# DEVELOPING CLINICAL PREDICTION MODELS

by

**CHAN WEI XIN**
*(B.Sc Biochemistry, University College London)*

**A THESIS SUBMITTED FOR THE DEGREE OF**

**DOCTOR OF PHILOSOPHY**

in

**COMPUTER SCIENCE**

in the

**SCHOOL OF COMPUTING**

of the

**NATIONAL UNIVERSITY OF SINGAPORE**

**2023**

Thesis Advisor:
Professor Wong Limsoon

Examiners:
Professor Wynne Hsu
Professor Zhang Louxin

# Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Chan Wei Xin

31 March 2023

*For my parents and wife Kavita,*
*without whom I would not be who I am today.*

## Acknowledgements

I would like to express my deepest gratitude to my thesis advisor Prof. Wong Limsoon, for his time and effort in guiding me through my Ph.D. journey. I have learnt tremendously under his guidance, and am grateful for his patience in discussing research ideas. I am truly grateful to Prof. Wilson Goh, for his invaluable advice, encouragement and useful feedback. I am also appreciative of him for providing us with the Westlake proteomics data set.

Many thanks to past and present members of Prof. Wong's group, Neamul Kabir, Xie Luyu, Stefano Perna, Lakshmi Alagappan, Xu Weinan, Jin Wenhao and Yong Chern Han for the interesting discussions and for all the fun times that we had. I would also like to sincerely thank my lab mates, Ramesh, Rizki, Fatir, Haroon and Siddharth for the laughter that we shared and the useful advice that they have provided along this journey. I am also grateful for my friends, who have provided much fun and laughter in my life, and who have been a constant source of precious memories.

Special thanks to my collaborators from NTU, Kai Peng, Harvard, and Ser Xian, for their indispensable help and useful discussions. I am also grateful to Prof. Allen Yeoh, Li Zhenhua and Chiew Kean Hui for their help, and for providing extra metadata on the Ma-Spore ALL data set.

Finally, I thank my parents, who have supported and encouraged me in this journey, and who have provided me with more than I could ever wish for in life. Last but not least, I would like to thank my wife, Kavita, for being my companion in life, and for all the fun, laughter, encouragement and good advice that she has given me throughout this journey.

# Contents

**Acknowledgements** **ii**

**Abstract** **vii**

**List of Figures** **ix**

**List of Tables** **xi**

**1 Introduction** **1**

1.1 Motivation . . . . . 3

1.2 Scope . . . . . 5

1.3 Thesis contributions and organisation . . . . . 5

**2 Obstacles to effective model deployment** **8**

2.1 Introduction . . . . . 8

2.2 Poor reproducibility and replicability . . . . . 9

2.3 Unfairness in prediction models . . . . . 11

2.4 Underlying causes . . . . . 12

2.4.1 Improper development of models . . . . . 13

2.4.2 Improper evaluation of models . . . . . 15

2.5 Key takeaways . . . . . 16

**3 RVP: Quantifying batch effects** **18**

3.1 Introduction . . . . . 18

3.2 Background . . . . . 20

3.2.1 Characteristics of batch effects . . . . . 20

3.3 Related work . . . . . 21

3.3.1 gPCA $\delta$ . . . . . 21

3.3.2 PVCA . . . 22
3.4 Recursive variance partitioning . . . 23
3.4.1 Case 1: Samples belong to the same class . . . 24
3.4.2 Case 2: Samples belong to different classes . . . 25
3.5 Methods . . . 26
3.5.1 Data sets . . . 26
3.5.2 Runtime and memory analysis . . . 30
3.6 Results . . . 31
3.6.1 Simulated RNA-seq count data . . . 31
3.6.2 Microarray and proteomics data . . . 40
3.7 Discussion . . . 42

**4 Accounting for treatment differences 44**
4.1 Introduction . . . 44
4.2 Methods in handling treatment differences . . . 45
4.2.1 Ignoring treatment . . . 46
4.2.2 Restricting analysis . . . 46
4.2.3 Composite outcome . . . 47
4.2.4 Modelling treatment . . . 47
4.2.5 Hypothetical prediction . . . 47
4.3 Prediction estimands . . . 48
4.4 Case study: Ma-Spore ALL data set . . . 48
4.5 Model evaluation: Incorporating treatment information . . . 49
4.6 Benefits of incorporating treatment information . . . 52
4.7 Handling complex treatment differences . . . 53
4.8 Closing remarks . . . 54

**5 Subtype-specific treatment outcome prediction 56**
5.1 Introduction . . . 56
5.1.1 Contributions . . . 59
5.2 Background . . . 60
5.2.1 Acute lymphoblastic leukaemia . . . 60
5.2.2 Related work . . . 63

5.3 Subtype-specific model . . . . . 67
5.3.1 Probe set selection . . . . . 68
5.3.2 Transcriptomic features . . . . . 69
5.3.3 Estimation of probability of long-term remission . . . . . 71
5.4 Methods . . . . . 72
5.4.1 Ma-Spore ALL data set . . . . . 72
5.4.2 Data preprocessing . . . . . 73
5.4.3 Ma-Spore ALL data set: Train-test split . . . . . 74
5.4.4 Receiver operating characteristic analysis . . . . . 75
5.4.5 Other methods . . . . . 75
5.4.6 Statistical analysis . . . . . 76
5.5 Results . . . . . 77
5.5.1 Probe set selection mitigates batch effects . . . . . 77
5.5.2 Transcriptomic features are predictive of treatment outcome in homogeneous subtypes . . . . . 80
5.5.3 Prediction probabilities from the subtype-specific model are significantly associated with treatment outcome in homogeneous subtypes . . . . . 81
5.5.4 Subtype-specific model outperforms other methods in treatment outcome prediction . . . . . 85
5.5.5 Subtype-specific model identifies low risk patients with high precision . . . . . 88
5.5.6 Analysis of treatment intensity recommendations . . . . . 90
5.5.7 Subtype-specific model outperforms MRD in treatment outcome prediction on novel DUX4-rearranged subtype . . . . . 91
5.5.8 Validation of biological hypothesis . . . . . 94
5.6 Discussion . . . . . 99
5.7 Conclusion . . . . . 100

**6 Conclusion** **102**
6.1 Future work . . . . . 104

**Bibliography** **105**

**Publications during PhD Study** **120**

# Abstract

Clinical prediction models are developed to estimate the absolute risk of clinically important outcomes in patients. These models are often designed with the purpose of guiding clinical decision making. Recently, there has been a deluge of publications regarding clinical prediction models due to the resurgence of interest in artificial intelligence. However, very few of these models end up being deployed in the real-world.

In this thesis, we discuss the main obstacles facing effective model deployment in healthcare. We identify improper development and evaluation of clinical prediction models as the principal cause behind some of the main obstacles. Clinical prediction models are particularly susceptible to improper development and evaluation due to the inherent heterogeneities in clinical data. We discuss two of the most prevalent heterogeneities in clinical data in further detail in this thesis.

Batch effects are a common heterogeneity in high-dimensional biological data, such as gene expression microarray data. Failure to properly account for batch effects during the development of prediction models often leads to poor generalisation ability. Very few quantitative batch effects metrics have been proposed for use in small data sets. The accuracy of these metrics are reduced when used to quantify batch effects in data where different batches contain different class proportions. We propose recursive variance partitioning (RVP), a novel metric for quantifying batch effects. We show that RVP is able to accurately estimate the proportion of total variance attributable to batch effects in data, over a range of magnitudes of batch effects. RVP exhibits similar performance even in data with severe batch-class imbalance.

Another common heterogeneity that exists in clinical data is when it is made up of patients who receive different treatments. This heterogeneity complicates model development and evaluation as different patient treatments affect patient outcome, which is often the prediction target, to varying degrees. We propose a scoring scheme for use in evaluating clinical prediction models, which incorporates patient treatment information. We use the Malaysia-Singapore acute lymphoblastic

leukaemia (ALL) data set as a case study to demonstrate the use of the proposed scoring scheme. Evaluating models in this manner would help to avoid errors that arise due to treatment differences.

In this thesis, we develop a subtype-specific prediction model for treatment outcome in ALL patients. Our subtype-specific model incorporates the use of transcriptomic features engineered from patient gene expression profiles (GEPs) at different time-points of treatment. Our model outperforms other methods in classifying patients who achieved continuous complete remission and patients who relapsed, in homogeneous ALL subtypes. Our subtype-specific model is designed to be robust to small sample sizes. We also present the biological hypothesis behind our model: GEPs measure the average gene expression of all leukaemic and normal cells in a sample, and patients who are more responsive towards treatment will experience a faster decrease in proportion of leukaemic cells and hence exhibit a greater shift in their GEP. We validate the hypothesis by estimating B-cell abundance in patient samples using various methods.

# List of Figures


3.1 Plots of the measured percentage of batch effects in the simulated RNA-seq data sets against the theoretical percentage of batch effects computed from data simulation hyper-parameters. . . . . 34
3.2 Plots of different batch effects metrics against the theoretical percentage of batch effects. . . . . 35
3.3 Plots of the measured variance due to batch effects against the theoretical estimate of variance due to batch effects in simulated RNA-seq data sets 37
3.4 Plots of batch effects variance estimated by the different batch effects metrics, against the theoretical batch effects variance computed from the hyper-parameters used to simulate the data. . . . . 38
3.5 Comparison of runtime and peak memory usage of batch effects metrics. 40

4.1 Risk-adapted treatment plan employed in Ma-Spore acute lymphoblastic leukaemia 2003 and 2010 studies. . . . . 50

5.1 Illustration of the effective response metric ratio, absolute response metric ratio and reorientation ratio. . . . . 70
5.2 Three-dimensional PCA plot of the Ma-Spore ALL data set. . . . . 77
5.3 Top four principal components of TEL-AML1 samples from the Ma-Spore ALL data set before and after probe set selection. . . . . 78
5.4 PCA plot of transcriptomic features (viz. ERM ratio, ARM ratio and $\phi$) of TEL-AML1 patients from the Ma-Spore ALL data set. . . . . 80
5.5 Distribution of features among long-term remission and relapse patients for each individual subtype model in the Ma-Spore ALL data set. . . . 83

5.6 Distribution of prediction probabilities and scatter plot of log-transformed minimal residual disease against prediction probabilities estimated by the subtype-specific model. . . . . 84
5.7 Receiver operating characteristic analysis comparing treatment outcome prediction performance. . . . . 87
5.8 Kaplan-Meier survival curves of patients belonging to homogeneous subtypes in the Ma-Spore ALL data set. . . . . 89
5.9 Frequency of scores awarded to two different treatment intensity prediction models. . . . . 91
5.10 Identification of DUX4-rearranged samples in the Ma-Spore ALL data set. 93
5.11 Quantity of B-lineage cell population in GEPs from the Ma-Spore ALL data set. . . . . 95
5.12 Gene expression of CD19 and average gene expression of CD19, CD38, CD72, CD79A and CD79B at multiple time points after treatment. . . 97
5.13 Gene expression of CD19 and average gene expression of CD19, CD38, CD72, CD79A and CD79B of paired Day 0 and Day 8 samples. . . . . 98

# List of Tables


3.1 Distribution of samples across batches and classes in simulated RNA-seq count data without and with batch-class imbalance. . . . . 27
3.2 Distribution of samples across batches and classes in four different subsamples of the MAQC-I microarray data set. . . . . 29
3.3 Distribution of samples across batches and classes in four different subsamples of the Ma-Spore ALL microarray data set. . . . . 30
3.4 Distribution of samples across batches and classes in five different subsamples of the Westlake proteomics data set. . . . . 31
3.5 Values of batch effects metrics when used to quantify batch effects in different subsets of the Ma-Spore ALL data set, MAQC-I data set and Westlake proteomics data set. . . . . 41

4.1 Scoring scheme for treatment intensity recommendations for paediatric ALL patients. . . . . 51

5.1 Distribution of patients across subtypes and treatment outcomes in the Ma-Spore ALL data set. . . . . 74
5.2 RVP of data before and after probe set selection across all subtypes. . . 79

# Chapter 1

# Introduction

Advances in high-throughput technologies and the rapid digitalisation of healthcare have led to an explosion in the amount of medical data freely available for research. The increased availability of medical data, coupled with recent advancements in machine learning, have led to a resurgence of interest in the use of predictive modelling in healthcare. This has resulted in an exponential increase in the number of publications on prediction models in healthcare (Bohr & Memarzadeh, 2020; Guo et al., 2020; Weissler et al., 2021). However, despite the copious publications on prediction models in healthcare, the number of models deployed in clinical practice still remains fairly limited (Perel et al., 2006; Wyatt & Altman, 1995).

The central theme of this thesis is the development of clinical prediction models. In this thesis, we discuss the main obstacles that impede the deployment of clinical prediction models in real life, and their underlying causes. Clinical prediction models are especially susceptible to improper development, due in large part to inherent heterogeneities in clinical data. Failure to account for certain heterogeneities leads to improper model development; it is imperative to identify and handle these heterogeneities. In particular, we discuss in greater detail two common heterogeneities in clinical data that impede proper model development: batch effects and patient treatment differences.

Batch effects are systematic errors in measurements between different batches of samples, and are an extremely prevalent problem in many different types of high-dimensional biological data (Goh et al., 2017; Leek et al., 2010). Failure to account for batch effects in data used to develop clinical prediction models often leads to models that show poor generalisation performance. In this thesis, we propose a quantitative batch effects metric that is able to accurately estimate the proportion

of total variance attributable to batch effects, even in data where different batches have different proportions of classes (i.e. data with batch-class imbalance). We term this metric recursive variance partitioning (RVP). We believe that RVP can serve as a tool to help modelling practitioners handle batch effects in their data. For example, RVP can assist practitioners in deciding which batch effects correction method would best mitigate batch effects.

In many diseases, risk-adapted treatment strategies are used to treat patients, where patients receive different treatments based on their individual risk level. Developing prediction models where patient outcome is the prediction target is problematic when using data where patients receive different treatments. This is because different treatments have different impacts on patient outcome. Failure to deal with patient treatment differences when developing prediction models results in inaccurate models that are unable to generalise (Groenwold et al., 2016; Sperrin et al., 2018). There are many nuances when attempting to account for differences in patient treatment; we discuss some of the subtleties in this thesis. We also propose a scoring scheme that incorporates patient treatment information in order to achieve more detailed and accurate evaluation of prediction models.

Acute lymphoblastic leukaemia (ALL) is a cancer that affects blood cells that originate from lymphoid progenitor cells. ALL is broadly classified according to the type of lymphocyte that has become malignant, namely B-cell precursors, T-cells, and mature B-cells. In addition to the above classification, ALL can be classified into different subtypes based on the type of chromosomal or genetic aberration that can be observed in leukaemic blasts. In this thesis, we propose a subtype-specific model that predicts treatment outcome in paediatric acute lymphoblastic leukaemia (ALL) patients, which incorporates the use of transcriptomic features computed from microarray gene expression data. During the development of our model, we utilised our proposed batch effects metric RVP to ascertain that our feature selection and feature engineering methodology had sufficiently mitigated batch effects in the data. As patients in the paediatric ALL data set received different treatment intensities, we use our proposed scoring scheme to evaluate treatment recommendations given by our subtype-specific prediction model. The use of our proposed scoring scheme, which incorporates the use of patient treatment information, allows us to achieve a more nuanced and accurate evaluation of predictions from our model.

## 1.1 Motivation

Despite the massive increase in research on machine learning applications in healthcare, the number of predictive models that are adopted in practice still remain fairly limited (Mateen et al., 2020; Steyerberg et al., 2013). In a recent review by Wynants et al., 2020, the authors found that most COVID-19 diagnosis models that were proposed in preprint or published articles suffered from high risk of bias, opaque reporting and reported performance measures that were probably inflated.

The main focus of this thesis is on the development of clinical prediction models. Our objective is to bring attention to challenges that are unique to the development of prediction models using clinical data. Clinical prediction models are extremely susceptible to improper development due to several heterogeneities that are particularly prevalent in clinical data. Two of these heterogeneities in clinical data are batch effects and differences in patient treatment.

Failure to handle heterogeneities in clinical data properly will result in prediction models with poor generalisation performance. In this thesis, we aim to highlight the most common obstacles to effective model deployment and their underlying causes, so that modelling practitioners will be better able to identify and handle these issues. In addition, we also discuss some of the subtleties involved in dealing with these issues. We hope that this would facilitate practitioners in developing prediction models that can be effectively deployed in a real-world setting.

One of the most prevalent heterogeneities in clinical data is batch effects. Developing a prediction model without accounting for batch effects in clinical data will result in a model that is unable to generalise to future batches of data. Batch effects are conventionally diagnosed qualitatively, through the use of principal component analysis (PCA) plots or uniform manifold approximation and projection (UMAP) plots. However, qualitative assessments of the magnitude of batch effects can be quite subjective. Quantitative batch effects metrics provide a more precise and objective way of measuring the magnitude of batch effects. Unfortunately, most existing metrics are inaccurate when used on data with different proportions of classes in different batches (i.e. batch class imbalance). In this thesis, we propose a quantitative batch effects metric that is accurate even when used on data with batch-class imbalance, which we call RVP. Our goal is to provide a tool for modelling

practitioners to use when attempting to mitigate batch effects in their data. For example, modelling practitioners will be able to use RVP to aid their decision on which batch effects correction method to use.

Differences in patient treatment occur frequently in clinical data, and will affect model performance if not dealt with properly during model development. In this thesis, we discuss the nuances behind common methods that have been used to deal with differences in patient treatment when developing prediction models. We also highlight that the specific differences in patient treatment vary between different diseases. Our objective is to help modelling practitioners better understand the nuances behind these methods, to enable them to select (and possibly tailor) the method to handle treatment differences that best suits their needs. We also seek to encourage modelling practitioners to incorporate information regarding treatment differences during evaluation of prediction models in order to improve evaluation accuracy.

ALL is the most prevalent form of paediatric cancer, making up approximately one fifth of paediatric malignancies (Inaba et al., 2013; Pui & Evans, 2006). Although modern risk-adapted therapy is able to achieve a long-term event-free survival rate of approximately 80%, relapse in B-cell precursor ALL still remains as one of the most common cause of cancer related deaths in children (Hunger & Mullighan, 2015). More accurate estimation of the risk of relapse would allow for improved treatment intensity recommendations, which would result in decreased number of deaths due to relapse and allow for de-intensification of chemotherapy in patients with good prognosis. The above reasons serve as our main motivation behind developing a prediction model for treatment outcome in paediatric ALL.

Previous work has demonstrated that accurate assignment of subtype in ALL can be achieved through the use of microarray gene expression profiles (Alizadeh et al., 2000; E.-J. Yeoh et al., 2002). Furthermore, there have been instances where ALL subtyping using gene expression profiling has led to identification of patients who would otherwise have been misclassified through conventional analysis (E.-J. Yeoh et al., 2002). Predicting treatment outcome solely through the use of microarray gene expression data would enable accurate risk stratification of ALL patients using a single technology platform. This would eliminate the need for multiple laboratory procedures, thus lowering the costs for risk stratification of ALL patients.

## 1.2 Scope

In this section, we define the scope of the main topics in our thesis. We first discuss the main obstacles facing the deployment of clinical prediction models in real-world settings, and their underlying causes. Subsequently, we investigate in greater detail two inherent heterogeneities in clinical data that if not accounted for, will lead to improper model development: batch effects and patient treatment differences. In our study of batch effects, we constrain our scope to batch effects that occur in high-dimensional bulk transcriptomics and proteomics data (i.e. we exclude single-cell data). Only batch effects metrics designed for use on such data are considered to be within our scope of study. In our study where we develop a clinical prediction model, we focus specifically on the task of predicting treatment outcome in paediatric ALL patients. For our discussion on patient treatment differences, we highlight common methods that are used to handle such differences. We demonstrate the value of incorporating treatment information during the evaluation of prediction models by presenting a scoring scheme specifically designed for the paediatric ALL treatment outcome prediction model.

## 1.3 Thesis contributions and organisation

In Chapter 2, we discuss the main obstacles facing effective model deployment, and investigate their underlying causes. We identify the main underlying cause behind these obstacles - improper development and evaluation of prediction models. Clinical prediction models are more susceptible to improper development and evaluation due to the inherent heterogeneities in clinical data. Two of the most important heterogeneities that have to be dealt with are batch effects and patient treatment differences. We also highlight other sources of heterogeneity and bias in clinical data that should be dealt with in order to develop accurate models. However, heterogeneities in clinical data are seldom reported due to privacy concerns or because they are deemed to be irrelevant. This further hinders our ability to deal with these heterogeneities. We offer suggestions to help overcome some of the obstacles facing effective model deployment. This chapter contains work that has been published in Chan and Wong, 2023.

In Chapter 3, we propose a novel quantitative batch effects metric, which we term RVP. RVP was designed to be robust to batch-class imbalance, and to be suitable for use in both small and large data sets. We demonstrate that RVP is able to accurately estimate the proportion of total variance attributable to batch effects over a range of batch effect magnitudes, using simulated RNA-seq data. RVP is able to do so even in data with batch-class imbalance. In addition, we show that RVP outperforms other metrics when evaluated on various types of real data, such as quantitative transcriptomics and proteomics data. Comparison of runtime and peak memory usage of the batch effects metrics on a data set with 8000 samples reveal that RVP is two orders of magnitude faster, and has a peak memory usage that is at most half that of other batch effects metrics.

In Chapter 4 we discuss the nuances of accounting for differences in patient treatment during the development or evaluation of clinical prediction models. We identify the most common methods used to handle differences in patient treatments and discuss certain caveats associated with each method. We use data from the Malaysia-Singapore ALL 2003 and 2010 studies (A. E. J. Yeoh et al., 2012; A. E. J. Yeoh et al., 2018) as a case study to demonstrate the complexities associated with differences in patient treatment. In addition, we present a plausible scoring scheme that incorporates treatment information to achieve more detailed evaluation of prediction models. The material presented in this chapter has been published in Chan and Wong, 2022.

In Chapter 5, we tackle the problem of predicting treatment outcome in paediatric ALL patients. We verify that feature selection is sufficient to mitigate batch effects, using our proposed batch effects metric RVP and PCA plots. We propose three transcriptomic features that are computed from patient gene expression profiles (GEPs) at different treatment time-points. We show that these features are associated with treatment outcome in homogeneous subtypes of ALL. We present a subtype-specific model that incorporates these features to predict treatment outcome in homogeneous ALL subtypes. We demonstrate that our subtype-specific model is able to discriminate well between patients who relapse and patients who achieve continuous complete remission (CCR), even when trained on small data sets. We evaluate our subtype-specific model in a more detailed manner by using our proposed scoring scheme, which incorporates patient treatment information. In addition, we

identify patients belonging to the recently discovered DUX4 subtype (Yasuda et al., 2016; Zhang et al., 2016) through clustering, and validate the performance of our model on the DUX4 subtype. The DUX4 subtype had not been discovered at the time of the Ma-Spore 2003 and 2010 studies; patients in the subtype were classified under the "B-Other" subtype. Lastly, we present our biological hypothesis behind the subtype-specific model, and substantiate it by inferring B-cell abundance in patient samples through various methods.

# Chapter 2

# Obstacles to effective model deployment

## 2.1 Introduction

Despite an exponential increase in publications on clinical prediction models over recent years, the number of models deployed in clinical practice remains fairly limited. In this chapter, we identify common obstacles facing effective model deployment in healthcare, and investigate their underlying causes. One of the most common obstacles is the lack of reproducibility and replicability (A. E. Johnson et al., 2017). Failure to replicate the performance of a prediction model on independent data sets casts doubts on the model's ability to perform effectively when deployed. Another common obstacle facing effective model deployment is the high risk of unfairness displayed by models. We provide examples of unfairness exhibited by prediction models deployed in real-life in this chapter.

We observe that a key underlying cause behind the various obstacles is the improper development and evaluation of prediction models. Healthcare/clinical prediction models are especially susceptible to improper development and evaluation due to the inherent heterogeneities in clinical data. We highlight common heterogeneities in clinical data and sources of biases that should be dealt with in order to avoid improper development and evaluation of models. Improper development of models often leads to inaccurate models that show poor generalisability, while improper evaluation of models may lead to overly optimistic results that cannot be replicated in independent data sets.

The purpose of this chapter is to familiarise modelling practitioners with the most

common obstacles facing effective model deployment, and their underlying causes. By being aware of the potential problems that may occur during development and evaluation of prediction models, practitioners would be better able to identify and deal with these issues. This would increase their chances of developing a model that can be effectively deployed.

## 2.2 Poor reproducibility and replicability

A core tenet of the scientific discovery process lies in the ability of the scientific community to either confirm or refute previous discoveries through independent studies. Scientific results and inferences that are replicable through independent studies have a higher likelihood of being true, and confidence in their reliability is built through repeated independent validation (National Academies of Sciences, Engineering, and Medicine, 2019). To avoid the many inconsistent definitions of reproducibility and replicability in scientific literature, we follow the definitions set out by National Academies of Sciences, Engineering, and Medicine, 2019:

> Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.
>
> Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

The lack of reproducibility is one of the most common and problematic issues found in clinical prediction models, and stems from incomplete and unclear reporting (A. E. Johnson et al., 2017; Wynants et al., 2020). To ensure that researchers are able to reproduce results from the original data, clear and complete reporting of all aspects concerning the prediction model should be performed. Firstly, the prediction task and clinical setting that a model is to be deployed in has to be clearly defined. This includes defining the prediction target and deployment population clearly. Secondly, population characteristics of patient data used to train and validate the prediction model should be described in detail, so that researchers have the necessary context to interpret the results. Thirdly, data pre-processing steps that were taken to transform the data should be clearly described, such as feature selection methodology.

In addition, predictors in the model should be clearly described, including details regarding how and when they were measured (if applicable). Fourthly, model details such as its type, parameters and architecture should be reported, along with design choices made regarding the model. Lastly, details regarding evaluation methodology should be clearly stated. The above-mentioned details are not exhaustive, and any other detail necessary to reproduce the original results should also be clearly reported. Modelling practitioners are encouraged to standardise reporting according to suggested guidelines in order to avoid incomplete and unclear reporting (Beam et al., 2020; A. E. Johnson et al., 2017; Weissler et al., 2021; Wynants et al., 2020).

Another major obstacle facing effective model deployment is the poor replicability of clinical prediction models. Models that do not show consistent results on independent data sets are deemed to be less reliable and often fail to gain the confidence of the scientific community. Even so, a majority of publications on clinical prediction models does not perform external validation (Ho et al., 2020; Siontis et al., 2015). This is compounded by the fact that publications that do perform external validation often report worse model performance (Siontis et al., 2015). For instance, A. E.-J. Yeoh et al., 2018 performed external validation of three treatment outcome prediction models (Bhojwani et al., 2008; Holleman et al., 2004; Meyer et al., 2011) that were developed using their own respective cohort of acute lymphoblastic leukaemia (ALL) patients. All of the models were unable to discriminate between treatment outcomes when evaluated on an independent data set of ALL patients.

There are two possible reasons why the prediction models mentioned above showed good internal validation while having poor external validation. Firstly, the data sets used to develop these models were small in size, and had low proportions of patients with the event of interest (i.e. relapse). Performing internal validation involves partitioning the small data sets into even smaller training and test sets, which may result in unstable results during evaluation (Steyerberg & Harrell, 2016). Secondly, the development data set and the data set used for external validation may consist of patients from different sub-populations (e.g. from different geographic regions). If that were the case, it would be more accurate to describe the models as having poor generalisability instead of poor replicability.

External validation of prediction models should be performed whenever possible

to demonstrate its replicability (Ho et al., 2020). Demonstrating the replicability of a model helps to build confidence in its reliability and improves its chances of clinical deployment. In cases where external validation is not possible, the next best alternative would be temporal validation, where the development data set can be split into training and test sets according to time (Ramspek et al., 2021).

## 2.3 Unfairness in prediction models

Another key reason behind the limited number of prediction models that are deployed in healthcare is due to the high risk of unfairness in prediction models. More recently, researchers have become aware of biases exhibited by prediction models that have been deployed in real-world applications. Most prediction models derive their function by recognising implicit patterns in the data that they are trained on. These models tend to learn the hidden biases that exist in the training data. As as result, these models make biased predictions that result in unfairness against certain groups or individuals. On the other hand, some prediction algorithms make biased predictions even when trained on data that is devoid of biases.

There are many different sources of biases in both data and algorithm. The introduction of these biases is mostly unintentional; often we are only made aware of their presence upon the discovery of errors or unfairness in model predictions. Procedures or design choices that seem innocuous are often responsible for introducing biases to data or algorithm. It is extremely hard to ensure that data is free of biases, as doing so would require pre-empting every possible source of bias, which requires an inordinate amount of care. A more reasonable goal would be to minimise the number of biases by learning the most frequent sources of biases that have recurred in previous literature. We enumerate below a few of the most prevalent sources of biases in prediction models deployed specifically in healthcare. For a more comprehensive summary of the sources of biases in machine learning, please refer to Mehrabi et al., 2021.

Omitted variable bias occurs when variables which have an impact on the dependent variable (i.e. prediction target) are omitted from the model. This may cause the model to attribute the effect of the omitted variable to variables that are included in the model. Usually, the omitted variable is excluded, or in some cases not

recorded, due to oversight. In some cases, variables may be omitted due to privacy concerns. An example of omitted variable bias was observed when a rule-based learning algorithm was trained to predict a patient's risk of dying from pneumonia (Caruana et al., 2015). One of the rules the model learned was that patients with a history of asthma had a lower risk of mortality from pneumonia. The rule was counter-intuitive, and on closer inspection it was discovered that patients with a history of asthma who contracted pneumonia were sent to the intensive care unit (ICU). Patients admitted into the ICU received intensive care which greatly lowered their risk of mortality. This resulted in the model erroneously associating lowered risk of mortality with asthma, instead of the omitted variable, ICU admission.

Representation bias occurs when the sampled data is not representative of the underlying population. When prediction models are trained on sampled data lacking in representation of certain subpopulations, they may exhibit poor prediction performance when deployed on these subpopulations. An example was when the Framingham risk score for cardiovascular disease, which was developed on a data set that was predominantly White and male, was found to be inaccurate when deployed on Black populations (Gijsberts et al., 2015). To maximise the prediction performance of a model on the target population or group, the model should be trained on data that contains a sizeable number of patients representative of the target group.

Measurement bias is the systematic error that arises during improper measurement of data. An example of measurement bias can be observed in the case of the algorithm used to facilitate COVID-19 relief funding allocation (Kakani et al., 2020). COVID-19 infection rates may be subject to measurement bias as it may be affected by differing diagnostic testing coverage between poorer and wealthier counties. Wealthier counties may receive higher diagnostic testing coverage, leading to a larger number of cases detected and thus higher COVID-19 infection rate.

## 2.4 Underlying causes

There are countless possible causes that may impede effective model deployment. We mention several of the causes above, but they are by no means exhaustive. Out of the many causes, we highlight an underlying cause that recurs frequently: improper

development and evaluation of models. We elaborate on how improper development and improper evaluation of prediction models impedes model deployment through the use of real-life examples in the sections below.

### 2.4.1 Improper development of models

Throughout the development process of prediction models, there are many actions that constitute improper development. These actions often result in models that do not function well or exhibit unfairness when deployed on the target population. One such action is when an incorrect proxy for the prediction target is used to train a prediction model; we show why this leads to a biased model using a real-life example.

This example concerns a commercial prediction model that was developed to predict the health risk of primary care patients (viz. risk of onset of common chronic illnesses). The model was used to identify patients who would benefit from high-risk care management programs. Developing the model using healthcare cost as a proxy to health risk resulted in a biased model, as healthcare cost was not an accurate proxy for health risk. Obermeyer et al., 2019 highlighted that for the same risk score predicted by the model, Blacks had a higher number of chronic illnesses than Whites. This reflected an inherent bias in the development data - for the same number of chronic illnesses, healthcare costs of Blacks were lower than that of Whites.

Other examples of improper model development include the sources of biases mentioned above, such as developing models on samples that are not representative of the deployment population, or not ensuring the uniformity of measurements across different sub-populations when collecting development data.

#### 2.4.1.1 Heterogeneities in data

Clinical prediction models are especially susceptible to improper development and evaluation due to the many possible heterogeneities in clinical data. It is important to deal with these heterogeneities when they arise, in order to avoid improper model development and evaluation.

The most important heterogeneity in clinical data that has to be accounted for is the differences in patient treatment. This is because different treatments have different magnitudes of effect on patient outcome. Failing to account for treatment

differences when developing models that predict patient outcome results in models that produce biased risk estimates. We elaborate on the differences in patient treatment and how to account for them during model development and evaluation in Chapter 4.

