.

#### 4.1.2.3 Laplace-MH

The second Bayesian model also had a Laplace prior but is approximated by Metropolis-Hastings. We chose MH sampling because, compared to VI, which approximates a full distribution, it approximates the posterior by directly attempting to sample from its distribution using Markov Chains Monte-Carlo techniques. Moreover, MH sampling did not require gradient computation in the chains and was, therefore, suitable for large feature dimensions compared to other MCMC-based sampling methods [44, 45].

#### 4.1.2.4 Horseshoe-MH

Finally, the third Bayesian model had a Horseshoe+ prior [40] and is approximated by MH. We selected the Horseshoe+ prior because it has already proven successful in inducing feature sparsity in the past [41, 55, 56]. In our experiments, we used a parametrisation of the Horseshoe+ prior, proposed by Piironen and Vehtari [57], as it was more robust for sampling than the Distributions 3.30-3.32, with the following hyperpriors and priors:

$$r_i^{\text{local}} \sim \mathcal{N}(0,1) \tag{4.1}$$

$$\rho_i^{\text{local}} \sim \Gamma^{-1}(\frac{1}{2},\frac{1}{2}) \tag{4.2}$$

$$r^{\text{global}} \sim \mathcal{N}(0,1) \tag{4.3}$$

$$\rho^{\text{global}} \sim \Gamma^{-1}(\frac{1}{2},\frac{1}{2}) \tag{4.4}$$

$$z \sim \mathcal{N}(0,1) \tag{4.5}$$

$$\lambda_i = r_i^{\text{local}}\sqrt{\rho_i^{\text{local}}} \tag{4.6}$$

$$\tau = r^{\text{global}}\sqrt{\rho^{\text{global}}} \tag{4.7}$$

$$\theta_i = z\lambda_i\tau \tag{4.8}$$

where $\Gamma^{-1}$ is the inverse Gamma distribution.

### 4.1.3 Model Fitting & Hyperparameter Selection

The penalty parameter of the frequentist LASSO was determined using 10-fold cross-validation. In contrast, the posterior distributions of the Bayesian models were determined by delineating 20% of the training set for validation. Cross-validation would have been computationally infeasible. The Laplace VI model was trained with the ADVI [54] algorithm in 3,000 optimisation steps. The MH sampling models had 2,000 samples to "burn-in" and 2,000 samples to approximate the posterior distribution. All Bayesian models sampled 10,000 data points per prediction to approximate the predictive distribution.

### 4.1.4 Predictive Evaluation

#### 4.1.4.1 Discrimination

We first compared the predictive performance of the frequentist LASSO with the Bayesian models based on their discrimination and calibration. Since Bayesian models provided a predictive distribution, we used the mean (expected value) of the distribution as the final risk prediction ($\bar{y}$ in Equation 3.13). We first evaluated our models by their discriminative performance with the following discrimination metrics:
**Area under the Receiver Operator Characteristic Curve (AUROC).** The AUROC, also known as C-statistic or concordance, is a decision threshold agnostic metric, as it summarises the performance across all possible thresholds. It plots the false-positive rate against the true-positive rate (equivalent to sensitivity) over all decision thresholds and calculates the area under this plot. The AUROC indicates the probability that a patient with ACU will have a higher risk score than a patient without ACU. A random classifier achieves an AUROC of 0.5.
**Area Under the Precision-Recall Curve (AUPRC).** Another threshold agnostic metric is the AUPRC, which, as its name implies, calculates the area under the precision-recall curve over all the decision thresholds. In contrast to the AUROC, this metric is especially suited for imbalanced datasets [58]. The event rate in the dataset defines the baseline for a random classifier.
**Log-Loss.** The log-loss, also known as cross-entropy loss, summarises how close the risk predictions are to the ground truth label. Mathematically, if $y_i \in \{0, 1\}$ is the $i$th label, and $\bar{y}_i \in [0, 1]$ the $i$th risk prediction is:

$$\mathcal{L}_{CE} = \sum_{i=1}^{n} y_i \log(\bar{y}_i) + (1 - y_i) \log(1 - \bar{y}_i) \tag{4.9}$$

Like the previous two metrics, cross-entropy is not dependent on a decision threshold.

#### 4.1.4.2 Calibration

**Expected Calibration Error.** To assess the model calibration, we quantified the calibration by calculating the expected calibration error (ECE) [59, 60]. It was calculated by partitioning predictions into $M \in \mathbb{N}$ equally-spaced bins and taking the weighted average of the bins' accuracy/risk difference

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \text{acc}(B_m) - \text{risk}(B_m) \Big| \tag{4.10}$$

where $n$ is the number of all data points, and $B_m$ is the bin containing the relevant predictions.
**Flexible Calibration Curves.** We also examined their flexible calibration curves [61], as these provide information about risk over and underestimation. Calibration curves show the observed proportion of events associated with the predicted risk of a model. Ideally, the observed proportions in the validation group correspond to the predicted risks, resulting in a diagonal line in the graph. As the observed proportions per predicted risk level cannot be directly observed, we follow Austin and Steyerberg [62]'s approach to fit a flexible, non-linear calibration curve with a locally weighted running line smoother (in the literature, also known as LOESS).

#### 4.1.4.3 Empirical Confidence Intervals

To calculate the confidence intervals (CI) of the discrimination and calibration metrics, we perform 1,000-fold bootstrapping on the test dataset. For every bootstrap, 80% of the test data was randomly and independently sampled, and the metrics were calculated. This yielded 1,000 different values for a single metric. To calculate the empirical, non-parametric confidence interval, the 2.5% and 97.5% percentile were taken as lower and upper thresholds, respectively. We judged a metric to be statistically significantly better than another if its 95%-CI did not overlap another's CI.

#### 4.1.4.4 Clinical Utility

Finally, we assessed the initial clinical utility of these four models through a Decision Curve Analysis (DCA) [63]. A DCA plots the Net Benefit across a range of decision thresholds and quantifies the number of true positives penalised for false positives. The curves for the prediction models were compared to two alternative clinical strategies: treat all (everyone is treated as if they will have an ACU event) and treat none (nobody is treated as if they will not have an ACU event).

### 4.1.5 Uncertainty Evaluation

The uncertainty performance was compared across the different Bayesian models. Unless otherwise stated, we defined the uncertainty of a prediction as the standard deviation of the predictive distribution, denoted $\sigma$ from Equation 3.13, as it shows how much dispersion from the mean exists (it indicates a "typical" deviation from the mean predicted risk probability). However, our choice was based on our personal preference, and uncertainty can also be quantified in other ways, e.g. by using specific credible intervals, such as $\gamma$ in Equation 3.15..

#### 4.1.5.1 Risk and Uncertainty of Individual Predictions

To illustrate how Bayesian logistic LASSOs predictions differ at the patient level, we focused on the predictive distributions of three individual patients (low risk, medium risk, high risk) and how these can be interpreted. This use case is inspired by the experiments of Mathiszig-Lee et al. [25]. We visualised the samples from the predictive distributions as histograms (bar plots) and the risk estimates as vertical lines. With this experiment, we aimed to show how Bayesian logistic LASSOs predictions differ from each other and how they can analyse predictions for individual patients.

#### 4.1.5.2 Predicted Risk vs Uncertainty

We compared the models in terms of their uncertainties by examining the distribution of quantified uncertainties compared to the predicted ACU risk score. To do this, we created a scatter plot of the test set's risk predictions on the $x$-axis, while the corresponding uncertainties lie on the $y$-axis. Using this plot, we visually examined how the risk predictions and uncertainties correlate, depending on prior and sampling methods

#### 4.1.5.3 Uncertainty vs Classification

We designed an experiment for a Bayesian model deployed in clinical practice for treatment classification. We focused on automatic treatment classification because we argue that ML models are often deployed for clinical decision-making to make processes more efficient. By setting an arbitrary decision threshold for risk probability (i.e. $t = 0.16$, the event rate), we found that some BLLR predictions have a risk estimate below the decision threshold. However, their quantified uncertainty ranges exceeded the decision threshold and vice versa. This means that these risk predictions could likely have been classified differently. This was not possible with the frequentist LASSO, as it had no uncertainty range around its predictions. Here we introduced a new metric called *coverage*, i.e. the ratio of certain classifications to the entire test dataset. By iterating over different decision thresholds, we observed how the coverage of the models changes compared to the classification performance (F1 score, recall, precision). As decision thresholds, we selected 0.1, 0.3, and 0.5, as well as the event rate of the train set (0.16), as these were conservative estimates to avoid false negatives, which is crucial in clinical modelling. This experiment allowed us to determine which model can automatically classify the most data while achieving good classification performance. The optimal position in the plots is the top right corner, where all the data is classified (maximal coverage), and the classification metrics are maximised.

We repeated this experiment by focusing on the predictive distributions of a single model (Horshoe-MH) and how the choice of uncertainty definition influences coverage and classification performance. Because uncertainty is not clearly defined, it is up to the modeller to quantify it from the predictive distribution by using statistical tools indicating variation and probability of the predictions. We inspected the behaviour with uncertainty defined as $\sigma$, $2\sigma$, 95%-credible interval, and 99%-credible interval.

### 4.1.6 Variable Distribution

An additional advantage of BLLRs over frequentist logistic LASSO is that the posterior is a distribution for each weight, not just a single point estimate. This means that in addition to determining the uncertainty of the risk prediction for a single patient, we can also quantify the uncertainty of different parameters. In our experiments, we examined the posterior distributions of the Horseshoe-MH and their 95%-credible intervals. We pointed out the most credibly positively or negatively correlated features by filtering out posterior distributions whose upper or lower boundary of the 95%-credible interval value did not exceed the zero thresholds. Our motivation behind this was to illustrate how the posterior distribution helps examine the most credible parameters to potentially support the choice for parameter reduction. With these credible intervals, we could say that a feature is with 95% probability of either being positively or negatively correlated with the label. We displayed the median in the density plots, as well as the coefficient weights of the Frequentist LASSO for comparison. In this plot, we omitted features representing diagnoses that, at the time of the thesis, had non-identifiable ICD-9, ICD-10 (international classification of diseases).

### 4.1.7 Evaluation of Disparities in the Predictive Uncertainty

Since Peterson et al. [1] has reported unfair algorithmic results for ACU prediction, we investigated whether uncertainty estimations could be affected too. We analysed the differences in the uncertainty of the predictions by dividing the test patient cohort into different groups and plotting the uncertainty distributions of a single model (Horshoe-MH) against each other. We examined demographic values (i.e. race, ethnicity, insurance type) and tumour characteristics (i.e. cancer type and stage) by plotting the box and whisker plot [64] the estimated uncertainties of the subgroups. We compared their medians with the Kruskal-Wallis [65] test to examine if these are significantly different from each other.

## 4.2 Results

The study cohort included 8,439 patients, of whom the mean age at the start of chemotherapy was 60.4 ($\pm$ 14.5), and 50.4% were female. A total of 1,306 patients (15.5%) met the primary criteria of having at least one OP-35 event within the first 30 days of starting chemotherapy. The majority of patients in the cohort were White (n=4,630; 54.9%), followed by Asian patients (n=1,897; 22.5%), and the least represented were Black patients (n=233; 2.8%). The most common cancer types were breast (n=1,383; 16.4%), lymphoma (n=1,175; 13.9%), and Pancreas (n=980; 11.6%), making up approximately a third of all the data. ACU events were most prevalent for lymphoma tumour (n=364; 26.5%) and least for prostate cancer (n=12; 0.9%). Most chemotherapy patients had a stage IV tumour (n=2,318; 27.5%). The most common insurance type in the cohort was Medicare (n=3,236; 38.3%) and private health insurance (n=3,049; 36.1%). Cohort characteristics are summarised in Table 4.1.

### 4.2.1 Discriminative Performance and Calibration

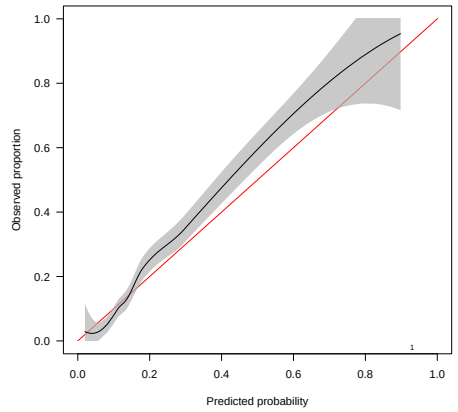
Table 4.2 lists the AUROC, AUPRC, log-loss and ECE values, including the bootstrapped 95%-confidence intervals of the three Bayesian models compared to the frequentist LASSO. The Horseshoe-MH model performed best on the AUROC (0.807, 95% CI: 0.793 - 0.821), while the frequentist LASSO had the best

| **Patient Characteristic** | **Total Cohort (N = 8,439)** | **Patients With OP-35 Events (n = 1,306, 15.5%)** | **Patients Without OP-35 Events (n = 7,133, 84.5% )** |
|---|---|---|---|
| **Age, mean ± sd** | | | |
| At diagnosis | 58.7 ± 14.4 | 56.23 ± 15.8 | 59.1 ± 14.1 |
| At first chemotherapy | 60.4 ± 14.5 | 57.9 ± 15.8 | 60.8 ± 14.2 |
| **Sex, No. (%)** | | | |
| Female | 4,250 (50.4) | 619 (47.4) | 3631 (50.9) |
| **Race, No. (%)** | | | |
| White | 4,630 (54.9) | 653 (50.0) | 3,977 (55.8) |
| Asian | 1,897 (22.5) | 299 (22.9) | 1,598 (22.4) |
| Black | 233 (2.8) | 51 (3.9) | 182 (2.6) |
| Other or Unknown | 1,679 (19.9) | 303 (23.2) | 1,376 (19.3) |
| **Ethnicity, No. (%)** | | | |
| Non-Hispanic or non-Latino | 7,231 (85.7) | 1,091 (83.5) | 6,140 (86.1) |
| Hispanic or Latino | 1,094 (13.0) | 208 (15.9) | 886 (12.4) |
| N/A | N/A | N/A | N/A |
| **Cancer type, No. (%)** | | | |
| Breast | 1,383 (16.4) | 125 (9.6) | 1,258 (17.6) |
| Lymphoma | 1,175 (13.9) | 346 (26.5) | 829 (11.6) |
| Pancreas | 980 (11.6) | 141 (10.8) | 839 (11.8) |
| Gastrointestinal | 949 (11.2) | 121 (9.3) | 828 (11.6) |
| Thoracic | 825 (9.8) | 127 (9.7) | 698 (9.8) |
| Genitourinary | 596 (7.1) | 99 (7.6) | 497 (7.0) |
| Head and neck | 697 (8.3) | 100 (7.7) | 597 (8.4) |
| Prostate | 569 (6.7) | 12 (0.9) | 557 (7.8) |
| Gynecologic | 562 (6.7) | 80 (6.1) | 482 (6.8) |
| Other | 703 (8.3) | 155 (11.9) | 548 (7.6) |
| **Cancer stage, No. (%)** | | | |
| Stage I | 1,432 (17.0) | 177 (13.6) | 1,255 (17.6) |
| Stage II | 1,679 (19.9) | 175 (13.4) | 1,504 (21.1) |
| Stage III | 1,168 (13.8) | 192 (14.7) | 976 (13.7) |
| Stage IV | 2,318 (27.5) | 486 (37.2) | 1,832 (25.7) |
| Unknown | 1,842 (21.8) | 276 (21.1) | 1,566 (22.0) |
| **Insurance, No. (%)** | | | |
| Medicare | 3,236 (38.3) | 429 (32.8) | 2,807 (39.4) |
| Private | 3,049 (36.1) | 512 (39.2) | 2,537 (35.6) |
| Medicaid | 719 (8.5) | 170 (13.0) | 549 (7.7) |
| Other or Unknown | 1,435 (17.0) | 195 (14.9) | 1,240 (17.4) |

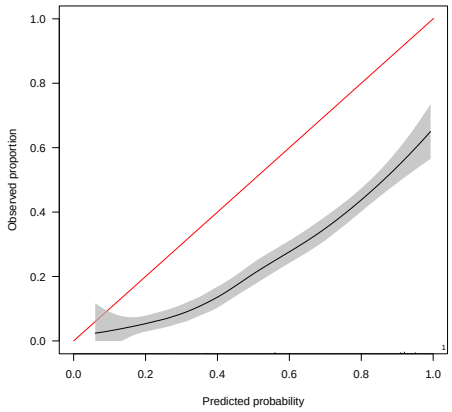
Table 4.1: Information about the complete patient cohort eligible for the OP-35 metric for 30-day prediction. "No." stand for number and "sd" for standard deviation. "N/A" was too small to report due to patient privacy concerns.

| **Model** | **AUROC** | **AUPRC** | **Log-Loss** | **ECE** |
|---|---|---|---|---|
| Frequentist LASSO | 0.806 | **0.511** | 0.357 | 0.045 |
| $b = 0.03$ | (0.792, 0.820) | (0.477, 0.543) | (0.344, 0.370) | (0.031, 0.058) |
| Laplace-VI | 0.774 | 0.437 | 0.539 | 0.242 |
| | (0.757, 0.789) | (0.406, 0.471) | (0.526, 0.551) | (0.233, 0.253) |
| Laplace-MH | 0.769 | 0.452 | 0.38 | 0.032 |
| | (0.754, 0.785) | (0.420, 0.484) | (0.363, 0.396) | (<0.001, 0.042) |
| Horseshoe-MH | **0.807** | 0.498 | **0.355** | **0.006** |
| | (0.793, 0.821) | (0.466, 0.528) | (0.340, 0.368) | (<0.001, 0.030) |

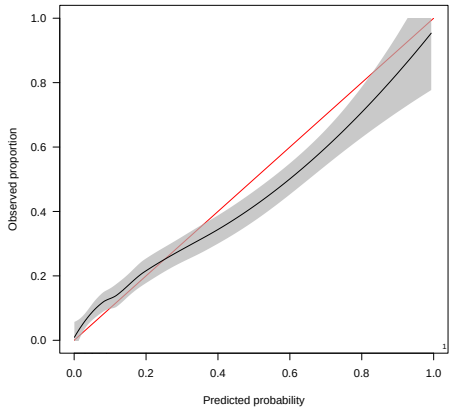
Table 4.2: Resulting metrics on the test set of the frequentist LASSO and the BLLRs. We report the 95%-confidence intervals of the metric estimates that have been calculated with 1,000-fold bootstrap in the brackets: (2.5%-CI, 97.5%-CI). The best-performing metrics for every label type per metric are marked in bold. The inverse regularisation parameter for the LASSO is denoted as $b$.



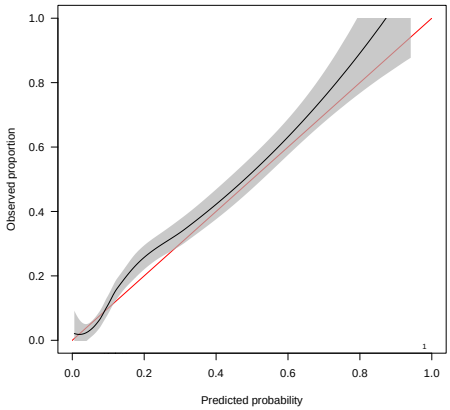

(a) Frequentist LASSO




(b) Laplace-VI




(c) Laplace-MH




(d) Horseshoe-MH

Figure 4.1: Calibration curves of the frequentist logistic LASSO and the three BLLR models. The red line indicates an ideally calibrated model. The red line indicates ideal calibration, while the black line is the flexible calibration with the 95%-confidence interval [61].