We use the MIMIC-III electronic health records (EHR) data set (A. E. Johnson et al., 2016) to demonstrate how failure to account for differences in treatment results in the development of models that show sub-optimal performance. This example also emphasises the importance of complete reporting, especially of factors that have a causal effect on the prediction target. MIMIC-III is a real-world EHR data set that comprises of data on patients who stayed in the ICU of the Beth Israel Deaconess Medical Center (in Boston) between 2001 and 2012. Patient year of care is randomised in the data set for privacy reasons. However, this prevents modelling practitioners from accounting for differences in treatment year of care when developing prediction models.

In a study by Nestor et al., 2019, the authors obtained a license to access the year of care of patients in the MIMIC-III data set; they highlighted two heterogeneities in the data set related to patient year of care. Firstly, clinical measurements were recorded in a different manner after 2008 due to a change in the data management system. Secondly, care practices and population demographics evolved through the years, resulting in temporal drift in the data. The authors developed models trained only on prior year data, by taking into account patient year of care. These models showed better discriminative performance than models developed without accounting for year of care.

Other heterogeneities that occur in clinical data include batch effects, which commonly arise due to the processing of patient data in batches because of, for example, limited patient availability at each point in time.

The overwhelming amount of heterogeneous details in clinical data makes it hard to ascertain which essential details have to be dealt with when developing prediction models. As a result, these essential details are frequently unreported and unaccounted for, which impedes the proper development and evaluation of models. Great care has to be taken to identify and handle these details to achieve proper model development and evaluation.

### 2.4.2 Improper evaluation of models

Clinical prediction models are susceptible to improper evaluation, mainly due to issues with data availability. Improper evaluation of models often produces overly optimistic model assessments which are not reflective of model performance under actual deployment. Healthcare decision makers are less likely to trust the authenticity of evaluation results if they observe evidence of improper evaluation. Performing proper evaluation is a key foundation for successful model deployment. In this section, we provide examples of improper evaluation.

The most common flaw when evaluating clinical prediction models is not performing external validation (Ho et al., 2020; Siontis et al., 2015). External validation on an external cohort of patients (i.e. not from the development cohort) should be performed whenever possible. A model may perform well on a hold-out test set partitioned from the development cohort but perform badly on an external cohort (i.e. patients from another study). Performing well during external validation is essential to prove a model's replicability and generalisability (Ho et al., 2020).

Other flaws in evaluation are less discernible; we present a flaw in the evaluation of a deep learning model (Coudray et al., 2018) that is used to predict gene mutation from small cell lung cancer histopathology images. In the original study, histopathology slide images are randomly split into training, validation and test sets. However, Oner et al., 2020 discovered that some of the slides in the training and test sets were highly correlated as they were derived from the same patient, which led to overly optimistic results during evaluation. Proper evaluation involved splitting slide images at the patient level, so that slides from the same patient can only be present together in either the training, validation or test set. The model showed significantly worse performance when properly evaluated. This example highlights the importance of proper evaluation, in giving an accurate assessment of a model's ability to generalise to new patients during real-world deployment.

Evaluating the performance of prediction models usually entails a few established procedures and metrics, such as plotting the receiver operating characteristic (ROC) and precision-recall curves, and calculating the c-statistic (i.e. area under the ROC curve), calibration and net benefit. Nonetheless, there is no one-size-fits-all approach in evaluating prediction models, with the most appropriate metrics to use differing

according to the intended use cases of the models (Weissler et al., 2021). Evaluating models using unsuitable metrics will give inconsequential results that offer little indication of how models would perform during real-world deployment. In addition, modelling practitioners should be aware that metrics such as accuracy and precision vary according to the proportion of positive and negative samples in the test set, even when model performance remains the same.

In cases where either the prediction task or data set is more complex than usual, we propose that evaluation should be customised accordingly. The aim is to evaluate the performance of the model on its prediction task as accurately as possible. Traditional metrics are often suited for standard tasks, but lack the finesse to accurately evaluate model performance on more specialised tasks. We provide an example of a customised evaluation method that incorporates differences in patient treatment in Section 4.5.

## 2.5 Key takeaways

Heterogeneities in data and sources of biases during model development differ according to each data set and situation. It is best for modelling practitioners to familiarise themselves with common heterogeneities in data and common sources of biases during model development. The best way to handle heterogeneities and biases is highly dependent on the characteristics of each individual situation. It is also crucial for practitioners to be aware of the intricacies in handling heterogeneities and biases.

Certain heterogeneities in clinical data are often unreported due to privacy and other reasons. Hence, it is important to anticipate, in particular, unreported heterogeneities that have a causal effect on the prediction target. A frequent example is treatment differences between patients. Permission to access patient treatment information for model development should be requested from data providers, as it is imperative to account for treatment differences during model development and evaluation. Failure to account for treatment differences would be equivalent to implicitly assuming that all patients receive the same treatment.

Proper evaluation of models helps to build confidence in their reproducibility and replicability. Firstly, external validation should be performed whenever possible,

to ensure fair assessment of model performance. Secondly, care has to be taken to ensure that the evaluation methodology is free of errors (e.g. data leakage). Lastly, the evaluation metrics used should be focused on assessing model performance in achieving the prediction objective.

Complete and clear reporting of all aspects of model development and evaluation is a simple yet often overlooked factor that improves the reproducibility, replicability and generalisability of models. All details regarding the prediction model, such as the prediction objective, target deployment population, data characteristics and composition, data pre-processing, model specification and evaluation methodology, should be provided. As a rule of thumb, all details required to reproduce the original results should be reported when presenting the prediction model.

# Chapter 3

# RVP: Quantifying batch effects

## 3.1 Introduction

Batch effects are systematic errors in measurements between different batches of samples, and have been widely reported in many types of high-dimensional biological data (Goh et al., 2017; Leek et al., 2010). Examples include quantitative data produced using microarray, RNA-seq, scRNA-seq and liquid chromatography/mass spectrometry (LC/MS) technologies. Batch effects may result from a variety of causes, such as differences in experimental laboratories, technology platforms, measurement times or laboratory personnel. Batch effects often result from an amalgamation of different contributing factors.

Batch effects in data impairs downstream analysis in various ways. For example, batch effects result in the identification of spurious differentially expressed genes during differential expression analysis (Leek et al., 2010). Also, batch effects frequently affect the generalisation ability of clinical prediction models. Significant batch effects often exist between future batches of data that clinical prediction models are deployed on, and batches of data used to train the model.

Batch effects are generally identified qualitatively, with the most popular method being principal component analysis (PCA). Frequently, variation between batches is captured in the first or second principal components (PCs), and samples from different batches can be easily distinguished from each other in PCA plots (Lazar et al., 2012). At times, variation between batches may be even bigger than biological variation between different classes (Belorkar & Wong, 2016). Other popular dimension reduction techniques that are used to visualise batch effects in high-dimensional data include uniform manifold approximation and projection (McInnes et al., 2018) and

t-distributed stochastic neighbour embedding (Van der Maaten & Hinton, 2008).

Very few quantitative batch effects metrics have been proposed for data with small sample sizes. Most of the metrics proposed recently were developed specifically for scRNA-seq data, which typically consist of thousands of samples. Some examples include $k$-nearest-neighbour batch-effect test (Büttner et al., 2019) and cell-specific mixing score (Lütge et al., 2021). The core concept behind these metrics is the same: to detect batch effects by examining the characteristics of the $k$-nearest neighbours ($k$-NNs) of all samples in the data. These metrics are not suited for small data sets as $k$-NN methods do not work well in small data sets; suitable values of $k$ values are limited in small data sets. More critically, these metrics are not robust and can be misleading when there are batch-class imbalances (Goh et al., 2022). Batch-class imbalance refers to data in which different batches have different class proportions. Most of the quantitative batch effects metric designed for small data sets, such as guided PCA (gPCA; Reese et al., 2013) and principal variance component analysis (PVCA; Scherer, 2009) are also inaccurate when used on data with batch-class imbalance.

We propose a novel quantitative batch effects metric, recursive variance partitioning (RVP), which estimates the proportion of variance in data due to batch effects. RVP was designed to be robust to batch-class imbalance, and to be suitable for use in both small and large data sets. We evaluate the performance of RVP against two batch effects metric designed for small data sets, gPCA and PVCA (Reese et al., 2013; Scherer, 2009). We show that RVP most accurately estimates the proportion of total variance in data due to batch effects across different magnitudes of batch effects, using simulated RNA-seq data. RVP is able to give accurate estimates even in data with batch-class imbalance. RVP also outperforms the other metrics in real-world microarray and proteomics data sets. Most quantitative batch effects metrics designed for small data sets involve PCA, either as a pre-processing step or as part of their model (Lütge et al., 2021; Reese et al., 2013; Scherer, 2009). We show that the runtime of RVP increases linearly with respect to the number of samples in the data, while the runtime of gPCA and PVCA increases polynomially with respect to the number of samples. Furthermore, we demonstrate that peak memory usage of RVP is lesser than that of gPCA and PVCA when benchmarked on the same data sets.

## 3.2 Background

Batch effects can be characterised as systematic errors in measurements between different batches of samples. Chen et al., 2011 defines a batch as a group of samples that have been collected at a single experimental site over a short period of time. Although batch effects are more commonly associated with high-dimensional data, they have been reported in low-dimensional data as well (Goh et al., 2017; Leek et al., 2010). Causes of batch effects include differences in experimental laboratories, technology platforms, measurement times or laboratory personnel. Batch effects observed in real-world data are often the result of multiple contributing factors.

Batch effects are not only present in high-dimensional biological data. For example, batch effects can be found in near-infrared spectrums of edible oils as well. Batch effects in different types of data may have different characteristics. We restrict our scope in this thesis to batch effects in high-dimensional biological data.

### 3.2.1 Characteristics of batch effects

In this section, we describe various characteristics of batch effects that have been observed in high-dimensional biological data. Batch effects in microarray gene expression data are generally modelled either as additive or multiplicative batch effects, or both (Lazar et al., 2012). The popular microarray batch effects correction method ComBat (W. E. Johnson et al., 2007) models batch effects as both additive and multiplicative batch effects as follows:

$$Y_{ijk} = X_{ijk} + \delta_{ik} + \gamma_{ik}\epsilon_{ijk}$$

where $Y_{ijk}$ is the measured log-transformed expression value for gene $i$ of sample $j$ from batch $k$, $X_{ijk}$ is the theoretical gene expression value, $\epsilon_{ijk}$ represents random Gaussian noise with mean zero and variance $\sigma_i^2$, and $\delta_{ik}$ and $\gamma_{ik}$ represents additive and multiplicative batch effects, respectively. Following the above model, additive batch effects can be visualised intuitively as the distance separating different batches of samples in a PCA plot, while multiplicative batch effects can be visualised as different batches having different levels of dispersion among their own samples.

Batch effects correction methods that perform location-scale adjustment (e.g. ComBat) assume batch effects to be constant across all samples in a batch, regardless

of biological class. In other words, batch effects in microarray data are assumed to be global, and correction of batch effects are performed uniformly across all samples in a batch. On the other hand, batch effects in scRNA-seq data have been characterised to be local, i.e. batch effects are assumed to vary between different samples in a batch (Haghverdi et al., 2018; Hie et al., 2019). However, variation in local batch effects between samples in a batch are expected to be small.

## 3.3 Related work

In this section, we describe two quantitative batch effects metrics that were developed for use in microarray data, gPCA $\delta$ and PVCA. We provide a brief summary on how each metric estimates the magnitude of batch effects in data. Both gPCA $\delta$ and PVCA incorporate the use of PCA when quantifying batch effects.

### 3.3.1 gPCA $\delta$

Guided PCA (gPCA) is a variation of PCA that incorporates the use of a batch factor that encodes the batch a sample is assigned to (Reese et al., 2013). Both gPCA and PCA are used to compute the test statistic $\delta$ that is used to quantify batch effects.

PCA is commonly implemented by performing singular value decomposition (SVD) on a centered matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with $n$ samples and $p$ features:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

where $\mathbf{V} \in \mathbb{R}^{p \times p}$ is the orthogonal feature loadings matrix and $\mathbf{Z} = \mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times p}$ represents the transformed PC score matrix.

In gPCA, SVD is performed on a batch matrix $\mathbf{W} = \mathbf{Y}^\top\mathbf{X} \in \mathbb{R}^{m \times p}$ with $m$ batches and $p$ features instead, where $w_{kj}$ is the aggregated expression value in batch $k$ of feature $j$, and $\mathbf{Y} \in \mathbb{R}^{n \times m}$ is the design matrix whose elements are given by

$$y_{ik} = \begin{cases} 1 & \text{if sample } i \text{ is in batch } k, \\ 0 & \text{otherwise.} \end{cases}$$

Performing SVD on the batch matrix $\mathbf{W}$ results in

$$\mathbf{W} = \mathbf{U}'\mathbf{D}'\mathbf{V}'^\top$$

where $\mathbf{V}' \in \mathbb{R}^{p \times p}$ is the gPCA feature loadings matrix. The column vectors of $\mathbf{V}'$ are known as the right singular vectors of $\mathbf{W}$.

To compute the test statistic $\delta$, the data is first projected onto the space spanned by the right singular vectors of $\mathbf{W}$ (i.e. PCs obtained from gPCA of $\mathbf{W}$), by multiplying the original data matrix $\mathbf{X}$ by the gPCA feature loadings matrix $\mathbf{V}'$ to obtain the resulting gPCA-transformed PC score matrix $\mathbf{Z}' \in \mathbb{R}^{n \times p}$:

$$\mathbf{Z}' = \mathbf{X}\mathbf{V}'$$

The test statistic $\delta$ is the ratio of the variance of the scores from the first PC of the gPCA-transformed score matrix, to the variance of the scores from the first PC of the PCA-transformed score matrix:

$$\delta = \frac{\text{Var}(\mathbf{Z}'_{\cdot 1})}{\text{Var}(\mathbf{Z}_{\cdot 1})}$$

where the column vectors $\mathbf{Z}'_{\cdot 1} \in \mathbb{R}^p$ and and $\mathbf{Z}_{\cdot 1} \in \mathbb{R}^p$ are the score vectors from the first PC of the gPCA-transformed matrix $\mathbf{Z}'$ and the PCA-transformed matrix $\mathbf{Z}$, respectively. Larger $\delta$ values (near one) imply that batch effects are larger.

In addition to the test statistic $\delta$, a corresponding $p$-value is also reported which shows the probability of obtaining a value greater than the observed test statistic value $\delta_0$ by chance. A Monte Carlo permutation test is used to approximate the $p$-value. The null distribution is generated by permuting the batch factor for a total of $S$ times, repeating gPCA each time to obtain the test statistic value $\delta_s$ of the $s$th permutation. The $p$-value is then approximated by

$$p = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\{\delta_s > \delta_0\}$$

where $\mathbf{1}\{\delta_s > \delta_0\}$ is the indicator variable that is equals to one if $\delta_s$ is larger than $\delta_0$ and zero otherwise.

### 3.3.2 PVCA

Principal Variance Component Analysis (PVCA) quantifies batch effects by estimating the proportion of variance in the data that can be attributed to batch-related factors (Scherer, 2009). PVCA takes in different factors of interest (e.g. batch, class) and estimates the proportion of total variance explained by these factors and their two-way interactions terms.

The PVCA algorithm consists of three main steps. First, PCA is performed on the data, and only the top PCs are retained. The number of PCs retained depends on the minimum number of PCs required for total variance of the retained PCs to be above a target percentage of total variance of the data. However, the number of retained PCs is limited at a maximum of 10.

Second, each PC was fitted separately using a linear mixed model. PC scores were modelled as the dependent variable while all factors of interest were modelled as random effects. Variance components due to the random effects are estimated using restricted maximum likelihood (REML). The linear mixed model partitions the total variance of each PC into variance components belonging to the factors and their two-way interaction terms, as well as the model residual.

Third, a weighted average of the variance components of each factor is calculated across PCs, using the respective eigenvalue of each PC as its weight. The weighted average of variance components of each factor is subsequently normalised by dividing by their sum (i.e. normalised so that they sum to unity). This normalised value is used to represent the proportion of total variance explained by each factor.

## 3.4 Recursive variance partitioning

We propose recursive variance partitioning (RVP), a quantitative batch effects metric which estimates the proportion of total variance in data that can be attributed to batch effects. To use our metric, users provide a single batch factor specifying the batch for each sample, along with the data matrix. Users should provide class-related factors (e.g. disease condition) if present, as this would allow for more accurate estimation of batch effects, particularly when data is batch-class imbalanced (i.e. when different batches have different class proportions).

Our method is based on an important statistical concept known as the partition of sums of squares, which is commonly used to attribute the proportion of variation in data to different sources. This concept is used in the statistical model, analysis of variance (ANOVA), and is also applied to calculate the coefficient of determination $R^2$ of linear regression models.

In RVP, the proportion of total variance in the data due to batch effects is estimated as follows. The sum of squares that can be attributed to the batch factor

is calculated for each feature, and subsequently summed across all features. The resulting sum is divided by the total sum of squares of the data to arrive at our estimate of the proportion of total variance in data that is attributable to batch effects.

To elaborate, we consider a two factor model with a batch factor $b$ and class factor $c$. Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be the data matrix with $p$ features and $n$ samples, whose elements $x_{ijkg}$ denote the expression value of feature $i$ in the $j$-th sample from batch $k$ and class $g$. The total sum of squares of the data matrix, which we denote as $\mathrm{SS}_T^{(\cdot\cdot)}$, can be obtained by summing the total sum of squares of all its features:

$$\mathrm{SS}_T^{(\cdot\cdot)} = \sum_i \mathrm{SS}_T^{(i\cdot)}$$

where $\mathrm{SS}_T^{(i\cdot)}$ is the total sum of squares of feature $i$. $\mathrm{SS}_T^{(\cdot\cdot)}$ is proportional by a factor of $n-1$ to the total sample variance of the data matrix. The same relationship applies between the total sum of squares and total variance of each feature vector in the data matrix as well.

### 3.4.1 Case 1: Samples belong to the same class

We first consider estimating the proportion of variance due to batch effects in a simpler case where all samples belong to the same class. We start by partitioning the total sum of squares of each feature $i$ into the sum of squares between batches $\mathrm{SS}_{B(b)}^{(i\cdot)}$ and the sum of squares within batches $\mathrm{SS}_{W(b)}^{(i\cdot)}$:

$$\mathrm{SS}_T^{(i\cdot)} = \mathrm{SS}_{B(b)}^{(i\cdot)} + \mathrm{SS}_{W(b)}^{(i\cdot)}$$

This is equivalent to the partition of sum of squares in a one-way ANOVA with batch factor $b$. The total sum of squares of feature $i$ is the sum of the squared deviations of all samples from the overall mean $\bar{x}_{i\cdots}$ and is defined as

$$\mathrm{SS}_T^{(i\cdot)} = \sum_{k,g,j} (x_{ijkg} - \bar{x}_{i\cdots})^2$$

The sum of squares between batches (encoded by batch factor $b$) of feature $i$ is calculated by

$$\mathrm{SS}_{B(b)}^{(i\cdot)} = \sum_k n_k (\bar{x}_{i\cdot k\cdot} - \bar{x}_{i\cdots})^2$$

where $n_k$ is the number of samples in batch $k$ and $\bar{x}_{i\cdot k\cdot}$ is the mean of feature $i$ across all samples belonging to batch $k$. The sum of squares within batches of feature $i$ is calculated by

$$\mathrm{SS}_{W(b)}^{(i\cdot)} = \sum_k \sum_{g,j} (x_{ijkg} - \bar{x}_{i\cdot k\cdot})^2$$

We estimate the proportion of total variance in the data that can be attributed to batch effects by summing up across features all the sum of squares between batches, and dividing it by the total sum of squares of the data matrix

$$\mathrm{RVP} = \frac{\sum_i \mathrm{SS}_{B(b)}^{(i\cdot)}}{\sum_i \mathrm{SS}_{T}^{(i\cdot)}}$$

### 3.4.2 Case 2: Samples belong to different classes

In the case where samples come from different classes, we use a class factor $c$ with levels $g$ to denote the classes that samples belong to. If more than one class-related factor is provided, all possible factor level combinations are considered as levels of a single class factor. For example, if disease condition and gender are provided as class-related factors, the four possible factor level combinations: (Disease, Male), (Disease, Female), (Normal, Male) and (Normal, Female) will be treated as individual levels of a single class factor.

The two-way ANOVA can be used to estimate the proportion of variance attributable to batch. However, the calculation of sum of squares of the batch factor $b$ in a two-way ANOVA is equivalent to the calculation of sum of squares between batches according to a one-way ANOVA when samples are assumed to be from the same class (as in section 3.4.1).

Disregarding information regarding the class factor when class effects are present leads to biased estimates of the proportion of variance attributable to batch in data with batch-class imbalance. In batch-class imbalanced data where a feature can be affected by both batch and class effects, variation due to class will be erroneously attributed to batch if the class factor is disregarded, resulting in an over-estimation of batch effects.

To ensure that our estimate of the proportion of total variance due to batch effects remains unbiased even in data with batch-class imbalance, we partition

samples according to their class before calculating the proportion of variance within each class that can be attributed to batch effects. This is performed by recursively partitioning the total sum of squares of feature $i$ using the class factor $c$ first, followed by the batch factor $b$:

$$\begin{aligned}
\mathrm{SS}_T^{(i\cdot)} &= \mathrm{SS}_{B(c)}^{(i\cdot)} + \mathrm{SS}_{W(c)}^{(i\cdot)} \\
&= \mathrm{SS}_{B(c)}^{(i\cdot)} + \sum_g \sum_{k,j} (x_{ijkg} - \bar{x}_{i\cdot g})^2 \\
&= \mathrm{SS}_{B(c)}^{(i\cdot)} + \sum_g \mathrm{SS}_T^{(ig)} \\
&= \mathrm{SS}_{B(c)}^{(i\cdot)} + \sum_g (\mathrm{SS}_{B(b)}^{(ig)} + \mathrm{SS}_{W(b)}^{(ig)})
\end{aligned}$$

where $\bar{x}_{i\cdot g}$ is the mean of feature $i$ across all samples belonging to class $g$, and $\mathrm{SS}_T^{(ig)}$, $\mathrm{SS}_{B(b)}^{(ig)}$ and $\mathrm{SS}_{W(b)}^{(ig)}$ are the total sum of squares, sum of squares between batches, and sum of squares within batches, respectively, of feature $i$ across all samples belonging to class $g$.

We estimate the proportion of total variance in the data that is due to batch effects by summing across all features and classes, the sum of squares between batches in each class $g$ for feature $i$. Subsequently, we divide the above sum by the total sum of squares of the data matrix to obtain our estimate:

$$\mathrm{RVP} = \frac{\sum_i \sum_g \mathrm{SS}_{B(b)}^{(ig)}}{\sum_i \mathrm{SS}_T^{(i\cdot)}}$$

By first partitioning samples into their respective classes before calculating the sum of squares between batches in each class, we manage to attribute to batch effects the variance that would have been attributed in a two-way ANOVA to the interaction between the batch factor and class factor.

## 3.5 Methods

### 3.5.1 Data sets

We evaluate the performance of all batch effects metrics on four data sets from different high-throughput technologies. We evaluate their accuracy in estimating

batch effects of different magnitudes, using simulated RNA-seq count data. We also evaluate their performance on real data sets, such as gene expression microarray and quantitative proteomics data sets. For all data sets, different subsamples were taken to represent data with and without batch-class imbalance.

#### 3.5.1.1 Simulated RNA-seq data set

RNA-seq count data was simulated using the *BatchQC* 1.22.0 package (Manimaran et al., 2016) in *R* (R Core Team, 2018). Data was simulated using a two-batch and two-class design, with equal batch and class sizes. Two sets of ten count matrices with different magnitudes of batch effects were simulated. The first set of data matrices were simulated to have severe batch-class imbalance while the second set of matrices represent data with a balanced batch-class design (see Table 3.1).

| Batch | A | B |
|---|---|---|
| 1 | 20 | 20 |
| 2 | 20 | 20 |

(a) Balanced

| Batch | A | B |
|---|---|---|
| 1 | 10 | 30 |
| 2 | 30 | 10 |

(b) Severe imbalance

Table 3.1: Distribution of samples across batches and classes in simulated RNA-seq count data for (a) the set of data matrices without batch-class imbalance and (b) the set of data matrices with batch-class imbalance.

*BatchQC* simulates counts $X_{ijkg}$ for feature $i$ in the $j$-th sample from batch $k$ and class $g$ by sampling from a negative binomial distribution parametrised by the dispersion parameter $r_k$ and mean parameter $\mu_{ijkg}$:

$$X_{ijkg} \sim \text{NB}(r_k, \mu_{ijkg})$$

The mean and dispersion parameters are defined in terms of user-specified hyper-parameters:

$$\mu_{ijkg} = \alpha + \phi_i + \delta_k + \psi_g + \epsilon Y_{ijkg}\,, \qquad r_k = \beta + \gamma_k$$

where $\alpha$ is the overall base mean, $\phi_i$ is the feature effect, $\delta_k$ is the batch effect, $\psi_g$ is the class effect, $\epsilon$ is the extra dispersion, the random variable (r.v.) $Y_{ijkg} \sim \text{Beta}(2, 2)$ is the multiplier term, $\beta$ is the base dispersion and $\gamma_k$ is an additional dispersion term specific to batch $k$.

We simulated data with additive batch effects using a three-step process. First, we simulated a data matrix $\mathbf{X} \in \mathbb{Z}_{\geq 0}^{p \times n}$ with no batch effects consisting of two classes using the following hyper-parameters: $n = 80$, $p = 8000$, $\alpha = 5000$, $\epsilon = 2000$, $\beta = 2000$, $\delta_k = 0$ and $\gamma_k = 0$ for all $k$, class effect vector $\boldsymbol{\psi} = (\psi_1, \psi_2) = (0, 10000)$, and feature effect vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p) = (0, 100, \ldots, 799900)$. Second, we simulated additive batch effect terms $\omega_{ijkg}$ for feature $i$ in the $j$-th sample from batch $k$ and class $g$ by sampling from a Poisson distribution:

$$\omega_{ijkg} \sim \text{Pois}(\delta_k)$$

where $\delta_k$ represents the magnitude of batch effects in batch $k$. Third, we added the batch effect terms to their respective counts to obtain counts with additive batch effects:

$$X'_{ijkg} = X_{ijkg} + \omega_{ijkg}$$

To simulate ten data matrices with different magnitudes of batch effects, we set $\delta_1 = 0$ and $\delta_2 = 1000, 2000, \ldots, 10000$ when simulating additive batch effect terms for each data matrix, respectively.

#### 3.5.1.2 MAQC-I microarray data set

The Microarray Quality Control I (MAQC- I) data set was generated to analyse the reproducibility of microarray measurements across different microarray platforms and test sites (Shi et al., 2006). Raw microarray gene expression data were downloaded from the Gene Expression Omnibus (GEO) repository (Edgar et al., 2002) under the accession number GSE5350. We restricted our analysis to microarrays from the Affymetrix HG-U133 Plus 2.0 platform. To ensure consistency of the RNA samples, two different classes of commercially available RNA samples were assayed, the 100% Universal Human Reference RNA (UHRR) samples and the 100% Human Brain Reference RNA (HBRR). A total of 60 samples were assayed at six different laboratories, representing six different batches. There were ten samples per batch, with five technical replicates from each class.

Different subsamples of the MAQC-I data set were used to simulate data with different degrees of batch-class imbalance, and data with no batch effects. To simulate data with no batch effects, samples from the same batch were artificially assigned to

different batches. Batch and class distributions of the different subsamples used in our experiments are in Table 3.2.

| Batch | UHRR | HBRR |
|---|---|---|
| 1 | 2 | 2 |
| $1'$ | 2 | 2 |

(a) Balanced (no batch effects)

| Batch | UHRR | HBRR |
|---|---|---|
| 1 | 3 | 1 |
| $1'$ | 1 | 3 |

(b) Severe imbalance (no batch effects)

| Batch | UHRR | HBRR |
|---|---|---|
| 1 | 4 | 4 |
| 2 | 4 | 4 |

(c) Balanced

| Batch | UHRR | HBRR |
|---|---|---|
| 1 | 3 | 5 |
| 2 | 5 | 3 |

(d) Moderate imbalance

Table 3.2: Distribution of samples across batches and classes in four different subsamples of the MAQC-I microarray data set. Samples in batch $1'$ refer to samples from batch 1 that have been artificially assigned to be from a different batch.

#### 3.5.1.3 Ma-Spore ALL microarray data set

The Ma-Spore acute lymphoblastic leukaemia (ALL) data set consists of time-series gene expression microarrays of childhood ALL patients at different treatment time points, specifically at Day 0 (diagnosis) and Day 8 (Stary et al., 2014; A. E. J. Yeoh et al., 2012; A. E. J. Yeoh et al., 2018). It is publicly available at the GEO repository under accession number GSE67684. Sample microarrays were scanned over a period of time ranging from 2002 to 2015, and grouped into ten different batches according to their scan dates. Patient subtypes are diagnosed through the use of cytogenetic analysis, immunophenotyping and molecular screening. Patients either achieve continuous complete remission or suffer from relapse.

Only samples from the TEL-AML1 subtype belonging to patients who achieved remission were used for experiments in this chapter. Different subsets of the above-mentioned samples were used to simulate data with different degrees of batch-class imbalance, and data with no batch effects. Samples from the same batch were artificially assigned to different batches to simulate data with no batch effects. Table 3.3 shows the batch and class distributions of the various subsets used in our experiments.

| Batch | Day 0 | Day 8 |
|---|---|---|
| 2 | 4 | 4 |
| $2'$ | 4 | 4 |

(a) Balanced (no batch effects)

| Batch | Day 0 | Day 8 |
|---|---|---|
| 2 | 6 | 2 |
| $2'$ | 2 | 6 |

(b) Severe imbalance (no batch effects)

| Batch | Day 0 | Day 8 |
|---|---|---|
| 2 | 4 | 4 |
| 9 | 4 | 4 |

(c) Balanced

| Batch | Day 0 | Day 8 |
|---|---|---|
| 2 | 5 | 3 |
| 9 | 3 | 5 |

(d) Moderate imbalance

Table 3.3: Distribution of samples across batches and classes in four different subsamples of the Ma-Spore ALL microarray data set. Samples in batch $2'$ refer to samples from batch 2 that have been artificially assigned to be from a different batch.

#### 3.5.1.4 Westlake proteomics data set

The Westlake data set is a quantitative proteomics data set that was derived from LC/MS measurements. It was generated to study batch effects in proteomics data. Protein expression of two different cancer cell lines, K562 and A549, were measured using two different mass spectrometers. Samples measured using the same mass spectrometer are considered to be from the same batch. For each class, a total of three biological replicates were obtained, and four technical replicates were performed for each biological replicate (giving 12 samples per class for a single batch). This was repeated using two different mass spectrometers to give a total of 48 samples. Different subsets of samples were used to simulate data with different degrees of batch-class imbalance, and data with no batch effects. Samples from the same batch were artificially assigned to different batches to simulate the absence of batch effects. Details regarding the batch and class distributions of samples in the different subsamples can be found in Table 3.4.

### 3.5.2 Runtime and memory analysis

Runtime was measured using the *Sys.time* function in *R* (R Core Team, 2018), while peak memory usage was measured using the *Linux* utility *time*. We simulated five RNA-seq count matrices with $2000, 4000, \ldots, 10000$ samples using the *BatchQC* package in *R*. All matrices were simulated using the same hyper-parameters stated

| Batch | K562 | A549 |
|---|---|---|
| 1 | 4 | 4 |
| $1'$ | 4 | 4 |

(a) Balanced (no batch effects)

| Batch | K562 | A549 |
|---|---|---|
| 1 | 2 | 6 |
| $1'$ | 6 | 2 |

(b) Severe imbalance (no batch effects)

| Batch | K562 | A549 |
|---|---|---|
| 1 | 8 | 8 |
| 2 | 8 | 8 |

(c) Balanced

| Batch | K562 | A549 |
|---|---|---|
| 1 | 6 | 10 |
| 2 | 10 | 6 |

(d) Moderate imbalance

| Batch | K562 | A549 |
|---|---|---|
| 1 | 4 | 12 |
| 2 | 12 | 4 |

(e) Severe imbalance

Table 3.4: Distribution of samples across batches and classes in five different subsamples of the Westlake proteomics data set. Samples in batch $1'$ refer to samples from batch 1 that have been artificially assigned to be from a different batch.

in section 3.5.1.1, but with a fixed batch effects parameter of $\boldsymbol{\delta} = (0, 5000)$. We ran PVCA using the *pvcaBatchAssess* function from the *pvca* package in *R*, with the minimum percentage of variance parameter set to 60%. We modified the implementation of the *gPCA.batchdetect* function in order to run gPCA with zero number of permutations in the permutation test, and to solve memory issues due to inefficient coding design. The permutation test is not required for the calculation of the gPCA $\delta$ metric. All experiments were run on a personal workstation with 3.40 GHz Intel Core i7-4770 CPU and 8 gigabytes (GB) of RAM.