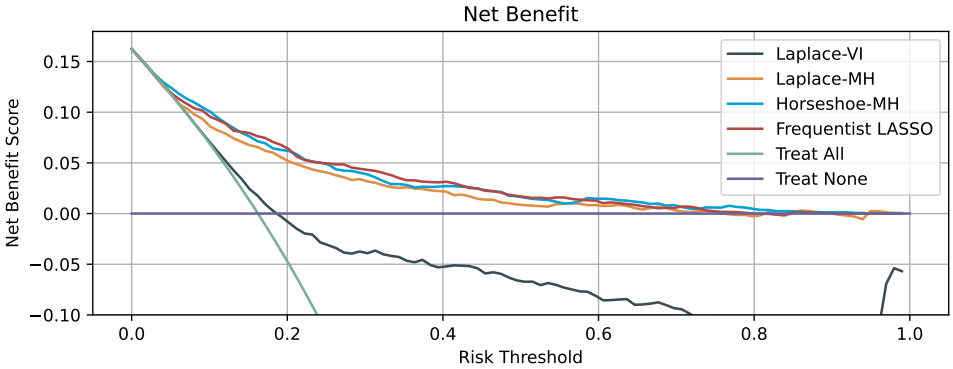



Figure 4.2: Net benefit curves of the frequentist logistic LASSO and the three BLLR models. The green curve indicates the benefit of all the patients treated. The purple curve indicates the benefit if no patient is treated.

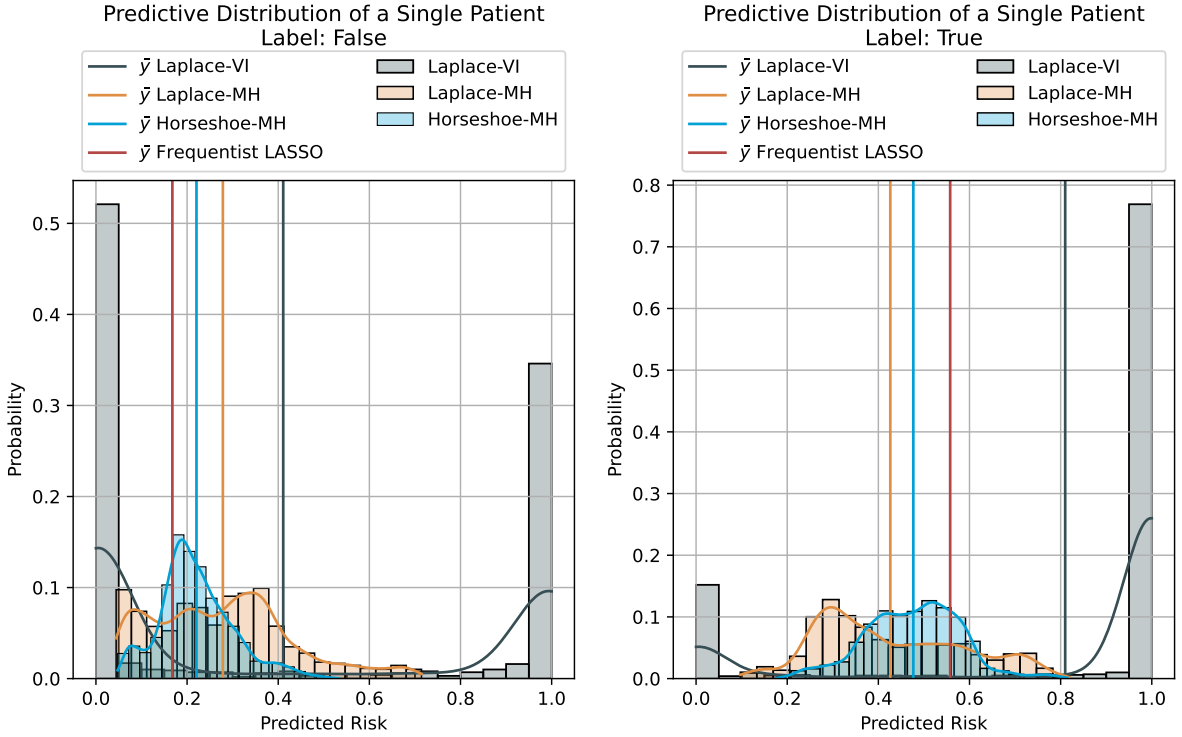
AUPRC value (0.511, 0.95% CI: 0.477 - 0.543), compared to the test event rate of 0.16. The Horseshoe-MH model also achieved the lowest negative log-likelihood (0.355, 95%-CI: 0.34 - 0.368) and had the lowest expected calibration error (0.006, 95%-CI: <0.001 - 0.03).
Calibration curves demonstrated that the Laplace-VI model often overestimates the ACU risk (Figure 4.1). In addition, the Decision Curve Analysis showed that the Net Benefit of the Laplace-VI model was consistently lower than the other models (Figure 4.2). It had a negative Net Benefit value if a decision threshold over 0.19 was chosen, while the other models persistently had positive Net Benefit scores for decision thresholds under 0.7.

### 4.2.2 Uncertainty Prediction for Individual Patients

Figure 4.3 shows the distribution and the expected value of the prediction models for ACU for three individual patients. The frequentist LASSO had no predictive distribution, only a single-point estimate of the probability. The predictive distributions of the Laplace-VI model had most of their predictions either at high values near 1.0 or low values near 0.0 probability values. In comparison, the predictive distributions of the models sampled with Metropolis-Hastings only spanned over a limited range of probabilities. In all three cases, the predictive distribution of the Laplace-MH model had a more extensive spread than the Horseshoe-MH model.

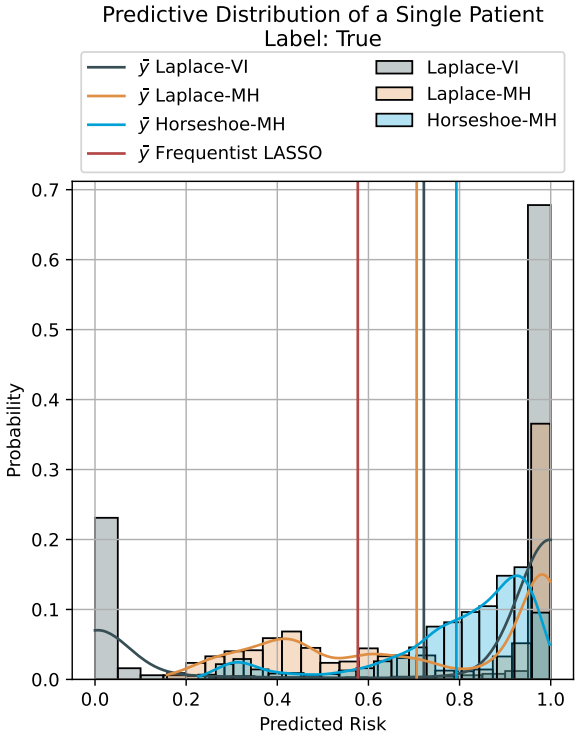
### 4.2.3 Uncertainty Prediction across Cohort

Figure 4.4 displays predicted risks of the test data (on the $x$-axis) and the corresponding uncertainty (on the $y$-axis). The quantified uncertainty of all Bayesian models was most prominent when the probability of ACU was $\sim 0.5$. Moreover, the Laplace-VI points follows an elliptical structure, while Laplace-MH and Horseshoe-MH have a more sparsely structured representation. For a given risk probability, the uncertainty of the Laplace-VI predictions was, in almost all cases, higher than the uncertainties of the MH samples.
The expected risk predictions ($\bar{y}$) of the Horseshoe-MH model were sorted from lowest to highest risk and presented in Figure 4.5. In addition, the predictive uncertainty is shown around the risk predictions. The arbitrary decision threshold for automatic treatment classification was set at the event rate ($t = 0.16$), meaning that patients with risk predictions above $t$ are allocated for treatment and vice-versa. The coverage for this particular use case was 0.72, which means that 28% of the patients were considered too uncertain to be automatically classified for treatment as the predictive uncertainty crossing the threshold. We repeated this exercise for the case where we define uncertainty as the 95%-credible interval (Appendix Figure A.1).
Figure 4.6 plots the coverage score of the four models at different decision thresholds and uncertainty against the F1 score, the sensitivity score (recall) and the positive predictive value score (PPV/precision). Frequentist LASSO always had coverage of 1.0 because each prediction was a point estimate; thus, the quantified uncertainty cannot lie on either side of the decision threshold. The Laplace-VI model had the highest F1 and recall scores over the thresholds but could confidently predict at most 20.9% of the data at $t = 0.1$. The Horseshoe-MH had a higher F1 score and recall than the Laplace-MH and frequentist LASSO models at $t \in \{0.1, 0.16\}$ and a higher precision for $t \in \{0.16, 0.3, 0.5\}$. The results with uncertainty defined as the 95%-credible interval are presented in supplement (Appendix Figure A.2).
When analysing the different quantified uncertainties of the Horseshoe-MH predictions, $\sigma$-uncertainty had the highest coverage, followed by a 95%-credible interval, then $2\sigma$ and finally 99%-credible $\forall t$. The coverage ranged from 0.28 ($t = 0.1$, 99%-credible interval) to 0.93 ($t = 0.5$, $\sigma$) (Figure 4.7). For F1 score and sensitivity, the $\sigma$-quantified uncertainty had combined values across all thresholds closest to the optimum in the upper right-hand corner.




(a) Low-risk Prediction

(b) Mid-risk Prediction




(c) High-risk Prediction

Figure 4.3: Predictive distributions of risk of acute care within 30 days after the start of chemotherapy for three individual patients with beliefs of risk: low-risk (a), mid-risk (b), high-risk (c). The histograms indicate the predictive distributions, while the lines in the respective colours are the distributions' expected values ($\bar{y}$).

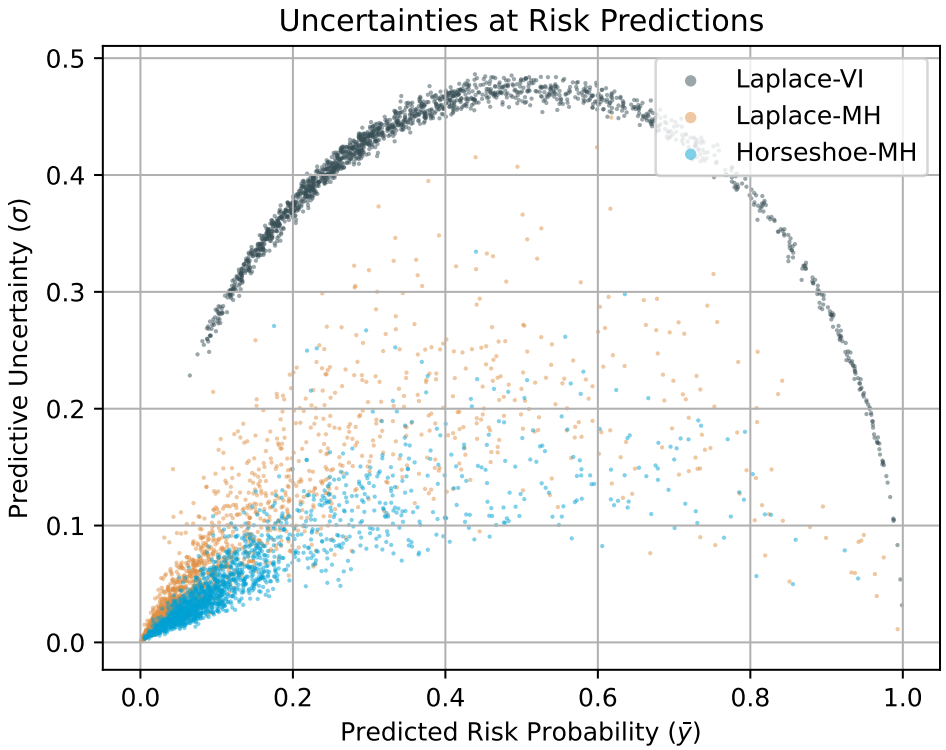



Figure 4.4: Scatter plot for the final risk prediction in the test set made by the three Bayesian logistic LASSOs, with the expected value of the predictive distribution ($\bar{y}$) on the $x$-axis, and the standard deviation of the risk prediction ($\sigma$) on the $y$-axis.

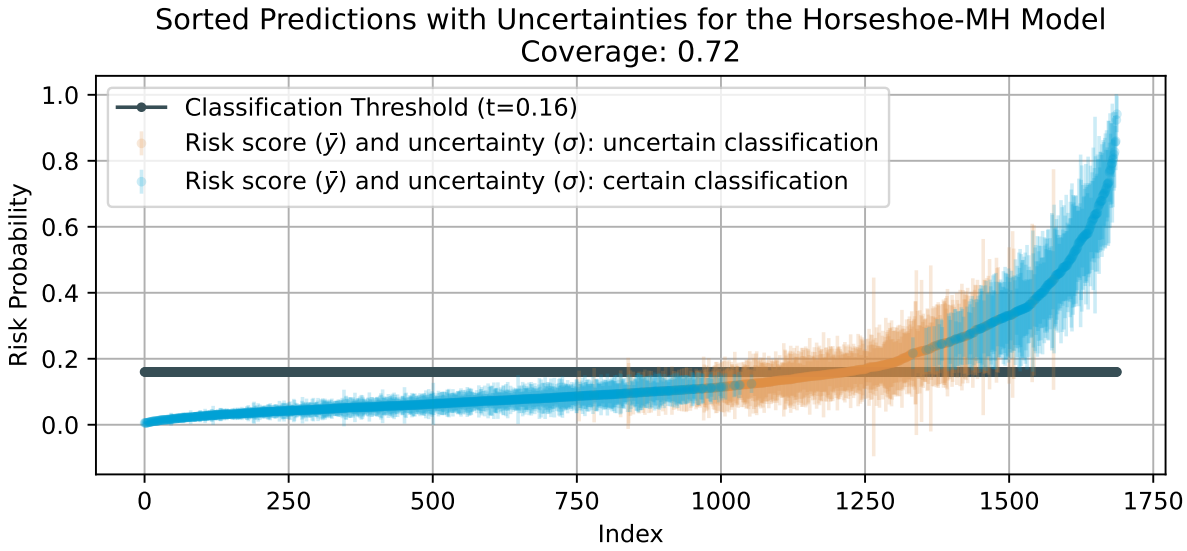
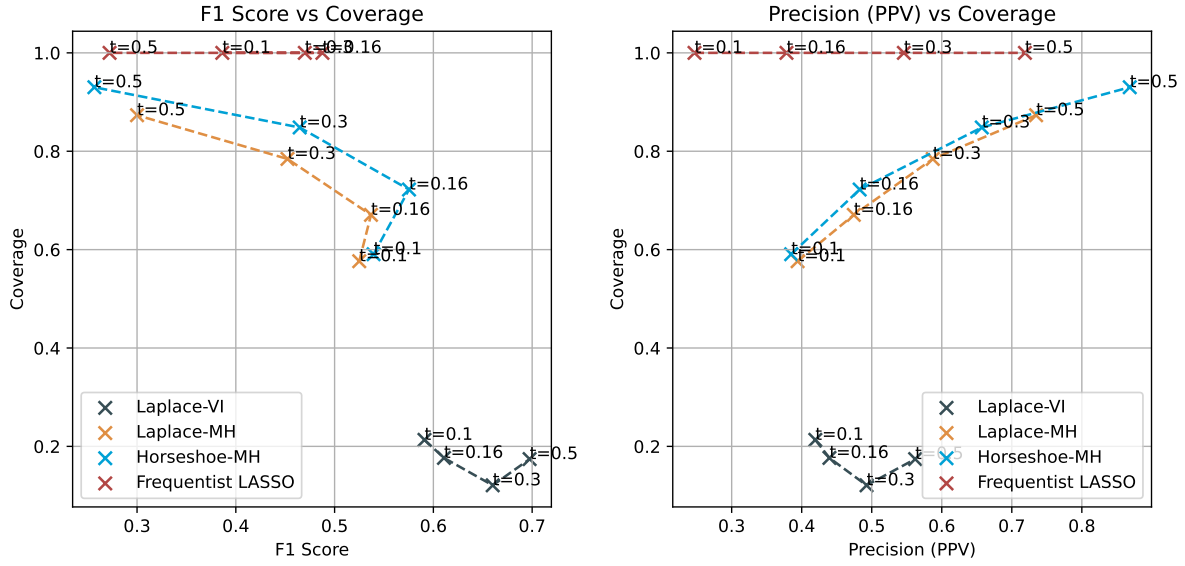



Figure 4.5: Sorted final risk predictions (mean of the predictive distribution, $\bar{y}$) with uncertainty range ($\pm\sigma$) for the Horseshoe-MH model. The predictions whose uncertainty does not exceed the decision threshold (certain classifications) are coloured blue, and those that do (uncertain classifications) are coloured orange. The dark grey line is our chosen classification threshold at 0.16, the event rate.

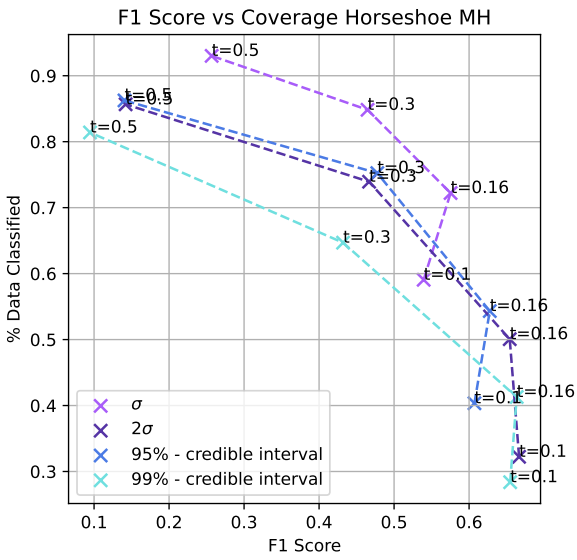



(a) F1 Score

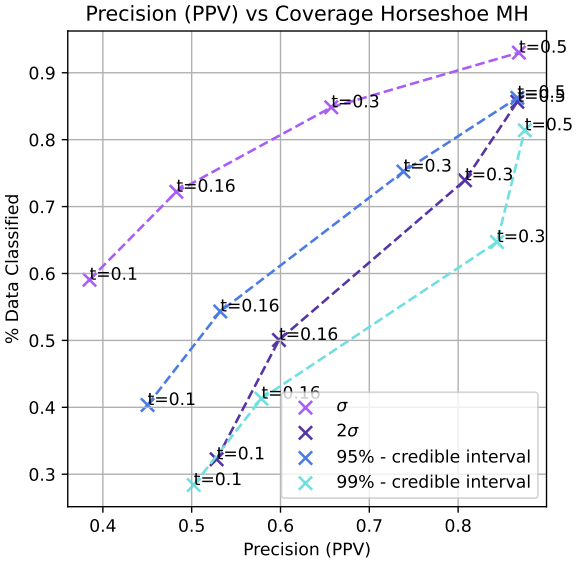
(b) Precision / PPV

(c) Recall / Sensitivity

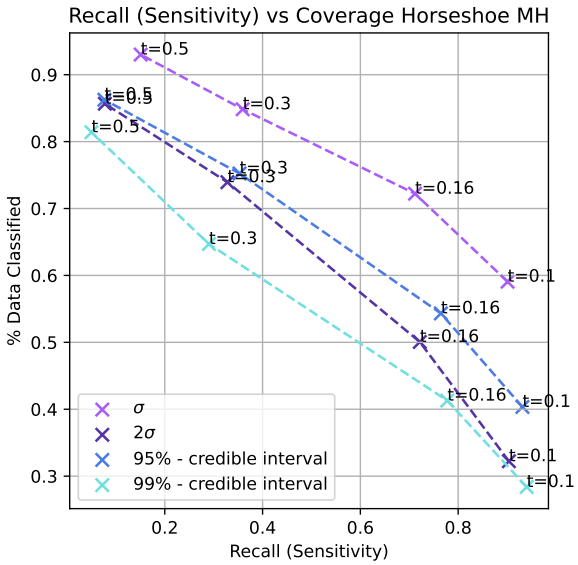
Figure 4.6: Coverage (ratio of automatically classified predictions) compared to F1-score (a), sensitivity (b), and PPV (c), over risk decision thresholds set at 0.1, 0.16, 0.3, 0.5. The uncertainty is defined as the predictive distribution's standard deviation ($\sigma$).