## 3.6 Results

### 3.6.1 Simulated RNA-seq count data

In order to evaluate the ability of RVP to accurately quantify batch effects of different magnitudes, we simulated RNA-seq count data with different magnitudes of batch effects using the *BatchQC* package. We simulated data consisting of two batches and two classes, with equal batch and class sizes. Two sets of data were simulated, one with batch-class imbalance and the other without.

We simulated data with additive batch effects through a three-step process. First, we simulated data without batch effects using *BatchQC*. Let $X_{ijkg}$ be the r.v. denoting the count of feature $i$ in the $j$-th sample from batch $k$ and class $g$. *BatchQC*

simulates counts by sampling from a negative binomial distribution parametrised by the dispersion parameter $r_k$ and mean parameter $\mu_{ijkg}$: $X_{ijkg} \sim \text{NB}(r_k, \mu_{ijkg})$. The parameters $r_k$ and $\mu_{ijkg}$ are defined by user-specified hyper-parameters, which control the step sizes of increment between different features, classes and batches. For more details on the hyper-parameters used to simulate the data, please refer to section 3.5.1.1. Second, we simulated additive batch effect terms $\omega_{ijkg}$ by sampling from a Poisson distribution: $\omega_{ijkg} \sim \text{Pois}(\delta_k)$, where $\delta_k$ represents the magnitude of batch effects in batch $k$. Third, we added the batch effect terms to their respective counts to obtain counts with additive batch effects: $X'_{ijkg} = X_{ijkg} + \omega_{ijkg}$. We simulated ten data matrices with different magnitudes of batch effects by setting $\delta_1 = 0$ and $\delta_2 = 1000, 2000, \ldots, 10000$ when simulating additive batch effect terms for each data matrix, respectively. The notation defined in this paragraph is used consistently throughout the following subsections.

To assess the accuracy of the batch effects metrics in quantifying batch effects of different magnitudes, we derived two different estimates of batch effects magnitude from the hyper-parameters used to simulate the data. The derived a priori estimates are 1) the percentage of batch effects in the data, and 2) the variance due to batch effects in the data. We use each estimate as the ground truth for the magnitude of batch effects in the data sets. We verify both a priori estimates by comparing them to measurements of the simulated additive batch effect terms used to generate the data.

#### 3.6.1.1 Percentage of batch effects in data

We estimate the percentage magnitude of batch effects in our simulated RNA-seq data, which we denote as $\pi$, by dividing the expected value of the total magnitude of additive batch effect terms, by the expected value of the total counts without

batch effects in the entire data matrix:

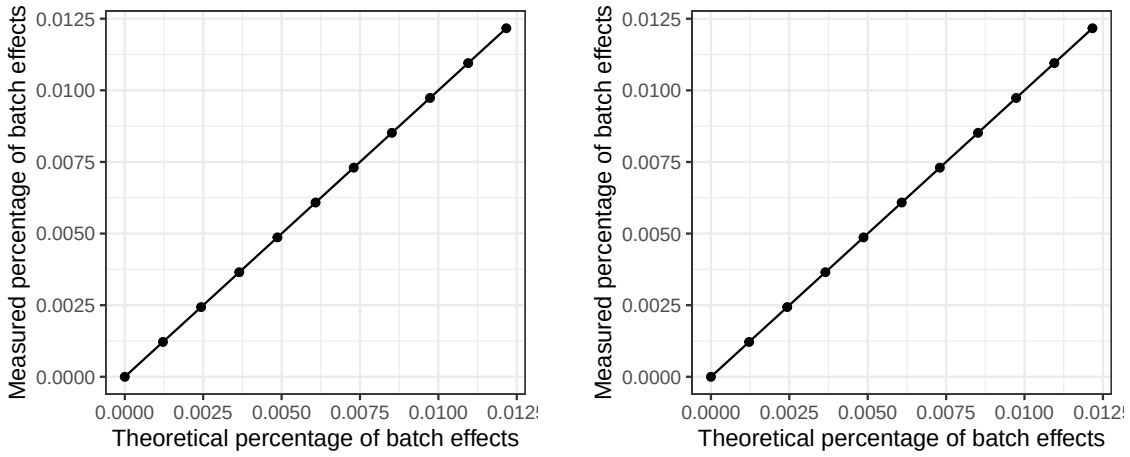
$$\begin{aligned}\pi &= \frac{\mathbb{E}\left(\sum_{i,k,g,j} X'_{ijkg} - X_{ijkg}\right)}{\mathbb{E}\left(\sum_{i,k,g,j} X_{ijkg}\right)} \\ &= \frac{\sum_{i,k,g,j} \mathbb{E}\left(\omega_{ijkg}\right)}{\sum_{i,k,g,j} \mathbb{E}\left(X_{ijkg}\right)} \\ &= \frac{\sum_{i,k,g,j} \delta_k}{\sum_{i,k,g,j} \alpha + \phi_i + \psi_g + \epsilon\mathbb{E}(Y_{ijkg})} \\ &= \frac{p\sum_k n_k\delta_k}{pn\alpha + 0.5pn\epsilon + p\sum_g n_g\psi_g + n\sum_i \phi_i}\end{aligned}$$

where $n_k$ is the number of samples in batch $k$, $n_g$ is the number of samples in class $g$, $n$ is the total number of samples and $p$ is the number of features. The r.v. $Y_{ijkg} \sim \text{Beta}(2,2)$ has an expected value of $\mathbb{E}(Y_{ijkg}) = 0.5$. Simplifications of summations over individual features, samples, batches and classes can be made as both batch effects $\delta_k$ and class effects $\psi_g$ are constant across all features.

In order to verify that the theoretical formula for percentage magnitude of batch effects is correct, we first obtained the matrix of additive batch effect terms and the original count data matrix (without batch effects) used to simulate each data set. Subsequently, we measured the percentage of batch effects in each data set by dividing the sum of all additive batch effect terms by the sum of all counts in the original data matrix. Figure 3.1 shows that the theoretical percentage of batch effects corresponds closely to the measured percentage of batch effects in all simulated data sets.

Figure 3.2 demonstrates that RVP is able to quantify batch effects accurately in both data with and without batch-class imbalance. RVP has the smallest absolute change in metric values between balanced and imbalanced data when the percentage of batch effects in the data is the highest (compare points at top end of curves in Figures 3.2c and 3.2f). Furthermore, the relationship between the percentage of batch effects in data and RVP most resembles a linear relationship out of the three metrics. However, RVP reports a small percentage of batch effects when there are no batch effects in the data.

PVCA estimates the percentage of batch effects more accurately when the percentage of batch effects in data is low (see Figures 3.2a and 3.2b). However, its




(a) (b)

Figure 3.1: Plots of the measured percentage of batch effects in the simulated RNA-seq data sets against the theoretical percentage of batch effects computed from data simulation hyper-parameters for data sets (a) with batch-class imbalance, and (b) without batch-class imbalance. The original data set without batch effects is included in the plot as well.

error increases exponentially as the percentage of batch effects in the data increases. gPCA $\delta$ fails to accurately and robustly quantify the percentage of batch effects in data with batch-class imbalance, at times reporting an increase in batch effects when the theoretical percentage of batch effects in the data is decreasing (see Figure 3.2d).

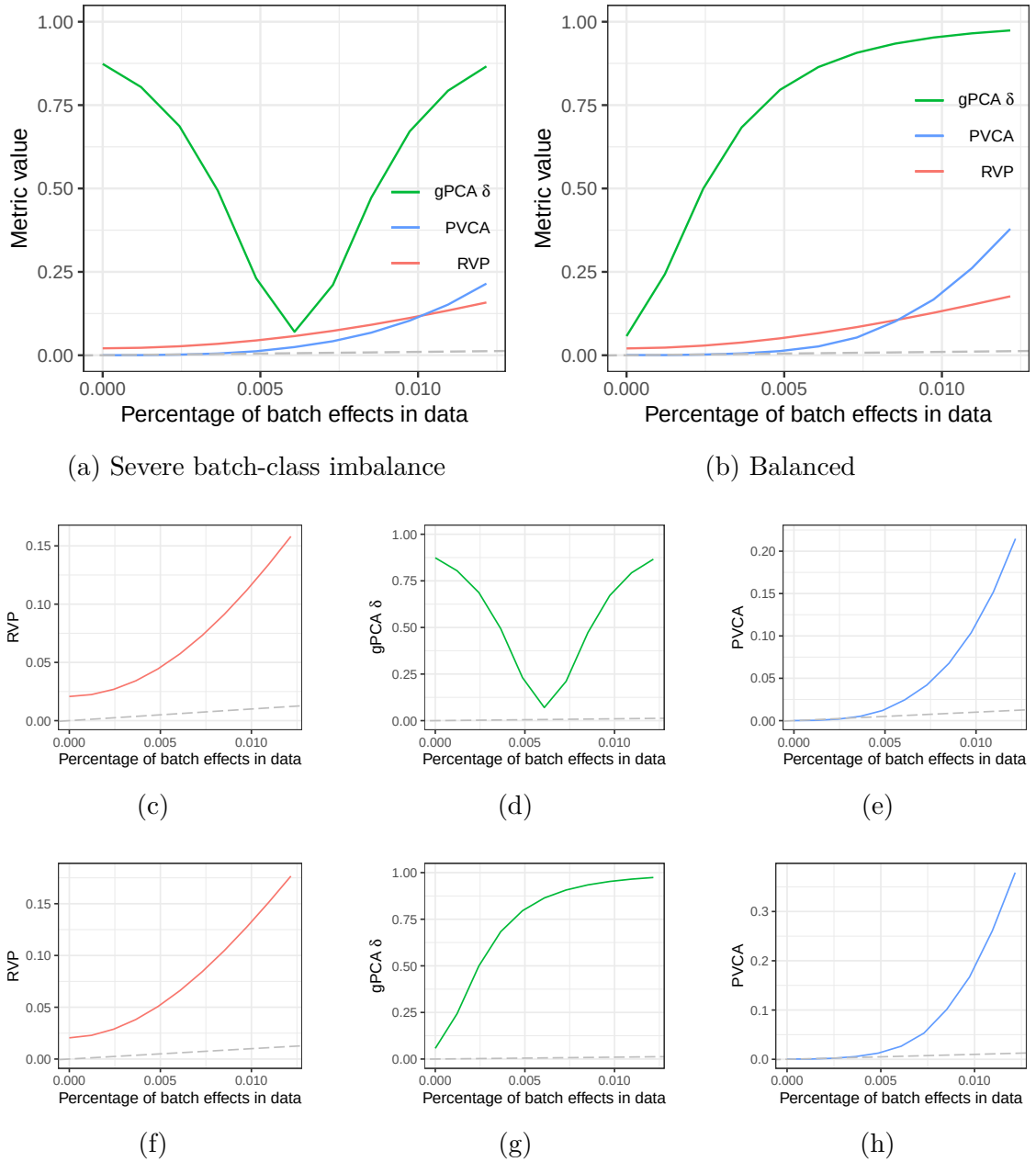
#### 3.6.1.2 Variance due to batch effects

We derive a theoretical estimate of variance due to batch effects in the simulated RNA-seq data, which is computed using the data simulation hyper-parameters. Let $\mathbf{W} \in \mathbb{Z}_{\geq 0}^{p \times n}$ be the count matrix with $p$ features and $n$ samples, whose elements $w_{ijkg}$ represent the expected value of additive batch effects in feature $i$ of the $j$-th sample from batch $k$ and class $g$:

$$w_{ijkg} = \mathbb{E}(X'_{ijkg} - X_{ijgk}) = \mathbb{E}(\omega_{ijkg}) = \delta_k$$

where $\delta_k$ is the mean parameter of the Poisson distribution used to simulate additive batch effect terms $\omega_{ijkg}$. The random variables $X'_{ijkg}$ and $X_{ijkg}$ denote counts with and without batch effects, respectively.

We use the total sample variance of $\mathbf{W}$, which we denote as $S^2_{\mathbf{W}}$, as our theoretical estimate of the variance in data that is attributable to batch effects. We compute




(a) Severe batch-class imbalance (b) Balanced

(c) (d) (e)

(f) (g) (h)

Figure 3.2: Plots of different batch effects metrics against the theoretical percentage of batch effects in (a) data with severe batch-class imbalance and (b) data without batch-class imbalance. (c-e) Individual magnified plots of each metric in (a). (f-h) Individual magnified plots of each metric in (b). RVP, gPCA $\delta$ and PVCA are coloured in red, green and blue, respectively. All metrics are percentage values, and are plotted as a decimal. For all metrics, higher values indicate stronger batch effects. Theoretical percentage of batch effects in data is plotted as a decimal. Dashed grey line (identity line) indicates the performance of a metric that correctly estimates the theoretical percentage of batch effects in the data with no error. Vertical distance between the curve and the dashed grey line represents the error at a specific batch effects percentage.

$S^2_{\mathbf{W}}$ by summing the sample variance of each row vector $\mathbf{w}_i$:

$$\begin{aligned} S^2_{\mathbf{W}} &= \sum_i S^2_{\mathbf{w}_i} \\ &= \sum_i \frac{1}{n-1} \sum_{k,g,j} (w_{ijkg} - \bar{w}_i)^2 \\ &= \frac{1}{n-1} \sum_i \sum_k n_k (\delta_k - \bar{w}_i)^2 \end{aligned}$$

where $n_k$ is the number of samples in batch $k$ and $\bar{w}_i$ is the mean of $\mathbf{w}_i$, which is given by:

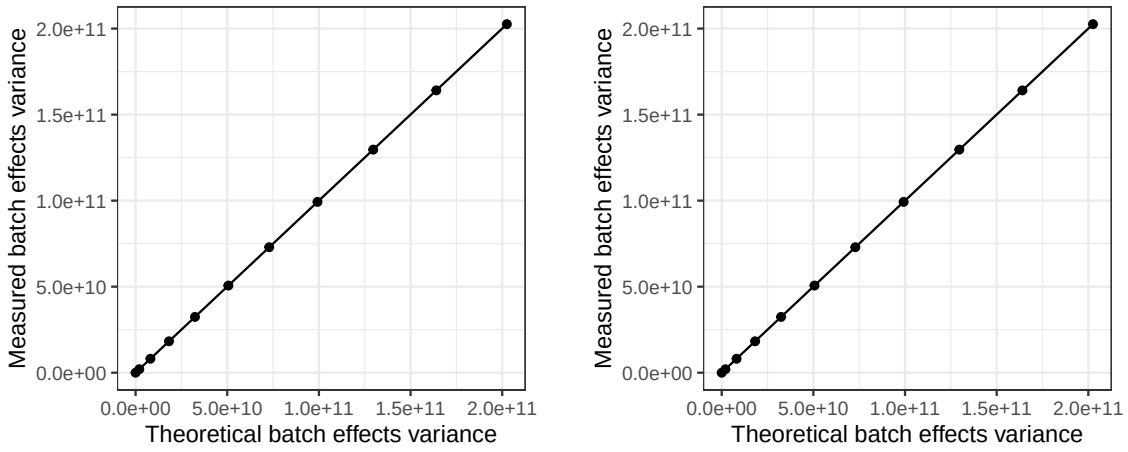
$$\bar{w}_i = \frac{\sum_{k,g,j} w_{ijkg}}{n} = \frac{\sum_k n_k \delta_k}{n}$$

As we simulated data following a two-batch and two-class design with equal batch and class sizes, the number of samples in each batch equals to half the total number of samples ($n_k = n/2$), and the batch effects magnitude of batches one and two can be defined as $\delta_1 = 0$ and $\delta_2 = \delta$. Thus, the above formula for $S^2_{\mathbf{W}}$ can be simplified to:

$$\begin{aligned} S^2_{\mathbf{W}} &= \frac{1}{n-1} \sum_i \sum_k \frac{n}{2} \left( \delta_k - \frac{\sum_k \frac{n}{2} \delta_k}{n} \right)^2 \\ &= \frac{pn}{2(n-1)} \sum_k \left( \delta_k - \frac{\delta}{2} \right)^2 \\ &= \frac{pn}{n-1} \left( \frac{\delta}{2} \right)^2 \end{aligned}$$

RVP estimates the percentage of variance attributable to batch effects in the data. By definition, it can also be used to estimate the variance due to batch effects in data by multiplying it with the observed sample variance of the data, which we denote as $S^2_{\mathbf{X}}$. Similarly, we can also multiply PVCA and gPCA $\delta$ by the observed sample variance of the data $S^2_{\mathbf{X}}$ to obtain their respective estimates of variance due to batch effects in data. This can be done as PVCA measures the proportion of variance attributable to batch effects, and gPCA $\delta$ is a ratio between two variances;

both PVCA and gPCA $\delta$ are percentages that when expressed as a decimal, range from zero to one.

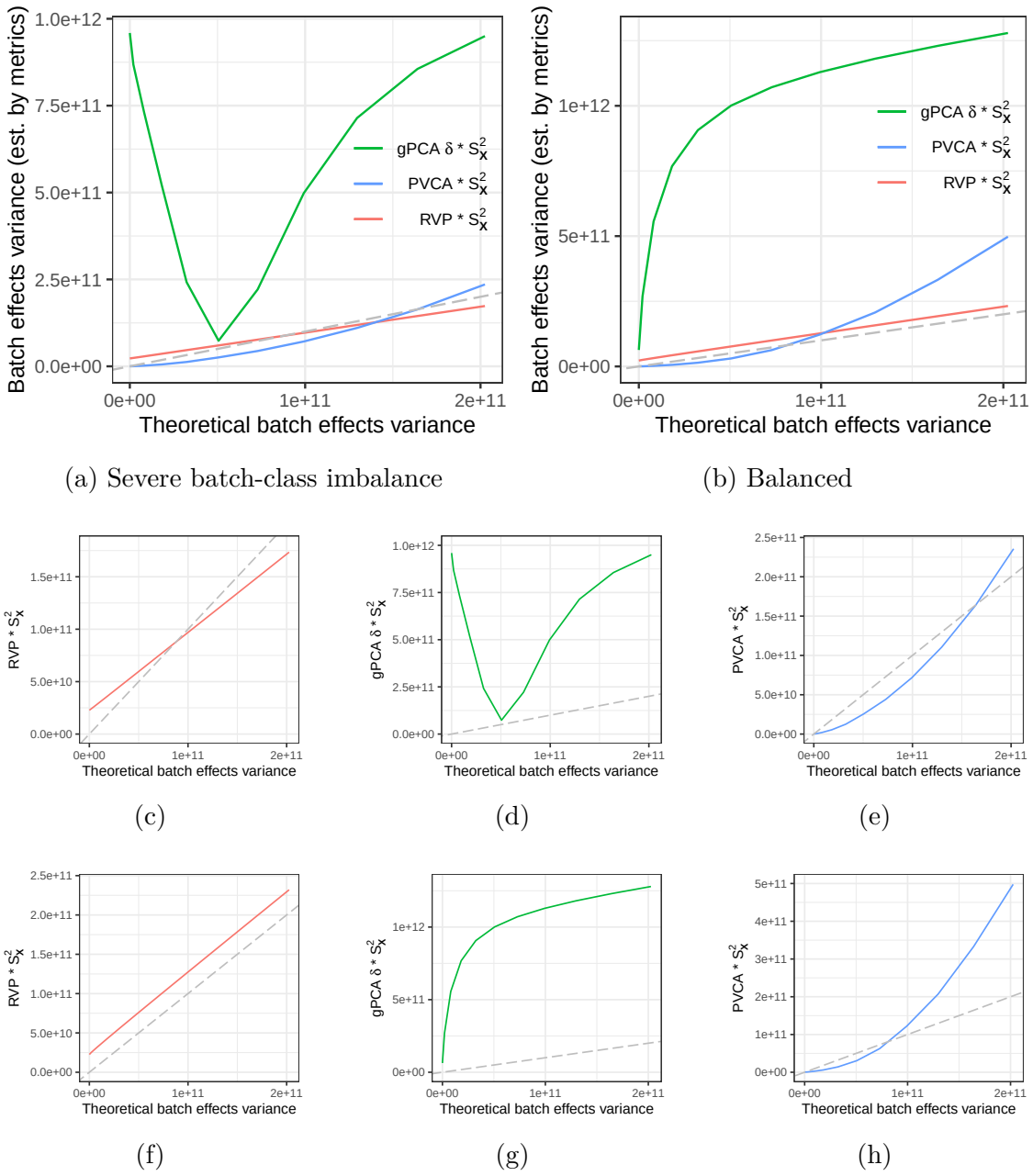



(a) (b)

Figure 3.3: Plots of the measured variance due to batch effects against the theoretical estimate of variance due to batch effects in simulated RNA-seq data sets (a) with batch-class imbalance, and (b) without batch-class imbalance. The original data set without batch effects is included in the plot as well.

We use the theoretical estimate of variance due to batch effects ($S^2_{\mathbf{W}}$) as the ground truth to evaluate the accuracy of each metric in quantifying the variance attributable to batch effects. In order to verify that the theoretical estimate of variance due to batch effects is correct, we first obtained the matrix of additive batch effect terms used to simulate each data set. Subsequently, we measured the total sample variance of the additive batch effects matrix for each data set. Comparison of the measured variance due to batch effects and the theoretical estimate of variance attributable to batch effects show that they correspond closely to each other in both data with and without batch-class imbalance (see Figure 3.3).

In Figure 3.4, we compare the batch effects variance estimated by each batch effects metric against the theoretical benchmark batch effects variance $S^2_{\mathbf{W}}$. RVP is able to accurately estimate variance due to batch effects in both data with and without batch-class imbalance (see Figures 3.4a and 3.4b). The batch effects variance estimated empirically using RVP has a linear relationship with the benchmark batch effects variance in both balanced and imbalanced data, and does not deviate much from the correct estimate (indicated by the dashed lines in Figures 3.4c and 3.4f).




(a) Severe batch-class imbalance

(b) Balanced

(c)

(d)

(e)

(f)

(g)

(h)

Figure 3.4: Plots of batch effects variance estimated by the different batch effects metrics, against the theoretical batch effects variance computed from the hyper-parameters used to simulate the data. (a) Comparison between metrics in data with severe batch-class imbalance and (b) Comparison between metrics in data without batch-class imbalance. (c-e) Individual magnified plots of each metric in (a). (f-h) Individual magnified plots of each metric in (b). RVP, gPCA $\delta$ and PVCA are coloured in red, green and blue, respectively. Dashed grey line (identity line) indicates the performance of a metric that correctly estimates batch effects variance in the data. Vertical distance between the curve and the dashed grey line represents error at a specific batch effects magnitude.

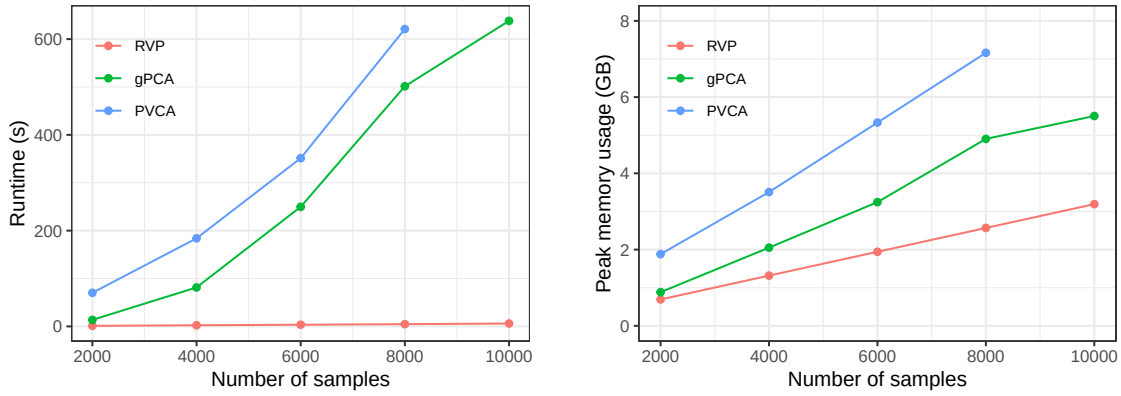
On the other hand, the batch effects variance estimated by PVCA increases exponentially as the benchmark batch effects variance increases, deviating more from the correct estimate in balanced data (indicated by the dashed line in Figure 3.4h). Batch effects variance estimated by gPCA $\delta$ deviates greatly from the benchmark batch effects variance in both balanced and imbalanced data. In imbalanced data, as the benchmark batch effects variance increases, the batch effects variance estimated by gPCA $\delta$ decreases at first before increasing again (see Figure 3.4d). Data with different batch effects variances may be erroneously estimated by gPCA $\delta$ to have equal batch effects variances. This undermines the reliability of gPCA $\delta$ in data with batch-class imbalance.

#### 3.6.1.3 Runtime and memory analysis

We evaluated the runtime and peak memory usage of RVP, gPCA and PVCA on five simulated RNA-seq data sets with different numbers of samples ($n = 2000, 4000, \ldots, 10000$). All data sets had 8000 features and did not have batch-class imbalance. Figure 3.5a shows that the runtime of RVP increases linearly with respect to the number of samples in the data set. On the other hand, the runtime of gPCA and PVCA increases polynomially with respect to the number of samples. We observe that RVP is two orders of magnitude faster than both gPCA and PVCA, taking 8.2 seconds to run on data with 8000 samples and 8000 features, as compared to 504 seconds for gPCA and 677 seconds for PVCA. PVCA failed to run on the data set with 10000 samples as it required more than 8 GB of RAM memory. We observe in Figure 3.5b that the peak memory usage of all the batch effects metrics increased linearly with respect to the number of samples, with RVP showing the smallest peak memory usage out of the three in all five data sets. For the data set consisting of 8000 samples, peak memory usage of RVP was roughly half that of gPCA and a third of the peak memory usage of PVCA.

RVP has a time complexity of $\mathcal{O}(mn)$ when used on a data matrix with $m$ features and $n$ samples. RVP performs basic arithmetic operations across $n$ samples independently for $m$ features, before summing up the intermediary results across all features. The number of basic arithmetic operations it performs is proportional to the number of entries in the data matrix $mn$. On the other hand, gPCA, PVCA and most batch effects metrics involve the use of PCA. Most implementations of PCA

perform SVD on the data matrix; the Jacobi SVD algorithm has a time complexity of $\mathcal{O}(m^2n + n^3)$ when performed on a $m \times n$ matrix (Golub & Van Loan, 2013). In the case of PVCA, in addition to performing PCA, PVCA fits a linear mixed model to PCA-transformed data using the REML method, which requires $\mathcal{O}(m^2n)$ time (Z. Tan et al., 2018).




(a) (b)

Figure 3.5: Comparison of (a) runtime and (b) peak memory usage of batch effects metrics when used to quantify batch effects in simulated RNA-seq data sets of different sample sizes. All data sets consist of 8000 features. PVCA failed to run on the data set with 10000 samples due to insufficient RAM memory.

### 3.6.2 Microarray and proteomics data

We compare the performance of RVP, PVCA and gPCA $\delta$ in quantifying batch effects in different microarray and proteomics data sets in Table 3.5. In particular, we investigate their performance in data with and without batch-class imbalance, both with and without batch effects. Different subsets of each data set are used to represent data with and without batch effects, and with different degrees of batch-class imbalance. All subsets that fall under the same category (having or not having batch effects) contain an equal number of samples and have equal batch and class sizes, to facilitate fair comparison.

gPCA $\delta$ is found to be highly unreliable in quantifying batch effects in data with batch-class imbalance. gPCA $\delta$ consistently reports strong batch effects in data with no batch effects when there is severe batch class imbalance. This corroborates with previous findings that gPCA $\delta$ over-estimates the magnitude of batch effects

in data with batch-class imbalance (Zhou et al., 2019). In comparison, both RVP and PVCA correctly report low magnitudes of batch effects in data without batch effects that have severe batch-class imbalance.

In data with batch effects, gPCA $\delta$ values increase along with the severity of batch-class imbalance, even though batch sizes and the total number of samples in the different subsamples of each data set remain the same. On the other hand, both RVP and PVCA report consistent values in the different subsamples of each data set with different degrees of batch-class imbalance.

| | No batch effects | | Batch effects | |
|---|---|---|---|---|
| | Balanced | Severe imbalance | Balanced | Moderate imbalance |
| gPCA $\delta$ | 0.345 | 0.545 (+0.20) | 0.701 (+0.36) | 0.692 (+0.35) |
| PVCA | 0.0353 | 0.0738 (+0.039) | 0.233 (+0.20) | 0.191 (+0.16) |
| RVP | 0.123 | 0.142 (+0.019) | 0.247 (+0.12) | 0.242 (+0.12) |

(a) Ma-Spore ALL data set

| | No batch effects | | Batch effects | |
|---|---|---|---|---|
| | Balanced | Severe imbalance | Balanced | Moderate imbalance |
| gPCA $\delta$ | 0.0605 | 0.858 (+0.80) | 0.122 (+0.062) | 0.469 (+0.41) |
| PVCA | 0.0312 | 0 (-0.031) | 0.101 (+0.070) | 0.0896 (+0.058) |
| RVP | 0.0879 | 0.0864 (-0.0015) | 0.116 (+0.028) | 0.112 (+0.024) |

(b) MAQC-I data set

| | No batch effects | | Batch effects | | |
|---|---|---|---|---|---|
| | Balanced | Severe imbalance | Balanced | Moderate imbalance | Severe imbalance |
| gPCA $\delta$ | 0.0499 | 0.942 (+0.89) | 0.416 (+0.37) | 0.584 (+0.53) | 0.834 (+0.78) |
| PVCA | 0.0335 | 0.0646 (+0.031) | 0.260 (+0.23) | 0.215 (+0.18) | 0.178 (+0.14) |
| RVP | 0.0532 | 0.0331 (-0.020) | 0.283 (+0.23) | 0.264 (+0.21) | 0.218 (+0.16) |

(c) Westlake proteomics data set

Table 3.5: Values of batch effects metrics when used to quantify batch effects in different subsets of the (a) Ma-Spore ALL data set, (b) MAQC-I data set and (c) Westlake proteomics data set. Different subsets represent data with varying degrees of batch-class imbalance, both with and without batch effects. Differences between each metric value and its value in data with no batch effects and no batch-class imbalance are included in brackets. Differences are comparable across metrics as they are all percentages expressed as decimals that range from zero to one.

The difference in metric values when measuring balanced data without batch

effects, and balanced data with batch effects, can be interpreted as the magnitude of a true positive when actual batch effects are detected (see bracketed values in Table 3.5). Similarly, the difference in metric values when measuring balanced data with no batch effects, and imbalanced data with no batch effects, can be thought of as the magnitude of a false positive that results from batch-class imbalance. The ratio of the magnitude of true positive, to the magnitude of false positive of a batch effects metric can be interpreted akin to its signal to noise ratio. RVP exhibits the highest ratio across all data sets, which highlights its reliability in quantifying batch effects in data with batch-class imbalance. However, it can be observed that RVP has a slight bias in estimating batch effects in balanced data with no batch-class imbalance, deviating slightly from the desired value of zero.

## 3.7 Discussion

Batch effects can be modelled either as additive effects, multiplicative effects or both additive and multiplicative effects (Lazar et al., 2012). An increase in magnitude of additive batch effects can be visualised as an increase in distance between batches in a PCA plot, while an increase in multiplicative batch effects can be visualised as an increase in dispersion of samples in any of the batches. RVP is designed to quantify additive batch effects, and may fail to capture variance caused by multiplicative batch effects. However, we believe that additive batch effects dominate multiplicative batch effects, especially after log-transformation of data.

In this chapter, we presented RVP, a novel quantitative batch effects metric suitable for use in small data sets. We demonstrated that RVP is able to accurately quantify the magnitude of batch effects across a wide range of magnitudes, using simulated RNA-seq data. We also showed that RVP outperforms other metrics in quantifying batch effects in real-world gene expression microarray and quantitative proteomics data sets, especially in data with batch-class imbalance. RVP is hardly affected by batch-class imbalance, and achieves similar estimates of magnitude of batch effects in both data with or without batch-class imbalance. We argued that RVP has a lower time complexity than PVCA and gPCA, as it does not involve computationally expensive operations such as PCA. We substantiated our claim by demonstrating that RVP runs orders of magnitude faster than gPCA and PVCA on

a data set with 8000 features and 8000 samples. In addition, we showed that the peak memory usage of RVP is less than both gPCA and PVCA when benchmarked on the same data sets.

# Chapter 4

# Accounting for treatment differences

## 4.1 Introduction

Clinical prediction models are often used for estimating the absolute risk of clinically important outcomes in patients. Prediction models can be employed to support clinical decision making, with estimated risks of clinically poor outcomes being used to guide treatment initiation in individuals. Examples of real-world applications include the Framingham risk score in cardiovascular disease (CVD) and Apgar score for the prognosis of newborns (Apgar, 1952; Wilson et al., 1998).

Differences in patient treatment occur frequently in clinical data as many diseases are treated using a risk-adapted strategy. Failure to handle treatment effects properly during the development of prediction models results in inaccurate models that have low generalisability (Groenwold et al., 2016; Sperrin et al., 2018).