(a) F1 Score




(b) Precision / PPV




(c) Recall / Sensitivity

Figure 4.7: Coverage of the Horseshoe-MH model compared to its F1-score (a), sensitivity (b), and PPV (c), over the risk decision thresholds set at 0.1, 0.16, 0.3 0.5. The predictive uncertainties are defined as $\sigma$, $2\sigma$, 95%-credible interval, and 99%-credible interval.

### 4.2.4 Posterior Distribution of Variables

In Figure 4.8, we display the posterior distributions of credible features for the Horseshoe-MH model. We see that Sarcoma cancer (median=0.086), Non-palliative patients (median=0.14), and previous hospitalisation days ("Hosp N", median=0.25) had their credible intervals clearly in the spectrum of positive correlation. On the other hand, the Albumin laboratory values ("LABS: ALB") seem to be have been negatively correlated (median=-0.25) with the predictions credibly. We observed that the Frequentist LASSO coefficients were within the credible intervals in 14 of the 19 cases. Eight features with unidentifiable IDC-9/-10 codes were omitted from the figure for readability.

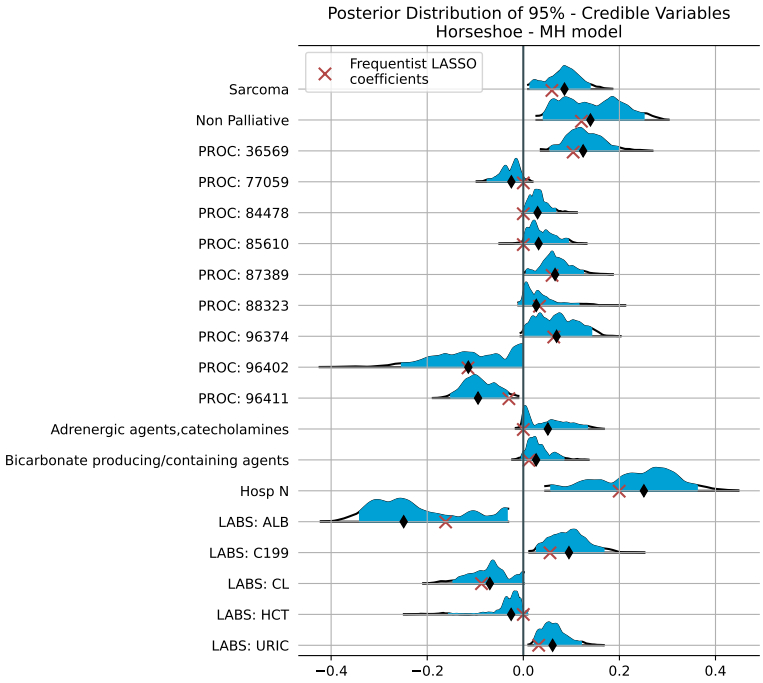
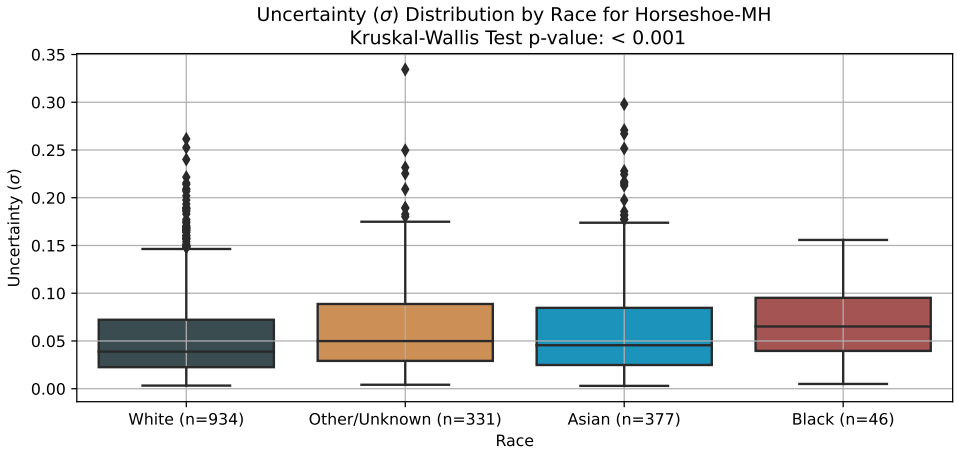


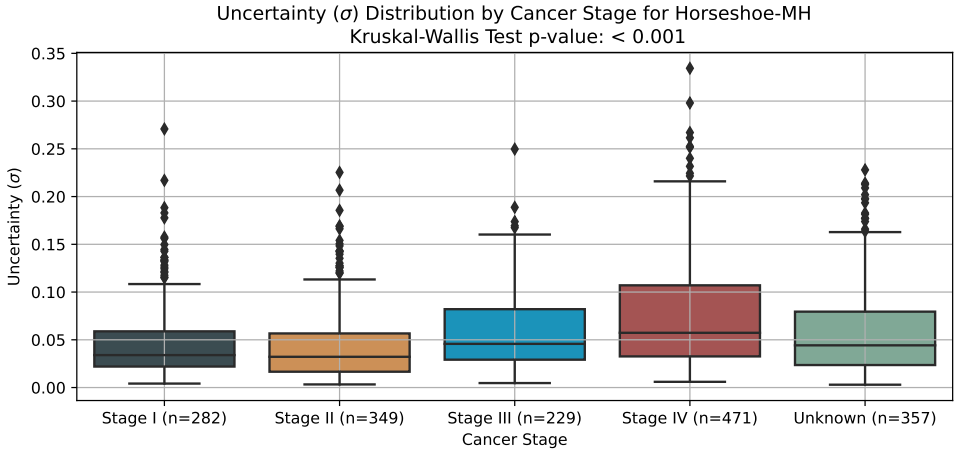

Figure 4.8: Posterior distribution of certain features, which have 5% probability of being sampled with a different sign. The black diamonds indicate the posterior median, while the red crosses are the Frequentist LASSO coefficients for comparison. A table with the description of the variables can be found in Appendix Table A.2

### 4.2.5 Sensitivity Analysis

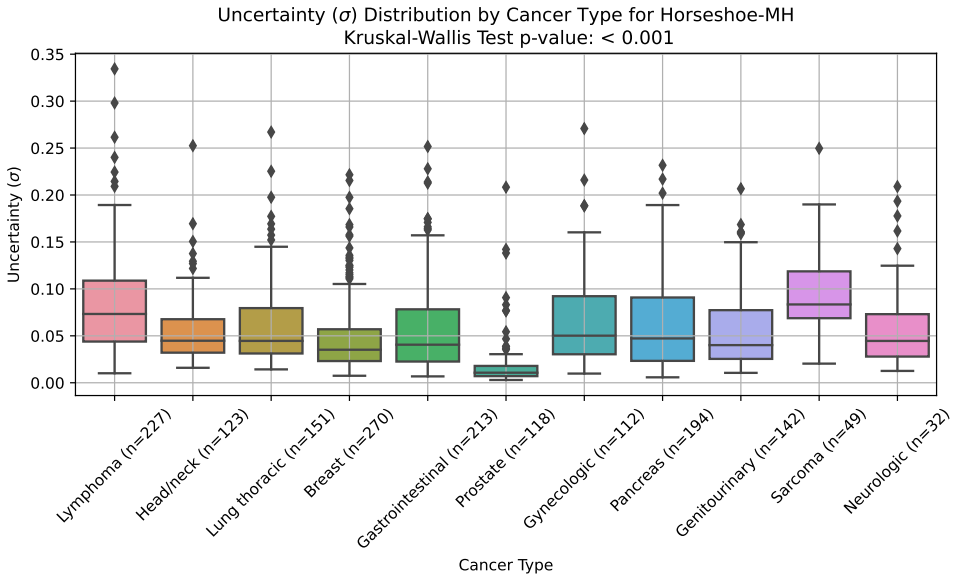
The distribution of quantified uncertainty ($\sigma$) of the Horseshoe-MH model by patient race demonstrates the median of the predictive uncertainties for Black patients (median=0.065) than for White patients (median=0.039), Asian patients (median=0.046) and other races (median=0.05) are significantly different (Kruskal-Wallis test: $p < 0.001$) (Figure 4.9a). This means that for the median Black patient, the typical error of its estimated risk is $\pm 6.5\%$, while for the median White patient, it is $\pm 3.9\%$.
Similarly, we examined the predictive uncertainties between the different cancer stages in Figure 4.9b and saw that their medians differ significantly (Kruskal-Wallis test: $p < 0.001$). We observe that stage IV cancer patients had the highest predictive uncertainty (median = 0.057) compared to the other stages (median < 0.05). In terms of tumour types, prostate tumour patients had the lowest median (0.011) of uncertainty, while sarcoma had the highest (0.083) (Figure 4.9c). Therefore, the median prostate tumour patient had a seven times smaller typical prediction error than the median sarcoma cancer patient. In supplement, we report the uncertainty distribution also for patients' gender, ethnicity, insurance status, and tumour type (Appendix Figure A.3).




(a) Race




(b) Cancer Stage




(c) Cancer Type

Figure 4.9: Distribution of $\sigma$-quantified uncertainty of the test set, stratified by race (a) and cancer stage (b), and cancer type (c). The Kruskal-Wallis statistic and significance can be found in the titles.

# Chapter 5

# Experiment II: Natural Language Processing to Predict ACU

This chapter demonstrates the methods and results for the second experiment. The goal was to analyse how NLP predicts the risk of ACU. Free-text notes were explored for ACU prediction in lieu of SHD or in combination. Additionally deep learning models were compared with manually engineered language features.

## 5.1 Methods

### 5.1.1 Dataset

Same as in Chapter 4, we based our initial patient population on Peterson et al. [1]'s work for ACU risk prediction at 30, 180 and 365 days after chemotherapy initiation. The OP-35 diagnostic codes were the supervised learning label and defined a positive event. In this experiment, we focused on all three label types after the start of chemotherapy.
For the SHD inputs, we used the original 760 features from Peterson et al. [1] extracted from the same EHR database, such as demographic, vital sign, procedural, and diagnostic data. It was generated 180 days before the first date of chemotherapy, resulting in a feature matrix $\mathbf{X}_{\text{tab}} \in \mathbb{R}^{n \times 760}$. For a detailed description of how the patient cohort was extracted, the inclusion and exclusion criteria for the OP-35 metric, and a full list of features, we refer again to the original paper [1].
Based on the above study population, we matched patients to their respective progress notes, and the history and physical (H&P) notes from the EHR database (Epic Systems Corp). We removed notes of less than 100 words, as these were mainly erroneous entries, and notes of more than 5,000 words often contained long copies of previous notes and laboratory analyses. We also removed history notes with mentions of clinical trial consents, as based on our review, these were copy-paste texts. Finally, we extract and aggregate the most recent clinical notes (at most three) created 180 days before the patient started chemotherapy, as in the SHD collection from Peterson et al. [1]. Patients with no clinical records in the EHR database were removed from the study population.
The cohort was previously randomly divided into a training set (80%) and a test set (20%) for modelling, and we, therefore, kept exactly these patient sets (except the ones without any clinical notes) to obtain comparable results.

### 5.1.2 Model Development

Five different risk prediction models were compared in this study: *Tabular LASSO*, *Language LASSO*, *Fusion LASSO*, *Language BERT* and *Fusion BERT*. The choice of these five models was based on previous works with medical NLP [1, 27, 32, 66].

#### 5.1.2.1 Tabular LASSO

Tabular LASSO is a logistic regression with an $\ell_1$-penalty. The inputs were the 760 structured health data points of the feature matrix $\mathbf{X}_{\text{tab}}$.

#### 5.1.2.2 Language LASSO

The language LASSO model is an $\ell_1$-penalised logistic LASSO regression with manually generated inputs from the clinical notes. The notes were preprocessed as follows: first, we removed special characters and personal, organisational, date and time entities using SpaCy's [67] part of speech tagging. Then we tagged negated terms with a "not_" using SpaCy's negator library. We removed auxiliary words, adpositions, determiners, interjections and pronouns, as we did not deem these useful for ACU prediction. Subsequently, we lemmatised [46] the remaining words. More detailed information about how these individual preprocessing procedures work can be found in the Theory section 3.2.1. Finally, we followed the method of Marafino et al. [27] by filtering out the $W \in \mathbb{N}$ most frequent terms of all the notes and weighting these words using the TF-IDF algorithm (see Theory section 3.2.2). The Language LASSO has $W$ input features corresponding to the $W$ most frequently occurring words, resulting in a language feature matrix $\mathbf{X}_{\text{nlp}} \in \mathbb{R}^{n \times W}$. In our experiments, we test the Language LASSO on $W \in \{500, 1000, 2000, 3000\}$ filtered words, based on Marafino et al. [27]'s original choice of 1,000 words. This yields four input feature matrices $\mathbf{X}_{\text{nlp,500}} \in \mathbb{R}^{n \times 500}$, $\mathbf{X}_{\text{nlp,1000}} \in \mathbb{R}^{n \times 1,000}$, $\mathbf{X}_{\text{nlp,2000}} \in \mathbb{R}^{n \times 2,000}$, and $\mathbf{X}_{\text{nlp,3000}} \in \mathbb{R}^{n \times 3,000}$ respectively.

#### 5.1.2.3 Fusion LASSO

The fusion LASSO is also a logistic regression LASSO model. This time it uses both, the tabular data and TF-IDF values, as input features $\mathbf{X}_{\text{fus}} = \left[\mathbf{X}_{\text{nlp}}^\top, \mathbf{X}_{\text{tab}}^\top\right]^\top \in \mathbb{R}^{n \times (W+760)}$. We combined these two to inspect if data extracted from the clinical notes has added value to SHD.

#### 5.1.2.4 Language BERT

The language BERT is a deep learning-based transformer [33, 34]. This model does not require manual feature engineering and can consume clinical notes with little preprocessing. As the input token sequence computationally limits transformer models, we decomposed the clinical notes into chunks of, at most, 25 sequences (to avoid GPU memory overflows), each 256 tokens. This chunks resulted in the input tensor $\mathbf{x} \in \mathbb{R}^{D \times 256}$ for one clinical note of $D$ chunks ($D \in \mathbb{N} : D \leq 25$). We aggregated the output CLS embeddings $\mathbf{e}_i \in \mathbb{R}^{786}$ of the transformers by averaging over the corresponding clinical note: $\bar{\mathbf{e}} = \frac{1}{D} \sum_{i=1}^{D} \mathbf{e}_i$. We connected $\bar{\mathbf{e}}$ linearly to one output neuron, $o = \mathbf{W}_{\text{out}} \bar{\mathbf{e}}$ with $\mathbf{W}_{\text{out}} \in \mathbb{R}^{1 \times 786}$. This single scalar was converted to the proportional odds of belonging to one of four classes, with a cumulative link layer [68]. In this layer we have three parametrised cut-off values $c_j \in \mathbb{R}, j = \{1, 2, 3\}$ that were adjusted during backpropagation. These four classes represented the probability distribution of an ACU event within the time intervals emanating from the different ground truth labels ($P(\text{ACU} \leq 30d)$,

$P(30d < \text{ACU} \leq 180d)$, $P(180d < \text{ACU} \leq 365d)$ and $P(\text{ACU} > 365d)$) and they were calculated by:

$$P(\text{ACU} > 365d) = f(c_1 - o) \tag{5.1}$$
$$P(180d < \text{ACU} \leq 365d) = f(c_2 - o) - f(c_1 - o) \tag{5.2}$$
$$P(30d < \text{ACU} \leq 180d) = f(c_3 - o) - f(c_2 - o) \tag{5.3}$$
$$P(\text{ACU} \leq 30d) = 1 - f(c_3 - o) \tag{5.4}$$

where $f$ is the sigmoid function, like in Equation 3.8. Since a patient that experienced an ACU event within the first 30 days is also eligible for an event within 180 days and 365 days, we added the corresponding probabilities. This corresponds to the original ground truth interpretation of an ACU within 30 days ($P(x \leq 30d)$), 180 days ($P(x \leq 180d)$), 365 days ($P(x \leq 365d)$) and not within 365 days ($P(x > 365d)$). Because of this nested structure, our cumulative output probabilities were calculated as follows:

$$P(\text{ACU} > 365d) = f(c_1 - o) \tag{5.5}$$
$$P(\text{ACU} \leq 365d) = 1 - f(c_1 - o) \tag{5.6}$$
$$P(\text{ACU} \leq 180d) = 1 - f(c_2 - o) \tag{5.7}$$
$$P(\text{ACU} \leq 30d) = 1 - f(c_3 - o) \tag{5.8}$$

Therefore, compared to the LASSO models, the BERT model was simultaneously trained on all ACU risk prediction times. An overview of the Language BERT model is found in Figure 5.1. We simultaneously analysed the added value of the cumulative link layer, trained on 30-day, 180-day, and 365-day ACU prediction, compared to a Language BERT trained on the three time intervals individually. The models trained on the labels individually had a single output neuron output $o_{\text{single}} \in \mathbb{R}$, that is passed through a sigmoid function (Equation 3.8) and optimised the cross-entropy loss (Equation 4.9). Furthermore, we compared three pre-trained encoding structures to determine the most appropriate. We chose distilBERT [69] because of its efficiency and network size, ClinicalBERT [32, 35] as it is pre-trained on clinical discharge notes, and LongFormer [70] encoder architecture because it processed longer token sequences (3 chunks $\times$ 1024 tokens) than the BERT models mentioned above.