This can be illustrated by a famous example presented by Caruana et al., 2015, where a machine learning model learns that patients with a history of asthma have a lower risk of dying from pneumonia. Information regarding intensive care unit (ICU) admission of asthmatic patients was not fed to the model, which resulted in the model erroneously learning that asthma is associated with lower risk. Deploying the model in a population where asthmatics are not admitted to the ICU would prove disastrous, as these patients would be predicted as low risk when they in fact require intensive treatment.

Despite the need to properly account for patient treatments, few prediction modelling studies report patient treatments in sufficient detail. Surveys on prediction modelling studies in CVD highlighted that a significant portion of studies do not mention treatment use, and that most studies fail to mention treatment initiated

after the time of prediction (Liew et al., 2011; Pajouheshnia et al., 2017).

Treatment strategies developed for diseases are highly specific and usually involve a combination of different treatments that may include both medical and non-medical interventions. In addition to treatment strategies differing between diseases, treatment strategies for the same disease evolve over time as well. In general, treatments differ in whether they are initiated prior to prediction (baseline treatment) or after the time of prediction (treatment drop-in). In complex cases, intensity levels of treatment are altered throughout the course of treatment.

As the treatment for each disease varies greatly, there is no one-size-fits-all approach in handling treatment effects during the development or evaluation of prediction models; the appropriate approach often depends on each individual situation. As a result, there are many subtleties to consider when accounting for treatment effects during the development or evaluation of prediction models in different settings.

In this chapter, we introduce common methods that are used to handle treatments during the development of prediction models, and discuss issues associated with these methods. When treatment strategies are the same in the development and deployment cohorts, ignoring treatment may be a viable approach. However, when treatment strategies are different between cohorts, patient treatments should be accounted for. We illustrate the benefits of incorporating treatment information during the evaluation of prediction models through a case study of the Ma-Spore ALL data set (A. E. J. Yeoh et al., 2012; A. E. J. Yeoh et al., 2018). We suggest a scoring scheme that incorporates patient treatment information to assess whether treatment intensity predictions are correct. We encourage examining individual cases in detail during the analysis and evaluation of clinical prediction models in order to properly account for the heterogeneities in clinical data.

## 4.2 Methods in handling treatment differences

Clinical prediction models are frequently employed to facilitate treatment initiation decisions, which requires estimating the risk of adverse outcomes in the absence of treatment (Groenwold et al., 2016; Hemingway et al., 2013). As clinical data often consists of a mix of treated and untreated patients, it is essential to properly account

for patient treatments when developing a model to support treatment initiation decisions. The most common methods used to handle treatments include ignoring treatment, restricting analysis to untreated patients, including treatment together with the adverse outcome as a composite outcome, modelling treatment, and hypothetical prediction (Groenwold et al., 2016; Sperrin et al., 2018; van Geloven et al., 2020). However, modelling treatments will only work if knowledge of treatment use is known at the time of prediction (i.e. will not work for treatment drop-ins).

### 4.2.1 Ignoring treatment

Many prediction modelling studies choose to ignore the fact that treated patients are present in the data being using to develop the models (Pajouheshnia et al., 2017). If the treatment was effective in reducing risk of adverse outcome in patients, treated patients would have a lower chance of suffering from the adverse outcome. Not accounting for treated patients in the data used to train the model would result in a model that under-estimates risk when deployed on untreated patients. The effectiveness of treatment and proportion of treated patients in the data would determine the extent to which the developed models would under-estimate risk (Groenwold et al., 2016).

Ignoring treatment also leads to a "treatment paradox" (Peek et al., 2017). Suppose a model predicts patients as high risk due to a specific predictor, and treatment is given to patients predicted as high risk. If the treatment is effective and lowers the patients' probability of having an adverse outcome, the predictor-outcome association would be attenuated. If the data is used to develop a model without accounting for treatment, the model would learn that the predictor is now associated with low risk instead, contradicting the original prediction.

### 4.2.2 Restricting analysis

Another commonly used method is to restrict analysis to untreated patients only. This method is susceptible to selection bias, as often patients deemed to be low risk are excluded from treatments (Sperrin et al., 2018). Developing models using data consisting only of untreated patients with a lower probability of the outcome might lead to an under-estimation of risk when the model is validated on the full range of

patients with different levels of risk (Groenwold et al., 2016).

### 4.2.3 Composite outcome

A method that can be used to handle treatment drop-ins is to combine the treatment together with the original outcome as a composite outcome. Essentially, we are estimating the risk of occurrence of the outcome or administration of treatment in a patient. One of the cases better suited for this method is when treatment most likely prevented the occurrence of the outcome (e.g. patients would most probably suffer from myocardial infarction if they did not undergo surgery).

### 4.2.4 Modelling treatment

Baseline treatments received by patients can be accounted for by explicitly including treatment as a predictor in the prediction model. However, as information about treatment drop-ins are unavailable at the time of prediction, this method does not apply to them.

### 4.2.5 Hypothetical prediction

The hypothetical prediction method accounts for treatments by performing a counterfactual prediction - predicting risk in a hypothetical world where treated patients are not given treatment instead. Causal inference methods, such as marginal structural models and g-formulas, are used in the estimation of hypothetical untreated risk (Dickerman et al., 2022; Sperrin et al., 2018). However, the validity of these estimates are conditioned on three key assumptions: exchangeability of treated and untreated patients given measured confounders, positivity (having a non-zero number of treated and untreated patients for all covariate patterns) and consistency, where a patient's hypothetical risk is equal to her observed risk in the real world (Dickerman et al., 2022; van Geloven et al., 2020). These assumptions are often unverifiable empirically in observational studies, where treatment is not randomised. In addition, it is challenging to 1) account for differences between interventional and conditional distributions, 2) identify all potential confounders and colliders, and 3) avoid model misspecification (Prosperi et al., 2020).

## 4.3 Prediction estimands

Accounting for treatment using different methods during the development of prediction models results in different models with subtly different estimands (van Geloven et al., 2020). Ignoring treatment effectively results in a model that estimates a patient's risk of the adverse outcome, given that patients are treated according to the current treatment strategy. Deployment of the model in a population with the same treatment strategy would yield legitimate predictions, as similar associations between predictors, treatment and outcome exist. If the deployed model predicted a good outcome for a patient, the patient should continue on the same treatment strategy; however, if the model predicted a poor outcome, the patient would do better not to follow the same treatment strategy. In cases where the deployment population does not have the same treatment strategy, it is best to account for treatment in order to avoid inaccurate predictions.

## 4.4 Case study: Ma-Spore ALL data set

Different diseases require different treatment strategies, each with their own set of issues and complexities. Treatment strategies should be reported in detail so that necessary actions can be taken to account for them during both the development and evaluation of prediction models. We use data from the Ma-Spore ALL 2003 and 2010 studies (A. E. J. Yeoh et al., 2012; A. E. J. Yeoh et al., 2018) as a case study to demonstrate some of the issues and complexities related to treatment, and offer suggestions on how to deal with them in a nuanced manner.

Patients in the Ma-Spore ALL data set underwent risk-adapted chemotherapy and were treated at different intensities at different times throughout the course of treatment. In addition, eligible patients in the high risk subtype BCR-ABL underwent a bone marrow transplant (BMT). Patient treatment information was not included in the public Ma-Spore ALL data set. We requested for patient treatment information from the authors and were granted information regarding, firstly, the final treatment intensity patients were treated at and, secondly, whether they underwent BMT (A. E. J. Yeoh et al., 2018).

However, details regarding patient treatment intensities at other points in time

during the therapy were not available. We inferred each patient's treatment intensities throughout the course of treatment from patient metadata and the set of decision rules in the risk-adapted treatment protocol. We concluded that the varying treatment intensities that patients received at different times throughout the course of treatment could be simplified down to four possible treatment routes. Patients undergo one out of four possible treatment routes, which we denote as standard risk (SR), intermediate risk (IR), high risk one (HR1) and high risk two (HR2).

We illustrate the treatment routes that patients can undergo in Figure 4.1. All patients start off on IR treatment intensity, and may be re-assigned to other treatment intensity levels at only two specific time points, Day 8 and post-induction (i.e. after the induction phase of chemotherapy). Patients who are escalated to HR treatment intensity will remain at HR treatment intensity for the remaining course of treatment. In the HR1 treatment route, patients are elevated to HR treatment intensity at Day 8. Patients who are assigned to the HR1 treatment route are primarily from high risk subtypes, namely the BCR-ABL, MLL and hypodiploid subtypes. In the HR2 treatment route, patients are elevated to HR treatment intensity post-induction. Patients who undergo the HR2 treatment route mostly have poor treatment response which is evidenced by their high Day 33 or Week 12 minimal residual disease (MRD) values. Patients who meet a set of stringent criteria, including having low Day 33 and Week 12 MRD values, will be assigned to the SR treatment route where patients will be treated on IR treatment intensity before being de-escalated to SR treatment intensity post-induction. In the IR treatment route, patients are treated on IR treatment intensity throughout the course of chemotherapy.

## 4.5 Model evaluation: Incorporating treatment information

The treatment strategy for ALL is more complicated than the simple case where patients are either treated or untreated. Not only are ALL patients treated at different intensities at different points in time throughout chemotherapy, eligible high risk patients also receive BMT. In this section, we propose a scoring scheme that incorporates the use of patient treatment information. We demonstrate why

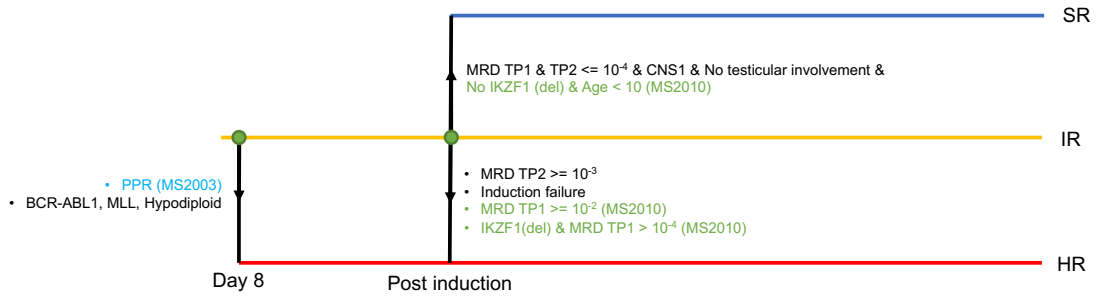



Figure 4.1: Risk-adapted treatment plan employed in Ma-Spore acute lymphoblastic leukaemia 2003 and 2010 studies (MS2003 and MS2010). All patients start off on the intermediate risk (IR) treatment arm, and treatment intensity may be altered at two decision time-points, namely Day 8 of chemotherapy and post induction. The criteria for escalation to high risk (HR) or de-escalation to standard risk (SR) is stated next to the arrowheads. Bullet points represent an "OR" relationship between criteria. Criteria specific to MS2003, MS2010 are highlighted in light blue and light green text, respectively. PPR: Poor prednisolone response, MRD: Minimal residual disease, TP1: Time-point one, TP2: Time-point two.

it is beneficial to utilise treatment information during the evaluation of prediction models.

Consider a prediction model that predicts the risk of relapse given that patients are treated according to the risk-adapted treatment protocol. The predicted risk of relapse is used to recommend the treatment intensity that patients should receive (SR, IR or HR). The correct treatment intensity has to achieve a balance between maximising a patient's probability of achieving continuous complete remission (CCR) and reducing the cytotoxic effects of chemotherapy. As achieving CCR is significantly more important than eliminating cytotoxic side-effects, the correct treatment intensity is essentially the lowest treatment intensity level at which a patient still achieves CCR.

Determining the correct treatment intensity patients should receive requires knowledge of individual treatment outcome under counterfactual treatments. However, there is no ground truth of counterfactual events as they do not occur in the real world. For example, if a patient who was on IR treatment was predicted to require SR treatment, we do not know for sure how the patient would respond if given SR treatment. We overcome the absence of ground truth by logically deducing whether a treatment recommendation is correct or not based on each patient's treatment information and treatment outcome. We deduce whether treatment

Table 4.1: Scoring scheme for treatment intensity recommendations for paediatric ALL patients.

| Outcome | Treatment | Prediction | Score |
|---|---|---|---|
| Remission | SR | SR | 1 |
| Remission | SR | IR | 0 |
| Remission | SR | HR | 0 |
| Remission | IR | SR | 1 |
| Remission | IR | IR | 1 |
| Remission | IR | HR | 0 |
| Remission | HR | SR | 0 |
| Remission | HR | IR | 1 |
| Remission | HR | HR | 1 |
| Relapse | SR | SR | 0 |
| Relapse | SR | IR | 1 |
| Relapse | SR | HR | 1 |
| Relapse | IR | SR | 0 |
| Relapse | IR | IR | 0 |
| Relapse | IR | HR | 1 |
| Relapse | HR | SR | 0 |
| Relapse | HR | IR | 0 |
| Relapse | HR | HR | 1 |

intensity recommendations are correct or not for all combinations of a patient's treatment outcome (relapse or CCR), treatment route (SR, IR or HR) and predicted treatment intensity recommendation (SR, IR or HR). We present a scoring scheme encompassing all cases in Table 4.1. All conceivably correct recommendations are awarded a score of one, while incorrect recommendations are given a score of zero.

We elaborate on how we logically deduced whether treatment intensity recommendations are correct or incorrect in the plausible framework we suggested above. For patients who relapsed, treatment intensity recommendations that are above the treatment intensities patients were treated at are deemed to be correct, while recommendations that are below or equal to the treatment intensities patients received are incorrect. This stems from the reasoning that patients who relapsed were treated at an insufficient treatment intensity level, and would benefit from a higher treatment intensity level. In the case that the patient relapses even when treated on the highest intensity level (HR), we take HR to be the correct treatment intensity. For patients who achieved CCR, recommendations that are equal to or less

intense (by one risk level) than the actual treatment intensity received are deemed to be correct, while a recommendation at a higher level than the treatment level received would be incorrect. This penalises cases of over-treatment, while awarding treatment de-escalation that will help reduce toxic side-effects of chemotherapy.

However, one subtlety regarding the above framework is that it relies on the implicit assumption that patients are able to complete the treatment intensity they were prescribed. In reality, some patients may discontinue treatment for various reasons, such as being unable to tolerate the side effects of the prescribed treatment intensity. Consequently, these patients suffer from relapse or treatment failure. For example, patients who receive HR treatment but are unable to tolerate the high dosage typically discontinue treatment, thereby ending up suffering from relapse or treatment failure. There has been sporadic evidence that such patients could potentially benefit from a lower dosage (B. K. J. Tan et al., 2021). These patients are different from patients who are able to complete treatment at the prescribed intensity but still relapse. In these cases (patients who relapsed previously but who might not be able to tolerate high intensity treatments), it is not straightforward to decide whether these recommendations of increased intensity are correct or incorrect.

## 4.6 Benefits of incorporating treatment information

To illustrate the benefits of incorporating treatment information during evaluation of prediction models, we consider a classifier $C$ that is able to predict treatment outcome labels perfectly. Originally, patients in the Ma-Spore ALL data set are treated using a risk-adapted treatment strategy, where decisions are made whether to alter patient treatment intensity at two time-points, based on individual patient features such as ALL subtype and MRD (see Figure 4.1). We look to adopt a new treatment strategy, where instead of deciding whether to alter patient intensity according to the original strategy, we use predictions from classifier $C$ to decide whether to escalate or de-escalate patient treatment intensity. Patients who are predicted to relapse would have their treatment escalated to HR, while patients predicted to achieve CCR would be de-escalated to SR treatment.

Evaluation of the new treatment strategy using traditional metrics such as

accuracy may give misleading results. If patients who relapsed are assumed to require HR treatment and patients who achieved CCR are assumed to be suitable for de-escalation to SR treatment, classifier $C$ would be deemed to have an accuracy of 100%. However, some patients who achieved CCR were treated on HR treatment intensity, and treatment recommendations of SR may not be sufficient for these patients, who may end up relapsing instead. Evaluation using our scoring scheme, which incorporates treatment information, would identify these recommendations as incorrect.

It is tricky to evaluate the above group of patients, as there are no ground truths available for these counterfactual claims. Patients who were treated at HR intensity and achieved CCR may either require HR treatment intensity to achieve CCR, or on the other hand may benefit from de-escalated treatment. A way to infer which of the two possibilities patients belong to is by examining a feature that is indicative of treatment response, such as the Day 33 MRD. Patients with good treatment response will be more likely to benefit from de-escalation of treatment.

## 4.7 Handling complex treatment differences

It is difficult to account for treatment differences in diseases with complex risk-adapted treatment plans (e.g. ALL). Developing models that estimate the risks of patients under the counterfactual assumption that they are untreated is best done by using causal inference methods. Although these methods work well when differences in treatment are simple (e.g. where patients either received treatment or not), they become unwieldy when differences in treatment are complex.

Instead of attempting to estimate the counterfactual risk of (hypothetically) untreated patients, we suggest a simple approach to deploy prediction models to guide treatment decisions when there are complex treatment differences between patients. This approach can only be used if patients undergo the same treatment plan in both the development and deployment cohort.

In this approach, the objective of the prediction model is to predict whether patients should continue to be treated on the original treatment plan, or be excluded from it. Prediction models that are developed without accounting for differences in patient treatment can be used, as the treatment strategy remains the same in

the deployment cohort. Using ALL as an example, a model that predicts whether a patient will relapse or achieve CCR can be used to guide treatment decisions in the following way: patients predicted to achieve CCR are recommended to continue with the existing treatment strategy, while patients predicted to relapse are recommended to be excluded from the existing treatment plan. If the latter group of patients are able to tolerate higher treatment intensity, they should be recommended to receive escalated treatment. However, if they are unable to tolerate higher treatment intensity, they may instead benefit from de-escalated treatment (B. K. J. Tan et al., 2021). In the worst case where they do not respond to treatment at all, taking patients off treatment would eliminate painful side effects of treatment and help in saving costs. A shortcoming of this approach is that we do not identify patients who might benefit from de-escalated treatment amongst those who are predicted to achieve CCR.

## 4.8 Closing remarks

Differences in patient treatments is a source of heterogeneity in clinical data that, if left unaccounted for, would lead to improper development of prediction models. Treatment information of patients should be reported in detail to facilitate the development and evaluation of clinical prediction models. We believe that proper handling of differences in patient treatment is crucial for the development of accurate and generalisable models. Incorporation of treatment information during evaluation of prediction models allows for more detailed and accurate evaluation.

Different treatment strategies are employed for different diseases. There is no one-size-fits-all approach in handling treatment differences, especially when patient treatment differences are of varying degrees of complexity in each disease. The best approach to handle differences in patient treatment is dependent on the characteristics of each situation.

In clinical data, patients are typically treated on a case by case basis, with clinicians prescribing treatment based on a patient's individual condition. Hence, examining individual patient details and taking consideration of individual patients in clinical data during the analysis and evaluation of clinical prediction models would lead to a more accurate and nuanced interpretation. As illustrated earlier,

evaluating clinical prediction models using a scoring scheme which incorporates patient treatment information allows for a more nuanced evaluation of treatment intensity recommendations.

# Chapter 5

# Subtype-specific treatment outcome prediction

## 5.1 Introduction

Acute lymphoblastic leukaemia (ALL) is the most common type of paediatric cancer (Inaba et al., 2013). Contemporary ALL treatment protocols adjust the intensity of chemotherapy according to a patient's risk of relapse. This is done to maximise the patient's chances of achieving continuous complete remission (CCR) while minimising cytotoxic side effects. The earliest prognostic factors used to assess risk were age and white blood cell (WBC) count at diagnosis. The National Cancer Institutes (NCI) risk criteria (Smith et al., 1996) assigned patients based on these prognostic factors as high risk (age $\geq 10$ or WBC count $\geq 50 \times 10^9/l$) or low risk (age $< 10$ and WBC count $< 50 \times 10^9/l$). As ALL is a heterogeneous disease consisting of various subtypes that respond differently to chemotherapy (Pui et al., 2008), genetic subtypes were also found to be good risk factors and incorporated into risk stratification strategies. Later studies discovered that early response during remission induction therapy was a good predictor of treatment outcome (Miller et al., 1989; Riehm et al., 1987). Early response was previously measured by a morphological count of blasts in peripheral blood or bone marrow. Technological advancements have enabled the measure of residual leukaemic cells that are present in morphologically undetectable amounts, through techniques such as flow cytometry and quantitative polymerase chain reaction (qPCR). These measurements are known as minimal residual disease (MRD), and have been shown to be the most powerful prognostic factor in ALL (Cavé et al., 1998; Coustan-Smith et al., 1998; van Dongen

et al., 1998).

Contemporary approaches to risk stratification rely on a combination of the above-mentioned clinical, cytogenetic and response features (A. E. J. Yeoh et al., 2012; A. E. J. Yeoh et al., 2018). However, a substantial percentage of patients identified by these approaches as low risk or intermediate risk still relapse (Conter et al., 2010). In addition, these approaches have a bias towards classifying patients as high risk. Patients who fulfil one of the high risk criteria, such as being of high risk subtype or having high risk MRD are classified as high risk, regardless of their other features. For example, a patient belonging to a high risk subtype would be classified as high risk, even if their MRD indicates imperceptible residual disease. Risk stratification performed by these approaches can be better improved.

In this work, we posit that different genetic subtypes of ALL have distinct responses to treatment, and propose the use of a subtype-specific prediction model. We explore the use of time-series gene expression profiles (GEPs) in predicting whether a patient will relapse or achieve long-term remission after chemotherapy. Each microarray GEP of a patient measures the average gene expression from all leukaemia and normal cells present in the patient sample. Chemotherapy kills leukaemia cells at a higher rate than normal cells, resulting in a change in the proportion of leukaemia and normal cells. In patients who are more responsive to chemotherapy, the difference between the killing rates of leukaemia and normal cells are larger. This results in a larger change in the proportion of leukaemia and normal cells.

We propose three transcriptomic features computed from a patient's Day 0 and Day 8 GEPs that measure the change in proportion of leukaemia and normal cells. Firstly, the effective response metric (ERM) ratio serves as a measure of the magnitude of a patient's treatment response along the direction of the leukaemia-to-normal trajectory. A larger ERM ratio corresponds to a larger shift between a patient's Day 0 and Day 8 GEPs along the leukaemia-to-normal trajectory. This implies that the treatment has a larger impact on a patient's GEP in the correct direction, shifting a leukaemia patient's GEP closer to that of a normal patient. Secondly, the absolute response metric (ARM) ratio measures the magnitude of a patient's treatment response, irrespective of direction. A larger ARM ratio corresponds to a larger absolute change between a patient's Day 0 and Day 8 GEP,

and indicates a larger treatment impact, whether it be in a positive or negative manner. Thirdly, the reorientation ratio, which we denote as $\phi$, is an alternative measure of a patient's treatment response from Day 0 to Day 8. The change in a patient's Day 0 to Day 8 GEP is captured in terms of a change in the direction of vectors which are anchored at the centroid of Day 0 leukaemia GEPs and point towards a patient's Day 0 and Day 8 GEPs. A larger reorientation ratio implies a larger difference between a patient's Day 0 and Day 8 GEPs, and hence a larger treatment response. These three transcriptomic features serve as different measures of a patient's treatment response. We hypothesise that for patients from the same subtype, having a larger ERM ratio, ARM ratio, or reorientation ratio are all indicative of a higher probability of long-term remission.

Aside from these transcriptomic features, the prognostic factor MRD still remains an effective measure of leukaemic cell clearance and hence treatment response. Larger MRD values indicate that patients have relatively larger proportions of residual leukaemia cells after chemotherapy (i.e. patients respond less to chemotherapy).

The ERM ratio and ARM ratio are derived from the original features ERM and ARM (A. E.-J. Yeoh et al., 2018), respectively. While these features are similar measures of treatment response, the fundamental difference between our proposal and A. E.-J. Yeoh et al., 2018 is that the former presents a subtype-specific prediction model, while the latter proposes the use of a global ALL model. In the global ALL model, both feature selection and feature computation are performed in a subtype-agnostic manner. All subtypes are assumed to have a similar leukaemia-normal trajectory; a common leukaemia centroid is calculated using patients from all subtypes. Patients with a higher ERM are taken to have a higher probability of long-term remission, regardless of their subtype. On the other hand, our subtype-specific prediction model is different in three main ways. Firstly, feature selection and feature computation are performed for each subtype individually. Secondly, the monotonic relationships between each of the patient's features (viz. ERM ratio, ARM ratio, $\phi$ and log-transformed MRD), and the probability of long-term remission are only assumed to exist between patients from the same subtype. Lastly, ERM ratio and ARM ratio expresses the magnitude of a patient's treatment response in terms of the remaining difference between a patient's Day 8 GEP and the centroid of normal patients' GEPs.

In this chapter, we present a subtype-specific model for treatment response prediction that is robust on small data sets. Our subtype-specific model fits a separate model for each subtype, and computes the transcriptomic features ERM ratio, ARM ratio and $\phi$ from time-series GEPs. These transcriptomic features are taken together with MRD for use in estimating a patient's probability of remission. We show that our subtype-specific model is better able to discriminate between relapse and long-term remission patients, and achieves more refined risk stratification than existing methods. In addition, our subtype-specific model is able to identify prime candidates for reduced-intensity chemotherapy with high precision.

### 5.1.1 Contributions

The main contributions of our work are as follows. First, we propose a subtype-specific model for treatment outcome prediction that estimates the probability of remission. Subtypes are considered individually, allowing for subtype-specific treatment responses to be modelled. Our subtype-specific model computes novel transcriptomic features using time-series transcriptomic data. These features reflect the change in the proportions of leukaemic and normal cells as a response to treatment. Our subtype-specific model is able to robustly estimate the probability of remission even when sample sizes within the individual subtypes are small. In addition, our subtype-specific model does not discretise MRD values, hence leading to no loss of information.

Second, we present our biological hypothesis and substantiate our hypothesis by demonstrating empirically that B-lineage cells decrease in long-term remission patients but not in relapse patients belonging to BCP-ALL subtypes. We infer the quantity of B-lineage cells using the deconvolution algorithm MCP-counter.

Third, we demonstrate that our subtype-specific model outperforms existing methods based on transcriptomic data and MRD when predicting treatment outcome in homogeneous ALL subtypes. In addition, we demonstrate the use of our subtype-specific model on the novel subtype DUX4-rearranged. Patients from DUX4-rearranged were identified through the use of transcriptomic data. PCR-based MRD erroneously classifies majority of DUX4-rearranged patients as intermediate or high risk, even though the DUX4-rearranged subtype has been shown to have a good

prognosis in general. In contrast, our subtype-specific model is able to discriminate well between long-term remission and relapse patients.

## 5.2 Background

### 5.2.1 Acute lymphoblastic leukaemia

Leukaemia is a cancer that affects blood cells, which originate from the bone marrow (BM). Leukaemia is classified according to the type of blood cells it affects. Lymphoblastic leukaemia affects cells that arise from lymphoid progenitor cells (i.e. lymphoid cell line), while myeloid leukaemia affects cells that arise from myeloid progenitor cells. In addition, acute leukaemia refers to leukaemia that progresses quickly, while chronic leukaemia refers to leukaemia that progresses slowly. Acute lymphoblastic leukaemia (ALL) is the most common form of leukaemia in children, while the most common form of leukaemia in adults is acute myeloid leukaemia (AML). In this paper, we specifically focus on predicting treatment outcome in paediatric ALL.

ALL can first be broadly classified into three groups according to the type of lymphocyte that has become malignant. Firstly, there is B-cell precursor ALL (BCP-ALL), which is characterised by the clonal proliferation of blast cells that resemble early B cell progenitors in the BM and/or peripheral blood (PB). Secondly, T-cell ALL (T-ALL) is characterised by T cells that are arrested at specific stages of development, with the underlying genetic abnormality often determining the stage of maturation arrest. Lastly, there is the mature B-cell ALL (B-ALL), which is also known as the Burkitt type ALL.

Immunophenotyping can be used to discriminate B-cells from T-cells, and to identify the different developmental stages of B-cells in a heterogeneous cell sample. A popular technique for immunophenotyping is flow cytometry, where the presence of cell surface antigens such as CD19 and CD20 are detected using fluorophore-conjugated antibodies. By analysing the BM and/or PB samples of ALL patients through flow cytometry, patients can be further classified according to their immunophenotype, which indicates the arrestment of B-cell development at specific differentiation stages.

In addition to classifying ALL into the three groups mentioned above, ALL can be further classified into individual subtypes based on the presence of specific chromosomal and genetic abnormalities found in leukaemic blasts. Chromosomal translocations are a common characteristic of many subtypes. These recurrent translocations in ALL frequently result in corresponding fusion genes that lead to expression of chimeric fusion oncoproteins in ALL. Notable subtypes of BCP-ALL that are characterised by specific translocations include t(9;22) / BCR-ABL1 (also known as Philadelphia chromosome ALL), t(1;19) / E2A-PBX1 (TCF3-PBX1), t(12;21) / TEL-AML1 (ETV-RUNX1) and t(v;11q23) / MLL (KMT2A) rearranged. A less commonly observed ALL subtype characterised by intra-chromosomal amplification of chromosome 21 is known as iAMP21 (Harewood et al., 2003). Another type of chromosomal aberration that occurs frequently in ALL is aneuploidy, where there is an increase or decrease in the number of chromosomes. Additional subtypes of BCP-ALL are characterised by high hyperdiploidy ($> 50$), low hyperdiploidy ($47 - 50$) or hypodiploidy ($< 45$). Patients from the hyperdiploid ($> 50$) subtype have been shown to have distinctly different GEPs from patients belonging to the low hyperdiploid subtype (E.-J. Yeoh et al., 2002). In general, patients with hyperdiploid ($> 50$) have a more favourable prognosis than patients from the low hyperdiploid and hypodiploid subtype. Even though the hyperdiploid ($> 50$) subtype has a distinct GEP, it is inherently heterogeneous as many hyperdiploid ($> 50$) patients have differing modal numbers of chromosome. Heerema et al., 2007 demonstrated a sequential pattern in the chromosomes that gain an extra copy, as modal number increases. There is a total of four different groups of chromosomal gain. Chromosomes 21, X, 14, 6, 18, 4, 17 and 10 are gained at modal numbers 51-54, chromosomes 8, 5, 11, and 12 at modal numbers 57–60, chromosomes 2, 3, 9,16, and 22 at modal numbers 63–67, chromosomes 1, 7, 13, 15, 19, and 20 at modal numbers 68–79. The gain of specific chromosomes has also been demonstrated to be of prognostic value in paediatric ALL (Harris et al., 1992; Heerema et al., 2000; Moorman et al., 2003; Sutcliffe et al., 2005).

Previously, the presence of these specific chromosomal and genetic abnormalities were determined through a combination of cytogenetic analysis, immunophenotyping, and molecular screening for fusion genes. BCP-ALL patients who did not exhibit any chromosomal or genetic abnormalities were all classified under the “B-other” subtype,

and they constituted approximately 25% of paediatric ALL patients (Lilljebjörn et al., 2016). The driver events of ALL for patients under the "B-other" group were unknown.

Characterisation of the long-term survival outcomes of the paediatric ALL subtypes have enabled their stratification into different risk groups. TEL-AML1 and hyperdiploid ($> 50$) are classified as low risk subtypes, while MLL-rearranged, BCR-ABL1, near haploid ($24 - 31$), low hypodiploid ($< 40$) and iAMP21 are classified as high risk subtypes. Both the E2A-PBX1 and "B-other" subtype are commonly classified as intermediate risk (Moorman et al., 2010; Schwab & Harrison, 2018; A. E. J. Yeoh et al., 2012).

However, recent work has identified novel subtypes that exist within the "B-other" subtype, enabled by the advancement in gene expression profiling and next-generation sequencing technology (Den Boer et al., 2009; Gu et al., 2016; Lilljebjörn et al., 2016; Mullighan et al., 2009; Yasuda et al., 2016; Zhang et al., 2016). These novel subtypes are characterised by distinct GEPs and/or specific genetic alterations that were detected through high-throughput genomic analyses. Most of the chromosomal rearrangements that resulted in these genetic alterations were undetectable by conventional cytogenetic analysis, with an example being the DUX4-rearranged subtype (Lilljebjörn et al., 2016; Yasuda et al., 2016; Zhang et al., 2016). Some of the novel subtypes have varying chromosomal aberrations which ultimately result in the alteration of a single gene (e.g. MEF2D-rearranged, ZNF384-rearranged, PAX5-altered) (Gu et al., 2016; Gu et al., 2019). There are also subtypes with distinct GEPs that were the result of different genetic alterations (e.g. BCR-ABL-like, TEL-AML1-like) (Den Boer et al., 2009; Lilljebjörn et al., 2016; Mullighan et al., 2009).