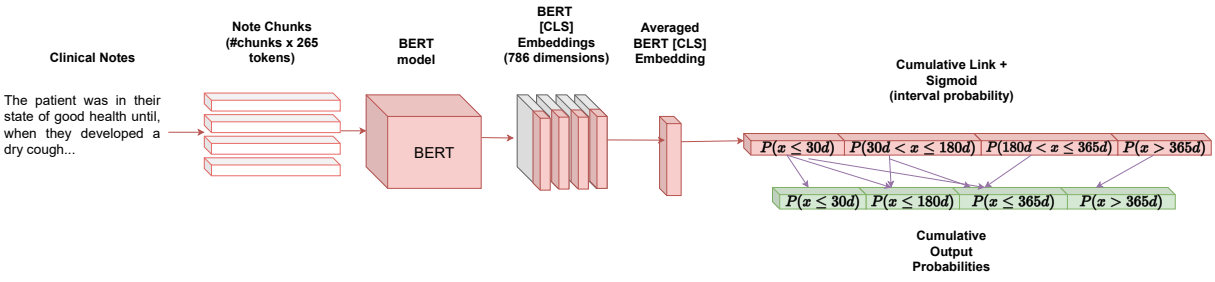



Figure 5.1: Overview of the Language BERT model

### 5.1.2.5 Fusion BERT

The fusion BERT model is the same as the language BERT model, except that the corresponding SHD were concatenated with the output embedding $\mathbf{e}_{\text{fus}} = \left[\bar{\mathbf{e}}, \mathbf{x}_{\text{tab}}\right] \in \mathbb{R}^{1,528}$. The newly-concatenated embedding was then linearly connected to the cumulative link layer. Figure 5.2a shows an overview of the fusion BERT with concatenation. In addition, this study compared concatenation fusion with the cross-modal attention mechanism, taken from Tsai et al. [71]. The idea was to have parallel attention layers:

one attended the tabular data to the language data, and vice versa. We calculated the crossmodal blocks as follows:

$$\mathbf{W}_{Q_{\text{tab}}} \in \mathbb{R}^{760\times760} \tag{5.9}$$

$$\mathbf{W}_{K_{\text{nlp}}}, \mathbf{W}_{V_{\text{nlp}}} \in \mathbb{R}^{768\times760} \tag{5.10}$$

$$\mathbf{Q}_{\text{tab}} = \mathbf{x}_{\text{tab}}\mathbf{W}_{Q_{\text{tab}}}, \quad \mathbf{K}_{\text{nlp}} = \bar{\mathbf{e}}\mathbf{W}_{K_{\text{nlp}}}, \quad \mathbf{V}_{\text{nlp}} = \bar{\mathbf{e}}\mathbf{W}_{V_{\text{nlp}}} \tag{5.11}$$

$$\mathbf{e}_{\text{tab}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{tab}}\mathbf{K}_{\text{nlp}}^{\top}}{\sqrt{760}}\right)\mathbf{V}_{\text{nlp}} \tag{5.12}$$

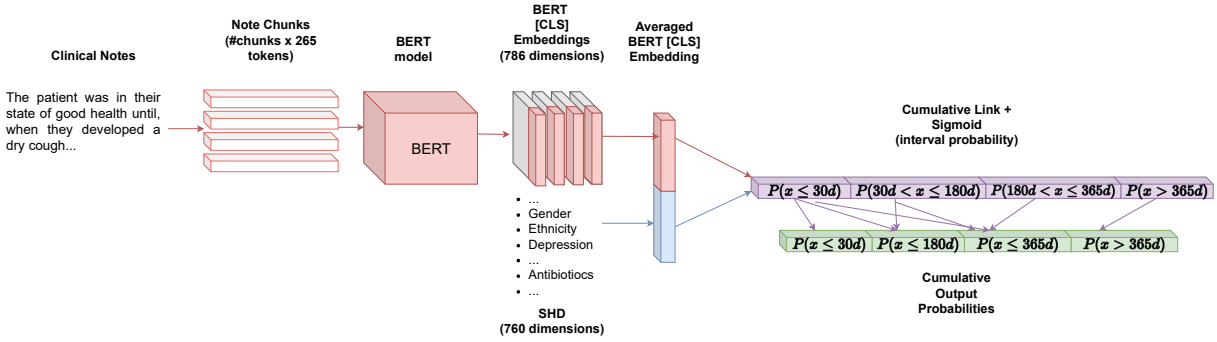
$$\mathbf{W}_{Q_{\text{nlp}}} \in \mathbb{R}^{786\times786} \tag{5.13}$$

$$\mathbf{W}_{K_{\text{tab}}}, \mathbf{W}_{V_{\text{tab}}} \in \mathbb{R}^{760\times768} \tag{5.14}$$

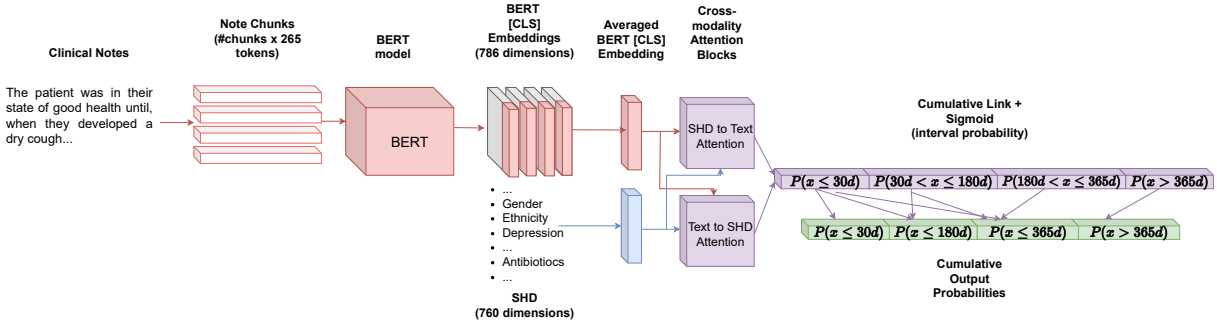
$$\mathbf{Q}_{\text{nlp}} = \bar{\mathbf{e}}\mathbf{W}_{Q_{\text{nlp}}}, \quad \mathbf{K}_{\text{tab}} = \mathbf{x}_{\text{tab}}\mathbf{W}_{K_{\text{tab}}}, \quad \mathbf{V}_{\text{tab}} = \mathbf{x}_{\text{tab}}\mathbf{W}_{V_{\text{tab}}} \tag{5.15}$$

$$\mathbf{e}_{\text{nlp}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{nlp}}\mathbf{K}_{\text{tab}}^{\top}}{\sqrt{768}}\right)\mathbf{V}_{\text{tab}} \tag{5.16}$$

Finally, the crossmodal attention outputs are concatenated $\mathbf{e}_{\text{fus}} = \left[\mathbf{e}_{\text{tab}}, \mathbf{e}_{\text{nlp}}\right] \in \mathbb{R}^{1,528}$. An overview of the Language BERT with a crossmodal attention mechanism can be found in Figure 5.2b.




(a) Concatenation




(b) Crossmodal Attention

Figure 5.2: Overview of the Fusion BERT model with concatenation (a) and crossmodal (b) fusion mechanism. The Language BERT only contains the upper (red) encoder architecture before leading into the output probabilities (purple and green).

### 5.1.3 Model Fitting & Hyperparameter Selection

A tenfold cross-validation grid search determined the regularisation hyperparameters of the LASSO models. In contrast, the hyperparameters of the two BERT models were determined by using 20% of the

training data as validation data. The LASSO models are individually trained on each time interval of the label (30d, 180d, and 365d). The Language BERT was trained on individual labels with the cross-entropy loss (Equation 4.9) and all the labels simultaneously with the cumulative link loss:

$$\mathcal{L}_{CL} = \begin{cases} -\log(f(c_1 - o)), & \text{if no ACU} \\ -\log(1 - f(c_1 - o)), & \text{if ACU within 365 days} \\ -\log(1 - f(c_2 - o)), & \text{if ACU within 180 days} \\ -\log(1 - f(c_3 - o)), & \text{if ACU within 30 days} \end{cases} \tag{5.17}$$

which was modified from the original [68] to ensure the nested ordinal regression structure[72], e.g. avoid cases where $P(\text{ACU} \leq 30d) > P(\text{ACU} \leq 180d)$. In this use case, any patient with an ACU event within 30 days was also marked to have had an ACU within 180 days. The Fusion BERT was only trained on cumulative link loss.

We used backpropagation with the ADAM [73] optimiser for ten epochs to train the transformer. The ADAM optimiser is currently a state-of-the-art optimiser and readily available in the PyTorch [74] ML framework. The ten epochs were chosen as the BERT model convergence was usually achieved already after fewer epochs. We aborted training if the validation loss did not improve for five epochs (early-stopping) and chose the model parameters at the best-performing epoch on the validation set in the hope of having the best generalisable performance. In every run through the neural network, we passed three clinical notes, equivalent to at most 75 batches ($3 \times 25$), as this was the highest number without causing a computational memory overflow on the GPU. The learning rate for the linear classifier $\mathbf{W}_{\text{out}}$ was set to $10^{-4}$, as the weights were randomly initialised. For the rest of the transformer encoder, the learning rate was $10^{-5}$, as it was already pre-trained on a large language corpus and required only fine-tuning. To increase convergence, the learning rates are reduced after five epochs by half. Additionally, we applied $10^{-3}$ weight decay during training to avoid large values of the model parameters.

### 5.1.4 Model Evaluation

To analyse the models' predictive performance, we used the previous experiment's methods, described in detail in section 4.1.4. We first evaluated the models with AUROC, the AUPRC, and the log-loss with a 1,000-fold bootstrap to obtain 95%-confidence intervals. To assess the model calibration, flexible calibration curves [61] were developed and the ECE (Equation 4.10) was calculated.

We reported the number of SHD used during risk prediction. We did this for the LASSO models by summing the number of non-zero model coefficients originating from SHD. For the Fusion BERT, we counted the number of connections of tabular features to the output neuron with values less than 0.001, as the backpropagation algorithm is not optimised for feature selection, unlike the LASSO.

Finally, we assessed the initial clinical utility of these four models through DCA [63]. To test the discriminatory power of the model in a setting similar to that in which it might be deployed at the point of care, the test cohort was stratified into high, medium and low-risk groups based on the tertile of predicted risk. Kaplan-Meier [75] survival curves for OP-35 events examined the separation between risk groups for language LASSO and language BERT on 180-day ACU risk prediction. In addition, the ten highest and lowest coefficients of the language LASSO model are presented. It helped us determine specific keywords' importance in the clinical notes.

Since Peterson et al. [1] have reported unfair algorithmic results for ACU prediction from structured data, we investigated whether language features might also be affected. We compared the empirical cumulative distributions of predicted risk score percentiles for subgroups to assess how the models predicted each subgroup's risk for OP-35 events. Specifically, we examined demographic values (i.e. race, ethnicity and insurance type) and tumour type and stage on the Language LASSO model for 180-day ACU risk prediction.

## 5.2 Results

This new study cohort included 6,938 patients with clinical notes, while the original included 8,439 patients. The mean age at chemotherapy initiation was 60.5 years ($\pm$ 14.4 years), and 52.7% were female. A total of 936 patients (13.5%) met the primary criteria of having at least one OP-35 event within the first 30 days of starting chemotherapy, 2,202 (31.7%) within the first 180 days and 2,704 (39.0%) within the first year. The majority of patients in the cohort were White (n=3,804; 54.8%), followed by Asian patients (n=1,619; 23.3%), then other and unknown races (1,327; 19.1%) and least represented were Black patients (n=188; 2.7%). The most common cancer type was breast cancer (n=1,321; 19.0%), gastrointestinal tumours (n=819; 11.8%), thoracic cancer (n=774; 11.2%) and Lymphoma (n=700; 10.1%), which accounted for more than half of all data. ACU events occurred most frequently in Lymphoma (30d: n=170, 18.2%; 180d: n=345, 15.7%; 365d: n=382, 14.1%) and least frequently in prostate cancer (30d: n=11, 1.2%; 180d: n=46, 2.1%; 365d: n=70, 2.6%) across all time periods. Most chemotherapy patients had a stage IV tumour (n=1,898; 27.4%), which was also most common in ACU events (30d: n=327, 34.9%; 180d: n=759, 34.5%; 365d n=937, 34.7%). The cohort's most common type of insurance was Medicare (n=2,863; 38.7%) and private health insurance (n=2,450; 35.3%). The cohort characteristics are summarised in Table 5.1.

### 5.2.1 Model Performance

In terms of vocabulary size $W$ for the Language LASSO, we found that the model trained on 3,000-word features ($\mathbf{X}_{\text{nlp,3000}}$) had the best-performing metrics, except for AUROC and log-loss for the 30-day prediction. In all cases, the 95% confidence intervals of the 2,000-word and 3,000-word models overlapped, with AUROC values always above 0.7. In our main results Table 5.2 we reported the Language LASSO with $W = 2,000$ because of its statistically insignificant inferior performance and the lower amount of model parameters. Further details can be found in Appendix Table B.1.

Comparing the cumulative link layer and single-label output for the Language BERT, our results showed that the ordinal regression model outperforms the individually trained models in terms of AUROC and AUPRC in all three time intervals. The single-label models achieved lower log-loss values (Appendix Table B.2). In the main results table we thus reported the Language BERT trained on all the time intervals.

The effect of the encoder for the Language BERT model showed that the ClinicalBERT and the distilBERT did not differ significantly in terms of AUROC. Both achieved higher values for AUROC and AUPRC than the LongFormer encoder. We chose to report the Language BERT with the distilBERT encoder in the main result table, because of its predictive performance, and computational efficiency. The detailed results can be found in the supplementary materials in Table B.3.

Comparing the concatenation and cross-modal attentional fusion mechanisms of the Fusion BERT, none of the fusion modes performed significantly better than the other (Appendix Table B.4). For simplicity, we therefore included the Fusion BERT with concatenation in our main result Table.

Finally, the main results Table 5.2 lists the AUROC, AUPRC, and log-loss scores, including the 95% confidence intervals of the five risk models for 30-day, 180-day and 365-day ACU prediction. For the 30-day acute care risk prediction, the Fusion LASSO model performed best on AUROC (0.778, 95%-CI: 0.760, 0.795) and log-loss (0.341, 95%-CI: 0.326, 0.356), using 73 SHD features. The highest AUPRC score had the Tabular LASSO (0.411, 95%-CI: 0.373, 0.447) compared to the event rate of 13.5%, using 83 tabular variables.

For 180-day ACU prediction, the Fusion LASSO model performed best in all metrics with 101 SHD features. The Language LASSO had a 0.730 (95%-CI: 0.717, 0.745) AUROC score, and the Language BERT achieved 0.702 (95%-CI: 0.688, 0.717), both of them without using any structured data.

In the full-year ACU prediction, we observed that the Fusion LASSO scores again had the highest C-statistic (0.770, 95%-CI:0.759, 0.782) and the lowest log-loss loss (0.553, 95%-CI:0.541, 0.563), using

| Patient Characteristic | Total Cohort (N=6,938) | OP-35 Events within 30 days (n=936, 13.5%) | OP-35 Events within 180 days (n=2,202, 31.7%) | OP-35 Events within 365 days (n=2,704, 39.0%) | No OP-35 Events within 365 days (n=4,234, 61.0%) |
|---|---|---|---|---|---|
| **Age, mean± sd** | | | | | |
| At diagnosis | 58.7±14.3 | 57.2±15.3 | 57.7±15.1 | 57.9±15.0 | 59.2±13.9 |
| At first chemotherapy | 60.5±14.4 | 58.9±15.2 | 59.4±15.1 | 59.6±15.0 | 61.0±14.0 |
| **Sex, No. (%)** | | | | | |
| Female | 3,659 (52.7) | 474 (50.6) | 1132 (51.4) | 1417 (52.4) | 2242 (53.0) |
| **Race, No. (%)** | | | | | |
| White | 3,804 (54.8) | 461 (49.3) | 1,113 (50.5) | 1,379 (51.0) | 2,425 (57.3) |
| Asian | 1,619 (23.3) | 226 (24.1) | 536 (24.3) | 649 (24.0) | 970 (22.9) |
| Black | 188 (2.7) | 42 (4.5) | 88 (4.0) | 100 (3.7) | 88 (2.1) |
| Other or unknown | 1,327 (19.1) | 207 (22.1) | 465 (21.1) | 576 (21.3) | 751 (17.7) |
| **Ethnicity, No. (%)** | | | | | |
| Non Hispanic/Latino | 5,989 (86.3) | 788 (84.2) | 1,867 (84.8) | 2,280 (84.3) | 3,709 (87.6) |
| Hispanic or Latino | 855 (12.3) | 142 (15.2) | 327 (14.9) | 414 (15.3) | 441 (10.4) |
| N/A | N/A | N/A | N/A | N/A | N/A |
| **Cancer type, No. (%)** | | | | | |
| Breast | 1,321 (19.0) | 113 (12.1) | 275 (12.5) | 346 (12.8) | 975 (23.0) |
| Gastrointestinal | 819 (11.8) | 93 (9.9) | 291 (13.2) | 366 (13.5) | 453 (10.7) |
| Thoracic | 774 (11.2) | 107 (11.4) | 258 (11.7) | 326 (12.1) | 448 (10.6) |
| Lymphoma | 700 (10.1) | 170 (18.2) | 345 (15.7) | 382 (14.1) | 318 (7.5) |
| Head and neck | 658 (9.5) | 90 (9.6) | 208 (9.4) | 238 (8.8) | 420 (9.9) |
| Pancreas | 585 (8.4) | 99 (10.6) | 214 (9.7) | 280 (10.4) | 305 (7.2) |
| Prostate | 520 (7.5) | 11 (1.2) | 46 (2.1) | 70 (2.6) | 450 (10.6) |
| Gynecologic | 513 (7.4) | 70 (7.5) | 176 (8.0) | 218 (8.1) | 295 (7.0) |
| Genitourinary | 461 (6.6) | 76 (8.1) | 184 (8.4) | 219 (8.1) | 242 (5.7) |
| Other | 587 (8.5) | 107 (11.4) | 205 (9.3) | 259 (9.6) | 328 (7.7) |
| **Cancer stage, No. (%)** | | | | | |
| Stage I | 1,099 (15.8) | 123 (13.1) | 281 (12.8) | 338 (12.5) | 761 (18.0) |
| Stage II | 1,415 (20.4) | 141 (15.1) | 336 (15.3) | 410 (15.2) | 1005 (23.7) |
| Stage III | 964 (13.9) | 131 (14.0) | 351 (15.9) | 429 (15.9) | 535 (12.6) |
| Stage IV | 1,898 (27.4) | 327 (34.9) | 759 (34.5) | 937 (34.7) | 961 (22.7) |
| Unknown | 1,562 (22.5) | 214 (22.9) | 475 (21.6) | 590 (21.8) | 972 (23.0) |
| **Insurance, No. (%)** | | | | | |
| Medicare | 2,683 (38.7) | 323 (34.5) | 788 (35.8) | 970 (35.9) | 1,713 (40.5) |
| Private | 2,450 (35.3) | 328 (35.0) | 747 (33.9) | 898 (33.2) | 1,552 (36.7) |
| Medicaid | 599 (8.6) | 130 (13.9) | 258 (11.7) | 307 (11.4) | 292 (6.9) |
| Other or unknown | 1,206 (17.4) | 155 (16.6) | 409 (18.6) | 529 (19.6) | 677 (16.0) |

Table 5.1: Information about the complete patient cohort for experiment II (train and test set) eligible for the OP-35 metric for 30-, 180-, and 365-day prediction. "No." stand for number and "sd" for standard deviation. "N/A" was too small to report due to patient privacy concerns.