Novel subtypes that make up a significant proportion of the "B-other" subtype are the PAX5 altered, PAX5 P80R, DUX4-rearranged, BCR-ABL1-like, TEL-AML1-like, ZNF384-rearranged and MEF2D-rearranged subtypes (Gu et al., 2019). The long-term survival outcomes of some of the novel subtypes have also been characterised (Gu et al., 2019), allowing for refinements to be made to existing risk stratification schemes. Previously, all patients under the "B-other" subtype were classified as intermediate risk. Novel subtypes like BCR-ABL1-like, PAX5-amplified, ZNF384-rearranged and MEF2D-rearranged subtypes are now classified as high risk, while

DUX4-rearranged is classified as low risk (Schwab & Harrison, 2018; Tran & Loh, 2016). Discovery of the driver genetic events of these subtypes allows for the development of targeted therapeutic strategies.

### 5.2.2 Related work

The earliest studies on paediatric ALL identified clinical and cytogenetic features for risk stratification. Clinical features include age, sex and white blood cell count at diagnosis, while cytogenetic features include the characteristic chromosomal or genetic abnormalities that define classical ALL subtypes. Subsequently, response features were identified to be good predictors of relapse (Miller et al., 1989; Riehm et al., 1987). In particular, MRD was found to be the best predictor of relapse in ALL (Cavé et al., 1998; Coustan-Smith et al., 1998; van Dongen et al., 1998). Many of the current risk stratification protocols classify paediatric ALL patients into three different risk groups (high risk, intermediate risk or standard risk), based on patient subtype, clinical features (viz. age and WBC count) and treatment response features (e.g. MRD) of a patient.

We identify three main flaws in current risk stratification protocols. Firstly, discretisation of continuous features such as age, WBC count and MRD in current protocols results in loss of information, which leads to a loss in statistical power. Secondly, patients belonging to high risk subtypes are generally classified as high risk regardless of their other clinical and response features, and given the corresponding chemotherapy regimen (A. E. J. Yeoh et al., 2012). As a result, even patients who exhibit good response (i.e. have low MRD) are given intensified chemotherapy. Thirdly, accurate assignment of subtype by itself is a labour-intensive and expensive process involving multiple laboratory tests (e.g. immunophenotyping, karyotyping, qPCR) that all require professional expertise. Measuring the WBC count at diagnosis or end-of-induction MRD also involve similar laboratory tests. As a result, measurement of these features may not be available in more remote hospitals in developing countries.

Previous work on treatment outcome prediction using transcriptomic data focused primarily on identifying gene expression signatures that are predictive of long-term treatment outcome (Bhojwani et al., 2008; Kang et al., 2010; Willenbrock et al.,

2004). These works share a general two-step framework: Firstly, univariate statistical tests are used to select probe sets or genes that have high statistical association with relapse-free survival. Secondly, a machine learning model is fitted on training data to predict the risk of relapse. In the case of Bhojwani et al., 2008 and Kang et al., 2010, generalised linear models were used for prediction (logistic regression and Cox proportional hazards model, respectively). Similar works identified gene expression signatures indicative of resistance to specific anti-leukaemia drugs, by incorporating data regarding drug response (e.g. $LC_{50}$) in addition to transcriptomic data (Holleman et al., 2004; Lugthart et al., 2005; Sorich et al., 2008). These works demonstrate that drug-resistant gene expression signatures could be used to predict treatment outcome.

A common trait these methods share is that feature selection is performed by identifying features with high statistical association with the prediction labels. Conducting multiple statistical tests on high-dimensional gene expression data inadvertently leads to a large number of false discoveries; a problem known as the multiple testing problem. As high-dimensional transcriptomic data contains up to tens of thousands of features, there will be a large number of statistically significant features that are simply a result of chance. These erroneous features will not be statistically significant when tested again in other data sets. Bhojwani et al., 2008 demonstrate a particular instance of such a problem, where out of the 24 features identified to be predictive of Day 7 bone marrow status in the training set, only eight features were found to be statistically significant in the test set. Most of the above prediction methods do not show good performance when evaluated on independent validation data sets (Bhojwani et al., 2008; Kang et al., 2010).

A. E.-J. Yeoh et al., 2018 was the first to explore the use of time-series GEPs in predicting treatment outcome. The authors proposed the transcriptomic features absolute response metric (ARM) and effective response metric (ERM), which are computed from a patient's GEPs at Day 0 and Day 8 of chemotherapy. A larger ARM represents a greater shift in a patient's GEP from Day 0 to Day 8, irregardless of direction. On the other hand, a larger ERM represents a greater shift in a patient's GEP from Day 0 to Day 8, in the direction along the leukaemia-to-normal response trajectory. This represents a greater treatment response, resulting in a patient's GEP shifting closer to resemble that of a normal patient. A more formal definition

of the features can be found in section 5.3.2. A drawback in their work is that the features were computed under the assumption that treatment response is the same across all subtypes. Previous work have demonstrated that different subtypes respond differently towards anti-leukaemic drugs (Aricò et al., 2000; Biondi et al., 2000; Raimondi et al., 1990). Also, ALL subtypes have been shown to have distinctly different GEPs (E.-J. Yeoh et al., 2002).

In the last decade, there have been numerous studies identifying a variety of features that are predictive of treatment outcome in ALL. The most common type of features identified are the presence of genetic or copy number alterations in specific genes, such as IKF1 deletions and TP53 deletions (Hof et al., 2011; Krentz et al., 2013; Mullighan et al., 2009). In addition, the continuous advancement of high-throughput technologies have allowed novel subtypes of ALL to be discovered and accurately characterised. In particular, the iAMP21 and BCR-ABL1-like subtypes have been shown to be of prognostic value (Harewood et al., 2003; Harvey et al., 2010), and have been newly included in the 2016 revision of the World Health Organisation (WHO) classification of lymphoid neoplasms (Swerdlow et al., 2016).

The development of single-cell isolation and barcoding technologies has enabled in-depth study of ALL cell populations at unprecedented single-cell resolution. Good et al., 2018 used single-cell mass cytometry to study B-lineage cell populations in BCP-ALL. The authors developed a cell classifier to assign each leukaemic cell based on its similarity to one of twelve developmental stages of B-cell differentiation in healthy B-lineage cells, or three populations of early progenitor, late progenitor or mature non-B cells. This study corroborates the finding by Greaves, 1986 that leukaemic cells are only slightly different from normal B-lineage cell populations. The study also provides evidence that leukaemic cells are arrested at early stages of B-cell differentiation, which results in leukaemic BM samples having a relatively large proportion of cells at early B-cell developmental stages when compared to BM samples from healthy patients. Comparing the proportions of different assigned cell subpopulations across different ALL subtypes has shown that the E2A-PBX1 subtype has smaller proportions of immature B cell populations, and that the CRLF2-rearranged subtype has smaller proportions of pro-BII and pre-BII subpopulations. However, other subtypes (viz. BCR-ABL1, TEL-AML1) did not show a difference in proportion of different assigned cell subpopulations. Most pertinently, Good

et al., 2018 demonstrate the use of an elastic-net-regularised Cox model for relapse prediction. The model selected six features out of a total of 352 candidate features for use in predicting relapse cases. Candidate features include age and WBC count at diagnosis, and for each of the early developmental subpopulations that were expanded in ALL: 1) the percentage of cells, 2) mean protein expression of 24 marker proteins, and 3) the percentage of cells with phosphorylated proteins for 9 different proteins under 5 states (basal state and in response to four different perturbations). Six features were selected from the above features, and are as follows:

1. % of cells with phosphorylated RPS6 (pRPS6) in pro-BII in the basal state
2. % change in cells with p4EBP1 in pro-BII in response to pervanadate ($PVO_4$)
3. % change in cells with pSYK in pre-BI in response to thymic stromal lymphopoietin
4. % change in cells with pCREB in pre-BI in response to $PVO_4$
5. % change in cells with pCREB in pre-BI in response to B-cell receptor cross-link (BCR-XL)
6. % change in cells with pRPS6 in pre-BI responding to BCR-XL

The authors concluded that these signalling proteins were activated in the pre-BCR and PI3K-mTOR pathways in early developmental subpopulations in patients who were at high risk of relapse. Validation of the model on a held-out test set gave good prediction performance that was independent of MRD risk status.

In a complementary study, Witkowski et al., 2020 investigated the bone marrow micro-environment in BCP-ALL, in particular the non-B lineage cell populations. As B-lineage leukaemic blasts dominate in BM samples of BCP-ALL patients, the authors enriched non-B lineage cells by mixing non-B lineage and B-lineage cells of patient samples in a five to one ratio, prior to analysis using single-cell RNA-seq. This allowed a more detailed and accurate study of non-B lineage cells. In the study, leukaemic samples were shown to have an increased composition of early developmental cell subpopulations among B-lineage cells, similar to findings in Good et al., 2018. In addition, the authors demonstrated that the non-classical monocyte subpopulation is expanded at diagnosis, but diminished during remission and re-emerged during relapse. The expansion of the non-classical monocyte subpopulation was accompanied by a decreased classical monocyte population. The authors showed

that monocyte abundance at diagnosis was a predictor of relapse in paediatric BCP-ALL, and suggested that monocytes facilitate the survival of BCP-ALL leukaemic blasts.

## 5.3 Subtype-specific model

We propose the use of a subtype-specific model for prediction of treatment outcome in ALL, based on our postulate that different ALL subtypes have distinct treatment responses. Our subtype-specific model fits individual prediction models for each ALL subtype.

The microarray GEP of a leukaemia patient measures the total gene expression summed from all leukaemia and normal cells present in the patient sample. Chemotherapy kills leukaemia cells at a higher rate than normal cells, which results in a decrease in the ratio of leukaemia to normal cells. Patients who are more responsive to chemotherapy would experience a greater increase in the ratio of normal to leukaemia cells, resulting in a greater change between a patient's Day 0 and Day 8 GEPs. We propose transcriptomic features that measure the change between a patient's Day 0 and Day 8 GEPs in terms of magnitude and direction. The effective response metric (ERM) ratio measures the magnitude of a patient's treatment response along the direction of the leukaemia-to-normal trajectory. A larger ERM ratio implies that the patient has responded more along the correct direction, resulting in the shifting of a patient's GEP closer to that of a normal patient. The ARM ratio is also a measure of the magnitude of a patient's treatment response, but does not take into account the direction of treatment response. A larger ARM ratio indicates that the treatment had a larger impact on the patient, regardless of whether the impact is positive or negative. We take a larger ARM ratio to indicate larger absolute treatment response. The reorientation ratio ($\phi$) is an alternative measure of the change between a patient's Day 0 and Day 8 GEPs. Having a larger $\phi$ indicates a larger treatment response.

These individual transcriptomic features serve as measures of leukaemic cell clearance and in essence measure a patient's treatment response. Based on the reasoning above, we hypothesise that between patients from the same subtype, having a larger ERM ratio, ARM ratio or $\phi$ are all indicative of a higher probability of long-

term remission. We build our subtype-specific model based on the assumption that the above hypothesis is true. Our subtype-specific model estimates the probability of a patient achieving long-term remission, based on the following features of a patient: genetic subtype, ERM ratio, ARM ratio, $\phi$ and $log_{10}$(MRD).

Our subtype-specific model is designed to be trained only using long-term remission patients, in order to mitigate problems commonly associated with clinical data sets. Clinical data sets are frequently small in size and often suffer from an imbalance in the number of positive and negative samples. In ALL data sets, there is a much smaller number of relapse patients than remission patients, with more than 80% of patients achieving long-term remission (Pui & Evans, 2006). A naive subtype-specific model exacerbates this problem by building an individual model for each subtype. This results in an extremely small number of relapse patients in each individual subtype prediction model. This prevents the use of machine learning algorithms, which will overfit on the small training data set. Instead, by training only on long-term remission patients, we reduce the chances of our model overfitting on the training data. We estimate the probability of long-term remission by averaging the percentile of each feature value in long-term remission patients in the training set. Building an individual prediction model for any particular subtype consists of the following steps: probe set selection, computation of transcriptomic features, and estimation of prediction probabilities. We describe each step in further detail in the sections below.

### 5.3.1 Probe set selection

For each subtype, we select probe sets that are differentially expressed between Day 0 and Day 8 of treatment according to the following criteria: Probe sets must 1) exhibit a $P$ value of below 0.05 in a paired $t$-test, 2) have an absolute log fold change above one, and 3) have a minimum average expression value of 6 in either of the classes (Day 0 or Day 8). We term these probe sets as "treatment response probe sets".

We performed probe set selection using only samples belonging to long-term remission patients. Both probe set selection and model training were performed using samples from long-term remission patients only. Hence, all relapse samples can

be treated as independent test samples to evaluate the generalisation performance of our model.

### 5.3.2 Transcriptomic features

We propose three transcriptomic features for use in treatment outcome prediction, namely the ERM ratio, ARM ratio and reorientation ratio ($\phi$). For each subtype, computation of transcriptomic features is performed in a reduced probe set space consisting of treatment response probe sets of the respective subtype.

Figure 5.1 provides a geometric visualisation of the transcriptomic features. The normal centroid $N$ is computed by taking the median of normal patient GEPs along each dimension, while the leukaemia centroid $L$ is computed by taking the median of Day 0 GEPs belonging to long-term remission patients along each dimension. A patient's Day 0 and Day 8 GEPs are denoted $D_0$ and $D_8$, respectively.

Let $\overrightarrow{D_0D_8}$ denote the vector from a patient's Day 0 GEP to a patient's Day 8 GEP, which we call a patient's treatment response vector, and $\overrightarrow{LN}$ denote the vector from the leukaemia centroid to the normal centroid, which we call the subtype treatment progression vector (or leukaemia-normal trajectory). The absolute response metric (ARM) is defined as the magnitude of the patient's treatment response vector, which can be written as:

$$\text{ARM} = \|\overrightarrow{D_0D_8}\|$$

The effective response metric (ERM) is defined as the scalar projection of the patient's treatment response vector onto the subtype treatment progression vector, and is hence:

$$\text{ERM} = \frac{\overrightarrow{D_0D_8} \cdot \overrightarrow{LN}}{||\overrightarrow{LN}||} \tag{5.1}$$

We propose the ARM ratio as measure of the magnitude of change between a patient's Day 0 and Day 8 GEP, irregardless of direction. The ARM ratio can be interpreted as the ARM of a patient expressed in terms of the remaining distance between his/her Day 8 GEP and the normal centroid, and is defined as:

$$\text{ARM ratio} = \frac{\|\overrightarrow{D_0 D_8}\|}{\|\overrightarrow{D_8 N}\|}$$

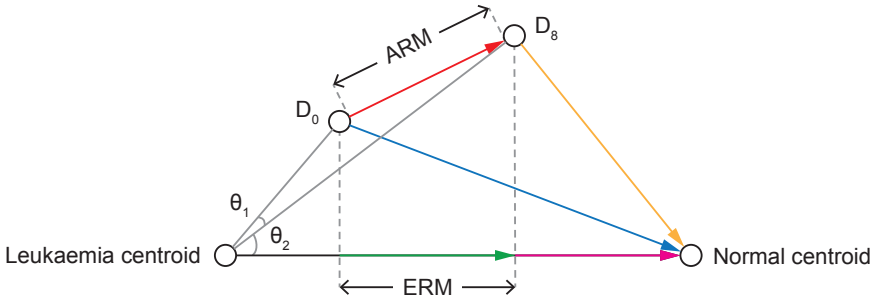



Figure 5.1: Illustration of the effective response metric (ERM) ratio, absolute response metric (ARM) ratio and reorientation ratio. The gene expression profiles (GEPs) of a patient at Day 0 and Day 8 of treatment are represented by the points $D_0$ and $D_8$, respectively. A patient's treatment response vector is represented by the red vector from $D_0$ to $D_8$. The ARM is the magnitude of a patient's response vector (red vector). The ARM ratio is computed by dividing the ARM by the magnitude of the orange vector. The ERM is the scalar projection of the patient's response vector (red vector) onto the vector from the leukaemia centroid to the normal centroid, and is represented by the magnitude of the green vector. The ERM ratio is computed by dividing the ERM by the magnitude of the magenta vector. The reorientation ratio is the ratio of $\theta_1$ to $\theta_2$.

Secondly, we propose the ERM ratio as a measure of the magnitude of change between a patient's Day 0 and Day 8 GEP, along the direction of the leukaemia-normal trajectory. The ERM ratio is the ERM of a patient expressed in terms of the projected distance between his/her Day 8 GEP and the normal centroid, onto the direction of the leukaemia-normal trajectory. It is formalised as:

$$\text{ERM ratio} = \frac{\text{ERM}}{proj_{\overrightarrow{LN}} \overrightarrow{D_8 N}}$$

Lastly, we propose the reorientation ratio as an alternative measure of the patient's treatment response from Day 0 to Day 8, normalised to the remaining response required for a patient to achieve full recovery (i.e. deviation remaining between a patient's Day 8 GEP and the normal centroid). Let $\theta$ be the function that measures the angle between vectors $\overrightarrow{A}$ and $\overrightarrow{B}$, which is defined by:

$$\theta(\overrightarrow{A}, \overrightarrow{B}) = cos^{-1}\left(\frac{\overrightarrow{A} \cdot \overrightarrow{B}}{\|\overrightarrow{A}\| \, \|\overrightarrow{B}\|}\right)$$

The reorientation ratio $\phi$ is the ratio of the angle between $\overrightarrow{LD_0}$ and $\overrightarrow{LD_8}$ to the angle between $\overrightarrow{LD_8}$ and $\overrightarrow{LN}$:

$$\phi = \frac{\theta(\overrightarrow{LD_0}, \overrightarrow{LD_8})}{\theta(\overrightarrow{LD_8}, \overrightarrow{LN})}$$

### 5.3.3 Estimation of probability of long-term remission

In our subtype-specific model, we assume that the monotonic relationships between each of the patient's features (viz. ERM ratio, ARM ratio, $\phi$ and log-transformed MRD) and the probability of long-term remission only exist between patients from the same subtype. For a patient from a specific subtype $s$, we estimate his/her probability of achieving long-term remission by comparing his/her feature values $\mathbf{x}$ with patients from the same subtype that achieve long-term remission. We estimate the probability of a patient achieving long-term remission by averaging the percentile each feature value achieves among long-term remission patients from the same subtype. The feature vector $\mathbf{x}$ consists of the ERM ratio, ARM ratio, $\phi$ and $log_{10}$(MRD).

A higher ERM ratio corresponds to a larger magnitude in the shift between a patient's Day 0 and Day 8 GEP along the leukaemia-normal trajectory. Similarly, a higher ARM ratio indicates a larger magnitude in the shift between a patient's Day 0 and Day 8 GEP, regardless of direction. A higher $\phi$ indicates a larger difference between a patient's Day 0 and Day 8 GEPs. A lower $log_{10}$(MRD) shows that there is a lower proportion of leukaemic cells remaining after treatment. Thus, a higher ERM ratio, higher ARM ratio, higher $\phi$ and lower $log_{10}$(MRD) are all indicators of a more favourable treatment response.

We hypothesise that a higher ERM ratio, higher ARM ratio, higher $\phi$ and lower $log_{10}$(MRD) are indicative of a higher probability of long-term remission. Our subtype-specific model is built on the assumption that the above hypothesis is true, and we validate that our hypothesis holds true in homogeneous ALL subtypes in section 5.5.2.

Within each subtype $s$, we estimate the probability of long-term remission given the observed value $x_i$ for each feature $i$ as the proportion of remission samples from

the subtype that has a less favourable or equal observed value than the current value $x_i$:

$$P(Remission|x_i, s) = \frac{n_s}{N_s}$$

where $N_s$ is the total number of remission patients belonging to the subtype $s$ and $n_s$ is the number of patients of that subtype with a less favourable or equal value than $x_i$. This is calculated for each feature independently, and subsequently averaged to give the probability of long-term remission given the feature vector $\mathbf{x_{D33}} = (x_{\text{ERM Ratio}},\ x_{\text{ARM Ratio}},\ x_\phi,\ x_{log(\text{MRD})})$:

$$P(Remission|\mathbf{x}, s) = \frac{1}{4}\sum_i P(Remission|x_i, s)$$

## 5.4 Methods

### 5.4.1 Ma-Spore ALL data set

The Ma-Spore ALL data set consists of children with *de novo* paediatric ALL treated in the Ma-Spore ALL 2003 (A. E. J. Yeoh et al., 2012), Ma-Spore ALL 2010 (A. E. J. Yeoh et al., 2018) and ALL-IC-BFM-2002 (Stary et al., 2014) studies, which ran from July 2002 to December 2014. Informed consent was obtained in accordance with the Declaration of Helsinki and approved by the respective review boards. Diagnosis of ALL was confirmed via bone marrow morphology and standard immunophenotyping. Subtype assignment was performed through the following procedures: Molecular screening for TEL-AML1, BCR-ABL, E2A-PBX1 and MLL fusion genes were carried out using quantitative polymerase chain reaction (qPCR). Karyotyping or flow cytometry (DNA index $\geq 1.16$) was used to determine hyperdiploidy ($> 50$). MRD of patients at Day 33 was measured via qPCR analysis of IGH/T-cell receptor gene rearrangements.

For gene expression profiling, bone marrow aspirates were taken from patients at Day 0 and Day 8 of treatment, from which mononuclear cells were isolated using Ficoll-Paque (GE Healthcare Life Sciences, Piscataway, NJ, USA) density gradient centrifugation. Total RNA was extracted using TRIzol© reagent (Invitrogen,

Carlsbad, CA, USA) and subsequently purified using RNeasy Mini Kit (Qiagen, Valencia, CA, USA). Affymetrix HG-U133A or HG-U133 Plus 2.0 microarrays (Affymetrix, Santa Clara, CA, USA) were used to measure Day 0 and Day 8 GEPs of all patients.

All patients were treated using the ALL Berlin-Frankfurt-Münster backbone. Patients in the Ma-Spore ALL 2003 and ALL-IC 2002 study received intrathecal methotrexate followed by 7 days of prednisolone. Patients in Ma-Spore ALL 2010 received intravenous vincristine instead of intrathecal methotrexate at Day 0. A 3-drug dexamethasone-based remission-induction therapy was performed for standard and intermediate risk patients, while high risk patients received additional daunorubicin in the Ma-Spore studies. Patients were defined as high risk if they satisfied any of the following criteria: Week 12 MRD $\geq 1 \times 10^{-3}$, belonging to high risk subtypes (viz. BCR-ABL, MLL and hypodiploid), infant CD10-negative ALL, poor prednisolone response or induction failure (A. E. J. Yeoh et al., 2012).

### 5.4.2 Data preprocessing

Raw microarray *CEL* data were processed using the *MAS5* algorithm (Affymetrix, 2002) from the R package *affy* (Gautier et al., 2004). Probe sets that had less than 10% "Present" calls in the entire dataset were filtered out, and expression values with "Marginal" or "Absent" calls were re-assigned a value of zero. As two different Affymetrix platforms were used in the Ma-Spore ALL data set, we only retain probe sets that are present in both platforms for that data set. Raw probe set intensities of each chip were normalised using trimmed-mean scaling as part of the *MAS5* protocol. We removed ambiguous probe sets (i.e. probe sets that map to multiple transcripts) from the data and only retained probe sets that were expressed in more than 70% of the patients at any time point. The probe set intensities were subsequently $log_2$-transformed to give their final expression values.

In the Ma-Spore ALL data set, outlier samples were identified through a principal component analysis (PCA) plot and removed along with its corresponding pair (belonging to the same patient). Day 0 and Day 8 sample pairs were removed if they came from different batches. These pairs were removed to avoid possible introduction of batch effects during computation of transcriptomic features from

| Subtype | Remission | Relapse |
|---|---|---|
| BCR-ABL | 4 | 5 |
| E2A-PBX1 | 3 | 0 |
| MLL | 2 | 2 |
| T-ALL | 7 | 1 |
| TEL-AML1 | 20 | 4 |

(a)

| Subtype | Remission | Relapse |
|---|---|---|
| BCR-ABL | 1 | 0 |
| E2A-PBX1 | 2 | 1 |
| MLL | 1 | 2 |
| T-ALL | 1 | 1 |
| TEL-AML1 | 9 | 1 |

(b)

Table 5.1: Distribution of patients across subtypes and treatment outcomes in the Ma-Spore ALL data set. (a) Training set consisting of patients from batches 1-7. (b) Test set consisting of patients from batches 8-10.

Day 0 and Day 8 samples. A total of 19 samples were removed from the original data set, resulting in 405 remaining samples.

### 5.4.3 Ma-Spore ALL data set: Train-test split

Three different versions of the Ma-Spore ALL data set were used for evaluation. All three versions of the Ma-Spore ALL data set do not contain patients with missing MRD values. Version 1 of the data set contains all subtypes aside from the hypodiploid subtype, which consists of only one patient. It was not split into training and test sets. Versions 2 and 3 contain all subtypes aside from the heterogeneous subtypes, viz. hypodiploid, hyperdiploid ($> 50$) and "B-other". Versions 2 and 3 of the data set contain the same set of patients, with the only difference being that version 3 was split into training and test sets while version 2 was not. Nonetheless, relapse samples in version 2 are not used in training the subtype-specific models; thus, they can serve as an independent test set for version 2 of the data set.

We evaluated the outcome prediction performance of the subtype-specific model against other methods on both versions 2 and 3 of the data set, as our subtype-specific model is only applicable to homogeneous subtypes. Training and test sets were split in version 3 of the data set according to the dates on which the batches were processed. Earlier batches (batches 1-7) were taken as the training set, while later batches (batches 8-10) were taken as the test set. Table 5.1 shows the distribution of patients across different subtypes and outcomes in the training and test set of data set version 3.

### 5.4.4 Receiver operating characteristic analysis

We standardise the partial area under the curve (pAUC) of receiver operating characteristic (ROC) curves using the method of McClish, 1989. Standardisation of pAUC is performed for easy interpretation, by providing a baseline for comparison; a random classifier will have a standardised pAUC of 0.5. Standardisation of pAUC is performed as follows:

$$A^* = 0.5\left(1 + \frac{A - min}{max - min}\right)$$

where $A$ is the unstandardised pAUC, $max$ is the maximum pAUC under the specified true positive rate (TPR) and/or false positive rate (FPR) limits, and $min$ is the pAUC of a random classifier, represented by the identity line. The standardised pAUC ($A^*$) is undefined when the unstandardised pAUC of the ROC curve is lesser than that of a random classifier (i.e. $A < min$).

### 5.4.5 Other methods

In this section, we briefly introduce other treatment outcome prediction methods that we evaluated against, and provide details on how each method was performed for ease of reproducibility.

**Minimal residual disease.** MRD is widely used in contemporary ALL treatment protocols for risk stratification, and has been shown to be the most powerful prognostic factor in ALL (Cavé et al., 1998; Coustan-Smith et al., 1998; van Dongen et al., 1998). It refers to the presence of small numbers of residual leukaemic cells after treatment. In the Ma-Spore ALL data set, MRD was measured at Day 33 using qPCR analysis of IGH/T-cell receptor gene rearrangements. The sensitivity of qPCR makes it possible to detect one leukaemia cell in 10000 normal cells ($1 \times 10^{-5}$).

**Global ERM (A. E.-J. Yeoh et al., 2018).** The authors propose the transcriptomic feature ERM, which is computed as part of a global model consisting of patients from every ALL subtype. This is different from our subtype-specific model, where transcriptomic features of a patient are computed in a patient's respective individual subtype model. We term ERM computed in a global model as "global

ERM" for the sake of clarity. Computing global ERM consists of two steps, both involving patients from every subtype. First, probe set selection is performed using a Wilcoxon signed rank test, where the top 1000 up- or down-regulated probe sets with the lowest $P$ values are chosen. Second, calculation of global ERM is performed according to the formula provided in Equation 5.1.

**Bhojwani et al., 2008.** The authors identified a 47 probe set signature that is predictive of long-term treatment outcome in ALL. Logistic regression (LR) with backward selection, LR with forward selection and LR with stepwise selection were used to build three different prediction models. Each of these models selected a different subset of variables from the 47 probe sets and clinical covariates (e.g. sex, age and WBC count). We use the best model reported by Bhojwani et al., 2008 for our evaluation. LR with backward selection was the best model, with the three probe sets "201472_at", "208687_x_at" and "212576_at" as its input features.

**Kang et al., 2010.** The authors present a 42 probe set classifier that is predictive of treatment outcome. Only 26 out of 42 probe sets are present in the Ma-Spore ALL data set. In our evaluation, we build the classifier using these 26 probe sets. Firstly, principal components analysis was performed on standardised values of the probe sets. Secondly, only the first principal component scores were used to fit a Cox proportional hazard regression model. The linear predictor from the Cox regression model, which is the product of the first principal component score and its coefficient, is used as a risk score to predict the risk of relapse.

### 5.4.6 Statistical analysis

Possible treatment outcomes in ALL patients are continuous complete remission (CCR), resistant disease, relapse after complete remission (any site) and death (regardless of cause). Event-free survival (EFS) was defined as time from diagnosis to relapse, resistant disease or death. EFS was estimated using the Kaplan-Meier estimator, and survival curves are compared using the log-rank test. For treatment outcome prediction, patients are labelled "relapse" if they achieve any outcome aside from CCR, and "long-term remission" if they are in CCR. The two-sided Wilcoxon-rank sum test was used to assess the statistical significance of differences

in feature values or prediction probabilities between relapse and long-term remission patients. All statistical analysis was performed using R 3.5.1 (R Core Team, 2018).

## 5.5 Results

### 5.5.1 Probe set selection mitigates batch effects

Microarray samples in the Ma-Spore ALL data set were processed on different dates ranging from 2002 to 2015. The samples were assigned to nine different batches, with each batch consisting of samples processed around the same period in time. Batch 1 consists of samples measured using the Affymetrix GeneChip Human Genome U133 Array, while samples in the rest of the batches were measured using the Affymetrix Human Genome U133 Plus 2.0 Array.

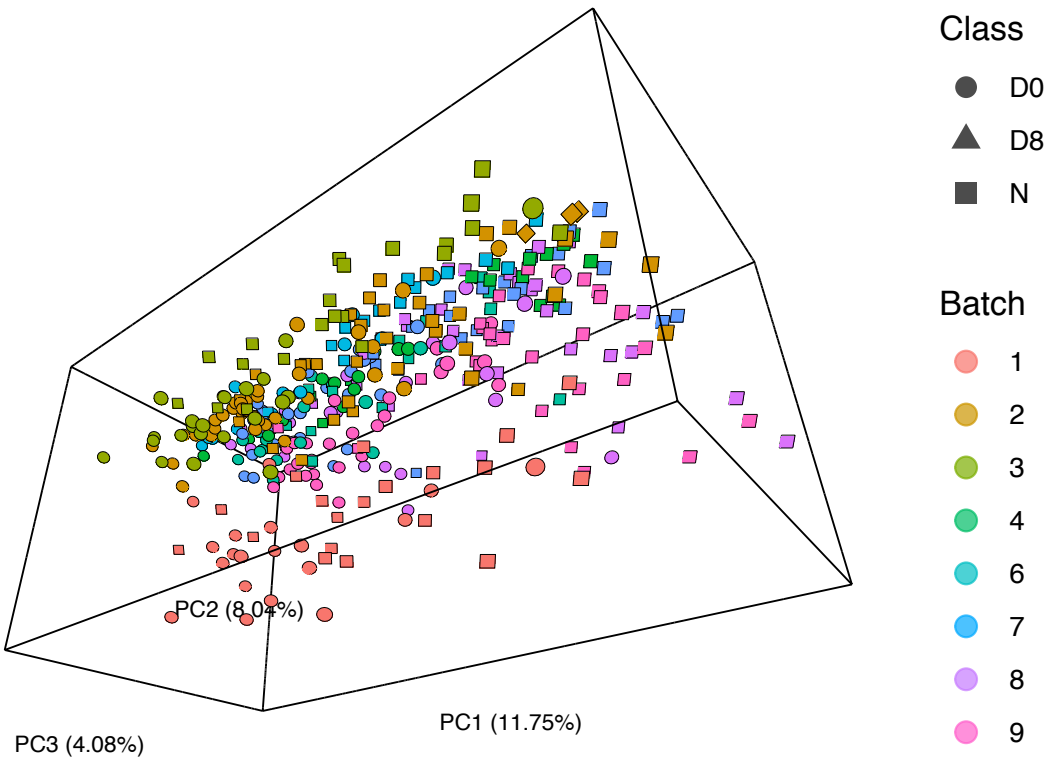
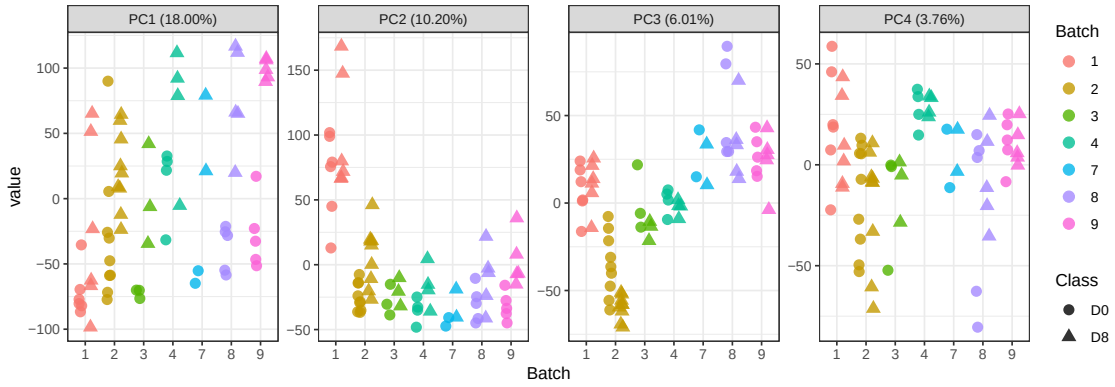


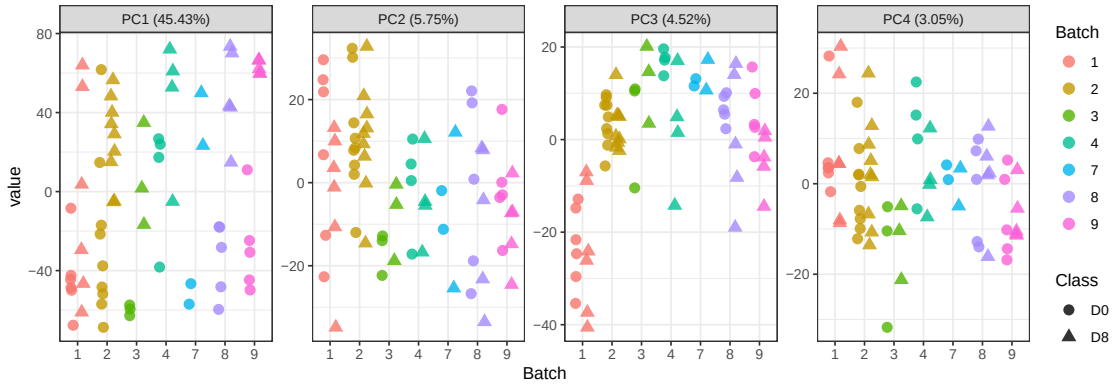

Figure 5.2: Three-dimensional PCA plot of the Ma-Spore ALL data set.

Figure 5.2 reveals significant batch effects in the Ma-Spore ALL data set, especially between batch 1 and the rest of the batches. To mitigate batch effects,

probe set selection was performed before the computation of transcriptomic features. We evaluate whether probe set selection is sufficient to prevent batch effects from affecting the prediction performance of our subtype-specific model in the following three ways.




(a) Before probe set selection




(b) After probe set selection

Figure 5.3: Top four principal components of TEL-AML1 samples from the Ma-Spore ALL data set (a) before and (b) after probe set selection.

Firstly, we assess qualitatively whether batch effects are mitigated in the reduced space consisting only of selected treatment response probe sets. We compare PCA visualisations of the top four PCs of samples in each subtype before and after probe set selection. We observe that probe set selection reduces most of the batch effects present in the original space. This is particularly evident in subtypes with more samples, such as TEL-AML1 (see Figures 5.3a and 5.3b). Before probe set selection, significant batch effects were observed in the second to fourth PCs, which account for a total of 20% of the total variance in the data (see Figure 5.3a). In

comparison, the first PC (which mainly captured class variation between Day 0 and Day 8 samples) only accounts for 18% of total variance in the data. After probe set selection, Figure 5.3b reports that PC1 (representing class variation) accounts for an increased proportion of total variance (45%), demonstrating that the ratio of class variation to batch variation has increased, and PC3 which still has observable batch effects now accounts for less than 5% of total variance. Furthermore, we can no longer observe significant batch effects in PC2 and PC4.

Secondly, we evaluate quantitatively whether batch effects are reduced by probe set selection, using the batch effects metric RVP proposed in Chapter 3. Table 5.2 shows that the proportion of variance attributable to batch effects is reduced after probe set selection in every subtype.

| Subtype | $\text{RVP}_{before}$ (%) | $\text{RVP}_{after}$ (%) | # probe sets | # samples | $p$-value |
|---|---|---|---|---|---|
| BCR-ABL | 75.9 | 53.9 | 21 | 20 | 0 |
| E2A-PBX1 | 82.3 | 48.9 | 239 | 12 | 0 |
| Hyperdiploid | 35.6 | 28.3 | 281 | 70 | 0 |
| MLL | 63.1 | 41.0 | 262 | 14 | 0 |
| B-other | 21.7 | 13.7 | 448 | 166 | 0 |
| T-ALL | 52.3 | 43.8 | 136 | 20 | 0 |
| TEL-AML1 | 31.2 | 23.2 | 505 | 68 | 0.001 |

Table 5.2: RVP of data before and after probe set selection across all subtypes. RVP estimates the percentage of total variance in the data that is attributable to batch effects. The number of treatment response probe sets selected for each subtype and number of samples in each subtype is indicated as well. $P$-value indicates the probability of observing a value lesser than $\text{RVP}_{after}$ by chance (i.e. by randomly selecting probe sets). It was approximated using 1000 Monte Carlo simulations.

We perform a Monte Carlo test to determine the statistical significance of the observed RVP value of data after probe set selection ($\text{RVP}_{after}$) in every subtype. This is done to assess whether a decrease in RVP value is due solely to a reduction in data dimensions. We define the $p$-value as the probability of observing a value lesser than $\text{RVP}_{after}$ after random probe set selection. To approximate the $p$-value, a total of $n = 1000$ Monte Carlo simulations were performed. In each simulation, probe sets are randomly selected before computing the RVP. The number of probe sets randomly selected is equal to the number of treatment response probe sets selected originally in each subtype. The $p$-value is calculated by

$$p = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\text{RVP}_i < \text{RVP}_{after}\}$$

where $\text{RVP}_i$ is the RVP value of the $i$th simulation and $\mathbf{1}\{\text{RVP}_i < \text{RVP}_{after}\}$ is the indicator variable that is equals to one if $\text{RVP}_i$ is lesser than $\text{RVP}_{after}$, and zero otherwise.

Table 5.2 reveals that the observed $\text{RVP}_{after}$ values for all subtypes are statistically significant ($P \leq 0.001$). This indicates that our method of probe set selection has effectively reduced the proportion of batch effects in the data, by selecting probe sets with high class variation.

Thirdly, we evaluate whether batch effects are present in the three transcriptomic features, ERM ratio, ARM ratio and $\phi$. Transcriptomic features of patients are computed using selected treatment response probe sets from their respective subtype. We perform PCA visualisations of the transcriptomic feature space in order to assess whether batch effects are present. We observe slight amounts of batch effects among the three PCs of the transcriptomic feature space of TEL-AML1 patients (see Figure 5.4).

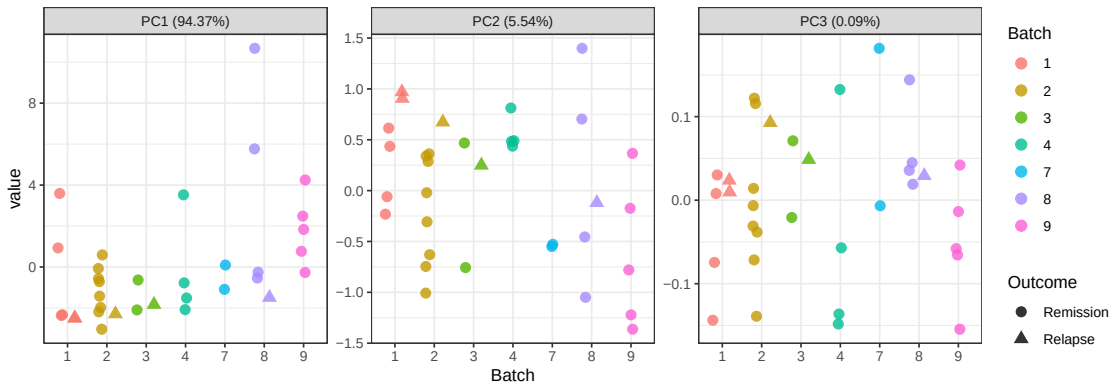



Figure 5.4: PCA plot of transcriptomic features (viz. ERM ratio, ARM ratio and $\phi$) of TEL-AML1 patients from the Ma-Spore ALL data set.

### 5.5.2 Transcriptomic features are predictive of treatment outcome in homogeneous subtypes

Figure 5.5 illustrates the relationship between each transcriptomic feature (viz. ERM ratio, ARM ratio and $\phi$) and treatment outcome in all subtype-specific models. Individual subtype-specific models are fitted with patients from the subtype in the

Ma-Spore ALL data set. We do not visualise transcriptomic features on a hold-out test set as performing a train-test split results in an inadequate number of samples in the training set for the building of accurate subtype-specific prediction models.

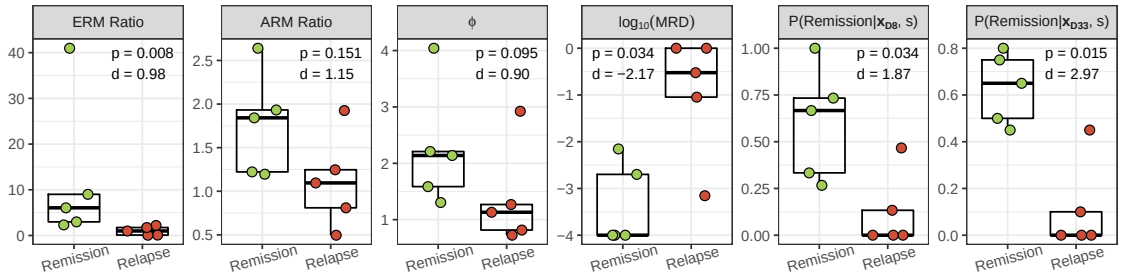
In general, we observe that long-term remission patients have higher ERM ratios, higher ARM ratios and higher $\phi$ than relapse patients, within homogeneous ALL subtypes (viz. BCR-ABL, E2A-PBX1, MLL, T-ALL and TEL-AML1). However, differences between transcriptomic features of remission and relapse patients are less distinct in the hyperdiploid ($> 50$) subtype, and are absent in the "B-other" subtypes (see Figures 5.5f and 5.5g). Both hyperdiploid ($> 50$) and "B-other" are heterogeneous subtypes.

This observation corroborates with our hypothesis that a higher ERM ratio indicates a larger effective treatment response, a higher ARM ratio indicates a larger absolute treatment response, and a higher $\phi$ represents a larger difference between patient's Day 0 and Day 8 GEPs, all of which imply a higher probability of achieving long-term remission. A likely reason for the diminished or absent differences between remission and relapse patients in heterogeneous subtypes is that patients within heterogeneous subtypes have significantly different genetic profiles and respond differently to treatment, and thus cannot be analysed collectively as a whole.

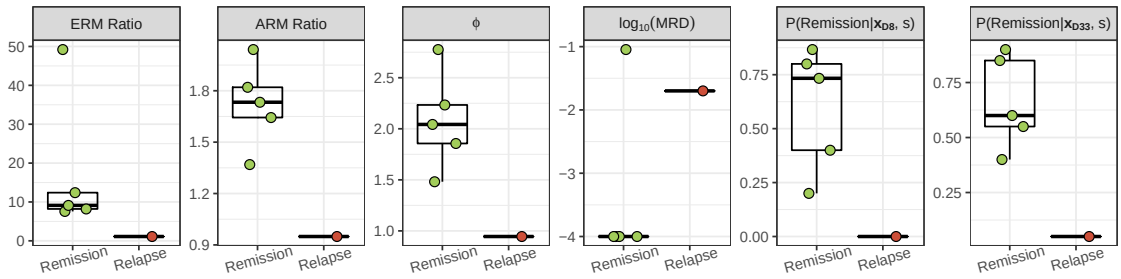
### 5.5.3 Prediction probabilities from the subtype-specific model are significantly associated with treatment outcome in homogeneous subtypes

We formally define the prediction probability estimated using our subtype-specific model as the conditional probability of a patient achieving long-term remission given his/her feature vector $\mathbf{x}$ and subtype $s$, which can be written as $P(Remission|\mathbf{x}, s)$.

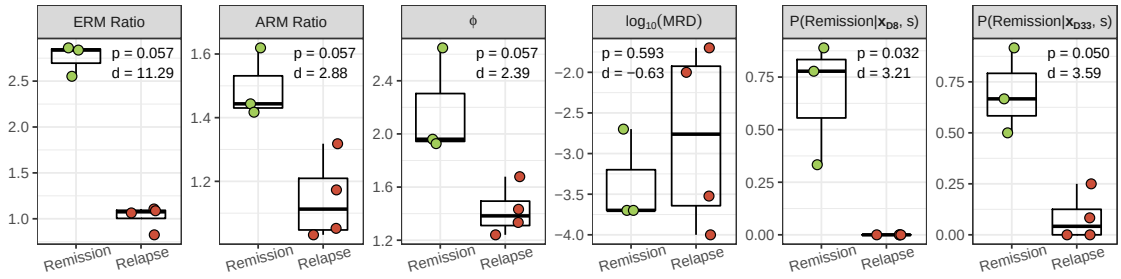
Through our subtype-specific model, we can estimate the conditional probability of achieving long-term remission using different sets of patient features. Figure 5.5 presents two sets of prediction probabilities that are estimated using Day 8 and Day 33 feature vectors. We use $\mathbf{x_{D8}}$ and $\mathbf{x_{D33}}$ to denote a patient's feature vector at Day 8 and Day 33, respectively. $\mathbf{x_{D8}}$ is made up of three features: ERM ratio, ARM ratio and $\phi$, while $\mathbf{x_{D33}}$ consists of four features, ERM ratio, ARM ratio, $\phi$ and log-transformed MRD. We log-transform MRD values as MRD has been shown to be log-normally distributed within each subtype in ALL (O'Connor et al., 2018).



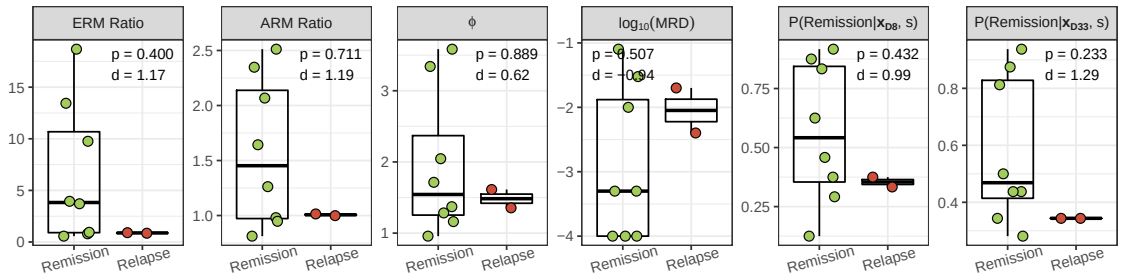

(a) BCR-ABL



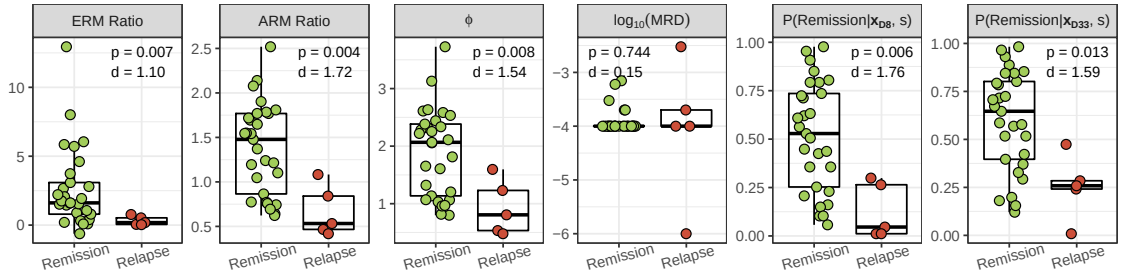

(b) E2A-PBX1



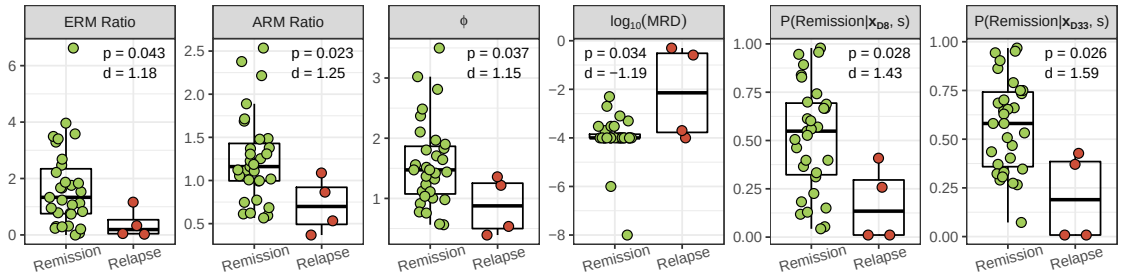

(c) MLL



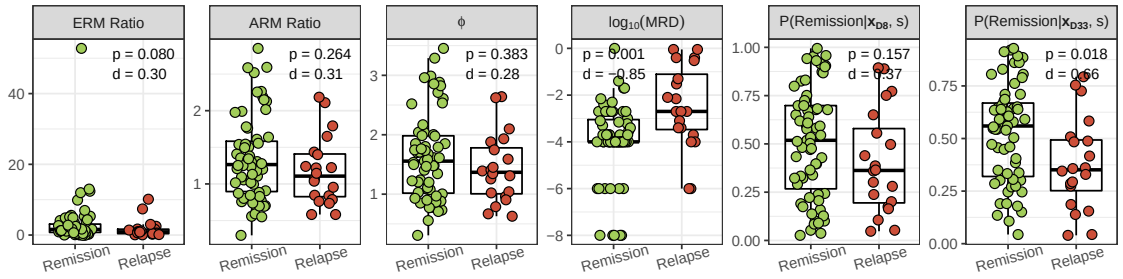

(d) T-ALL




(e) TEL-AML1



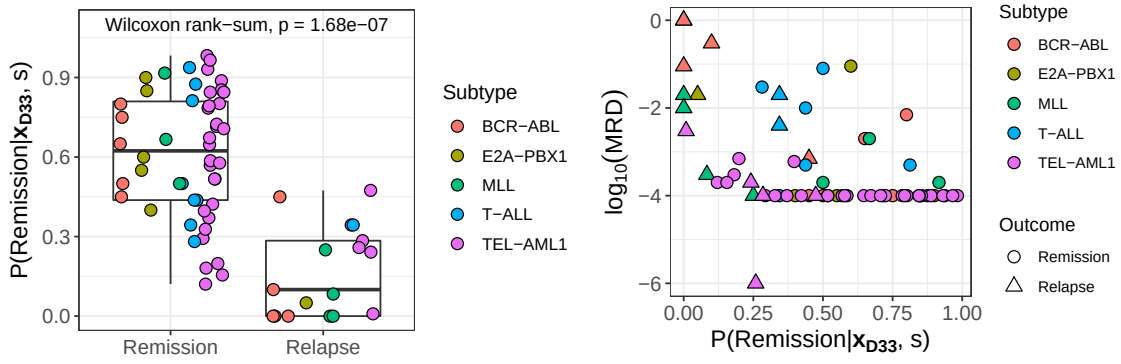

(f) Hyperdiploid ($> 50$)




(g) "B-other"

Figure 5.5: Distribution of features among long-term remission and relapse patients for each individual subtype model in the Ma-Spore ALL data set. Long-term remission patients are shown as green points while relapse patients are shown as red points. Prediction probabilities estimated with our subtype-specific model are conditioned on both a patient's subtype $s$ and his/her Day 8 feature vector $\mathbf{x_{D8}}$ or Day 33 feature vector $\mathbf{x_{D33}}$. Prediction probabilities are denoted as $P(Remission|\mathbf{x_{D8}}, s)$ or $P(Remission|\mathbf{x_{D33}}, s)$ based on the feature vector used to make the prediction. $\mathbf{x_{D8}}$ consists of three features, namely ERM ratio, ARM ratio, and $\phi$, while $\mathbf{x_{D33}}$ consists of four features, with $log_{10}$(MRD) as an additional feature. The Wilcoxon rank-sum test was performed to test for association between each individual feature and treatment outcome; $P$ values are reported for cases where the Wilcoxon rank-sum test can be performed. Effect sizes calculated using Cohen's $d$ formula (assuming unequal variance) are also reported.

We observe in Figure 5.5 that higher prediction probabilities are generally associated with long-term remission patients in homogeneous subtypes, viz. BCR-ABL, E2A-PBX1, MLL, T-ALL and TEL-AML1. There is a large effect size between remission and relapse patients for prediction probabilities estimated using either Day 8 or Day 33 feature vectors in homogeneous subtypes, with Cohen's $d > 1.6$ in all measurable cases. Effect sizes between remission and relapse patients are relatively smaller in the hyperdiploid (> 50) subtype, and smallest in the "B-other" subtype (see Figures 5.5f and 5.5g). J. Cohen, 2013 suggests the classification of Cohen's d effect sizes into small ($d = 0.2$), medium ($d = 0.5$) and large ($d \geq 0.8$). Although the effect sizes are smaller in both the hyperdiploid (> 50) and "B-other" subtypes, they still fall under the large and medium categories. Besides the large effect sizes, differences in prediction probabilities between long-term remission and relapse patients reach statistical significance in most homogeneous subtypes ($P < 0.05$, Wilcoxon rank-sum test). Absence of statistical significance in some cases may be due to small sample sizes, which limits the power of the statistical test or precludes the use of statistical analysis. In Figure 5.6a, we can see that the consolidated prediction probabilities from all homogeneous subtypes is significantly associated with treatment outcome ($P < 0.001$, Wilcoxon rank-sum test), with higher prediction probabilities associated with long-term remission patients.




(a) (b)

Figure 5.6: (a) Distribution of prediction probabilities estimated by the subtype-specific model in long-term remission and relapse patients from homogeneous subtypes in the Ma-Spore ALL data set. (b) Scatter plot of log-transformed minimal residual disease (MRD) against prediction probabilities estimated by the subtype-specific model in patients belonging to homogeneous subtypes in the Ma-Spore ALL data set.

The subtype-specific prediction model operates on the implicit assumption that long-term remission patients generally have a higher ERM ratio, higher ARM ratio and higher $\phi$ than relapse patients. We demonstrated in section 5.5.2 that ERM ratio, ARM ratio and $\phi$ are not significantly different between remission and relapse patients of heterogeneous subtypes. As the key assumption in our prediction model does not hold in heterogeneous subtypes, we advise against the use of our subtype-specific model to predict treatment outcome in heterogeneous subtypes.

Our subtype-specific model only utilises long-term remission patients in the training phase. Both probe set selection and probability estimation are performed without the use of relapse patients. Hence, all relapse patients from the data set essentially form a hold out test set. Despite not being trained on any relapse samples, our prediction model is able to correctly estimate that probabilities of long-term remission for relapse patients (which can be regarded as independent test set samples) are lower than long-term remission patients. In comparison, other prognostic factors such as $log_{10}$(MRD) and global ERM are not consistently different between relapse and long-term remission patients in all homogeneous subtypes. We report $P$ values from the Wilcoxon rank-sum test in Figure 5.5; $P$ values are omitted when it is not possible to perform a Wilcoxon rank-sum test.

### 5.5.4 Subtype-specific model outperforms other methods in treatment outcome prediction

We compare the outcome prediction performance of the subtype-specific model against MRD, global ERM (A. E.-J. Yeoh et al., 2018), and diagnostic gene expression classifiers proposed by Bhojwani et al., 2008 and Kang et al., 2010 through a receiver operating characteristic (ROC) analysis. Both the subtype-specific model and global ERM utilise Day 0 and Day 8 GEPs, while diagnostic gene expression classifiers are fitted on Day 0 GEPs. Although MRD is a continuous-valued measure of residual blasts, it is typically dichotomised or trichotomised into different MRD risk groups for use in clinical risk stratification. In our ROC analysis, we choose to use continuous MRD values. For the subtype-specific model, prediction probabilities from each individual subtype prediction model are consolidated. We evaluate the prediction performance of all methods on versions 2 and 3 of the Ma-Spore ALL

data set (see section 5.4.3 for details on the different data set versions). These two data set versions both contain only homogeneous subtypes, and are the same aside from the fact that version 3 was split into training and test sets while version 2 was not.

Figure 5.7a presents the prediction performance of all methods on version 2 of the data set. As version 2 of the data set is not split into training and test sets, training and evaluation was performed on the same set of patients. Nonetheless, as described in section 5.3, our subtype-specific model is trained only using long-term remission patients. Relapse patients were not used in both feature selection and probability estimation steps. We observe that the subtype-specific model is better at discriminating between long-term remission and relapse patients than all other methods, having achieved the highest standardised pAUC of 0.858 (a random classifier would show a standardised pAUC of 0.5.). Standardised pAUCs on the false positive rate (FPR) range from 0 to 0.2 are reported, as we are interested only in classification thresholds that give small FPRs.

We believe that our subtype-specific model does not overfit to the training data, as it does not utilise relapse patients (positive samples) during training. In contrast, methods by Bhojwani et al., 2008 and Kang et al., 2010 involve machine learning algorithms (namely logistic regression and Cox proportional hazards regression), which are likely to overfit on training data. These methods often present falsely inflated performance when evaluated on the same data used for training. The subtype-specific model is able to outperform these methods that are prone to overfitting even when evaluated on the same data used for training.

We evaluate the generalisation ability of all prediction models on version 3 of the Ma-Spore ALL data set. Version 3 of the data set is split into training and test sets, with batches processed at earlier dates taken as the training set and batches processed at later dates taken as the test set. The split was performed in this manner to reflect the typical process of training and deploying a real-life prediction model, where prior observed samples are used to train models that are subsequently deployed on new batches of samples. Figure 5.7c illustrates that the subtype-specific model, global ERM and MRD achieve similar prediction performance, all outperforming the diagnostic GEP classifiers (Bhojwani et al., 2008; Kang et al., 2010) on the test set. Both diagnostic GEP classifiers are built on genetic signatures identified

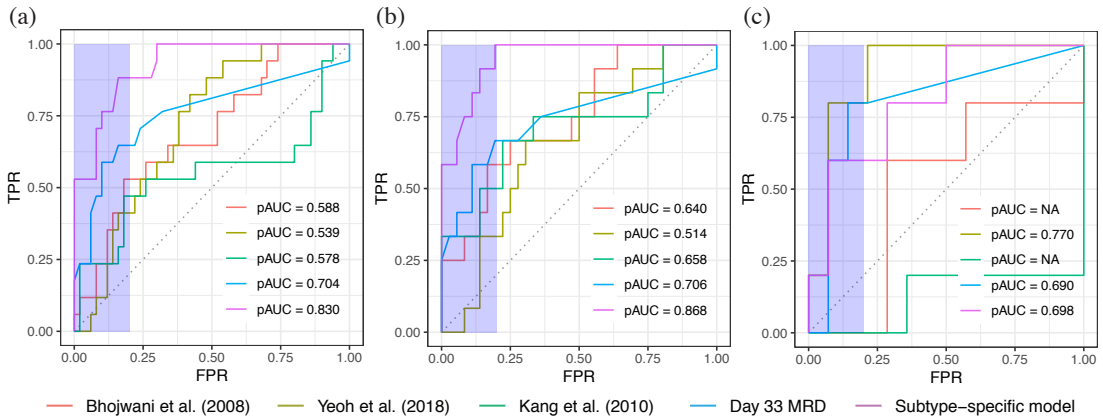



Figure 5.7: Receiver operating characteristic analysis comparing treatment outcome prediction performance. Only patients belonging to homogeneous subtypes in the Ma-Spore ALL data set are included in the analysis. Relapse patients are taken as positive samples. Standardised partial area under the curve (pAUC) values are reported for the false positive range of 0 to 0.2 (shaded in blue). pAUCs are standardised using the method of McClish, 1989. The identity line represents the ROC curve of a random classifier, which has a standardised pAUC value of 0.5. Global ERM is labelled as A. E.-J. Yeoh et al., 2018 (a) Version 2 of the Ma-Spore ALL data set with no train-test split (n = 67). (b) Training set of version 3 of the Ma-Spore ALL data set (n = 48). (c) Test set of version 3 of the Ma-Spore ALL data set (n = 19). Standardised pAUC is reported as NA for curves with a lower pAUC than that of a random classifier.

by their respective authors to be predictive of treatment outcome. Figure 5.7b shows the performance of the prediction models on the corresponding training set. Bhojwani et al., 2008 and Kang et al., 2010 show markedly improved prediction performance compared to their test set performance, demonstrating that diagnostic GEP classifiers overfit to the training data and fail to generalise to the test set.

We believe that our subtype-specific model outperforms other methods on version 2 but not version 3 of the data set because there are inadequate samples in the training set of data set version 3. The lack of training samples increases the chance of errors during probe set selection and probability estimation, which would result in the construction of inaccurate prediction models. The subtype-specific prediction model is most susceptible to a lack in samples, as it separates samples according to subtype before constructing a model for each subtype. In comparison, other subtype-agnostic models are trained on a larger number of patients from all subtypes. Hence, evaluating on version 3 of the data set puts the subtype-specific prediction

model at a disadvantage to subtype-agnostic methods.

### 5.5.5 Subtype-specific model identifies low risk patients with high precision

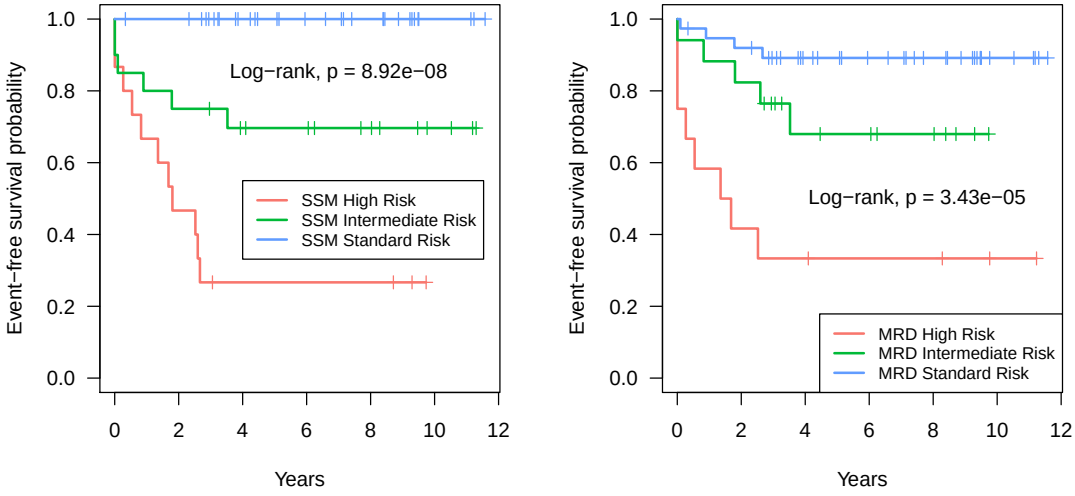
We stratify patients into different risk groups based on their subtype-specific model prediction probabilities at Day 33, $P(Remission|\mathbf{x_{D33}}, s)$. The patient feature vector at Day 33, $\mathbf{x_{D33}}$, consists of the ERM ratio, ARM ratio, $\phi$ and Day 33 MRD. Patients with $P(Remission|\mathbf{x_{D33}}, s) > 0.5$ are classified as standard risk, patients with $0.25 < P(Remission|\mathbf{x_{D33}}, s) \leq 0.5$ are classified as intermediate risk while patients with $P(Remission|\mathbf{x_{D33}}, s) \leq 0.25$ are classified as high risk.

Figure 5.6b demonstrates that the subtype-specific model is able to identify patients who have a low risk of relapse with high precision. All patients with $P(Remission|\mathbf{x_{D33}}, s) > 0.5$ were long-term remission patients. In contrast, out of the 38 patients classified as MRD standard risk ($\leq 1 \times 10^{-4}$), four of them relapse.

In Figure 5.8, we compare the effectiveness of risk stratification using our subtype-specific model to risk stratification using a patient's Day 33 MRD. Day 33 MRD was previously shown to be the most powerful predictor of treatment outcome (Cavé et al., 1998; Coustan-Smith et al., 1998; van Dongen et al., 1998). MRD risk stratification thresholds from A. E.-J. Yeoh et al., 2018 were used, with an MRD of $\leq 1 \times 10^{-4}$, between $1 \times 10^{-4}$ and $1 \times 10^{-2}$, and $\geq 1 \times 10^{-2}$ being classified as standard risk, intermediate risk and high risk, respectively.