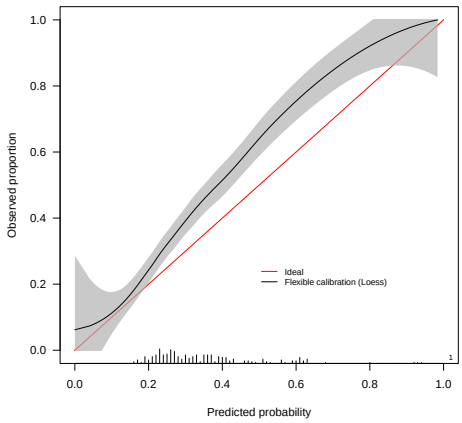
| Label | Model | No. SHD | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|---|
| 30 | Tabular LASSO $b = 0.02$ | 83 | 0.775 (0.757, 0.792) | **0.411** (0.373, 0.447) | 0.344 (0.329, 0.358) | 0.036 (0.015, 0.049) |
| | Language LASSO $b = 0.03$ | N/A | 0.726 (0.707, 0.744) | 0.294 (0.264, 0.323) | 0.363 (0.346, 0.379) | **<0.001** (<0.001, 0.021) |
| | Fusion LASSO $b = 0.02$ | 73 | **0.778** (0.760, 0.795) | 0.410 (0.372, 0.447) | **0.341** (0.326, 0.356) | **<0.001** (<0.001, 0.024) |
| | Language BERT | N/A | 0.710 (0.692, 0.729) | 0.259 (0.235, 0.282) | 0.435 (0.415, 0.455) | 0.131 (0.117, 0.145) |
| | Fusion BERT | 419 | 0.766 (0.749, 0.784) | 0.315 (0.286, 0.343) | 0.393 (0.377, 0.406) | 0.103 (0.089, 0.116) |
| 180 | Tabular LASSO $b = 0.03$ | 221 | 0.748 (0.735, 0.762) | 0.623 (0.600, 0.647) | 0.540 (0.527, 0.552) | 0.017 (<0.001, 0.039) |
| | Language LASSO $b = 0.02$ | N/A | 0.730 (0.717, 0.745) | 0.577 (0.555, 0.601) | 0.558 (0.546, 0.570) | **<0.001** (<0.001, 0.034) |
| | Fusion LASSO $b = 0.02$ | 101 | **0.765** (0.752, 0.779) | **0.632** (0.610, 0.655) | **0.530** (0.517, 0.543) | **<0.001** (<0.001, 0.025) |
| | Language BERT | N/A | 0.702 (0.688, 0.717) | 0.543 (0.517, 0.567) | 0.625 (0.603, 0.644) | 0.107 (0.093, 0.119) |
| | Fusion BERT | 419 | 0.753 (0.741, 0.767) | 0.620 (0.597, 0.644) | 0.548 (0.536, 0.558) | 0.038 (0.023, 0.059) |
| 365 | Tabular LASSO $b = 0.02$ | 150 | 0.763 (0.752, 0.775) | **0.704** (0.685, 0.724) | 0.559 (0.549, 0.569) | **<0.001** (<0.001, 0.035) |
| | Language LASSO $b = 0.02$ | N/A | 0.732 (0.730, 0.755) | 0.639 (0.637, 0.678) | 0.585 (0.567, 0.586) | **<0.001** (<0.001, 0.022) |
| | Fusion LASSO $b = 0.02$ | 115 | **0.770** (0.759, 0.782) | 0.702 (0.683, 0.722) | **0.553** (0.541, 0.563) | 0.041 (<0.001, 0.057) |
| | Language BERT | N/A | 0.709 (0.695, 0.723) | 0.617 (0.594, 0.640) | 0.666 (0.647, 0.683) | 0.135 (0.122, 0.148) |
| | Fusion BERT | 419 | 0.760 (0.748, 0.774) | 0.695 (0.675, 0.714) | 0.565 (0.554, 0.575) | 0.021 (<0.001, 0.041) |

Table 5.2: Resulting metrics on the test set of the tabular, language and fusion LASSO models, as well as the language and fusion BERT, trained on 30, 180 and 365 days ACU prediction. The best-performing metrics for every label type are marked in bold. We display the number of SHD used for prediction in the third column, where "N/A" means that SHD was used for prediction. The bootstrapped 95%-confidence intervals are reported in the brackets: (2.5%-CI, 97.5%-CI). The inverse regularisation parameter for the LASSO is denoted as $b$. The results were drawn from the inter-model comparison for the Language LASSO (Appendix Table B.1), Language BERT (Appendix Table B.2 and B.3), and Fusion BERT (Appendix Table B.4).

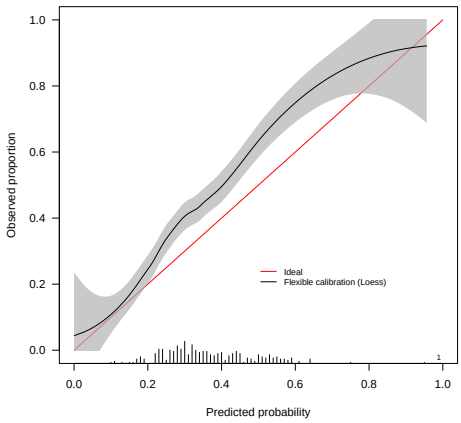
115 tabular features. At the same time, the Tabular LASSO had the highest AUPRC score (0.704, 95%-CI:0.685, 0.724), using 150 SHD points.
We show the flexible calibration curves for the 180-day models in Figure 5.3, where we observed a risk underestimation of the three LASSO models and underestimation of low-risk patients and overestimation of high-risk patients with the Language BERT model.
The Fusion BERT used the most SHD points (419 tabular inputs) for all three label types to make predictions.

### 5.2.2 Exploration of Clinical Usage of Language Models

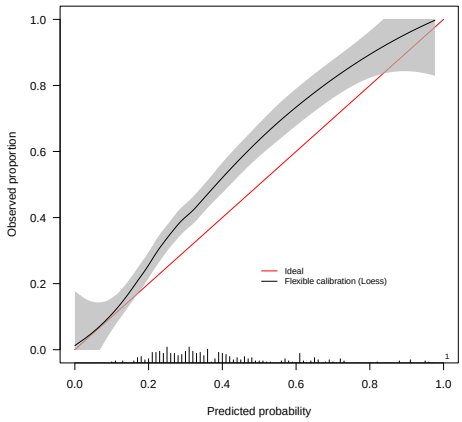
The Decision curve analysis for the 180-day ACU prediction showed that the net benefit of the Language BERT model yields a negative benefit when the decision threshold for treatment is chosen above 0.6 (Figure 5.4) and less or equal net benefit to treating every patient with a threshold below 0.19. The other models, including the Language LASSO model, had positive benefit values for decision thresholds



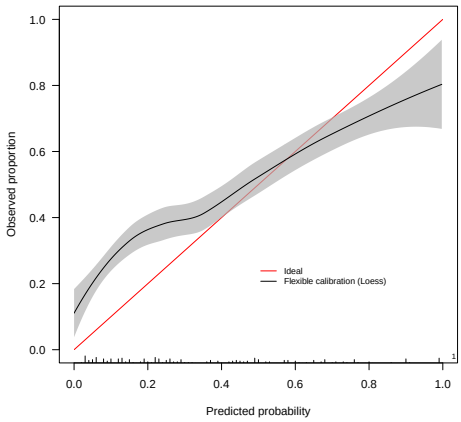

(a) Tabular LASSO



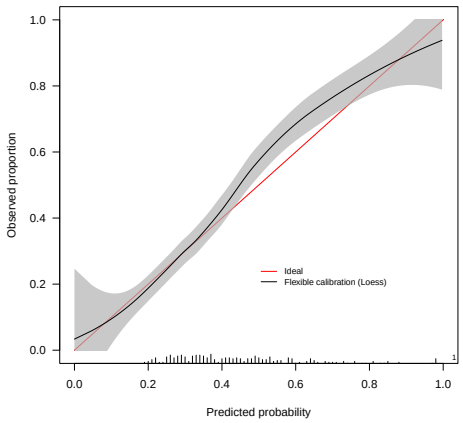

(b) Language LASSO




(c) Fusion LASSO




(d) Language BERT




(e) Fusion BERT

Figure 5.3: Calibration curves of the 180-day ACU risk prediction models. The red line indicates ideal calibration, while the black line is the flexible calibration with the 95%-confidence interval [61].

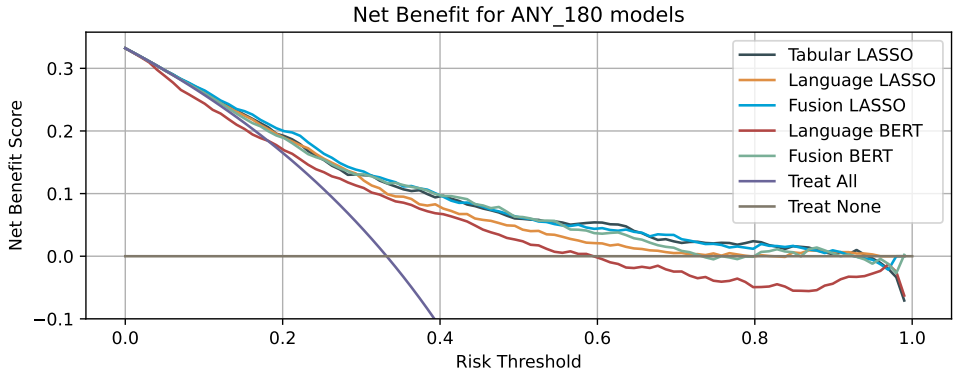
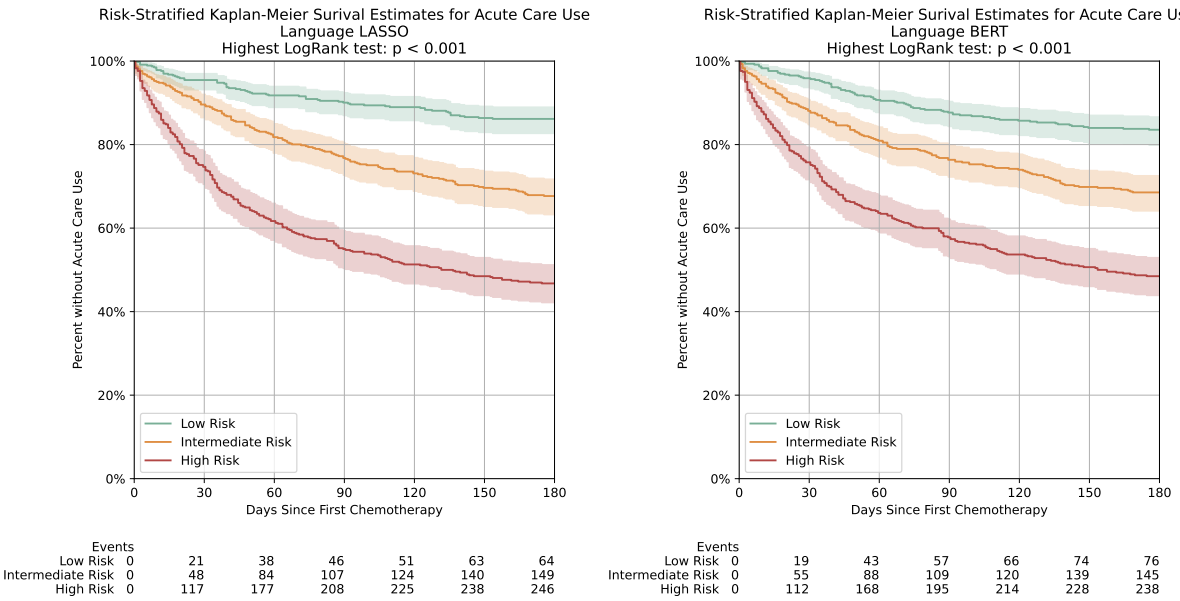



Figure 5.4: Net benefit curves of the tabular, language and fusion models. The purple curve indicates the benefit of all the patients being treated, whereas the grey curve indicates the benefit if no patient is treated

below 0.7. In the Appendix we demonstrate also the DCA for 30-day and 365-day ACU prediction (Figure B.1).
The Kaplan-Meier survival curves for OP-35 events showed good separation between risk groups (Figure 5.5, $p < 0.001$ for each group by log-rank test) for the two language-only models. By 180 days after the start of chemotherapy, 64 (13.9%) of the 462 low-risk patients in the language LASSO prediction had an OP-35 event and 76 (16.5%) in the language BERT prediction. On the other hand, 246 (53.2%) of the 462 high-risk patients had an event for the speech LASSO prediction and 238 (51.5%) for the language BERT prediction.
Figure 5.6 shows the relative importance of the ten highest and lowest coefficients of the language LASSO model for the 180-day prediction. The words "Admission", "Failure", "Pain", and "Palliative" were among the ten highest coefficients, while "Breast", "PSA (Prostate-specific antigen)", "Nourished", and "Prostate" were among the ten lowest coefficients. The supplementary materials also present the word importance for the Language LASSO on 30-day and 365-day prediction (Figure B.2).

### 5.2.3 Sensitivity Analysis

Figure 5.7a shows that Black patients were predicted to have a disproportionately higher risk than White, Asian or other-race patients. We note that the number of Black patients was at least seven times smaller than that of non-Black races. The cumulative risk by different insurance types is displayed in Figure 5.7b, where we note a risk overestimation of Medicaid patients.
In Figure 5.7c, we see that the risk predictions for patients with stage III and IV tumours were overestimated. In contrast, the predictions for patients with stage I, II and unknown stage cancer were underestimated.
We show the empirical cumulative risk distribution of the Language LASSO on for gender, ethnicity, and cancer type in the Appendix Figure B.3.




(a) Language LASSO (180d predictions)

(b) Language BERT (180d predictions)

Figure 5.5: Kaplan-Meier curves for ACU events for patients in the test cohort stratified by predicted risk. The shaded area represents the 95%-CIs.

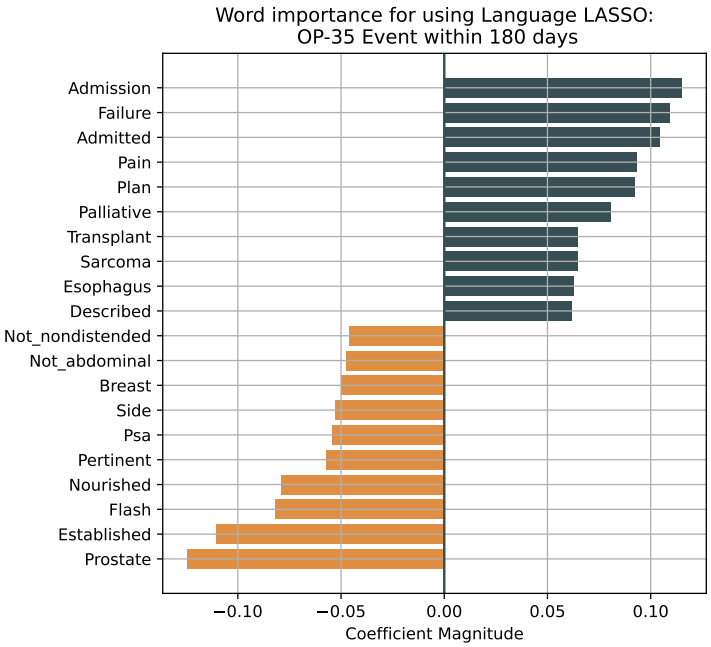
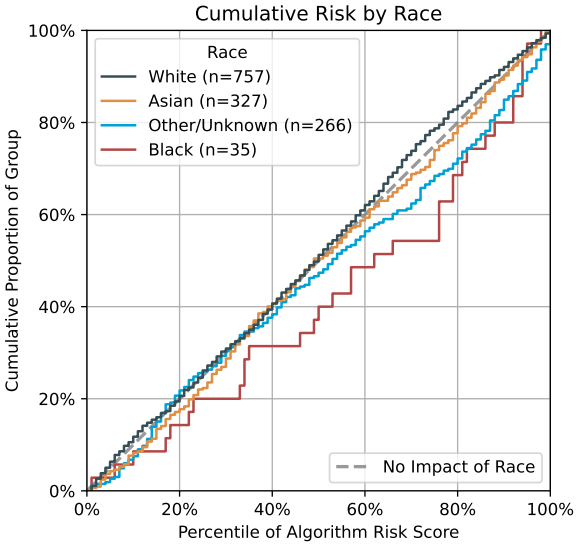
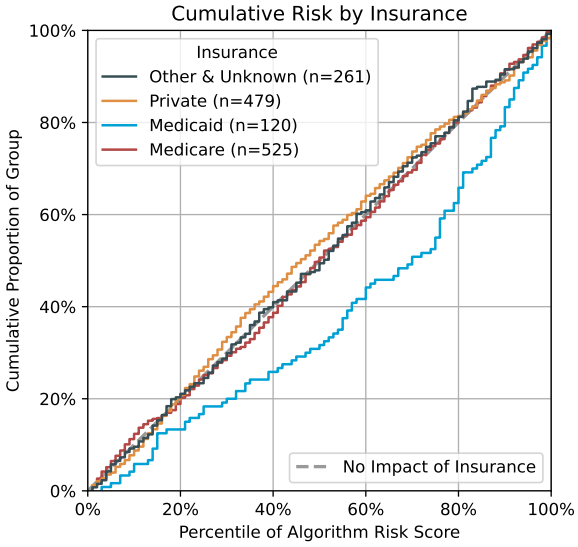


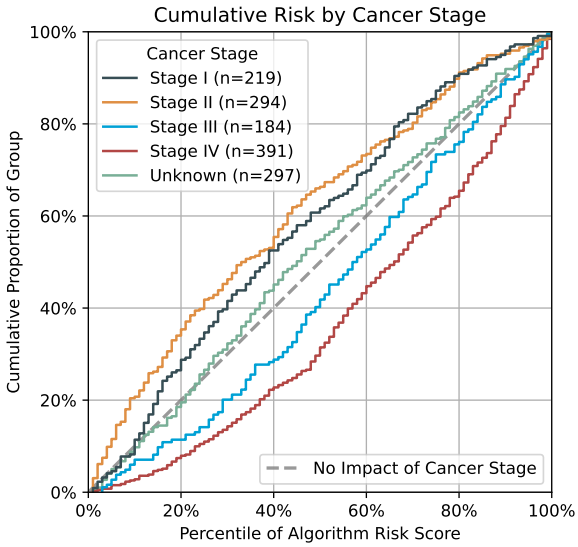

Figure 5.6: Coefficient magnitudes for the Language LASSO for 180-day ACU prediction, displaying the ten highest and ten lowest. The coefficients in this model are single words found in the clinical notes before the ACU event.




(a) Cumulative risk by race




(b) Cumulative risk by insurance




(c) Cumulative risk by cancer stage

Figure 5.7: Cumulative risk of the Language LASSO on 180 ACU prediction, stratified by race (a), insurance type (b) and cancer stage (c).

# Chapter 6

# Discussion

This thesis investigates how uncertainty estimation and NLP methods can help predict the risk of Acute Care Use (ACU) in oncology patients after starting chemotherapy. In this section, the results of each experiment are discussed, followed by the implications and limitations.

## 6.1 Uncertainty Estimation for ACU Prediction

In the first experiment, we presented an analysis of uncertainty estimation for cancer patients at risk of acute care utilisation following initiation of chemotherapy, based on high-dimensional SHD. We obtained four critical findings. First, BLLRs are a suitable alternative to the frequentist LASSO to predict the risk of ACU, with the addition that they quantify the predictive uncertainty, despite high-dimensional inputs. Second, the Horseshoe-MH and Laplace-MH models are superior to Laplace-VI in predicting risk for ACU with uncertainty. Third, this approach calculates the proportion of predictions in a dataset that is certain. Fourth, an algorithmic bias of the predictive uncertainties is detected in different patient groups.

### 6.1.1 BLLR vs Logistic LASSO

Based on our results, we argue that BLLRs can be a good alternative compared to the standard logistic LASSO. The results show that the Horseshoe-MH and Laplace-MH models achieved comparable results in predicting risk compared to a standard logistic LASSO, consistent with results of previous works [15, 56, 76]. In particular, the Horseshoe-MH approach achieves virtually the same, or better, results on the test set for all metrics, calibration, and Net Benefit. When comparing calibration, we can see that the Horseshoe-MH has significantly better ECE than the frequentist LASSO. The calibration curve underestimates the risk for patients less in the mid-section of probabilities. Comparing individual patient predictions demonstrates how the BLLR provides a predictive distribution rather than just a point estimate. Especially for the mid-risk patient, we observe a large dispersion for the Laplace-MH and Horseshoe-MH, indicative of high uncertainty. We cannot obtain the same information from the Frequentist LASSO estimate. This is also the case for the posterior distribution of the input features. While the Frequentist LASSO provides a single value for its weights, the BLLRs provide the posterior distribution, which often encloses their equivalent traditional counterparts. In this study, we used 760 SHD input features, which is a lot more than previous clinical works using Bayesian logistic regression [25, 76] (<30 features). Thus, the input dimension is not a limiting factor in using a BLLR instead of an ordinary logistic LASSO.

### 6.1.2 Differences within the BLLRs

Compared to Mathiszig-Lee et al. [25]'s work, we focus not solely on the predictive distributions for individual patients but also discuss how uncertainty can be computed and how it differs in the Bayesian models. Here, we demonstrate essential differences between different BLLRs. The Laplace-VI model has worse calibration and a significantly higher cross-entropy loss than the other models. Furthermore, by inspecting the DCA plot, we find that the Laplace-VI generates almost no Net Benefit, and with most decision thresholds, even yields negative Net Benefit. This underperformance is potentially because MCMC methods are asymptotically exact, while VI is not [43, 77]. Regarding uncertainty quantification, our results show that the Laplace-VI model has very high estimated uncertainties and low coverage compared to the MH-sampled Bayesian models. This is because the probability mass of the predictive distribution is at very high and very low values, which can be noticed in the individual patient predictions. This observation makes it difficult to quantify the uncertainty using classical statistical methods such as standard deviation and quantiles. We argue that this property makes the Laplace-VI model potentially infeasible for clinical use with its worse discrimination and calibration. For the choice of priors, our results show that the parameterised Horseshoe+ prior outperforms the Laplace prior in predictive power and provides more certain predictions when the exact definition for uncertainty is applied. We note that the coverage of the Horseshoe-MH model is always higher than that of the other two BLLRs at the respective decision thresholds.