Patients stratified by the subtype-specific model into standard, intermediate and high risk groups had 5-year event-free survival (EFS) of 100.0% (95% CI: 100.0-100.0%), 79.0% (95% CI: 62.6-99.6%), and 22.1% (% CI: 0.1-55.8%), respectively. In comparison, corresponding MRD risk categories had 5-year EFS of 89.2% (95% CI: 79.7-99.8%), 68.0% (95% CI: 47.9-96.5%) and 33.3% (95% CI: 15.0-74.2%), respectively. Figures 5.8a and 5.8b show that the difference between survival curves of subtype-specific model defined risk groups ($P = 3.72 \times 10^{-9}$, log-rank test) is larger than the difference between survival curves of MRD risk groups ($P = 3.43 \times 10^{-5}$, log-rank test). This implies that the subtype-specific model achieves more refined risk stratification than MRD.

Improved estimation of risk of relapse after treatment advances leukaemia treat-




(a) (b)

Figure 5.8: Kaplan-Meier estimates of event-free survival in patients belonging to homogeneous subtypes in the Ma-Spore ALL data set. Vertical dashes indicate censored observations. (a) Different risk groups are determined by the subtype-specific model (SSM). (b) Different risk groups are determined using Day 33 minimal residual disease (MRD).

ment in two ways. Firstly, having the ability to accurately identify patients with high risk of relapse allows for treatment intensification that will maximise a patient's chance of survival. Secondly, being able to identify patients with a high probability of long-term remission allows for a reduction in treatment intensity. Reduced chemotherapy will lessen side effects and improve patient quality of life. However, as we are dealing with counterfactual outcomes, we are not able to determine with full certainty whether reduced treatment intensity will lower the chance of survival of the patient. A patient might have achieved CCR due to receiving high treatment intensity. We discussed these nuances in greater detail in Chapter 4. Nonetheless, the ability to identify patients with a high probability of long-term remission is still of utility to clinicians, who may use it to support their decision to keep patients on their current course of treatment.

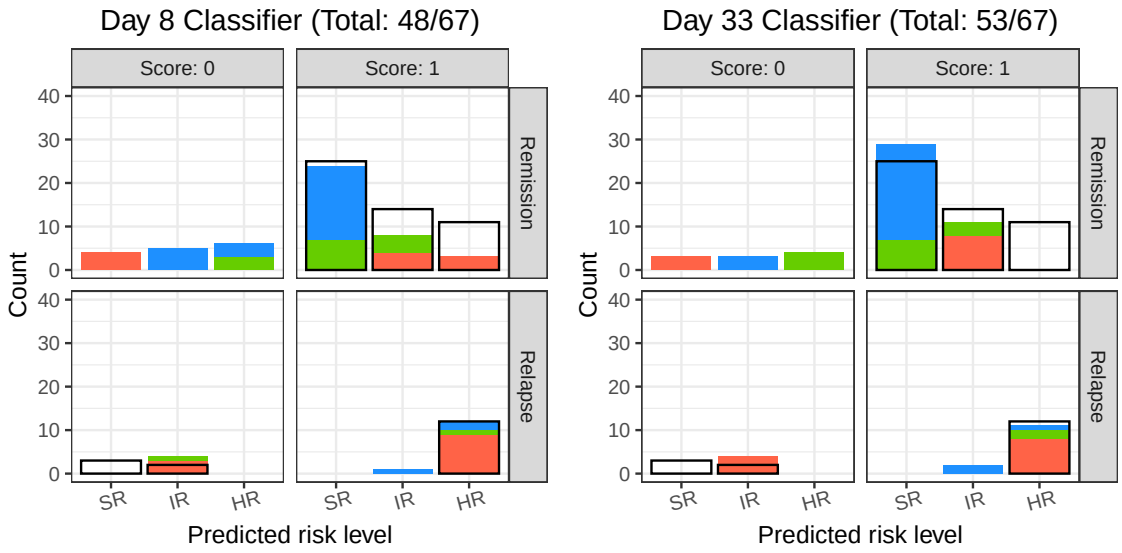
### 5.5.6 Analysis of treatment intensity recommendations

Our subtype-specific model can be used to guide the adjustment of patient treatment intensity during chemotherapy. In essence, the risk groups (SR, IR and HR) presented in the previous section are used as treatment intensity recommendations.

We consider two sets of treatment intensity recommendations: Day 8 and Day 33. Day 8 and Day 33 recommendations are classified based on two different sets of prediction probabilities estimated at Day 8, $P(Remission|\mathbf{x_{D8}}, s)$, and Day 33, $P(Remission|\mathbf{x_{D33}}, s)$, respectively. The probabilities are classified according to the same thresholds with HR: $\leq 0.25$, IR: $> 0.25$ *and* $\leq 0.5$, and SR: $> 0.5$. It is important to note that these treatment intensity recommendations are not the actual treatment intensities that patients were treated on.

We evaluate the two sets of treatment intensity recommendations using our proposed scoring scheme (see section 4.5 for details). Our scoring scheme allows us to incorporate actual treatment information in order to assess whether recommendations are correct or not. Figure 5.9 allows us to analyse the Day 8 and Day 33 treatment intensity recommendations in greater detail. Both sets of treatment intensity recommendations were mostly correct for patients, with an average of approximately 77% correct treatment recommendations. Correct recommendations can either be patients who achieved CCR being recommended the same treatment intensities that they underwent, or patients who relapsed being recommended higher treatment intensities than the intensity they were treated on.

Majority of incorrect recommendations for both Day 8 and Day 33 were recommendations that would result in over-treatment (cases where patients achieved CCR after undergoing actual treatment intensities that are lower than the recommended treatment). Although these recommendations are incorrect, they have a less detrimental impact than incorrect recommendations that result in under-treatment. In addition, most of the incorrect over-treatment recommendations only exceeded by a single risk level (e.g. patients who underwent IR treatment and achieved CCR were recommended HR treatment).




(a) (b)

Figure 5.9: Frequency of scores awarded to two different treatment intensity prediction models. Scores are awarded according to the proposed scoring scheme in section 4.5. Correct and incorrect recommendations are awarded a score of one and zero, respectively. Scores are aggregated for each prediction model, and the total score for each model expressed out of the maximum score attainable is presented in each chart title. Stacked bars represent the cumulative frequency of scores awarded to recommendations. The bars are coloured according to the actual treatment intensity received, with blue, green and red denoting standard risk (SR), intermediate risk (IR) and high risk (HR), respectively. Bar outlines represent the frequency of scores if the actual treatment intensities that patients received were used as treatment intensity recommendations. Each subfigure is divided into four panels, with patients in each panel having the same score and treatment outcome. Performance of prediction models that are variations of a subtype-specific prediction model, based on different sets of overlapping features that are available at various time-points throughout the course of treatment, namely a) Day 8 and b) Day 33.

### 5.5.7 Subtype-specific model outperforms MRD in treatment outcome prediction on novel DUX4-rearranged subtype

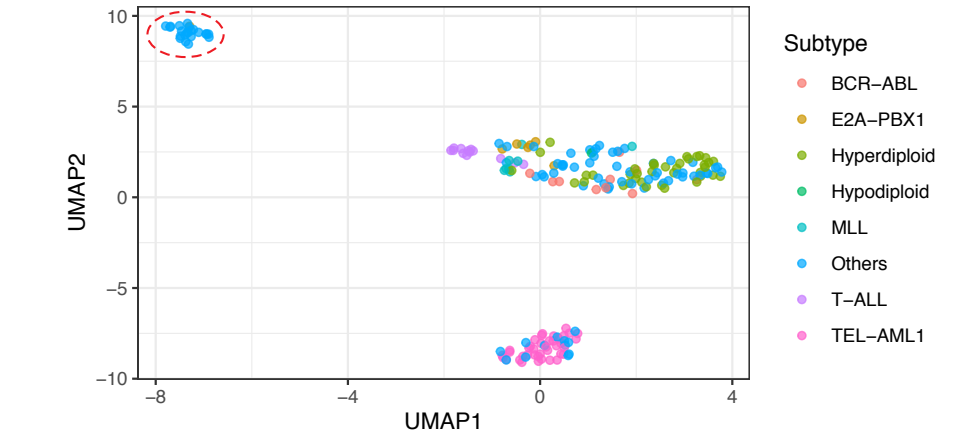
We have shown in Figure 5.5g that the subtype-specific model achieves poor treatment outcome prediction performance on the heterogeneous "B-other" subtype. It is known that the "B-other" subtype is made up of BCP-ALL patients who do not exhibit any of the characteristic cytogenetic abnormalities defining typical B-ALL subtypes. Recent studies have discovered novel subtypes within the "B-

other" subtype, with one of the most prevalent subtypes being the DUX4-rearranged subtype (Gu et al., 2019).

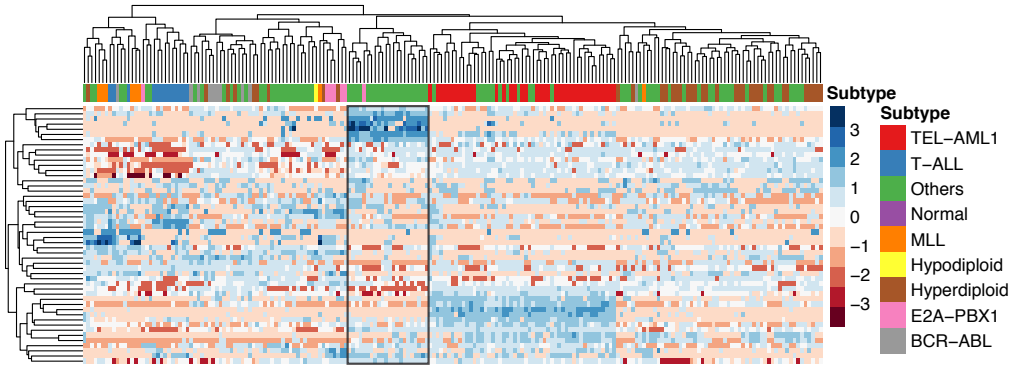
In order to identify DUX4-rearranged samples in the Ma-Spore data set, we utilised genes that were identified by E.-J. Yeoh et al., 2002 to be the gene expression signatures of the following subtypes: BCR-ABL1, E2A-PBX1, MLL, TEL-AML1, hyperdiploid ($> 50$), T-ALL and a novel subtype retrospectively identified as DUX4-rearranged. We then performed uniform manifold approximation and projection (UMAP) dimensionality reduction on all Day 0 samples in the data set, and identified an isolated cluster of "B-other" subtype Day 0 samples (see Figure 5.10a). We performed a hierarchical clustering and only selected for samples that clustered together within the isolated UMAP cluster (see Figure 5.10b). In addition, we verified that these samples were DUX4-rearranged by checking the presence of its gene expression signature.

Figure 5.10c shows that within the DUX4-rearranged subtype, long-term remission patients generally have a larger ERM ratio, ARM ratio and $\theta$ than relapse patients. This implies that patients who achieve long-term remission have both a larger effective and absolute response to treatment. Conversely, according to conventional MRD risk classification thresholds, many of the patients who achieve long-term remission would be falsely classified as intermediate or high risk. The tendency of PCR-based MRD to indicate high residual disease in patients of the DUX4-rearranged subtype has also been observed in Novakova et al., 2021. The authors reported that DUX4 leukaemic blasts underwent a monocytic switch, and retained their PCR-based MRD markers while losing CD19 proteins that act as marker proteins in flow-cytometry-based MRD. This causes PCR-based MRD to report high residual disease whilst flow-cytometry-based MRD indicates low residual disease. As a result, many patients are falsely identified by MRD to be intermediate or high risk even though the DUX4-rearranged subtype has been characterised as being a low risk subtype (Gu et al., 2019; Zaliova et al., 2019).

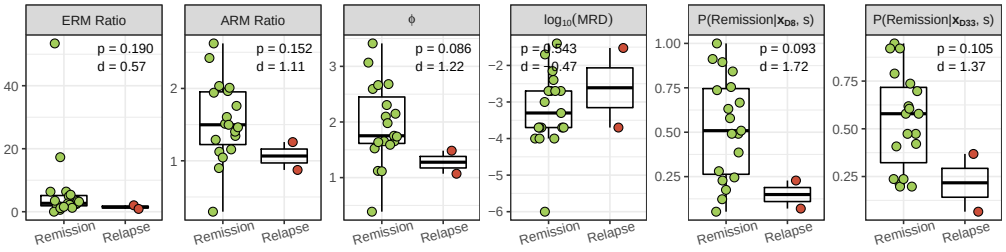
As the subtype-specific model does not trichotomise $log_{10}$(MRD) values into different pre-determined levels, it does not face the same problem. Only the relative rankings of $log_{10}$(MRD) values are used in computing final prediction probabilities. In general, the prediction probabilities estimated by the subtype-specific model on the DUX4-rearranged subtype is higher in long-term remission patients than in




(a)




(b)




(c)

Figure 5.10: Identification of DUX4-rearranged samples in the Ma-Spore ALL data set. (a) UMAP plot of the top 50 most variable genes in Day 0 samples. Cluster of predominantly DUX4-rearranged samples is circled in red. (b) Heat map of gene expression matrix with genes as rows and samples as columns. Gene expression values are row standardised. Unsupervised hierarchical clustering is performed across rows and columns, with cluster of predominantly DUX4-rearranged samples circled in black. (c) Distribution of features computed by the subtype-specific model in the DUX4-rearranged subtype. $P$ values from the Wilcoxon rank-sum test and effect sizes calculated using Cohen's $d$ formula (assuming unequal variance) are reported in the individual panels.

relapse patients.

### 5.5.8 Validation of biological hypothesis

In our biological hypothesis, we assume that the GEP of a patient is a measure of the average gene expression of all leukaemic and normal cells in the patient sample. A patient who has a greater response towards chemotherapy would experience a higher rate of decrease in leukaemic cells than in normal cells. This results in a faster decrease in the proportion of leukaemic cells over time. As a result, the Day 8 GEP of a leukaemic patient with a greater response is shifted further away from the patient's Day GEP, towards the GEP of a normal patient.

We validate our biological hypothesis in two ways. First, we use the deconvolution algorithm MCP-counter (Becht et al., 2016) to infer the quantity of different immune and stromal cell populations in patient samples. Figure 5.11 compares the quantities of B-lineage cell populations across different subtypes and at different time points. Quantities of different types of leukocytes, fibroblasts or endothelial cells estimated by MCP-counter can be compared across samples. However, these quantities cannot be used as an indication of intra-sample proportions of different cell populations.

In general, Day 0 samples belonging to BCP-ALL subtypes have increased B-lineage cell populations when compared to samples from normal patients. This is because BCP-ALL Day 0 samples have a large proportion of leukaemic blasts that resemble early B-cell progenitors (Good et al., 2018). We also observe a general decrease in B-lineage cell populations in Day 33 long-term remission samples that are more sporadic in relapse patients. This validates our biological hypothesis, as we expect the proportion of leukaemic B-lineage cells to decrease as treatment progresses, with long-term remission patients experiencing a greater decrease in leukaemic blasts than relapse patients. On the other hand, the quantity of B lineage cell populations in Day 8 samples do not show a significant difference to that of Day 0 samples. This suggests that the ERM ratio and ARM ratio may be a more sensitive indicator of the decrease in leukaemic cells than the MCP-counter algorithm.

Besides using MCP-counter to infer the quantities of cell populations in samples, we also infer the quantity of B-lineage cell populations directly from gene expression levels of various cluster of differentiation (CD) cell surface markers. These cell

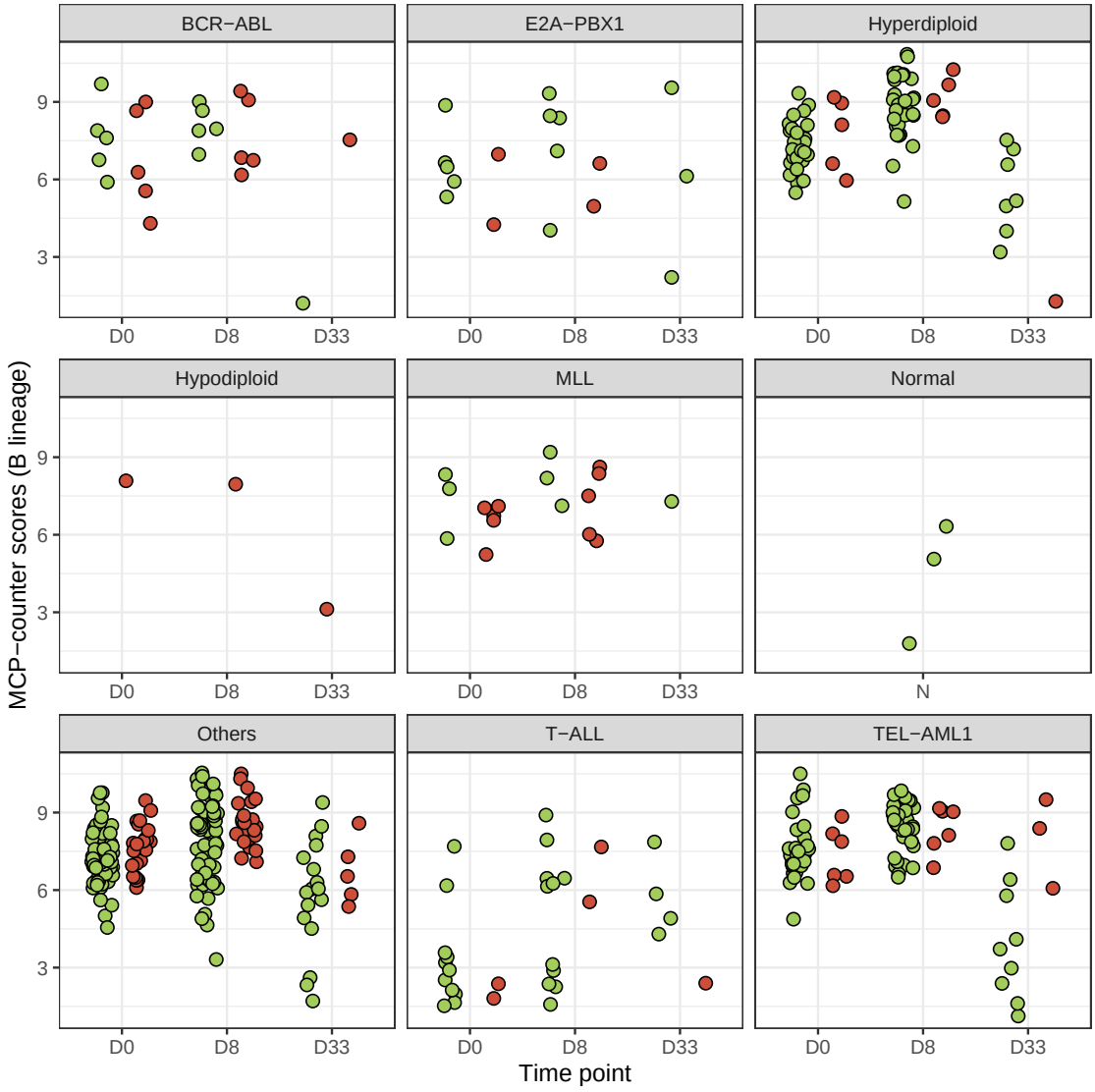
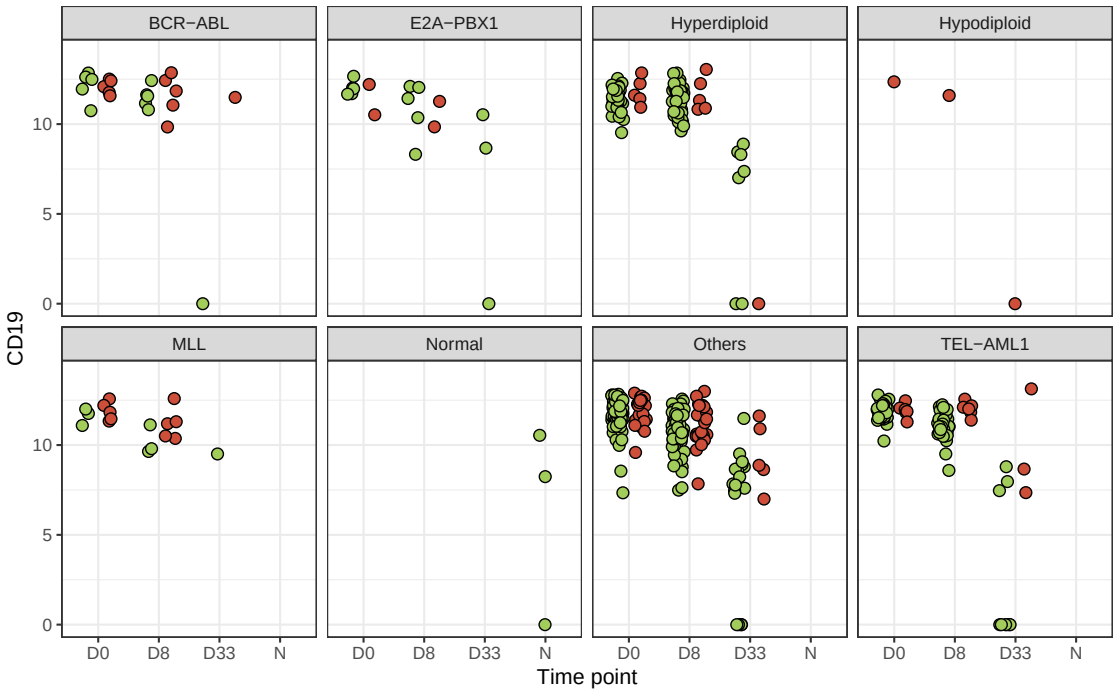



Figure 5.11: Quantity of B-lineage cell population in GEPs from the Ma-Spore ALL data set. MCP-counter (Becht et al., 2016) was used to estimate the quantity of the B-lineage cell population. Quantity is presented in arbitrary units and can be compared across samples. Each panel represents GEPs from a different subtype, with GEPs grouped according to their time points, Day 0 (D0), Day 8 (D8) and Day 33 (D33). GEPs of patients who achieve long-term remission are green in colour, while GEPs of relapse patients are red in colour.

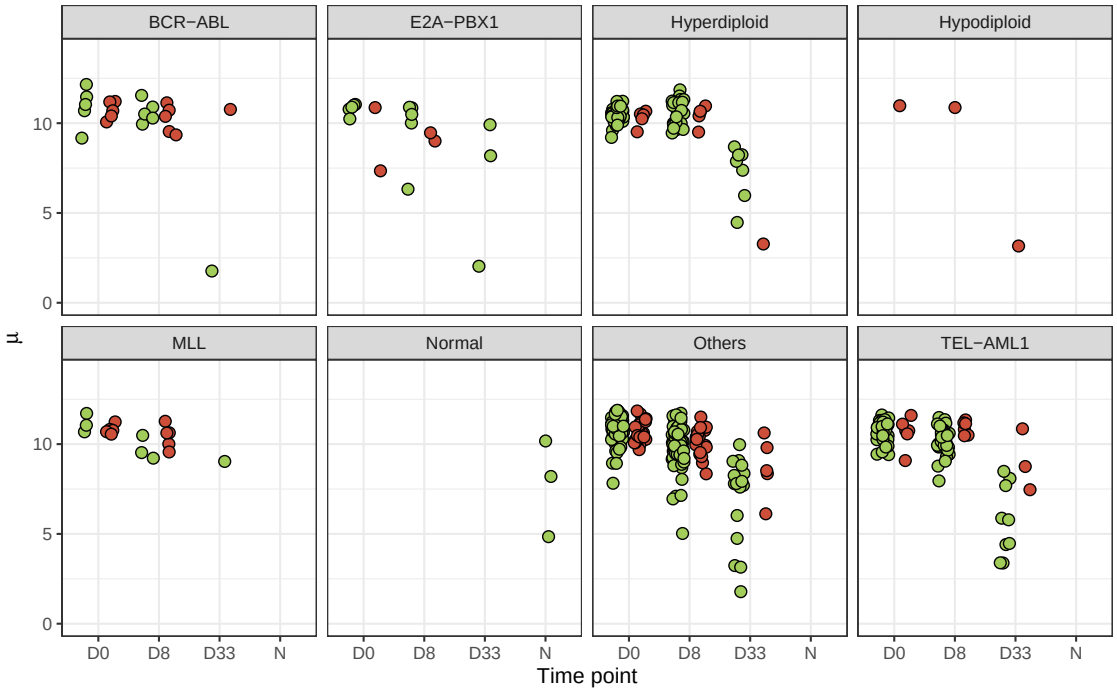
surface markers are frequently used in immunophenotyping ALL. The CD19 cell surface antigen is most commonly used as a marker for B-cells, as it is present in all developmental subpopulations of B-cells. In addition, we curated a list of other CD marker proteins that are commonly expressed in B-cells, namely CD19, CD38, CD72, CD79A and CD79B. We denote the arithmetic average of the gene expression levels of these marker proteins as $\mu$. We use $\mu$ as an estimate of the quantity of B-cells present, in order to get a less noisy estimate as opposed to using CD19 alone.

We observe in Figures 5.12a and 5.12b that both CD19 and $\mu$ values exhibit a decreasing trend as treatment progresses from Day 0 to Day 33 in all BCP-ALL subtypes, with the exception of Day 8 hyperdiploid ($> 50$) samples. The decreasing trend is more pronounced in long-term remission patients, with Day 33 long-term remission and relapse samples showing the largest difference. The decrease in CD19 and $\mu$ values as treatment progresses implies a decrease in proportion of B-cell populations in patient samples at successive time points. It can be observed especially in the TEL-AML1 and B-Others subtype that the decrease in B-cell populations from Day 8 to Day 33 is larger in long-term remission patients than relapse patients. This implies a larger decrease in the proportion of leukaemic cells in long-term remission patients, as leukaemic cells are known to resemble their normal haematopoietic cell counterparts closely (Lukk et al., 2010). This validates our biological hypothesis, which postulates that patients who respond better to treatment experience a faster decrease in proportion of leukaemic cells in their samples over time.

We evaluate the statistical significance of differences in CD19 and $\mu$ values between paired Day 0 and Day 8 samples for each BCP-ALL subtype using a one-tailed paired $t$-test ($H_a : \mu_{D0-D8} > 0$) and present $P$ values and effect sizes (mean difference) in Figures 5.13a and 5.13b. Patients are grouped according to treatment outcome, and paired $t$-tests are performed for each group separately. This is due to our hypothesis that relapse patients will exhibit an inferior response towards treatment, and thus may not show a prominent decrease in proportion of leukaemic cells. The paired differences for both CD19 and $\mu$ values are statistically significant ($P < 0.0005$) in the TEL-AML1 and "B-other" subtypes for long-term remission patients. However, a caveat for interpreting the results is that most of the other subtypes contain less than six pairs of samples. The decrease in CD19 and $\mu$ values

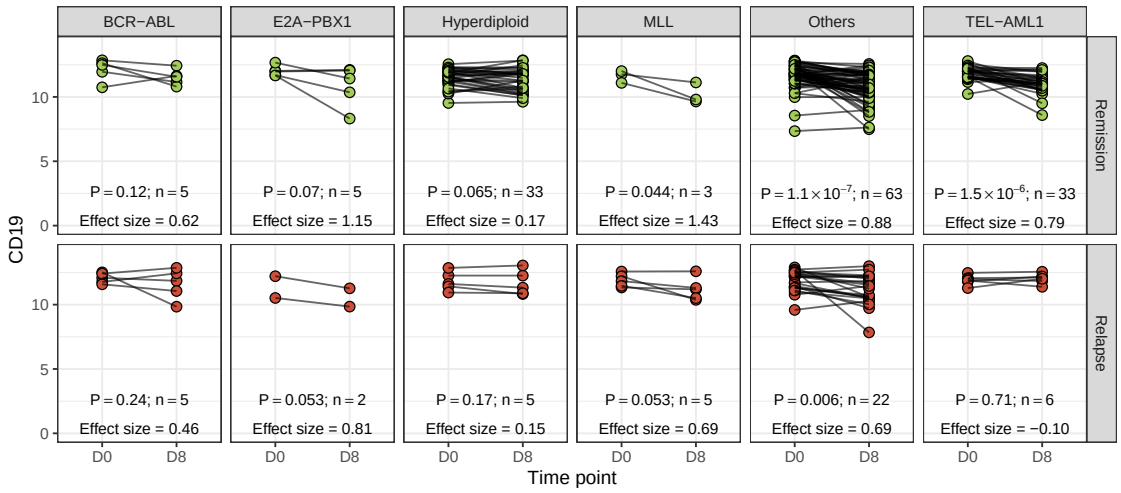



(a)

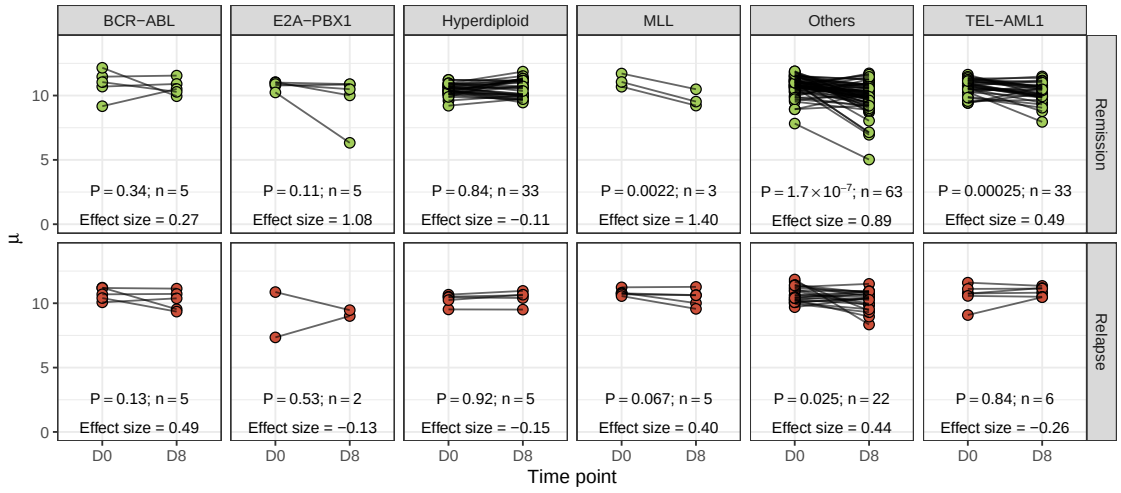



(b)

Figure 5.12: Gene expression of (a) CD19 and (b) average gene expression of CD19, CD38, CD72, CD79A and CD79B (denoted by $\mu$) at multiple time points after treatment (Day 0, Day 8 and Day 33). Long-term remission and relapse samples are represented as green and red dots, respectively.




(a)




(b)

Figure 5.13: Gene expression of (a) CD19 and (b) average gene expression of CD19, CD38, CD72, CD79A and CD79B (denoted by $\mu$) of paired Day 0 (D0) and Day 8 (D8) samples. Long-term remission and relapse samples are represented as green and red dots, respectively. Black lines between dots denote paired samples. $P$ values from a one-sided paired $t$-test ($H_a : \mu_{D0-D8} > 0$) is reported for each panel, along with the number of pairs of samples ($n$). Effect size is measured by the mean paired difference between D0 and D8 samples.

between a patient's sample at Day 0 and Day 8 provides further evidence for our biological hypothesis, which postulates that a response to therapy can be evidenced by a decrease in proportion of leukaemic cells in GEPs of a patient.

## 5.6 Discussion

We postulate that ALL subtypes have distinct treatment responses, with different subtypes having different basal rates of treatment response. Patients of some ALL subtypes may on average experience a higher rate of decrease in leukaemic cells during chemotherapy than patients from other subtypes. However, these patients may all share similar risk of relapse. O'Connor et al., 2018 showed that different ALL subtypes have different log-normal MRD distributions, and that patients from different subtypes with the same MRD level may have different risks of relapse (except for patients with extreme MRD levels). Based on their findings, the authors suggested the use of subtype-specific MRD thresholds for risk stratification. This corroborates with our postulate and supports the use of subtype-specific models for prediction.

Our subtype-specific prediction model was designed to be trained only on long-term remission patients. This helps to reduce over-fitting and enables robust outcome prediction on small data sets. This is essential as clinical ALL data sets are often limited in size. Small data sets pose a challenge for most machine learning algorithms. Firstly, these algorithms tend to over-fit to the small number of training samples. Secondly, it is difficult to assess the accuracy and generalisation performance of a model on a small test set. In spite of not being trained on relapse patients, our subtype-specific prediction model is able to distinguish between long-term remission and relapse patients.

Another important feature about our prediction model is that it estimates prediction probabilities using continuous values of MRD instead of dichotomised MRD risk group. This allows our model to account for different MRD distributions in different ALL subtypes more accurately. On the other hand, most contemporary risk stratification approaches dichotomise continuous features such as age and WBC count, aside from MRD. This may lead to a loss in power and exaggerate differences between similar values separated by an arbitrary threshold.

Previous approaches that identified candidates for reduced-intensity chemotherapy using the dual criteria of absence of high risk cytogenetics and having standard risk MRD were unsuccessful, with candidates suffering an increased risk of relapse after reduced-intensity treatment (Schrappe et al., 2018). In comparison, our subtype-specific model is able to identify with high precision, prime candidates for reduced-intensity chemotherapy in homogeneous ALL subtypes.

A limitation of our study is that the Ma-Spore ALL data set is limited in size. However, we are unable to find independent test sets with similar time-series gene expression data to perform external validation.

## 5.7 Conclusion

In this chapter, we presented our biological hypothesis for treatment response. We hypothesised that chemotherapy kills off leukaemia cells at a faster rate than normal cells in responsive patients. This results in a faster decrease in proportion of leukaemia cells, which can be observed as a greater shift in a patient's GEP from Day 0 to Day 8. We validated our biological hypothesis by estimating B-cell abundance using the deconvolution algorithm MCP-counter, and by examining the gene expression values of specific B-cell markers.

We proposed a subtype-specific model for prediction of treatment outcome in paediatric ALL patients that utilised transcriptomic features computed from microarray gene expression data. Feature selection was performed on gene expression data to select probe sets that are responsive towards treatment. We showed using our proposed batch effects metric RVP (see Chapter 3) and PCA plots that feature selection successfully mitigated batch effects. After feature selection, three transcriptomic features were computed from a patient's Day 0 and Day 8 GEPs. The three transcriptomic features are the ERM ratio, ARM ratio and reorientation ratio $\phi$. ERM ratio measures the magnitude of shift of a patient's GEP (from Day 0 to Day 8) in the direction of the general leukaemia-normal trajectory. Similarly, ARM ratio measures the magnitude of the shift, but does not take into account the direction of the shift. $\phi$ measures the difference between a patient's Day 0 and Day 8 GEPs.

We hypothesised that within patients of the same subtype, having a higher ERM ratio, higher ARM ratio or higher $\phi$ are all indicative of a higher probability

of long-term remission. We showed empirically that the above hypothesis is true in homogeneous ALL subtypes. In addition, we demonstrated that our subtype-specific model outperforms existing methods in treatment outcome prediction on homogeneous ALL subtypes. Our model is able to identify prime candidates for reduced-intensity chemotherapy with high fidelity. Successful reduction of chemotherapy in candidates will reduce cytotoxic side effects in patients, without increasing their risk of relapse. Our subtype-specific model can be used to guide adjustments of patient treatment intensity. Treatment intensity recommendations can be classified using thresholds on a patient's predicted probability of achieving CCR. Evaluation of treatment intensity recommendations using our proposed scoring scheme (see Chapter 4) showed that most of the recommendations were correct; majority of the incorrect recommendations were one intensity level too high (which is preferable to recommending an insufficient treatment intensity level).

# Chapter 6

# Conclusion

Over recent years, there has been massive interest in artificial intelligence and its potential applications in healthcare. As a result, there has been a deluge of clinical prediction models proposed in healthcare literature. However, few of these models end up being deployed in the real-world, as most of them show poor generalisation ability (Perel et al., 2006; Wyatt & Altman, 1995).

In this thesis, we discussed the key reasons behind why so few clinical prediction models are deployed in practice. We identified improper development and evaluation of clinical prediction models as one of the main reasons, and highlighted that clinical prediction models are especially susceptible to improper development and evaluation because of multiple inherent heterogeneities in clinical data. In particular, we delved deeper into two commonly encountered heterogeneities in clinical data that impede proper model development - batch effects and differences in patient treatment.

We discussed the main characteristics of batch effects present in high-dimensional biological data, and proposed a novel quantitative batch effects metric named RVP. RVP was developed to be robust to batch-class imbalance, and to be suitable for use in both small and large data sets. We demonstrated that RVP is able to achieve accurate estimation of the proportion of variance across different magnitudes of batch effects using simulated RNA-seq data, for both data with and without batch-class imbalance. We also showed that RVP is able to outperform existing metrics on real world data sets from different high-throughput technologies (viz. gene expression microarray and quantitative proteomics data). RVP has a faster time complexity and lower peak memory usage than other batch effects metrics that involve the use of PCA. We believe that RVP is a powerful tool that would facilitate modelling practitioners to account for batch effects when developing their prediction models.

Differences in patient treatment are common in clinical data due to the prevalence of risk-adapted treatment strategies. Treatment outcome of patients are affected to different degrees depending on the treatment received. In clinical prediction models, treatment outcome is often the prediction target. Failing to account for differences in patient treatment during model development will result in inaccurate models. In this thesis, we examined methods that have been proposed for handling treatment differences when developing prediction models, and discussed the nuances in using these methods. Using the Ma-Spore ALL data set as a case study, we proposed a plausible scoring scheme that incorporates treatment information to achieve more detailed evaluation of prediction models. We also demonstrated a way to visualise results from the scoring scheme that facilitates analysis of predictions made by a model.

In this thesis, we developed a subtype-specific prediction model for paediatric ALL patients that predicts whether patients will achieve long-term remission or relapse. Our model incorporates the use of patient GEPs at different time-points of treatment. We demonstrated using PCA plots and our proposed batch effects metric RVP that feature selection successfully mitigated batch effects in the data set. We proposed three transcriptomic features that are computed from patient GEPs at different time-points, and showed that these features are associated with treatment outcome in homogeneous subtypes of ALL. We demonstrated that our subtype-specific model is able to discriminate between patients who achieve long-term remission and patients who relapse, and is robust to small sample sizes. We also validated our model in a recently discovered ALL subtype, the DUX4 subtype. In addition, we utilised our proposed scoring scheme (which incorporates patient treatment information) to evaluate our subtype-specific model in greater detail.

Our biological hypothesis behind the subtype-specific model was that the GEP of a patient measures the average gene expression of all leukaemic and normal cells in the patient sample, and that patients responsive towards chemotherapy would show a bigger shift in their GEPs due to the faster decrease in proportion of leukaemic cells. We validated our biological hypothesis by estimating the abundance of B-cells in patient samples through various methods.

## 6.1 Future work

There are a few natural directions for further research that can be pursued from the works and findings presented in this thesis. Firstly, we seek to evaluate the performance of RVP in quantifying batch effects in larger data sets such as scRNA-seq data. The total sample size of scRNA-seq data is typically an order of magnitude larger than that encountered in our experimental data sets. We believe that RVP is able to scale efficiently with the increase in data size. We also seek to further evaluate RVP using simulated data with a more complicated design, with an example being a data set consisting of multiple batches and classes, but with certain missing batch-class combinations. Furthermore, as RVP implicitly de-convolutes the total variance in the data into feature-specific, class-specific and batch-specific variances, we are also interested in the possibility of deriving better batch effect correction methods by taking advantage of the de-convoluted variances produced in the course of computing RVP.

Secondly, the successful validation of our biological hypothesis through the estimation of B-cell abundance demonstrated the value of examining compositions of cell sub-populations as a way to assess treatment response in ALL. This provides evidence of the potential of using scRNA-seq technology to predict treatment response. Measuring samples at different time points of treatment using scRNA-seq would enable more accurate quantification of cell sub-populations. More importantly, the use of scRNA-seq technology may enable the identification of two different clonal evolution models behind relapse patients. The first model consists of leukaemic cells that are resistant from the start, while the second model involves clonal selection, where sensitive cells that formed the majority initially were killed off and replaced by a small subpopulation of resistant cells (Y. C. Cohen et al., 2021). We believe that a possible reason why our subtype-specific model incorrectly classifies some relapse patients is because those patients belong to the second model. Patients in the second model often show good initial treatment response, but end up relapsing. They are often responsible for the poor performance of models that base their predictions solely on initial patient treatment response.

# Bibliography

Affymetrix. (2002). Statistical algorithms description document.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511.

Apgar, V. (1952). A proposal for a new method of evaluation of the newborn. *Classic Papers in Critical Care*, *32*(449), 97.

Aricò, M., Valsecchi, M. G., Camitta, B., Schrappe, M., Chessells, J., Baruchel, A., Gaynon, P., Silverman, L., Janka-Schaub, G., Kamps, W., et al. (2000). Outcome of treatment in children with Philadelphia chromosome–positive acute lymphoblastic leukemia. *New England Journal of Medicine*, *342*(14), 998–1006.

Beam, A. L., Manrai, A. K., & Ghassemi, M. (2020). Challenges to the reproducibility of machine learning models in health care. *Journal of the American Medical Association*, *323*(4), 305–306.

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, *17*(1).

Belorkar, A., & Wong, L. (2016). GFS: Fuzzy preprocessing for effective gene expression analysis. *BMC Bioinformatics*, *17*(17), 540.

Bhojwani, D., Kang, H., Menezes, R. X., Yang, W., Sather, H., Moskowitz, N. P., Min, D.-J., Potter, J. W., Harvey, R., Hunger, S. P., et al. (2008). Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: a Children's Oncology Group Study. *Journal of Clinical Oncology*, *26*(27), 4376–4384.

Biondi, A., Cimino, G., Pieters, R., & Pui, C.-H. (2000). Biological and therapeutic aspects of infant leukemia. *Blood*, *96*(1), 24–33.

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial intelligence in healthcare* (pp. 25–60). Elsevier.

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, *16*(1), 43–49.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

Cavé, H., van der Werff Ten Bosch, J., Suciu, S., Guidal, C., Waterkeyn, C., Otten, J., Bakkus, M., Thielemans, K., Grandchamp, B., Vilmer, E., et al. (1998). Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia. *New England Journal of Medicine*, *339*(9), 591–598.

Chan, W. X., & Wong, L. (2022). Accounting for treatment during the development or validation of prediction models. *Journal of Bioinformatics and Computational Biology*, *20*(6), 2271001. https://doi.org/10.1142/S0219720023710014

Chan, W. X., & Wong, L. (2023). Obstacles to effective model deployment in healthcare. *Journal of Bioinformatics and Computational Biology*, *21*(2), 2371001. https://doi.org/10.1142/S0219720023710014

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PloS One*, *6*(2), e17238.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, Y. C., Zada, M., Wang, S.-Y., Bornstein, C., David, E., Moshe, A., Li, B., Shlomi-Loubaton, S., Gatt, M. E., Gur, C., et al. (2021). Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing. *Nature Medicine*, *27*(3), 491–503.

Conter, V., Bartram, C. R., Valsecchi, M. G., Schrauder, A., Panzer-Grümayer, R., Möricke, A., Aricò, M., Zimmermann, M., Mann, G., De Rossi, G., et al. (2010). Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia:

results in 3184 patients of the AIEOP-BFM ALL 2000 study. *Blood*, *115*(16), 3206–3214.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, *24*(10), 1559–1567.

Coustan-Smith, E., Behm, F. G., Sanchez, J., Boyett, J. M., Hancock, M. L., Raimondi, S. C., Rubnitz, J. E., Rivera, G. K., Sandlund, J. T., Pui, C.-H., et al. (1998). Immunological detection of minimal residual disease in children with acute lymphoblastic leukaemia. *The Lancet*, *351*(9102), 550–554.

Den Boer, M. L., van Slegtenhorst, M., De Menezes, R. X., Cheok, M. H., Buijs-Gladdines, J. G., Peters, S. T., Van Zutven, L. J., Beverloo, H. B., Van der Spek, P. J., Escherich, G., et al. (2009). A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: A genome-wide classification study. *The Lancet Oncology*, *10*(2), 125–134.

Dickerman, B. A., Dahabreh, I. J., Cantos, K. V., Logan, R. W., Lodi, S., Rentsch, C. T., Justice, A. C., & Hernán, M. A. (2022). Predicting counterfactual risks under hypothetical treatment strategies: An application to HIV. *European Journal of Epidemiology*, *37*(4), 367–376.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy — analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315.

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., et al. (2015). Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*, *10*(7), e0132321.

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology*, *35*(6), 498–507.

Goh, W. W. B., Yong, C. H., & Wong, L. (2022). Are batch effects still relevant in the age of big data? *Trends in Biotechnology*, *40*(9), 1029–1040.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*. The Johns Hopkins University Press.

Good, Z., Sarno, J., Jager, A., Samusik, N., Aghaeepour, N., Simonds, E. F., White, L., Lacayo, N. J., Fantl, W. J., Fazio, G., et al. (2018). Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nature Medicine*, *24*(4), 474–483.

Greaves, M. F. (1986). Differentiation-linked leukemogenesis in lymphocytes. *Science*, *234*(4777), 697–704.

Groenwold, R. H., Moons, K. G., Pajouheshnia, R., Altman, D. G., Collins, G. S., Debray, T. P., Reitsma, J. B., Riley, R. D., & Peelen, L. M. (2016). Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology*, *78*, 90–100.

Gu, Z., Churchman, M., Roberts, K., Li, Y., Liu, Y., Harvey, R. C., McCastlain, K., Reshmi, S. C., Payne-Turner, D., Iacobucci, I., et al. (2016). Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nature Communications*, *7*, 13331.

Gu, Z., Churchman, M. L., Roberts, K. G., Moore, I., Zhou, X., Nakitandwe, J., Hagiwara, K., Pelletier, S., Gingras, S., Berns, H., et al. (2019). PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nature Genetics*, *51*(2), 296–307.

Guo, T., Fan, Y., Chen, M., Wu, X., Zhang, L., He, T., Wang, H., Wan, J., Wang, X., & Lu, Z. (2020). Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19). *Journal of the American Medical Association Cardiology*, *5*(7), 811–818.

Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421.

Harewood, L., Robinson, H., Harris, R., Al-Obaidi, M. J., Jalali, G., Martineau, M., Moorman, A., Sumption, N., Richards, S., Mitchell, C., et al. (2003). Amplification of AML1 on a duplicated chromosome 21 in acute lymphoblastic leukemia: a study of 20 cases. *Leukemia*, *17*(3), 547–553.

Harris, M. B., Shuster, J. J., Carroll, A., Look, A. T., Borowitz, M. J., Crist, W. M., Nitschke, R., Pullen, J., Steuber, C. P., & Land, V. J. (1992). Trisomy of leukemic cell chromosomes 4 and 10 identifies children with B-progenitor cell acute lymphoblastic leukemia with a very low risk of treatment failure: a Pediatric Oncology Group study. *Blood*, *79*(12), 3316–3324.

Harvey, R. C., Mullighan, C. G., Wang, X., Dobbin, K. K., Davidson, G. S., Bedrick, E. J., Chen, I.-M., Atlas, S. R., Kang, H., Ar, K., et al. (2010). Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood*, *116*(23), 4874–4884.

Heerema, N. A., Raimondi, S. C., Anderson, J. R., Biegel, J., Camitta, B. M., Cooley, L. D., Gaynon, P. S., Hirsch, B., Magenis, R. E., McGavran, L., et al. (2007). Specific extra chromosomes occur in a modal number dependent pattern in pediatric acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer*, *46*(7), 684–693.

Heerema, N. A., Sather, H. N., Sensel, M. G., Zhang, T., Hutchinson, R. J., Nachman, J. B., Lange, B. J., Steinherz, P. G., Bostrom, B. C., Reaman, G. H., et al. (2000). Prognostic impact of trisomies of chromosomes 10, 17, and 5 among children with acute lymphoblastic leukemia and high hyperdiploidy ($>$ 50 chromosomes). *Journal of Clinical Oncology*, *18*(9), 1876–1887.

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G., Steyerberg, E. W., et al. (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *British Medical Journal*, *346*, e5595.

Hie, B., Bryson, B., & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, *37*(6), 685.

Ho, S. Y., Phua, K., Wong, L., & Goh, W. W. B. (2020). Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, *1*(8), 100129.

Hof, J., Krentz, S., Van Schewick, C., Körner, G., Shalapour, S., Rhein, P., Karawajew, L., Ludwig, W.-D., Seeger, K., Henze, G., et al. (2011). Mutations

and deletions of the TP53 gene predict nonresponse to treatment and poor outcome in first relapse of childhood acute lymphoblastic leukemia. *Journal of Clinical Oncology*, *29*(23), 3185–3193.

Holleman, A., Cheok, M. H., den Boer, M. L., Yang, W., Veerman, A. J., Kazemier, K. M., Pei, D., Cheng, C., Pui, C.-H., Relling, M. V., et al. (2004). Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *New England Journal of Medicine*, *351*(6), 533–542.

Hunger, S. P., & Mullighan, C. G. (2015). Acute lymphoblastic leukemia in children. *New England Journal of Medicine*, *373*(16), 1541–1552.

Inaba, H., Greaves, M., & Mullighan, C. G. (2013). Acute lymphoblastic leukaemia. *The Lancet*, *381*(9881), 1943–1955.

Johnson, A. E., Pollard, T. J., & Mark, R. G. (2017). Reproducibility in critical care: A mortality prediction case study. *Machine Learning for Healthcare Conference*, 361–376.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.

Kakani, P., Chandra, A., Mullainathan, S., & Obermeyer, Z. (2020). Allocation of COVID-19 relief funding to disproportionately black counties. *Journal of the American Medical Association*, *324*(10), 1000–1003.

Kang, H., Chen, I.-M., Wilson, C. S., Bedrick, E. J., Harvey, R. C., Atlas, S. R., Devidas, M., Mullighan, C. G., Wang, X., Murphy, M., et al. (2010). Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood*, *115*(7), 1394–1405.

Krentz, S., Hof, J., Mendioroz, A., Vaggopoulou, R., Dörge, P., Lottaz, C., Engelmann, J., Groeneveld, T., Körner, G., Seeger, K., et al. (2013). Prognostic value of genetic alterations in children with first bone marrow relapse of childhood B-cell precursor acute lymphoblastic leukemia. *Leukemia*, *27*(2), 295–304.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., & Nowé, A. (2012). Batch effect removal methods for microarray gene expression data integration: A survey. *Briefings in Bioinformatics*, *14*(4), 469–490.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739.

Liew, S., Doust, J., & Glasziou, P. (2011). Cardiovascular risk scores do not account for the effect of treatment: A review. *Heart*, *97*(9), 689–697.

Lilljebjörn, H., Henningsson, R., Hyrenius-Wittsten, A., Olsson, L., Orsmark-Pietras, C., Von Palffy, S., Askmyr, M., Rissler, M., Schrappe, M., Cario, G., et al. (2016). Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nature Communications*, *7*(1), 1–13.

Lugthart, S., Cheok, M. H., den Boer, M. L., Yang, W., Holleman, A., Cheng, C., Pui, C.-H., Relling, M. V., Janka-Schaub, G. E., Pieters, R., et al. (2005). Identification of genes associated with chemotherapy crossresistance and treatment response in childhood acute lymphoblastic leukemia. *Cancer Cell*, *7*(4), 375–386.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., & Brazma, A. (2010). A global map of human gene expression. *Nature Biotechnology*, *28*(4), 322–324.

Lütge, A., Zyprych-Walczak, J., Kunzmann, U. B., Crowell, H. L., Calini, D., Malhotra, D., Soneson, C., & Robinson, M. D. (2021). CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life Science Alliance*, *4*(6).

Manimaran, S., Selby, H. M., Okrah, K., Ruberman, C., Leek, J. T., Quackenbush, J., Haibe-Kains, B., Bravo, H. C., & Johnson, W. E. (2016). BatchQC: interactive software for evaluating sample and batch effects in genomic data. *Bioinformatics*, *32*(24), 3836–3838.

Mateen, B. A., Liley, J., Denniston, A. K., Holmes, C. C., & Vollmer, S. J. (2020). Improving the quality of machine learning in health applications and clinical research. *Nature Machine Intelligence*, *2*(10), 554–556.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*(3), 190–195.

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, *3*(29), 861.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115.

Meyer, L. H., Eckhoff, S. M., Queudeville, M., Kraus, J. M., Giordan, M., Stursberg, J., Zangrando, A., Vendramini, E., Möricke, A., Zimmermann, M., et al. (2011). Early relapse in ALL is identified by time to leukemia in NOD/SCID mice and is characterized by a gene signature involving survival pathways. *Cancer Cell*, *19*(2), 206–217.

Miller, D. R., Coccia, P. F., Bleyer, W. A., Lukens, J. N., Siegel, S. E., Sather, H. N., & Hammond, G. D. (1989). Early response to induction therapy as a predictor of disease-free survival and late recurrence of childhood acute lymphoblastic leukemia: A report from the children's cancer study group. *Journal of Clinical Oncology*, *7*(12), 1807–1815.

Moorman, A. V., Ensor, H. M., Richards, S. M., Chilton, L., Schwab, C., Kinsey, S. E., Vora, A., Mitchell, C. D., & Harrison, C. J. (2010). Prognostic effect of chromosomal abnormalities in childhood B-cell precursor acute lymphoblastic leukaemia: results from the UK Medical Research Council ALL97/99 randomised trial. *The Lancet Oncology*, *11*(5), 429–438.

Moorman, A. V., Richards, S. M., Martineau, M., Cheung, K. L., Robinson, H. M., Jalali, G. R., Broadfield, Z. J., Harris, R. L., Taylor, K. E., Gibson, B. E., et al. (2003). Outcome heterogeneity in childhood high-hyperdiploid acute lymphoblastic leukemia. *Blood*, *102*(8), 2756–2762.

Mullighan, C. G., Su, X., Zhang, J., Radtke, I., Phillips, L. A., Miller, C. B., Ma, J., Liu, W., Cheng, C., Schulman, B. A., et al. (2009). Deletion of IKZF1 and

prognosis in acute lymphoblastic leukemia. *New England Journal of Medicine*, *360*(5), 470–480.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

Nestor, B., McDermott, M. B., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Goldenberg, A., & Ghassemi, M. (2019). Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. *Machine Learning for Healthcare Conference*, 381–405.

Novakova, M., Zaliova, M., Fiser, K., Vakrmanova, B., Slamova, L., Musilova, A., Brüggemann, M., Ritgen, M., Fronkova, E., Kalina, T., et al. (2021). DUX4r, ZNF384r and PAX5-P80R mutated B-cell precursor acute lymphoblastic leukemia frequently undergo monocytic switch. *Haematologica*, *106*(8), 2066–2075.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

O'Connor, D., Enshaei, A., Bartram, J., Hancock, J., Harrison, C. J., Hough, R., Samarasinghe, S., Schwab, C., Vora, A., Wade, R., et al. (2018). Genotype-specific minimal residual disease interpretation improves stratification in pediatric acute lymphoblastic leukemia. *Journal of Clinical Oncology*, *36*(1), 34–43.

Oner, M. U., Cheng, Y.-C., Lee, H. K., & Sung, W.-K. (2020). Training machine learning models on patient level data segregation is crucial in practical clinical applications. *medRxiv*. https://doi.org/10.1101/2020.04.23.20076406

Pajouheshnia, R., Damen, J. A., Groenwold, R. H., Moons, K. G., & Peelen, L. M. (2017). Treatment use in prognostic model research: A systematic review of cardiovascular prognostic studies. *Diagnostic and Prognostic Research*, *1*, 15.

Peek, N., Sperrin, M., Mamas, M., Van Staa, T., & Buchan, I. (2017). Hari Seldon, QRISK3, and the prediction paradox. *British Medical Journal*, *357*, j2099.

Perel, P., Edwards, P., Wentz, R., & Roberts, I. (2006). Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making*, *6*, 38.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*(7), 369–375.

Pui, C.-H., & Evans, W. E. (2006). Treatment of acute lymphoblastic leukemia. *New England Journal of Medicine*, *354*(2), 166–178.

Pui, C.-H., Robison, L. L., & Look, A. T. (2008). Acute lymphoblastic leukaemia. *The Lancet*, *371*(9617), 1030–1043.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raimondi, S. C., Behm, F. G., Roberson, P. K., Williams, D. L., Pui, C.-H., Crist, W. M., Look, A. T., & Rivera, G. K. (1990). Cytogenetics of pre-B-cell acute lymphoblastic leukemia with emphasis on prognostic implications of the t (1; 19). *Journal of Clinical Oncology*, *8*(8), 1380–1388.

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: What, why, how, when and where? *Clinical Kidney Journal*, *14*(1), 49–58.

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., De Andrade, M., Kocher, J.-P. A., & Eckel-Passow, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics*, *29*(22), 2877–2883.

Riehm, H., Reiter, A., Schrappe, M., Berthold, F., Dopfer, R., Gerein, V., Ludwig, R., Ritter, J., Stollmann, B., & Henze, G. (1987). Corticosteroid-dependent reduction of leukocyte count in blood as a prognostic factor in acute lymphoblastic leukemia in childhood (therapy study ALL-BFM 83). *Klinische Padiatrie*, *199*(3), 151–160.

Scherer, A. (2009). *Batch effects and noise in microarray experiments: Sources and solutions*. John Wiley & Sons.

Schrappe, M., Bleckmann, K., Zimmermann, M., Biondi, A., Möricke, A., Locatelli, F., Cario, G., Rizzari, C., Attarbaschi, A., Valsecchi, M. G., et al. (2018). Reduced-intensity delayed intensification in standard-risk pediatric acute lymphoblastic leukemia defined by undetectable minimal residual disease: Results of an international randomized trial (aieop-bfm all 2000). *Journal of Clinical Oncology*, *36*(3), 244–253.

Schwab, C., & Harrison, C. J. (2018). Advances in B-cell precursor acute lymphoblastic leukemia genomics. *HemaSphere*, *2*(4), e53.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., De Longueville, F., Kawasaki, E. S., Lee, K. Y., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, *24*(9), 1151–1161.

Siontis, G. C., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34.

Smith, M., Arthur, D., Camitta, B., Carroll, A. J., Crist, W., Gaynon, P., Gelber, R., Heerema, N., Korn, E. L., Link, M., et al. (1996). Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *Journal of Clinical Oncology*, *14*(1), 18–24.

Sorich, M. J., Pottier, N., Pei, D., Yang, W., Kager, L., Stocco, G., Cheng, C., Panetta, J. C., Pui, C.-H., Relling, M. V., et al. (2008). In vivo response to methotrexate forecasts outcome of acute lymphoblastic leukemia and has a distinct gene expression profile. *PLoS Medicine*, *5*(4), e83.

Sperrin, M., Martin, G. P., Pate, A., Van Staa, T., Peek, N., & Buchan, I. (2018). Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*, *37*(28), 4142–4154.

Stary, J., Zimmermann, M., Campbell, M., Castillo, L., Dibar, E., Donska, S., Gonzalez, A., Izraeli, S., Janic, D., Jazbec, J., Konja, J., Kaiserova, E., Kowalczyk, J., Kovacs, G., Li, C.-K., Magyarosy, E., Popa, A., Stark, B., Jabali, Y., ... Schrappe, M. (2014). Intensive chemotherapy for childhood acute lymphoblastic leukemia: results of the randomized intercontinental

trial ALL IC-BFM 2002. *Journal of Clinical Oncology*, *32*(3), 174–184. https://doi.org/10.1200/JCO.2013.48.6522

Steyerberg, E. W., & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, *69*, 245–247.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., & Group, P. (2013). Prognosis research strategy (progress) 3: Prognostic model research. *PLoS Medicine*, *10*(2), e1001381.

Sutcliffe, M., Shuster, J., Sather, H., Camitta, B., Pullen, J., Schultz, K., Borowitz, M., Gaynon, P., Carroll, A., & Heerema, N. (2005). High concordance from independent studies by the Children's Cancer Group (CCG) and Pediatric Oncology Group (POG) associating favorable prognosis with combined trisomies 4, 10, and 17 in children with NCI Standard-Risk B-precursor Acute Lymphoblastic Leukemia: a Children's Oncology Group (COG) initiative. *Leukemia*, *19*(5), 734–740.

Swerdlow, S. H., Campo, E., Pileri, S. A., Harris, N. L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G. A., Zelenetz, A. D., et al. (2016). The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, *127*(20), 2375–2390.

Tan, B. K. J., Teo, C. B., Tadeo, X., Peng, S., Soh, H. P. L., Du, S. D. X., Luo, V. W. Y., Bandla, A., Sundar, R., Ho, D., et al. (2021). Personalised, rational, efficacy-driven cancer drug dosing via an artificial intelligence SystEm (PRECISE): a protocol for the PRECISE CURATE.AI pilot clinical trial. *Frontiers in Digital Health*, *3*, 635524.

Tan, Z., Roche, K., Zhou, X., & Mukherjee, S. (2018). Scalable Algorithms for Learning High-Dimensional Linear Mixed Models. *arXiv e-prints*, Article arXiv:1803.04431, arXiv:1803.04431. https://doi.org/10.48550/arXiv.1803.04431

Tran, T. H., & Loh, M. L. (2016). Ph-like acute lymphoblastic leukemia. *Hematology 2014, the American Society of Hematology Education Program Book*, *2016*(1), 561–566.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

van Dongen, J. J., Seriu, T., Panzer-Grümayer, E. R., Biondi, A., Pongers-Willemse, M. J., Corral, L., Stolz, F., Schrappe, M., Masera, G., Kamps, W. A., et al. (1998). Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *The Lancet*, *352*(9142), 1731–1738.

van Geloven, N., Swanson, S. A., Ramspek, C. L., Luijken, K., van Diepen, M., Morris, T. P., Groenwold, R. H., van Houwelingen, H. C., Putter, H., & le Cessie, S. (2020). Prediction meets causal inference: The role of treatment in clinical prediction models. *European Journal of Epidemiology*, *35*(7), 619–630.

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., et al. (2021). The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials*, *22*, 537.

Willenbrock, H., Juncker, A., Schmiegelow, K., Knudsen, S., & Ryder, L. (2004). Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia*, *18*(7), 1270–1277.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837–1847.

Witkowski, M. T., Dolgalev, I., Evensen, N. A., Ma, C., Chambers, T., Roberts, K. G., Sreeram, S., Dai, Y., Tikhonova, A. N., Lasry, A., et al. (2020). Extensive remodeling of the immune microenvironment in B cell acute lymphoblastic leukemia. *Cancer Cell*, *37*(6), 867–882.

Wyatt, J. C., & Altman, D. G. (1995). Commentary: Prognostic models: Clinically useful or quickly forgotten? *British Medical Journal*, *311*, 1539.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Dahly, D. L., Damen, J. A., Debray, T. P., et al. (2020). Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *British Medical Journal*, *369*, m1328.

Yasuda, T., Tsuzuki, S., Kawazu, M., Hayakawa, F., Kojima, S., Ueno, T., Imoto, N., Kohsaka, S., Kunita, A., Doi, K., et al. (2016). Recurrent DUX4 fusions in B

cell acute lymphoblastic leukemia of adolescents and young adults. *Nature Genetics*, *48*(5), 569–574.

Yeoh, A. E.-J., Li, Z., Dong, D., Lu, Y., Jiang, N., Trka, J., Tan, A. M., Lin, H. P., Quah, T. C., Ariffin, H., et al. (2018). Effective Response Metric: a novel tool to predict relapse in childhood acute lymphoblastic leukaemia using time-series gene expression profiling. *British Journal of Haematology*, *181*(5), 653–663.

Yeoh, A. E. J., Ariffin, H., Chai, E. L. L., Kwok, C. S. N., Chan, Y. H., Ponnudurai, K., Campana, D., Tan, P. L., Chan, M. Y., Kham, S. K. Y., et al. (2012). Minimal residual disease-guided treatment de-intensification for children with Acute Lymphoblastic Leukemia: Results from the Malaysia-Singapore Acute Lymphoblastic Leukemia 2003 Study. *Journal of Clinical Oncology*, *30*(19), 2384–2392.

Yeoh, A. E. J., Lu, Y., Chin, W. H. N., Chiew, E. K. H., Lim, E. H., Li, Z., Kham, S. K. Y., Chan, Y. H., Abdullah, W. A., Lin, H. P., et al. (2018). Intensifying treatment of childhood B-lymphoblastic leukemia with IKZF1 deletion reduces relapse and improves overall survival: results of Malaysia-Singapore ALL 2010 study. *Journal of Clinical Oncology*, *36*(26), 2726–2735.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*(2), 133–143.

Zaliova, M., Potuckova, E., Hovorkova, L., Musilova, A., Winkowska, L., Fiser, K., Stuchly, J., Mejstrikova, E., Starkova, J., Zuna, J., et al. (2019). ERG deletions in childhood acute lymphoblastic leukemia with DUX4 rearrangements are mostly polyclonal, prognostically relevant and their detection rate strongly depends on screening method sensitivity. *Haematologica*, *104*(7), 1407–1416.

Zhang, J., McCastlain, K., Yoshihara, H., Xu, B., Chang, Y., Churchman, M. L., Wu, G., Li, Y., Wei, L., Iacobucci, I., et al. (2016). Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nature Genetics*, *48*(12), 1481–1489.

Zhou, L., Sue, A. C.-H., & Goh, W. W. B. (2019). Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects? *Journal of Genetics and Genomics*, *46*(9), 433–443.

# Publications during PhD Study

Chan, W. X., & Wong, L. (2022). Accounting for treatment during the development or validation of prediction models. *Journal of Bioinformatics and Computational Biology*, *20*(6), 2271001. https://doi.org/10.1142/S0219720023710014

Chan, W. X., & Wong, L. (2023). Obstacles to effective model deployment in healthcare. *Journal of Bioinformatics and Computational Biology*, *21*(2), 2371001. https://doi.org/10.1142/S0219720023710014