### 6.1.3 Uncertainty Comparison for Clinical Classification

Our approach can quantify the proportion of predictions in a dataset that is certain and, therefore, potentially more trusted by the end user. A similar approach has been explored by Joshi and Dhar [18] with their uncertainty filtering method, where uncertain predictions are excluded from classification with their Bayesian neural network. In this work, however, we focus on the impact of the coverage ratio compared to classification metrics for different Bayesian models to determine their utility. We can select which Bayesian model is closest to the optimum based on our method. Furthermore, our analysis allows us to examine these effects for different decision thresholds and definitions of uncertainty. If their uncertainties are computed differently, two different Bayesian models may achieve the same coverage (and potentially classification results). Additionally, there are only scaling differences when computing uncertainty with either standard deviation or credible intervals when the predictive distributions approximately follow a Gaussian distribution. However, suppose the predictive distributions have a different shape, e.g. in the case of the Laplace-VI. In that case, the credible intervals are less suited, as they cover the whole risk probability space. Nevertheless, compared to $\sigma$-quantified uncertainty, the 95%-credible interval has a more intuitive interpretation: 95% of the predictions made by the BLLR lay within the uncertainty range for this patient. In a nutshell, this approach allows data scientists and clinicians to determine the best performing model, decision threshold and quantified uncertainty based on their guidelines and available resources.

### 6.1.4 Bias in Predictive Uncertainties

We showed how algorithmic fairness can also be analysed in terms of the predictive uncertainties of a model. Compared to other work in clinical informatics, we do not focus on the biases in the data or the risk probability scores [1]. We show how to visualise potential biases in quantified uncertainty stratified by different patient groups. Our results show that a classification model would disproportionately fail to classify Black, stage IV, and sarcoma cancer patients. The increased uncertainty for Black and Stage sarcoma patients may be due to the comparably low patient count. However, this is not the case for Stage IV patients. While this uncertainty bias might be reasonable for different tumour stages or types, we believe it should not be the case for demographic values such as race. In the results, the median Black

patient has a 67% higher typical error (i.e. $\sigma$) around its risk prediction than the median White patient. Our method extends the current analysis of algorithmic bias in biomedical informatics to the field of uncertainty estimation.

## 6.2 ACU Prediction with Free-Text Clinical Notes

Identifying methods to improve external generalisability is a priority for the medical informatics community [78]. This second experiment presents an analysis of NLP to identify patients at risk of seeking acute care at different time intervals. We compared models trained on SHD, free-text data, and both modalities combined. From the results of this experiment, we obtained three critical findings. First, the Tabular LASSO only slightly outperforms the Language LASSO, while transformer-based language models do not outperform the other models. Second, clinical notes can be exploited to predict the risk of ACU, which is scalable across sites because it does not require medical institutions to have a common data model or to convert their SHD. Third, despite not explicitly using demographic variables, we observe algorithmic bias in the risk prediction of language-based models.

### 6.2.1 Tabular vs Language vs Multimodal Models

Our results demonstrate Tabular LASSO outperforming Language LASSO and Language BERT at all three time intervals. Nevertheless, both language-based methods achieve good discriminative performance (AUROC $\geq 0.7$) even without SHD on all three risk intervals. On the other hand, combining the two input modalities (clinical notes and SHD) does not yield significant improvements over using tabular data alone. This is potentially because the information contained in the SHD is also reported in the clinical notes. In addition, the results show that the popular BERT-based classifier does not outperform $\ell_1$-penalised logistic regression with TF-IDF features. This result is contrary to Gatto et al. [37] results, as it favours simpler models with human features. The BERT models' performance is likely due to the aggregation of chunks of the lengthy clinical documents into a single output probability, containing a lot of noisy data, which makes deep learning training difficult. While deep learning models are powerful at extracting valuable features from unstructured text, they might fail to attribute a single label to a large input. Before starting to train a sophisticated transformer architecture, a modeller might first consider lexical-based features with a simple linear model. Training a BERT model on all three time intervals with the cumulative link loss leads to significantly increased discrimination performance instead of training them individually. Thus, our method optimises the training process considerably as the computationally intensive model training for the different labels is not required. This effect can potentially be accounted to a clinical note containing specific parts in texts that are common on 30-day, 180-day, and 365-day predictions and others that are important for the individual time frame.

### 6.2.2 Clinical Utility of Language Models

ACU risk prediction models for chemotherapy patients perform well using free-text data from the last (at most three) H&P and progress notes before chemotherapy. Apart from achieving good discrimination performance on the test set, the models can also be helpful in the clinical setting. This work demonstrates that language-based models can stratify patient risk groups based on their predictions. Based on the Kaplan-Meier curves, we see that the Language LASSO outperforms the Language BERT also in this task, as it can detect a higher percentage of high-risk patients. In the DCA, we observe that while the Language LASSO Net Benefit is almost always positive, the Language LASSO yields negative values if the decision threshold is larger than 0.6. Based on these two findings, we conclude that the Langauge LASSO is better suited to be deployed at the point of care. Moreover, we show that Language LASSO coefficients can help clinicians understand relative word meaning when interpreting a prediction

(although credible intervals would be even better). In contrast, the interpretability mechanism of the Language BERT model is limited to its attention maps, which are hard to interpret.

### 6.2.3 Risk Prediction Bias in Language Models

In the sensitivity analysis, we find that certain groups may be subject to risk bias, despite not using demographic values explicitly as inputs. Similar to Peterson et al. [1]'s work with SHD, our results demonstrate that Medicaid and Black patients are proportionately overestimated. For Black patients, this might potentially be due to a comparably low number of patients. Interestingly, we see that cancer stage I and II patients are more likely to have underestimated risk prediction, while cancer stage III and IV have overestimated risk scores. This might initially be intuitive, as patients in a more advanced stage could require ACU more likely [1]. However, we believe that because of this underestimation, stage I and II cancer patients require special attention from specialists on the risk predictions if a language model is deployed.

## 6.3 Implications

We have developed a method for estimating uncertainty that provides essential information about the certainty of the risk score and model weights for ACU prediction. Data scientists and clinicians can think not only about the risk probability of an event but also about the acceptable uncertainty of their predictions. Subsequently, the range of uncertainty can vary depending on the prediction problem and the resources available to medical institutions. For example, if an ML model is, e.g. developed to triage patients, predictions with an uncertainty range that exceeds the decision threshold will likely be misclassified and thus uncertain. We argue that clinicians should be aware of these cases rather than over-relying on point estimates of probabilities, in contrast to previous ACU studies [1, 6, 7, 8, 12]. The same applies to the weights when inspecting their significance and predictive power. Moreover, when deploying such a model, users ought to be aware of potential biases in the predicted uncertainty and its effects down the line.
ACU risk prediction models for chemotherapy patients perform well using clinical notes as inputs. This implies that NLP methods could be easily implemented across sites or facilities as they only require access to written medical notes without re-structuring or mapping structured data, potentially saving costs in feature collection. Additionally, we believe that language model users should be aware of these subgroup differences when interpreting the results of ML models. Our results suggest that Medicaid and Black patients and stage I and II tumour patients need close monitoring. These results suggest that clinicians should be aware of these subgroup differences when interpreting the results of ML models [79].

## 6.4 Limitations

The experiments and results of this thesis have limitations. For the first experiment, other, more sophisticated MCMC sampling techniques, such as the NUTS sampler [80], lead to more stable posterior approximations. However, these techniques require even more computations, especially for high-dimensional input features, which may affect the feasibility of model deployment at the point of care. In addtion, it is more difficult to compare the quantified uncertainties of models than their predictive capabilities, as there is no ground truth for uncertainty, unlike in supervised learning. Different definitions of uncertainty could lead to other results in our experiments. A modeller has to consider an additional aspect of choice with uncertainty rather than just the risk of an event. Further research is needed for the value assessment of uncertainty estimation for ML.
The second experiment also has limitations. The collected dataset of clinical notes might still contain erroneous entries. Furthermore, the notes' length, number and detail vary enormously per patient. Further

preprocessing and selection of the dataset could improve the discriminative performance of the language models. However, this is a very timely and costly procedure. Regarding the BERT models, the choice of hyperparameters of the neural networks is mainly motivated through trial-and-error on the validation set. Different hyperparameter combinations could have potentially yielded better results for the transformer models. However, this is also a lengthy process.
Finally, both experiments have been validated only on one dataset for risk prediction of acute care use. Testing these experiments on various medical problems and other care systems would be beneficial. Moreover, both were performed at a single academic institute and may not be generalisable to other healthcare settings. Nonetheless, our focus lies in providing an analysis of the methods.

# Chapter 7

# Conclusion

In conclusion, this thesis explored the application of different uncertainty estimation methods and NLP to help identify the risk of ACU for oncology patients. Our first experiment shows the importance of estimating uncertainty in a high-stakes environment such as medicine, as it provides additional information about the certainty of ML models. This uncertainty quantification helps increase trust in the model for all stakeholders. We show how BLLR can replace the logistic LASSO regression, as it performs equally well in prediction, despite the high number of input features, and provides the uncertainty of its predictions. A suitable prior choice is Horseshoe+, and the posterior can be sampled with Metropolis-Hastings. Overall, this work offers a paradigm shift in how we think about and use uncertainty estimates for risk scores in clinical decision support. Accounting for uncertainty increases the accuracy of predictions and trust in ML systems and allows clinicians to use the risk score in a more informed context. Using this uncertainty approach, we improve the capabilities of automated decision making [81] and identify cases where uncertainty is high, and ML cannot provide an accurate risk probability estimate. This is an advance over the current point estimation approach, where one gets a probability score of an event, and the uncertainty is unknown.

The second experiment demonstrates the utility of using free-text data to identify patients at risk of needing acute care once they have started chemotherapy. It is an alternative to structured health data, which may require significant preprocessing and may not be generalisable across settings. We show that the Language LASSO is a suitable model, especially for 180-day prediction, and is well interpretable. This work advances the knowledge of risk prediction models and provides an alternative for cross-site generalisability.

In this thesis, we have shown that the proposed methods offer numerous advantages in identifying patients at high risk of ACU, either by increasing end-user trust through quantifying predictive uncertainty or using free-text clinical data as an alternative to SHD.

# Chapter 8

# Outlook

The work presented in this thesis represents only a first step towards improving clinical decision support with uncertainty estimates and ACU predictions with NLP. To efficiently share the results of this work with the broader scientific community, efforts are underway to prepare a manuscript summarising the main findings for a peer-reviewed scientific publication for the first experiment (chapter 4). Ideally, this publication will contribute to developing clinical prediction models that quantify uncertainty, as this is crucial in an environment where the stakes are as high as in medicine. The second experiment (chapter 5) has already been submitted to a peer-reviewed clinical informatics conference. The paper aims to present researchers with a comparison of language models for ACU prediction. We hope this will provide the impetus for further research on ACU risk prediction based on easily obtainable free-text clinical data that can be generalised across all medical institutions.

Beyond the scope of these experiments, a natural next step would be to reproduce the methods and analyses presented for other medical problems and at other institutions with other patient datasets. To facilitate this, all code for the experiments will be published on GitHub/GitLab.

Another focus of future research would be the combination of uncertainty estimation and NLP and its impact on clinical decision support. Different methods for uncertainty estimation are currently being researched, especially for neural networks, so it is possible to compare the predictive uncertainty of language models in healthcare. This would allow further evaluation of the added value of language models compared to SHD once they are deployed at the point of care.

There are still many challenges in researching uncertainty assessment and natural language processing in biomedical informatics. While it may be impossible to develop a universal definition of uncertainty suitable for all problems, research efforts should focus on harmonising the multitude of different uncertainty concepts into a coherent analytical framework shared by the scientific community. Similar to comparing models through their predictive performance, it would be helpful to have tests and experiments that can analytically compare those models' uncertainty across a test set.

# Appendix A

# Supplementary Experiment I

## A.1 Data and Code Availability

Under the terms of the data-sharing agreement for this study, we cannot share the source data directly. Requests for anonymous patient-level data can be made directly to the authors.
All the experiments were implemented in Python [82], using the SciKit-Learn [83] library for the metrics and frequentist logistic regression LASSO model, PyMC3 [84] for the Bayesian models, and PyTorch [74] with the Huggingface [85] library for the transformer models. We used R [86] to create the calibration plots. The LASSO and Bayesian models were run locally on a computer laptop (Lenovo Yoga X1, Intel Core i7-8550U CPU, 1.80GHz, 1.99 GHz, 16.0 GB RAM), while the transformer models were trained on a dual-GPU (NVIDIA Tesla T4 $\times$ 2, 10.0 GB) on the google cloud platform of Stanford University. The (currently, still private) code for our models and analysis is available on :

- Bayesian logistic LASSO Regressions:
  https://code.stanford.edu/boussard-lab/acu-uncertainty-estimation
- NLP logistic LASSO Regressions:
  https://code.stanford.edu/boussard-lab/nlp-for-acu
- BERT models & training:
  https://code.stanford.edu/boussard-lab/claudio-master-thesis

## A.2 Additional Results

In Table A.1 we show the predictive performance frequentist LASSO and the BLLRs for 30-day, 180-day, and 365-day ACU prediction.
Figure A.1 visualises the sorted risk predictions for 30-day ACU when the 95%-credible interval quantifies uncertainty. We note that in this example, the coverage is 0.54, meaning that 46% of the predictions were too uncertain to be classified.
Figure A.2 demonstrates the coverage against the classification scores when uncertainty is quantified with the 95%-credible interval. We observe that the Laplace-VI model had low coverage.
In Figure A.3 we show the uncertainty distributions of the Horseshoe-MH model, stratified by gender, ethnicity, insurance type, and cancer type. We observe no significant difference between female and male ($p = 0.051$). Hispanic/Latino patients had a higher median uncertainty distribution than non-Hispanic patients ($p < 0.001$). Medicaid patients have the highest uncertainty median (median > 0.05, $p < 0.001$) compared to other insurance types (median < 0.05).

| Label | Model | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|
| 30 | LASSO C=0.03 | 0.806 (0.792, 0.820) | **0.511** (0.477, 0.543) | 0.357 (0.344, 0.370) | 0.045 (0.031, 0.058) |
| | Bayesian LASSO (Laplace - VI) | 0.774 (0.757, 0.789) | 0.437 (0.406, 0.471) | 0.539 (0.526, 0.551) | 0.242 (0.233, 0.253) |
| | Bayesian LASSO (Laplace - MH) | 0.769 (0.754, 0.785) | 0.452 (0.420,0.484) | 0.38 (0.363, 0.396) | 0.032 (0.000, 0.042) |
| | Bayesian Horseshoe (Horseshoe - MH) | **0.807** (0.793, 0.821) | 0.498 (0.466, 0.528) | **0.355** (0.340, 0.368) | **0.006** (0.000, 0.030) |
| 180 | LASSO C=0.02 | **0.794** (0.783, 0.805) | 0.719 (0.701, 0.738) | **0.515** (0.505, 0.526) | 0.03 (0.000, 0.048) |
| | Bayesian LASSO (Laplace - VI) | 0.793 (0.782, 0.804) | **0.72** (0.702, 0.736) | 0.526 (0.512, 0.539) | 0.06 (0.046, 0.073) |
| | Bayesian LASSO (Laplace - MH) | 0.781 (0.769, 0.792) | 0.696 (0.676, 0.717) | 0.537 (0.522, 0.551) | 0.039 (0.020, 0.052) |
| | Bayesian Horseshoe (Horseshoe - MH) | **0.794** (0.783, 0.804) | 0.716 (0.698, 0.736) | **0.515** (0.503, 0.528) | **0.015** (0.000, 0.030) |
| 365 | LASSO C=0.02 | 0.792 (0.781, 0.802) | 0.763 (0.748, 0.779) | 0.537 (0.528, 0.547) | 0.028 (0.000, 0.041) |
| | Bayesian LASSO (Laplace - VI) | 0.793 (0.782, 0.804) | 0.762 (0.746, 0.776) | 0.538 (0.525, 0.551) | 0.032 (0.000, 0.046) |
| | Bayesian LASSO (Laplace - MH) | 0.78 (0.769, 0.791) | 0.745 (0.729, 0.760) | 0.556 (0.542, 0.570) | 0.035 (0.007, 0.049) |
| | Bayesian Horseshoe (Horseshoe - MH) | **0.795** (0.785, 0.806) | **0.766** (0.751, 0.781) | **0.532** (0.521, 0.543) | **0.027** (0.000, 0.038) |

Table A.1: Resulting metrics on the test set of the frequentist LASSO and the Bayesian Logistic Regression, trained on 30, 180 and 365 days ACU prediction. We report the 95%-confidence intervals of the metric estimates in the brackets (2.5%-CI, 97.5%-CI), that have been calculated with 1,000-fold bootstrap. The best-performing metrics for every label type per metric are marked in bold.

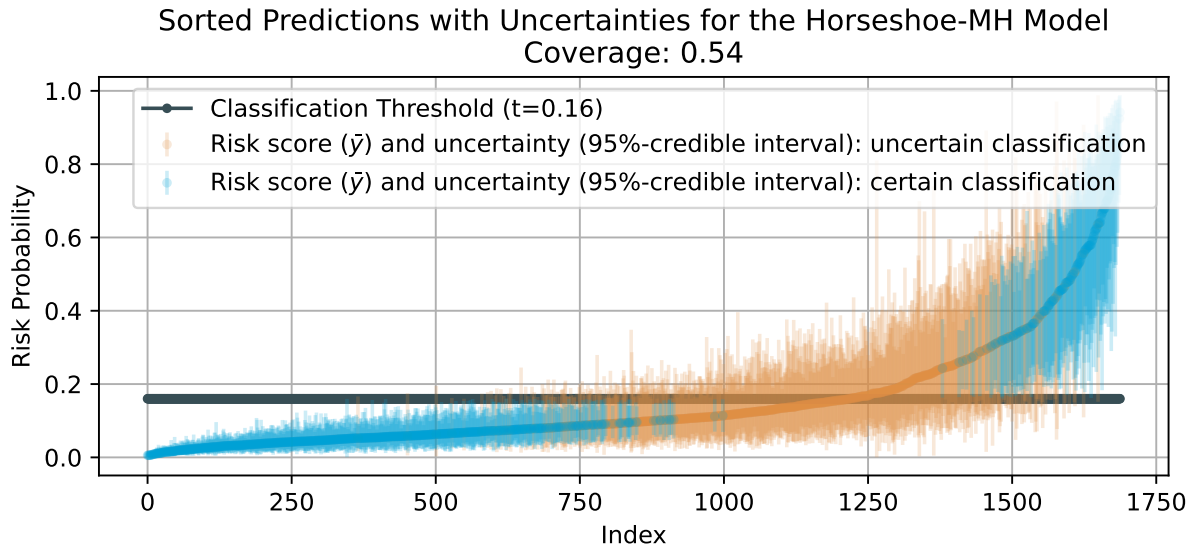
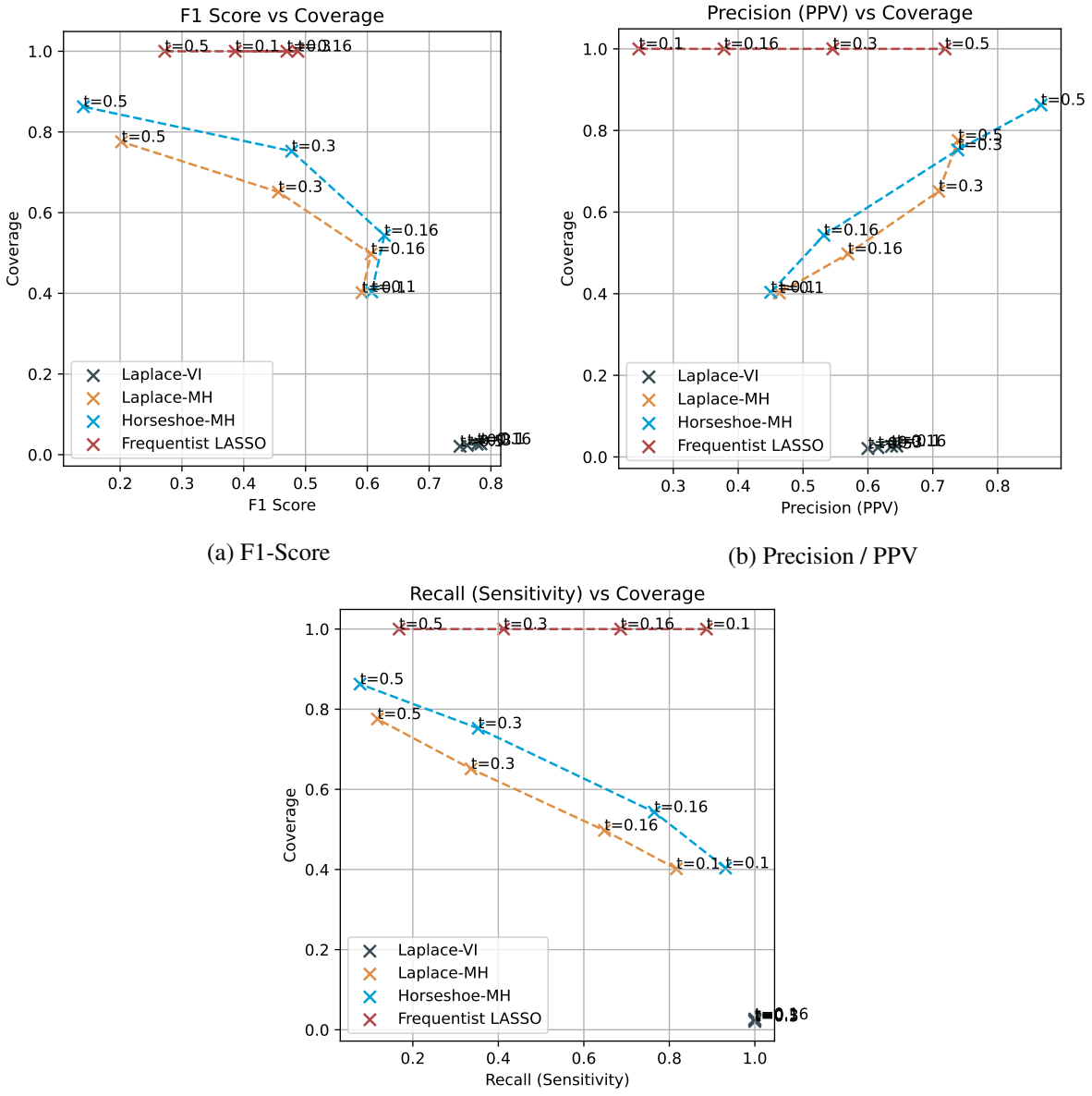



Figure A.1: Sorted final risk predictions (mean of the predictive distribution, $\bar{y}$) with uncertainty range (95%-credible interval) for the Horseshoe-MH model. The predictions whose uncertainty does not exceed the decision threshold (certain classifications) are coloured blue, and those that do (uncertain classifications) are coloured orange. The dark grey line is our chosen classification threshold at 0.16, the event rate.

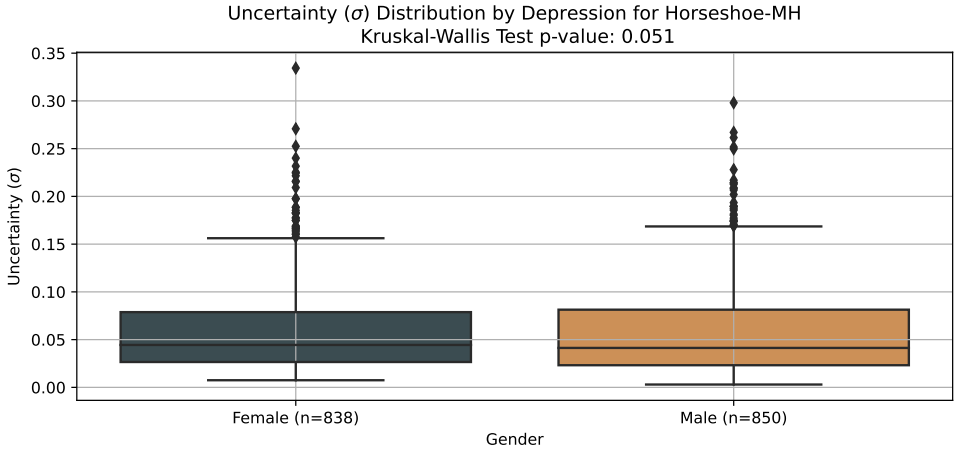



(a) F1-Score

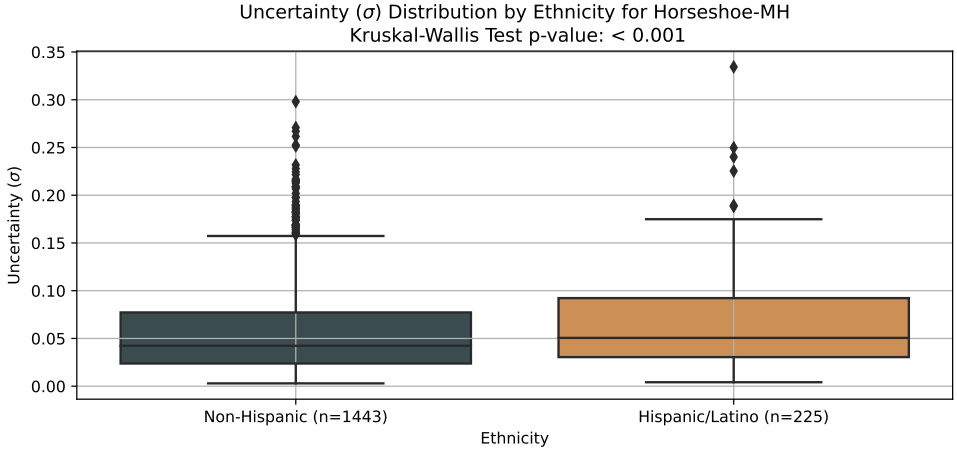
(b) Precision / PPV

(c) Recall / Sensitivity

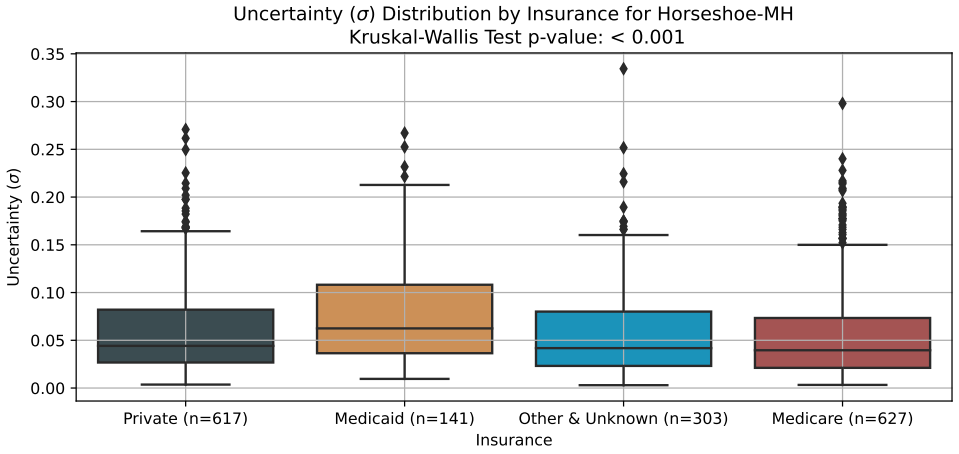
Figure A.2: Coverage (ratio of automatically classified predictions) compared to F1-score (a), sensitivity (b), and PPV (c), over four risk decision thresholds (0.1, the event rate 0.16, 0.3, 0.5) of all the models, with uncertainty quantified by 95% inverse cumulative density function.




(a) Gender




(b) Ethnicity




(c) Insurance Type

Figure A.3: Distribution of quantified uncertainty $\sigma$ of the test set, stratified by gender (a), ethnicity (b), insurance status (c). The Kruskal-Wallis statistic and significance can be found in the title.

| **Variable Code** | **Explanation** |
|---|---|
| Sarcoma | Sarcoma cancer patient |
| Non Palliative | Patient is currently in non-palliative care |
| PROC: 36569 | Provider places catheter through a vein in the upper extremity of a patient and passes the catheter into one of the major veins carrying the blood to the heart or directly into the right atrium, without placing a subcutaneous port or pump |
| PROC: 77059 | MRI scan of both breasts with contrast (deleted) |
| PROC: 84478 | Amount of triglycerides in the patient specimen |
| PROC: 85610 | Chemicals like calcium and tissue factor are added to the plasma sample and then the time is noted when the plasma clots. |
| PROC: 87389 | test to screen for human immunodeficiency virus, called HIV |
| PROC: 88323 | Pathologist, provides consultation and a report on referred material, such as a tissue block, and prepares and stains slides |
| PROC: 96374 | Single medication or other substance rapidly into a vein to treat, prevent, or diagnose a condition. |
| PROC: 96402 | Chemotherapy either subcutaneously or intramuscularly |
| PROC: 96411 | Additional chemotherapy drug using an intravenous push technique |
| Adrenergic agents, catecholamines | Nerve stimulating hormones and drugs are prescribed |
| Bicarbonate producing/containing agents | Bicarbonate producing/containing agents are prescribed |
| Hosp N | Number of days of previous hospitalisation |
| LABS: ALB | Albumin values through laboratory test |
| LABS: C199 | Amount of a protein called CA 19-9 (cancer antigen 19-9) |
| LABS: CL | Amount of chloride in blood |
| LABS: HCT | Percentage of red blood cells in blood |
| LABS: URIC | Amount of uric acid in blood or urine |

Table A.2: Explanation of credible variables for the Horseshoe-MH model. The descriptions for the procedures (PROC) are taken from https://www.aapc.com/codes.

# Appendix B

# Supplementary Experiment II

## B.1 Additional Results

In Figure B.1 we show the DCA plots for the net benefit of the ACU prediction models for the second experiment on 30-day and 365-day acute care use prediction. we observe on 30-day prediction (Figure B.1a) that both BERT models yield negative net benefit, if the decision threshold is over 0.38. On 365-day prediction, we observe a negative net benefit for the Language BERT for decision thresholds over 0.6.
Figure B.2 shows the word importance of the Language LASSO for 30-day and 365-day ACU prediction. Finally, in Figure B.3, we report the empirical cumulative risk for patients stratified by gender, ethnicity and cancer type. We observe some risk overestimation for Hispanic/Latino patients (Figure B.3a). In Subfigure B.3c, we observe overestimation for Sarcoma tumours and underestimation for prostate tumours.

| Label | Model | No. Vocabulary | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|---|
| 30 | Language LASSO | 500 | 0.696 (0.677, 0.715) | 0.269 (0.242, 0.297) | 0.372 (0.355, 0.389) | **<0.001** (<0.001, 0.021) |
| | | 1000 | 0.698 (0.680, 0.718) | 0.294 (0.265, 0.327) | 0.370 (0.352, 0.387) | 0.019 (0.000, 0.043) |
| | | 2000 | **0.726** (0.707, 0.744) | 0.294 (0.264, 0.323) | **0.363** (0.346, 0.379) | **<0.001** (<0.001, 0.021) |
| | | 3000 | 0.717 (0.697, 0.734) | **0.296** (0.264, 0.326) | 0.365 (0.348, 0.382) | **<0.001** (<0.001, 0.023) |
| 180 | Language LASSO | 500 | 0.705 (0.692, 0.720) | 0.547 (0.523, 0.570) | 0.571 (0.558, 0.584) | **<0.001** (<0.001, 0.044) |
| | | 1000 | 0.719 (0.705, 0.734) | 0.573 (0.551, 0.597) | 0.562 (0.549, 0.574) | 0.013 (0.000, 0.047) |
| | | 2000 | 0.730 (0.717, 0.745) | 0.577 (0.555, 0.601) | 0.558 (0.546, 0.570) | **<0.001** (<0.001, 0.034) |
| | | 3000 | **0.734** (0.721, 0.748) | **0.584** (0.561, 0.607) | **0.555** (0.542, 0.566) | **<0.001** (<0.001, 0.028) |
| 365 | Language LASSO | 500 | 0.716 (0.703, 0.729) | 0.624 (0.603, 0.647) | 0.596 (0.586, 0.605) | **<0.001** (<0.001, 0.022) |
| | | 1000 | 0.718 (0.705, 0.732) | 0.635 (0.616, 0.657) | 0.592 (0.581, 0.602) | 0.025 (0.000, 0.041) |
| | | 2000 | 0.732 (0.719, 0.745) | 0.639 (0.618, 0.661) | 0.585 (0.575, 0.595) | **<0.001** (<0.001, 0.022) |
| | | 3000 | **0.742** (0.730, 0.755) | **0.657** (0.637, 0.678) | **0.576** (0.567, 0.586) | **<0.001** (<0.001, 0.039) |

Table B.1: Resulting metrics on the test set of the different vocabulary sizes for the Language LASSO trained on 30, 180 and 365 days ACU prediction. The best-performing metrics for every label type are marked in bold. We report the 95%-confidence intervals of the metric estimates in the brackets (2.5%-CI, 97.5%-CI).

| Label | Model | Output | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|---|
| 30 | Language BERT | Cumulative Link | **0.710** (0.692, 0.729) | **0.259** (0.235, 0.282) | 0.435 (0.415, 0.455) | 0.131 (0.117, 0.145) |
| | | Single Labels | 0.638 (0.616, 0.659) | 0.209 (0.189, 0.230) | **0.390** (0.372, 0.408) | **0.006** (0.000, 0.027) |
| 180 | Language BERT | Cumulative Link | **0.702** (0.688, 0.717) | **0.543** (0.517, 0.567) | 0.625 (0.603, 0.644) | **0.107** (0.093, 0.119) |
| | | Single Label | 0.665 (0.650, 0.680) | 0.494 (0.470, 0.516) | **0.620** (0.605, 0.633) | 0.103 (0.088, 0.116) |
| 365 | Language BERT | Cumulative Link | **0.709** (0.695, 0.723) | **0.617** (0.594, 0.640) | 0.666 (0.647, 0.683) | 0.135 (0.122, 0.148) |
| | | Single Label | 0.681 (0.667, 0.696) | 0.593 (0.571, 0.614) | **0.621** (0.610, 0.631) | **<0.001** (<0.001, 0.036) |

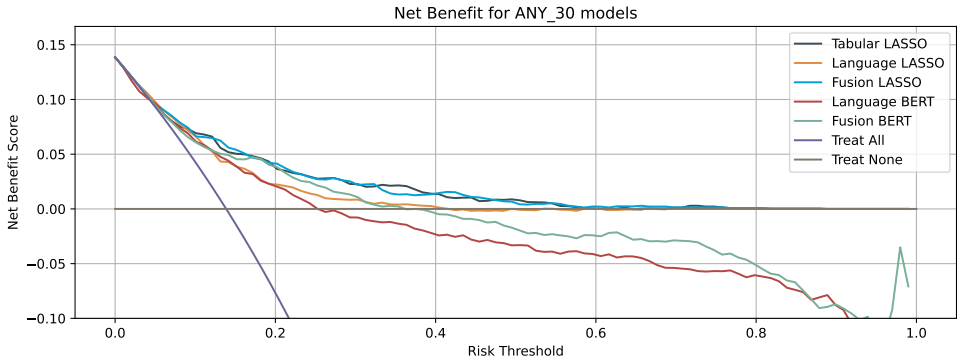
Table B.2: Resulting metrics on the comparison of the single sigmoid output network (trained three times individually) and the ordinal regression output with the modified cumulative link layer (trained on the three times simultaneously). The best-performing metrics for every label type are marked in bold. We report the 95%-confidence intervals of the metric estimates in the brackets (2.5%-CI, 97.5%-CI).

| Label | Model | Encoder | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|---|
| 30 | Language BERT | distilBERT | **0.710** (0.692, 0.729) | 0.259 (0.235, 0.282) | 0.435 (0.415, 0.455) | 0.131 (0.117, 0.145) |
| | | ClinicalBERT | **0.710** (0.691, 0.730) | **0.261** (0.238, 0.286) | 0.427 (0.414, 0.440) | 0.139 (0.128, 0.153) |
| | | LongFormer | 0.659 (0.639, 0.678) | 0.212 (0.192, 0.231) | **0.389** (0.372, 0.405) | **0.030** (0.000, 0.045) |
| 180 | Langage BERT | distilBERT | 0.702 (0.688, 0.717) | 0.543 (0.517, 0.567) | 0.625 (0.603, 0.644) | 0.107 (0.093, 0.119) |
| | | ClinicalBERT | **0.709** (0.695, 0.725) | **0.547** (0.523, 0.571) | 0.611 (0.598, 0.623) | 0.118 (0.104, 0.131) |
| | | LongFormer | 0.674 (0.661, 0.689) | 0.500 (0.476, 0.524) | **0.600** (0.586, 0.613) | **0.035** (0.009, 0.059) |
| 365 | Language BERT | distilBERT | 0.709 (0.695, 0.723) | **0.617** (0.594, 0.640) | 0.666 (0.647, 0.683) | 0.135 (0.122, 0.148) |
| | | ClinicalBERT | **0.713** (0.699, 0.727) | 0.615 (0.593, 0.638) | **0.614** (0.602, 0.625) | **0.048** (0.025, 0.064) |
| | | LongFormer | 0.679 (0.665, 0.694) | 0.573 (0.551, 0.597) | 0.643 (0.632, 0.653) | 0.090 (0.077, 0.109) |

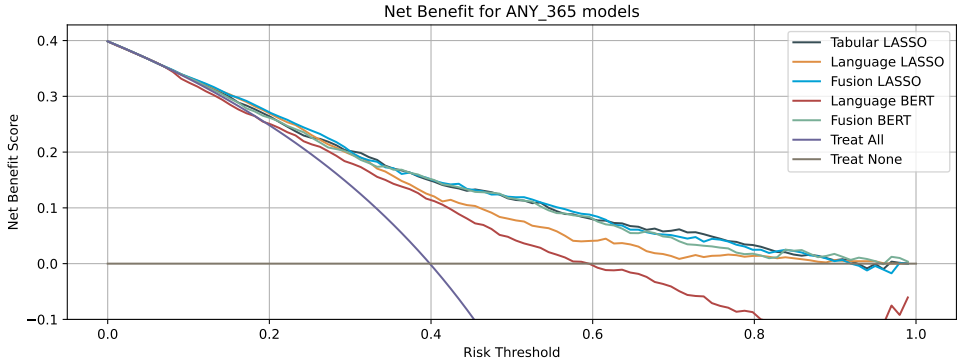
Table B.3: Resulting metrics on the test set of the various transformer encoders for language BERT trained on 30, 180 and 365 days ACU prediction. The best-performing metrics for every label type are marked in bold. We report the 95%-confidence intervals of the metric estimates in the brackets (2.5%-CI, 97.5%-CI).

| Label | Model | Fusion Mechanism | AUROC | AUPRC | Log-Loss | ECE |
|---|---|---|---|---|---|---|
| 30 | Fusion BERT | Concatenation | **0.766** (0.749, 0.784) | **0.315** (0.286, 0.343) | **0.393** (0.377, 0.406) | **0.103** (0.089, 0.116) |
| | | Cross-Attention | 0.757 (0.740, 0.775) | 0.310 (0.280, 0.341) | 0.460 (0.436, 0.482) | 0.158 (0.143, 0.173) |
| 180 | Fusion BERT | Concatenation | 0.753 (0.741, 0.767) | 0.620 (0.597, 0.644) | **0.548** (0.536, 0.558) | **0.038** (0.023, 0.059) |
| | | Cross-Attention | **0.754** (0.741, 0.768) | **0.629** (0.606, 0.651) | 0.560 (0.542, 0.575) | 0.066 (0.055, 0.082) |
| 365 | Fusion BERT | Concatenation | **0.760** (0.748, 0.774) | 0.695 (0.675, 0.714) | **0.565** (0.554, 0.575) | **0.021** (<0.001, 0.041) |
| | | Cross-Attention | 0.759 (0.747, 0.773) | **0.697** (0.677, 0.715) | 0.591 (0.576, 0.605) | 0.087 (0.073, 0.103) |

Table B.4: Resulting metrics on the test set of the various multimodal fusion strategies for the fusion BERT trained on 30, 180 and 365 days ACU prediction. The best-performing metrics for every label type are marked in bold. We report the 95%-confidence intervals of the metric estimates in the brackets (2.5%-CI, 97.5%-CI).

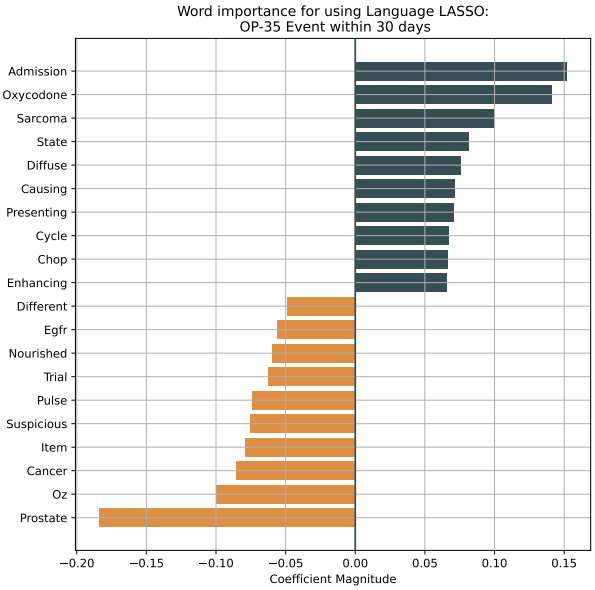



(a) 30-day ACU prediction

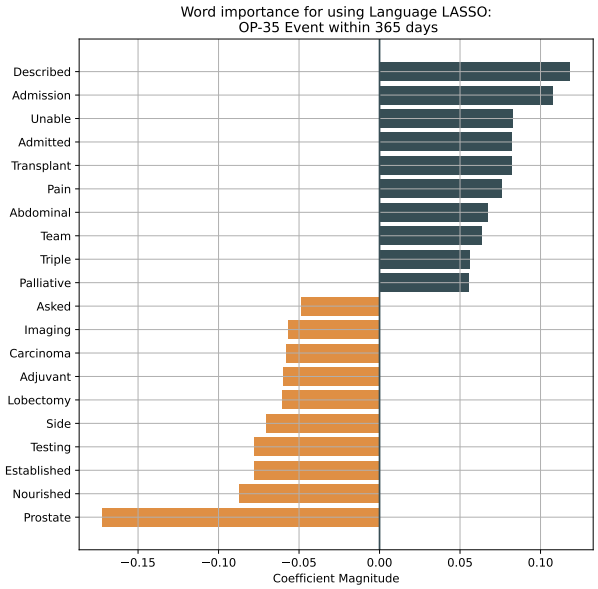



(b) 365-day ACU prediction

Figure B.1: Net benefit curves of the tabular, language and fusion models for 30-day and 365-day prediction. The purple curve indicates the benefit of all the patients treated, whereas the grey curve indicates the benefit is no patient is treated

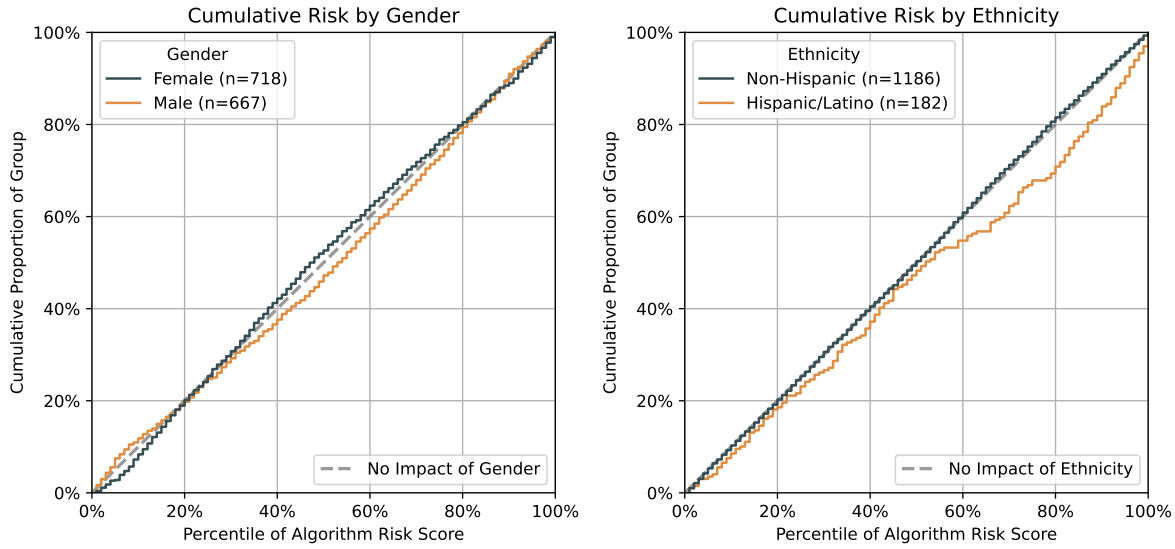



(a) 30-day ACU prediction




(b) 365-day ACU prediction

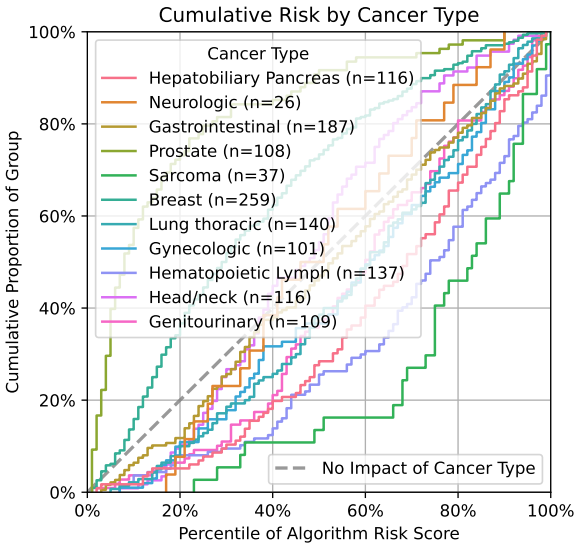
Figure B.2: Coefficient magnitudes for the Language LASSO for 30-day and 365-da ACU prediction, displaying the ten highest and ten lowest. The coefficients in this model are single words found in the clinical notes before the ACU event.




(a) Cumulative risk by race




(b) Cumulative risk by insurance




(c) Cumulative risk by cancer stage

Figure B.3: Cumulative risk of the Language LASSO on 180 ACU prediction, stratified by gender (a), ethnicty (b) and over cancer type (c).

# Bibliography


[1] Dylan J. Peterson, Nicolai P. Ostberg, Douglas W. Blayney, James D. Brooks, and Tina Hernandez-Boussard. Machine learning applied to electronic health records: Identification of chemotherapy patients at high risk for preventable emergency department visits and hospital admissions. *JCO Clinical Cancer Informatics*, (5):1106–1126, 2021. doi: 10.1200/CCI.21.00116. URL https://doi.org/10.1200/CCI.21.00116. PMID: 34752139.

[2] Kerin B. Adelson, Vanna Dest, Salimah Velji, Richard Lisitano, and Rogerio Lilenbaum. Emergency department (ed) utilization and hospital admission rates among oncology patients at a large academic center and the need for improved urgent care access. *Journal of Clinical Oncology*, 32 (30_suppl):19–19, 2014. doi: 10.1200/jco.2014.32.30_suppl.19. URL https://doi.org/10.1200/jco.2014.32.30_suppl.19. PMID: 28141471.

[3] Hajime Uno, Joseph O. Jacobson, and Deborah Schrag. Clinician assessment of potentially avoidable hospitalization in patients with cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 32 30_suppl:4, 2014.

[4] Gabriel A Brooks, Ling Li, Hajime Uno, Michael J Hassett, Bruce E Landon, and Deborah Schrag. Acute hospital care is the chief driver of regional spending variation in medicare patients with advanced cancer. *Health Aff. (Millwood)*, 33(10):1793–1800, October 2014.

[5] K Robin Yabroff, Elizabeth B Lamont, Angela Mariotto, Joan L Warren, Marie Topor, Angela Meekins, and Martin L Brown. Cost of care for elderly cancer patients in the united states. *J. Natl. Cancer Inst.*, 100(9):630–641, May 2008.

[6] Gabriel A. Brooks, Ankit J. Kansagra, Sowmya R. Rao, James I. Weitzman, Erica A. Linden, and Joseph O. Jacobson. A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy. *JAMA Oncology*, 1(4):441–447, 07 2015. ISSN 2374-2437. doi: 10.1001/jamaoncol.2015.0828. URL https://doi.org/10.1001/jamaoncol.2015.0828.

[7] Robert C. Grant, Rahim Moineddin, Zhan Yao, Melanie Powis, Vishal Kukreti, and Monika K. Krzyzanowska. Development and validation of a score to predict acute care use after initiation of systemic therapy for cancer. *JAMA Network Open*, 2, 2019.

[8] Robert M Daly, Dmitriy Gorenshteyn, Lior Gazit, Stefania Sokolowski, Kevin Nicholas, Claire Perry, Lynn Adams, Abigail Baldwin, Jessie Holland, Alice Zervoudakis, Han Xiao, Rori Salvaggio, Yeneat O. Chiu, Lauren Katzen, Margarita Rozenshteyn, Diane L. Reidy, Brett A. Simon, Wendy Perchick, and Isaac Wagner. A framework for building a clinically relevant risk model. *Journal of Clinical Oncology*, 2019.

[9] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Med.*, 4(1):4, January 2021.

[10] Office of the national coordinator for health information technology: National trends in hospital and physician adoption of electronic health records. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records.

[11] Lukas Golder, Tatjana Grez, Thomas Burgunder, and Roland Rey. *Swiss eHealth Barometer 2022: Bericht zur Befragung der Gesundheitsfachpersonen und Akteure des Gesundheitswesens*. 2022.

[12] Gabriel A. Brooks, Hajime Uno, Erin J Aiello Bowles, Alex R. Menter, Maureen O'Keeffe-Rosetti, Anna N.A. Tosteson, Debra P. Ritzwoller, and Deborah Schrag. Hospitalization risk during chemotherapy for advanced cancer: Development and validation of risk stratification models using real-world data. *JCO clinical cancer informatics*, 3:1–10, 2019.

[13] 2019 chemotherapy measure facts admissions and emergency department (ED) visits for patients receiving outpatient chemotherapy hospital outpatient quality reporting (OQR) program (op-35). https://qualitynet.cms.gov/files/5dcc6762a3e7610023518e23?filename=CY21_OQRChemoMsr_FactSheet.pdf.

[14] Bradley P Carlin, Hwanhee Hong, Tatyana A Shamliyan, François Sainfort, and Robert L Kane. *Case Study Comparing Bayesian and Frequentist Approaches for Multiple Treatment Comparisons*. Agency for Healthcare Research and Quality (US), Rockville (MD), March 2013.

[15] Arianna Dagliati, Alberto Malovini, Pasquale Decata, Giulia Cogni, Marsida Teliti, Lucia Sacchi, Carlo Cerra, Luca Chiovato, and Riccardo Bellazzi. Hierarchical bayesian logistic regression to forecast metabolic control in type 2 DM patients. *AMIA Annu. Symp. Proc.*, 2016:470–479, 2016.

[16] Wiktor Beker, Agnieszka Wołos, Sara Szymkuć, and Bartosz A Grzybowski. Minimal-uncertainty prediction of general drug-likeness based on bayesian neural networks. *Nat Mach Intell*, 2(8): 457–465, August 2020.

[17] Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916115. doi: 10.1145/168304.168306. URL https://doi.org/10.1145/168304.168306.

[18] Prasoon Joshi and Riddhiman Dhar. EpICC: A bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci. Rep.*, 12(1):14628, August 2022.

[19] Charlotte Syrykh, Arnaud Abreu, Nadia Amara, Aurore Siegfried, Véronique Maisongrosse, François X Frenois, Laurent Martin, Cédric Rossi, Camille Laurent, and Pierre Brousset. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. *NPJ Digit. Med.*, 3(1):63, May 2020.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

[21] Christian F. Baumgartner, Kerem Can Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio F. Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *MICCAI*, 2019.

[22] Lotta Meijerink, Giovanni Ciná, and Michele Tonutti. Uncertainty estimation for classification and risk prediction in medical settings. *ArXiv*, abs/2004.05824, 2020.

[23] Dae Y Kang, Pamela N DeYoung, Justin Tantiongloc, Todd P Coleman, and Robert L Owens. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *NPJ Digit. Med.*, 4(1):142, September 2021.

[24] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656, 2001.

[25] Jakob F Mathiszig-Lee, Finneas J R Catling, S Ramani Moonesinghe, and Stephen J Brett. Highlighting uncertainty in clinical risk prediction using a model of emergency laparotomy mortality risk. *NPJ Digit. Med.*, 5(1):70, June 2022.

[26] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc.*, 27(3):457–470, March 2020.

[27] Ben J Marafino, Miran Park, Jason M Davies, Robert Thombley, Harold S Luft, David C Sing, Dhruv S Kazi, Colette DeJong, W John Boscardin, Mitzi L Dean, and R Adams Dudley. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw. Open*, 1(8):e185097, December 2018.

[28] Iraklis A Klampanos. Manning christopher, prabhakar raghavan, hinrich schütze: Introduction to information retrieval. *Inf. Retr. Boston.*, 12(5):609–612, October 2009.

[29] Ben J Marafino, Jason M Davies, Naomi S Bardach, Mitzi L Dean, and R Adams Dudley. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *J. Am. Med. Inform. Assoc.*, 21(5):871–875, September 2014.

[30] W W Chapman, W Bridewell, P Hanbury, G F Cooper, and B G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.*, 34(5): 301–310, October 2001.

[31] Tak Sung Heo, Yu Seop Kim, Jeong Myeong Choi, Yeong Seok Jeong, Soo Young Seo, Jun Ho Lee, Jin Pyeong Jeon, and Chulho Kim. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J. Pers. Med.*, 10(4):286, December 2020.

[32] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://www.aclweb.org/anthology/W19-1909`.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. URL `https://arxiv.org/pdf/1706.03762.pdf`.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[35] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*, 2019.

[36] Ashish Sarraju, Jean Coquet, Alban Zammit, Antonia Chan, Summer Ngo, Tina Hernandez-Boussard, and Fatima Rodriguez. Using deep learning-based natural language processing to identify reasons for statin nonuse in patients with atherosclerotic cardiovascular disease. *Communications Medicine*, 2, 2022.

[37] Joseph Gatto, Parker Seegmiller, Garrett Johnston, and Sarah Masud Preum. Identifying the perceived severity of patient-generated telemedical queries regarding COVID: Developing and evaluating a transfer learning-based solution. *JMIR Med. Inform.*, 10(9):e37770, September 2022.

[38] Yarin Gal, Petros Koumoutsakos, Francois Lanusse, Gilles Louppe, and Costas Papadimitriou. Bayesian uncertainty quantification for machine-learned models in physics. *Nat Rev Phys*, 4(9): 573–577, August 2022.

[39] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *CoRR*, abs/1703.04977, 2017. URL `http://arxiv.org/abs/1703.04977`.

[40] Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon T. Willard. The horseshoe+ estimator of ultra-sparse signals. *arXiv: Statistics Theory*, 2015.

[41] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/carvalho09a.html`.

[42] Martin J. Wainwright and Michael Irwin Jordan. Graphical models, exponential families, and variational inference. [electronic resource] / wainwright, martin j. 2008.

[43] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859 – 877, 2016.

[44] N. Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, A. H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[45] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL `https://doi.org/10.1093/biomet/57.1.97`.

[46] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41, 1992.

[47] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[48] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, 2016.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[50] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[52] Jürgen Schmidhuber. Deep learning in neural networks: an overview. *Neural Netw.*, 61:85–117, January 2015.

[53] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

[54] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18:14:1–14:45, 2017.

[55] Tomi Peltola, Aki S. Havulinna, Veikko Salomaa, and Aki Vehtari. Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop - Volume 1218*, BMAW'14, page 79–88, Aachen, DEU, 2014. CEUR-WS.org.

[56] Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *AISTATS*, 2017.

[57] Juho Piironen and Aki Vehtari. Projection predictive variable selection using stan+r. *arXiv: Methodology*, 2015.

[58] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143874. URL `https://doi.org/10.1145/1143844.1143874`.

[59] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

[60] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[61] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.*, 74:167–176, June 2016.

[62] Peter C Austin and Ewout W Steyerberg. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat. Med.*, 33(3):517–535, February 2014.

[63] Andrew Julian Vickers, Ben van Calster, and Ewout Willem Steyerberg. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*, 3, 2019.

[64] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[65] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

[66] Swaraj Khadanga, Karan Aggarwal, Shafiq R. Joty, and Jaideep Srivastava. Using clinical notes with time series data for ICU management. *ArXiv*, abs/1909.09702, 2019.

[67] Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings 2021*, (in press, n.d.). URL http://arxiv.org/abs/2106.07799.

[68] Fabian Pedregosa, Francis R. Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *J. Mach. Learn. Res.*, 18:55:1–55:35, 2017.

[69] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019.

[70] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[71] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 7 2019. Association for Computational Linguistics.

[72] Ethan Rosenthal. Spacecutter: ordinal regression models in pytorch, Dec 2018. URL https://www.ethanrosenthal.com/2018/12/06/spacecutter-ordinal-regression.

[73] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[75] E L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53(282):457, June 1958.

[76] Abolfazl Hosseinnataj, Mohammad RezaBaneshi, and Abbas Bahrampour. Mortality risk factors in patients with gastric cancer using bayesian and ordinary lasso logistic models: a study in the southeast of iran. *Gastroenterol. Hepatol. Bed Bench*, 13(1):31–36, 2020.

[77] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/salimans15.html.

[78] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*, 2(9):e489–e492, September 2020.

[79] Tina Hernandez-Boussard, Selen Bozkurt, John P. A. Ioannidis, and Nigam Haresh Shah. Minimar (minimum information for medical ai reporting): Developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association : JAMIA*, 27: 2011 – 2015, 2020.

[80] Matthew D. Homan and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, jan 2014. ISSN 1532-4435.

[81] Steven R Steinhubl and Eric J Topol. Digital medicine, on its way to being just plain medicine. *NPJ Digit. Med.*, 1(1):20175, January 2018.

[82] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[83] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[84] John Salvatier, Thomas V. Wiecki, and Christopher J Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Comput. Sci.*, 2:e55, 2016.

[85] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: state-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[86] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.